Visualisieren von Korrelation mit Verteilungsdiagrammen

von Erich Neuwirth, Universität Wien
Bearbeitung : Bernd Wollring, Universität Münster

Zusammenfassung

Bei der Diskussion der Korrelation im Unterricht sind Beispiele von Datensätzen nützlich, die eine gegebene Korrelation besitzen. Dieser Beitrag beschreibt eine Methode, mit Hilfe von Verteilungsdiagrammen (spreadsheeds) die zugehörigen Punktwolken zu zeichnen und "Konfidenzgebiete" zu erhalten.

Einführung

In diesem Beitrag stellen wir eine Methode vor, um Punktwolken mit einem gegebenen Korrelationskoeffizienten aus Zufallszahlen zu erhalten, so wie sie gewöhnlich mit Verteilungsdiagramm-Programmen zu enthalten sind. In Fortführung dieser Methode können wir ebenfalls Graphen von zugehörigen Konfidenzgebieten einzeichnen, die analytisch abgeleitet sind und den Zufallsdaten überlagert werden können.

Betrachten wir den Graphen in Bild 1. Versuchen Sie bitte ohne vorher weiterzulesen, den Korrelationskoeffizienten der dort gezeichneten Punktwolke zu schätzen. Die Lösung wird am Ende des Beitrages verraten. War Ihre Schätzung einigermaßen brauchbar?

Wir hoffen, dieses Beispiel zeigt, daß es ein wichtiges Ziel ist, Schüler und Studenten zu befähigen, Korrelationkoeffizienten zu schätzen, und zwar ohne irgendeine weitere rechnerische Tätigkeit.

Benötigte mathematische Ergebnisse

Beginnen wir mit dem einfachsten Fall. Wir möchten eine Stichprobe bivariater Daten haben, wobei die beiden Zufallsvariablen nicht korreliert sind. Um die Dinge so natürlich und einfach wie möglich zu gestalten, möchten wir die beiden Zufallsvariablen normalverteilt haben. Da wir den Begriff der Korrelation verdeutlichen wollen, und zwar mit möglichst wenig Pa-

rametern, wählen wir standardisierte Normalverteilungen, d.h. $\mu = 0$ und $\sigma^2 = 1$.

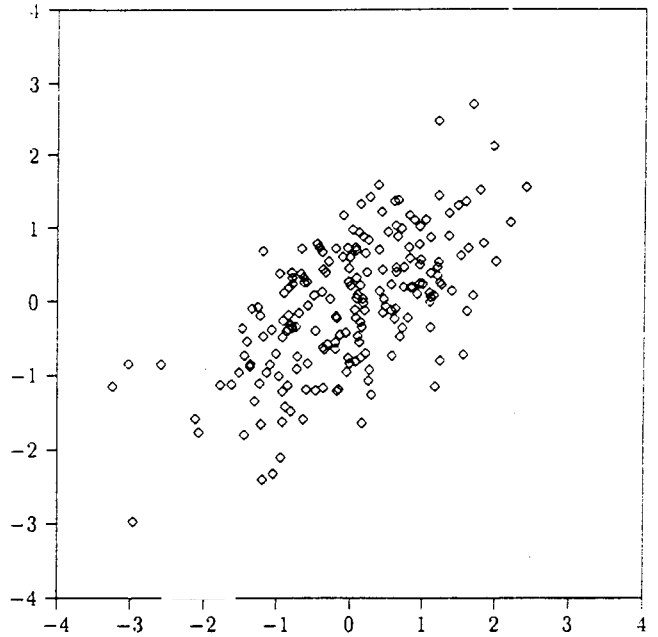

Bild 1 : P nktwc ke , Korrelationskoeffizient bitte raten !

Wir verwenden eine der Sta ıdardr ethoden, um standardisierte normalverteilte Zufallsgrößen zu erhalten: Wir addieren z völ Z fallszahlen, die aus einer Gleichverteilung auf dem Intervall [0,1] stammen. Wie l ekannt hat diese Standard-Gleichverteilung die Varianz $\sigma^2 = 1/12$, so daß die Summe be ein di : Varianz $\sigma^2 = 1$ besitzt. Um $\mu = 0$ zu erhalten, haben wir von dieser Summe 6 zu subtra ıieren. All dies kann sehr leicht in einem Verteilungsdiagramm-Programm durchge ührt w :rden, das einen Zufallsgenerator besitzt, und zum Unterrichten von Statistik sollte nan nu - solche Programme verwenden .

Wir wollen nun die Ausleg ıngsmode Verteilungsdiagrammes kurz zurückstellen und zunächst die benötigten theoretische Ergel nisse darlegen.

Wir können also mit Hilfe des Zufallsgenerators eine Stichprobe von zwei nichtkorrelierten Zufallsgrößen erzeugen. Im nächsten Schritt wollen wir eine dritte Zufallsgröße konstruieren, die eine gegebene Korrelation mit der ersten Zufallsgröße besitzt. Die Grundidee besteht darin, zu diesem Zweck eine Linearkombination der ersten und der zweiten Zufallsgröße zu verwenden. So beginnen wir mit den standardisierten normalverteilten Variablen X und Y mit der Korrelation 0 und betrachten $Z = \alpha X + \beta Y$. Da sowohl X als auch Y den Erwartungswert 0 haben, gilt dies auch für Z . Für die Varianz von Z gilt :

$$\sigma^2(Z) = \alpha^2 \sigma^2(X) + \beta^2 \sigma^2(Y) + 2 \alpha \beta \operatorname{cov}(X,Y)$$

Wegen $\sigma^2(X) = \sigma^2(Y) = 1$ und $\operatorname{cov}(X;Y) = 0$ folgt :

$$\sigma^2(Z) = \alpha^2 + \beta^2$$

Ferner gilt :

$$\operatorname{cov}(X,Z) = \alpha \operatorname{cov}(X,X) + \beta \operatorname{cov}(X,Y) = \alpha$$

bestimmen. Um alle Zeichnungen, die wir herstellen wollen, vergleichbar zu machen, wählen wir α und β so, daß $\sigma^2(Z) = 1$ entsteht. Daher müssen wir $\beta = \sqrt{(1 - \alpha^2)}$ wählen.

Unser Endergebnis besteht darin, daß wir für zwei standardisierte nicht korrelierte Zufallsgrößen X und Y die neue Zufallsgröße $Z = \rho X + \sqrt{(1 - \rho^2)}\, Y$ erhalten, die ebenfalls standardisiert ist und mit der Zufallsgröße X den Korrelationskoeffizienten ρ besitzt.

Auslegung des Verteilungsdiagramms

Der wichtigste Teil unseres Verteilungsdiagramms sind die Spalten der Tafel mit den Werten der Zufallsgrößen X , Y und Z . Für jeden Wert von X bzw. Y benötigen wir zwölf Zufallszahlen aus dem Intervall [0,1] . Diese Zahlen sind nur Zwischenergebnisse für unsere Zwecke, so daß sie nicht unbedingt sichtbar in dem Diagramm auftreten müssen. Die zwei ersten Spalten können die Werte für X und Y enthalten, wie wir sie aus diesen Zufallszahlen bestimmen, die dritte Spalte kann die Werte für Z enthalten. Wir benötigen ebenfalls die Angaben von ρ und $\sqrt{(1 - \rho^2)}$, so daß unser Verteilungsdiagramm wie Tabelle 1 aussehen kann.

Correlation visualised		
rho		0.65
sqrt (1 − rho ^ 2)		0.76
X	Y	Z
− 1.14	− 0.64	− 1.23
.	.	.
.	.	.
.	.	.

Tabelle 1 : Beispiel zur Auslegung des Verteilungsdiagramms

Der Term $\sqrt{(1 - \rho^2)}$ sollte als Formel dargestellt sein, so daß seine Beziehung zu ρ unmittelbar zu sehen ist. Die Darstellung der Spalten X und Y entsteht aus den nicht dargestellten Teilen des Verteilungsdiagramms , den 24 Zufallszahlen in jeder Zeile. Die Felder mit den Werten ρ und $\sqrt{(1 - \rho^2)}$ sollten mit entsprechenden Namen gekennzeichnet sein, und die Formeln zu der Z-Spalte sollten indirekt auf die Spalten X und Y bezogen sein und direkt auf ρ und $\sqrt{(1 - \rho^2)}$.

Die Spalten X und Z sollten als die Datenbereiche für die Punktwolke bezeichnet werden. im Verteilungsdiagramm-Programm oft bezeichnet als X-Y-Zeichnung (!) , wobei in der Graphik nur Symbole benutzt werden sollten, die die Datenpunkte verbinden, und keine Linien. Um vergleichbare Graphiken zu erhalten, sollten die Minima und Maxima der Werte für die X- und Y-Bereiche von Hand gesetzt werden.

Die Zahl der benutzten Datenpunkte ist sehr wichtig, um einen Eindruck davon zu vermitteln, daß man wirklich mit einem statistischen Ensemble arbeitet und nicht nur mit ein paar Punkten. Man sollte mindestens 100 Punkte benutzen. Auf regulären Personal-Computern kann man mit Hilfe von Verteilungsdiagramm-Programmen ohne Probleme 200 Punkte erstellen. Alle Graphiken in diesem Beitrag sind mit Hilfe von 200 Datenpunkten und der angegebenen Methode erstellt.

Das hier verwendete Programm für die Verteilungsdiagramme ist so organisiert, daß jeweils nach Eingabe des betreffenden Wertes für ρ und Verabschieden durch Tastendruck das Programm eine Punktwolke mit dieser gegebenen Korrelation zeichnet.

Wenn wir die theoretischen Ergebnisse aus dem letzten Abschnitt nicht verwenden wollen, können wir ebenso den empirischen Korrelationskoeffizienten mit den Funktionen des Verteilungsdiagrammes berechnen. Wir können auf diese Weise den Mechanismus zur Berechnung von Z gewissermaßen verstecken und empirisch prüfen, wie die Daten für einen gegebenen Korrelationskoeffizienten aussehen. Es sollte nicht zu schwierig sein, unser Verteilungsdiagramm-Programm für die gegebene empirische Korrelation aufzubauen. In dem Beitrag von Lageard (1988) wird gezeigt, wie dies durchzuführen ist.

Betrachten wir nun zwei Beispiele, die Bilder 2 und 3.

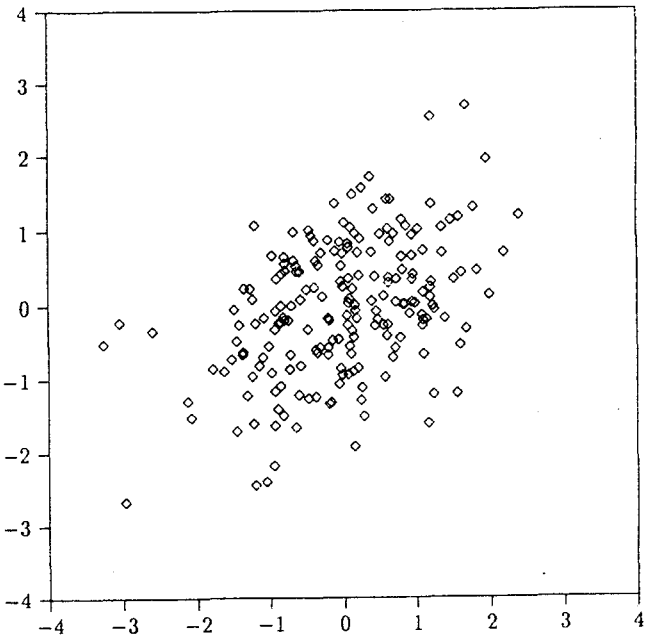

Bild 2 : Punktwolke mit $\rho = 0.5$

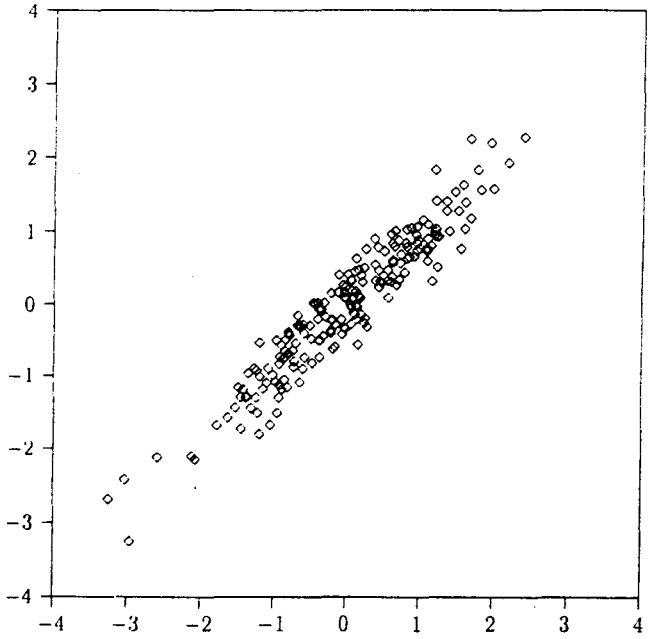

Bild 3 : Punktwolke mit $\rho = 0.95$

Diese Beispiele zeigen, daß eine Korrelation von $\rho = 0.5$ nicht immer zu einer Zeichnung gehört, die man ohne weiteres als funktionale Abhängigkeit interpretieren würde. Beim Experimentieren mit verschiedenen Werten von ρ kann man die graphische Darstellung von beliebig vielen Datenmengen untersuchen, und man findet eine Menge verschiedener Graphiken fest, die zeigen, welche unterschiedlichen Interpretationen von Korrelationskoeffizienten möglich sind.

An dieser Stelle ist auf einen wichtigen Punkt hinzuweisen : Zufallszahlen-Generatoren in den meisten üblichen Programmen auf Personal-Computern sind nicht so gut, wie Statistiker sie gerne hätten. Besonders dann, wenn es um Probleme mit Dimensionen größer als 1 geht, zeigen sich diese Schwächen. So ist es auch bei unserem Programm zum Verteilungsdiagramm. (Wir haben das Programm ASEASYAS 3.01a benutzt.) Die Korrelation zwischen

den Daten X und Y, die ja als unkorreliert angenommen werden, erwies sich als - 0.14 . Man kann dieses Problem allerdings überwinden, indem man den empirischen Korrelationskoeffizienten zwischen X und Z berechnet und dann das theoretisch vorgegebene ρ ändert, um so den angestrebten Wert für die empirische Korrelation zu erhalten. Andererseits ist der Unterschied zwischen dem theoretischen und dem empirischen Wert hinreichend klein, um unsere Programme zu den Verteilungsdiagrammen für die genannten Zwecke nach wie vor als brauchbar anzusehen.

Überlagern der Daten mit Konfidenz-Ellipsoiden

Um eine Beziehung zwischen den "Zufallsdaten" und den theoretischen Verteilungen darzustellen, ist es ganz nützlich, ein Konfidenz-Ellipsoid für die zweidimensionale Normalverteilung darzustellen und dieses der "Punktwolke" zu überlagern, um so ihre relativen Lagen miteinander zu vergleichen. Die Berechnung dieses Ellipsoids ist recht einfach, falls für die theoretische Verteilung $\rho = 0$ ist. In diesem Fall ist das Gebiet einfach ein Kreis, und es ist ebenfalls leicht, den Radius dieses Kreises zu berechnen. Wenn X und Y unkorrelierte Standardnormalverteilungen sind, dann folgt $X^2 + Y^2$ einer χ^2-Verteilung mit zwei Freiheitsgraden. So ist das Konfidenzgebiet für das Konfidenzniveau γ durch die Ungleichung

$$x^2 + y^2 \leq C_\gamma$$

gegeben, wobei C_γ das γ-Quantil der χ^2-Verteilung mit zwei Freiheitsgraden ist.

Da bekanntlich die χ^2-Verteilung mit zwei Freiheitsgraden eine Exponentialverteilung mit Mittelwert 2 ist, kann man leicht zeigen, daß das Radiusquadrat C_γ des Konfidenzbereiches durch folgende Formel gegeben ist :

$$C_\gamma = - 2 \ln (1 - \gamma)$$

So erhalten wir für $\gamma = 0.95$ den Wert $C_\gamma = 5.99$. Mit Hilfe dieser Gleichung und durch Einzeichnen der Datenpunkte und des Kreises erhalten wir schließlich Bild 4.

Bei genauerer Betrachtung dieses Bildes kann man prüfen, wie ähnlich sich das Gebiet, das

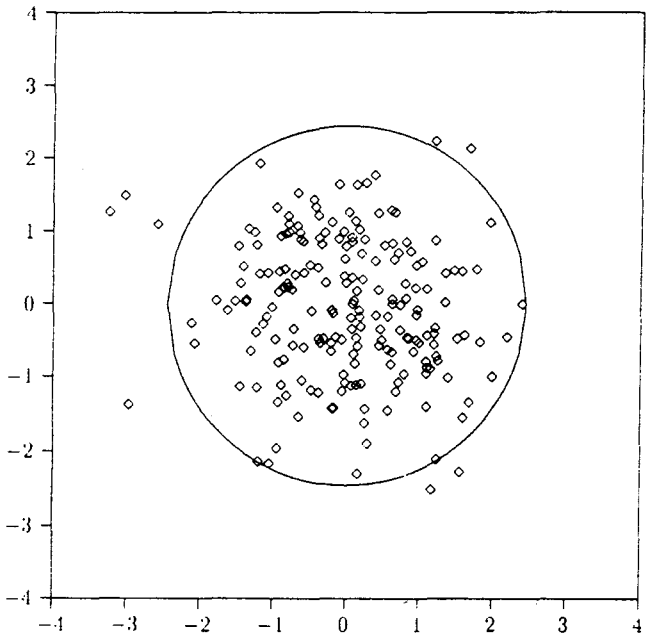

Bild 4 : Punktwolke mit $\rho = 0.0$

die meisten Punkte überdeckt, und das Ellipsoid sind. Man kann ferner den Anteil der Datenpunkte prüfen, die außerhalb des Ellipsoides liegen, das in diesem Fall ein Kreis ist, und ihn mit dem theoretischen Wert 5 % vergleichen.

Das Berechnen der Ellipse wird ein bißchen schwieriger für den Fall, daß die Korrelation ungleich 0 ist. Mit Hilfe einiger einfachen Rechnungen kann man jedoch die benötigten Ergebnisse herleiten.

Nehmen wir also an, wir haben standardisierte normalverteilte Zufallsgrößen X und Y mit der Korrelation ρ . Wir definieren die Zufallsgröße W wie folgt :

$$W = (X - \rho Z) / \sqrt{(1 - \rho^2)}$$

Dann ist W nicht mit Z korreliert, denn es gilt :

$$\mathrm{cov}(W,Z) = (1/\sqrt{(1-\rho^2)})\ \mathrm{cov}(X,Z) - (\rho/\sqrt{(1-\rho^2)})\ \mathrm{cov}(Z,Z) = 0$$

Daher hat der Konfidenzkreis für die aus W und Z zusammengesetzte Verteilung den Radius $w^2 + z^2 \leq C_\gamma$.

Mit Hilfe der Definition von W können wir diesen Konfidenzkreis für die zusammengesetzte Verteilung (W,Z) in ein Konfidenzellipsoid für die zusammengesetzte Verteilung (X,Z) transformieren. Es gilt :

$$((x - \rho z)^2 / (1 - \rho^2)) + z^2 \leq C_\gamma$$

$$(x - \rho z)^2 \leq (1 - \rho^2)\ (C_\gamma - z^2)$$

$$x \leq \rho z \pm \sqrt{((1 - \rho^2)\ (C_\gamma - z^2))}$$

Wir können diese Gleichung verwenden, um das Konfidenzellipsoid in unser Verteilungsdiagramm einzuzeichnen.

Dazu führen wir eine neue Spalte ein, die äquidistante Werte x_1 von $- C_\gamma$ bis $+ C_\gamma$ enthält und eine zweite Spalte mit den Werten

$$x_2 = \rho x_1 - \sqrt{((1 - \rho^2)\ (C_\gamma - x_1^2))} ,$$

die den unteren Rand dieses Ellipsoides ergeben. Dann setzen wir die erste Spalte fort, indem wir von $+ C_\gamma$ auf $- C_\gamma$ zurückgehen, die entsprechenden x_1-Werte wieder äquidistant auftragen und dazu die zweite Spalte jetzt mit

$$x_2 = \rho x_1 + \sqrt{((1 - \rho^2)\ (C_\gamma - x_1^2))} ,$$

die den oberen Rand des Ellipsoides ergeben.

Diese Punkte können wir verwenden, um die Ellipse zu zeichnen. Wir haben lediglich die erste Spalte als x-Werte zu markieren und die zweite Spalte als y, um einen (x,y)-Graphen zu erhalten und mit Hilfe des Graphikprogrammes die Punkte dieses Graphen miteinander zu verbinden, ohne Symbole an diesen Punkten zu zeichnen. Wenn wir die explizite Formel für C_γ verwenden, können wir sogar Konfidenzellipsen für verschiedene Konfidenzniveaus zeichnen.

Schließlich können wir den Ellipsengraphen und die Punktwolken überlagern. Führt man alle notwendigen Schritte für $\rho = 0{,}65$ und $\gamma = 0.25, 0.50, 0.75$ und 0.90 durch, so entsteht Bild 5.

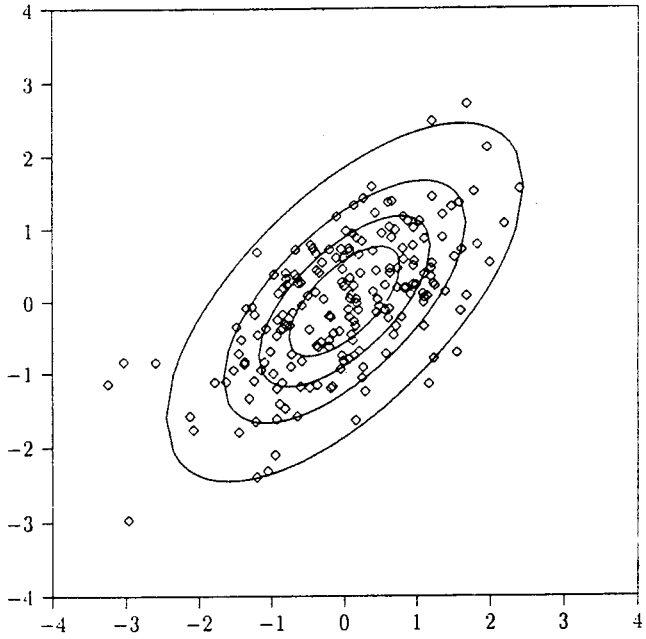

Bild 5 : Punktwolke mit $\rho = 0.65$

Didaktische Bemerkungen

Es ist ganz klar, daß einige Teile der theoretischen Herleitung jenseits des Anspruchs liegen,

der für einen großen Teil der Schüler und Studenten realisierbar ist, für die das Visualisieren von Korrelationskoeffizienten eine sinnvolle Übung wäre. Diese Tatsache schließt allerdings nicht ein, daß das Verteilungsdiagramm wie hier beschrieben nicht mit solchen Schülern oder Studenten zusammen verwendet werden kann. Es ist im wesentlichen eine Sache des Lehrers zu entscheiden, welche theoretischen Konzepte und Rechnungen er zur Konstruktion dieses Verteilungsdiagrammes vorstellt, welche er mit den Schülern bearbeitet und welche er nicht erwähnt. Man kann sich sehr verschiedene Gebrauchsformen dieses Verteilungsdiagrammes vorstellen. Die einfachste Methode würde sein, geradewegs den Rechner zu benutzen und eine Projektion herzustellen, um so interaktiv mit der Zeichnung der zweidimensionalen Verteilung zu arbeiten. In diesem Fall könnte man die Theorie vollständig verstecken und nichts weiter zeigen, als eine Folge von Zeichnungen, die zu bestsimmten Korrelationswerten gehören. Die am meisten fortgeschrittene Methode könnte darin bestehen, all die hier vorgestellten Begriffe und Formeln herzuleiten, um auf diese Weise zu zeigen, wie statistische Theorie zur Konstruktion statistischer Hilfsmittel verwendet wird.

Der Hauptvorteil der Verwendung von Verteilungsdiagrammen zum Visualisieren von Korrelationen ist der, daß das Verteilungsdiagramm letztlich - so sollte es wenigstens sein - ein fertig verwendbares Werkzeug ist, wenn man Statistik unterrichtet. Der Autor wollte ebenfalls darlegen, daß Verteilungsdiagramme nicht nur zur Analyse bereits bestehender Daten verwendet werden können, sondern ebenso zur Herstellung von Daten, die bestimmte wichtige statistische Zusammenhänge erläutern können.

Und nun abschließend die Antwort auf die Frage zu Beginn des Beitrages. Die in Bild 1 gezeichnete Datenmenge hat den Wert $\rho = 0.65$.

Literatur

LAGEARD, J. L. : Using a Spreadsheet to Teach Coding of Bivariate Data .- Teaching Statistics, 10 (1988) , vol.1 , S. 20 - 22