EIN HOCH AUF PYTHAGORAS

nach Alan Sykes, University College of Swansea
Originaltitel in 'Teaching Statistics' Vol.7(1985), Nr.1:
Three Cheers for Pythagoras
Übersetzung: Bernd Wollring

Zwei wichtige Überlegungen stehen über jedem Unterrichtsvorhaben:

(1) Darstellen der Beziehungen zwischen verwandten Disziplinen und
(2) Korrespondieren mit den langfristigen Zielen der eigenen Fachsystematik

In 'Teaching Statistics' 6(1984) Nr.1 führt Anna HART an, es sei schwer zu rechtfertigen, weshalb wir σ^2 und entsprechend σ schätzen, indem wir **quadratische Abweichungen** benutzen. (Gemeint ist hier die Standardabweichung als Parameter der für die Grundgesamtheit angenommenen Normalverteilung. Anm. d. Ü., siehe HART) Darf ich einen möglichen Ansatz vorschlagen, der zugleich die oben genannten Prinzipien illustriert ?

Nicht informierte Studenten untersuchen als Einführung in dieses Problem ein einfaches Beispiel eines **linearen statistischen Modells.** Dabei begegnen sie im Anfängerkurs (A-level-course) der Technik des t-Tests für zwei Stichproben und der linearen Regression, ohne direkt wahrzunehmen, daß diese Dinge in enger Verbindung zu dem stehen, womit sie sich befassen. Erst im zweiten Jahr des Statistik-Kurses aud der Universität werden sie vermutlich die einheitliche Theorie hinter diesen Modellen wahrnehmen und erkennen, wie wichtig diese für den praktizierenden Statistiker und Datenanalytiker ist. Da eine fundamentale Frage gestellt ist, so ist es vielleicht nicht überraschend, daß zu ihrer

Stochastik in der Schule, Heft 3, Jg. 5 (1985)

Lösung ein fundamentaler mathematischer Satz herangezogen wird: der Satz des Pythagoras.

Nehmen wir den einfachen Fall zweier Beobachtungen Y_1 und Y_2, die zusammen eine Stichprobe vom Umfang 2 aus einer normalverteilten Grundgesamtheit mit unbekanntem Mittelwert μ und unbekannter Varianz σ^2 bilden. Das lineare statistische Modell für diese Situation betrachtet den Zufallsvektor (Y_1,Y_2), im Bild durch OA dargestellt, dessen unbekannter Mittelwert der Punkt C = (μ,μ) auf der Geraden $y_1 = y_2$ durch den Ursprung ist.

Wenn sie dem Problem in dieser Form gegenüberstehen, so schätzen die Studenten den Mittelwert μ ohne Zögern mit Hilfe des Punktes B auf der Geraden $y_1 = y_2$, der zu A die kürzeste Entfernung hat und freuen sich, daß man diesen Punkt finden kann, indem man OA auf diese Linie projiziert, so daß das rechtewinklige Dreieck OAB für den Punkt B die Koordinaten $((Y_1 + Y_2)/2 , (Y_1 + Y_2)/2)$ ergibt.

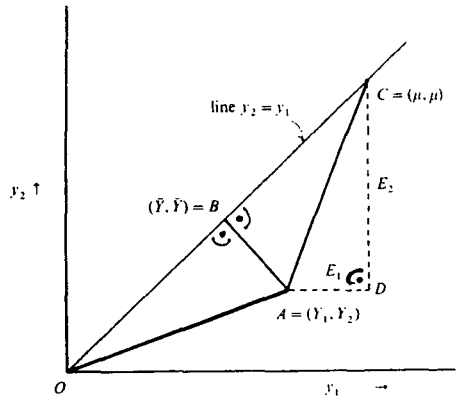

Bild 1

Soviel zur Schätzung des Mittelwertes. Was ist mit σ^2 ? Würden wir σ^2 kennen, so hätten die Werte $Y_1 - \mu$ und $Y_2 - \mu$, dargestellt durch die Zufallsgrößen E_1 und E_2 mit $N(0,\sigma^2)$-Verteilung, in dem Diagramm die Eigenschaft $AC^2 = E_1^2 + E_2^2$ (Wieder der Satz des Pythagoras, diesmal im Dreieck ADC). Da E_1 und E_2 die Varianz σ^2 haben, folgt, daß der Erwartungswert von AC^2 gleich σ^2 ist. Für eine Stichprobe vom Umfang n ist der entsprechende Erwartungswert $n\,\sigma^2$, und daher erhält man die passende Schätzfunktion, indem man die Quadratsumme der Abweichungen vom wahren Mittelwert durch n dividiert.

Kennen wir aber den Mittelwert μ nicht, so können wir lediglich den Punkt C mit Hilfe des Punktes B schätzen, somit wird die Länge AC durch AB ersetzt. Aber AB^2 ist natürlich stets kleiner als AC^2, daher kann der Erwartungswert nicht $2\,\sigma^2$ sein.

Rettung bringt wieder der Satz des Pythagoras im Dreieck ABC:

$$AC^2 = AB^2 + BC^2$$

Somit gilt:

$$E(AB^2) = 2\,\sigma^2 - E(BC^2)$$

Nun ist aber:

$$BC^2 = ((Y_1 + Y_2)/2 - \mu)^2 + ((Y_1 + Y_2)/2 - \mu)^2$$
$$= 2\,(\bar{Y} - \mu)^2$$

Und weil $Var(\bar{Y}) = \sigma^2/2$ ist, folgt $E(BC^2) = \sigma^2$. Das gilt zudem für jeden Stichprobenumfang.

Also ist $E(AB^2) = 2\sigma^2 - \sigma^2$, im allgemeinen Fall $(n-1)\sigma^2$. Folglich schätzen wir σ^2 durch die Quadratsumme der Abweichungen vom Mittelwert der Stichprobe, dividiert durch (n-1).

Dieser einfache Ansatz trifft den Kern linearer statistischer Modelle, wobei wie oben erwähnt, t-Test und lineare Regression eingeschlossen sind. Bei jedem solchen Modell erfolgt die Analyse in folgenden zwei Schritten:

(1) Abschätzen des Mittelwertes oder der Regressionsparameter durch Projektionstechniken und

(2) Abschätzen von σ^2 durch das Längenquadrat des 'Residual-Vektors' AB, dividiert durch die 'Zahl der Freiheitsgrade', die im allgemeinen der Zahl der Beobachtungen minus der Zahl der geschätzten Lageparameter entspricht.

Es scheint mir, daß dieser Ansatz einem intelligenten Studenten mit Erfolg anzubieten ist, denn er beruht auf einem wichtigen Satz aus der Mittelstufe und weist darauf hin, wie dieser Satz später weiterverwendet wird, wen die Technik der Vektoren und Matrizen entwickelt ist.

Literatur

HART,A.E.: How Should We Teach the Standard Deviation ? .- Teaching Statistics 6(1984) Nr.1 ; Übersetzung in diesem Heft