BESCHREIBENDE STATISTIK: Zentrale Tätigkeiten, lokale Bedeutungen, globale Ideen

Werner Peschek, Universität Klagenfurt

1. Zum Thema

Es ist weitgehend unbestritten, dass die beschreibende Statistik in vielen gesellschaftlichen Lebensbereichen angewandt wird, in den Wissenschaften ebenso wie in der Wirtschaft und Technik, in der Politik, im Sport und in vielen kulturellen Bereichen; die Ergebnisse statistischer Erhebungen begegnen uns nicht nur in Fachzeitschriften, sondern fast täglich auch in Zeitungen und Journalen, im Radio und im Fernsehen, in politischen Argumentationen wie in Festtagsreden, auf Werbeplakaten wie auch in amtlichen Formularen oder Bescheiden. Die Bedeutung der Statistik in vielen beruflichen und außerberuflichen Lebenssituationen (auch von Schülern/innen) ist ein Argument, das für eine unterrichtliche Behandlung dieses mathematischen Gebiets spricht.

Manchem/r Lehrer/in erscheint die *beschreibende Statistik* jedoch als *zu einfach und elementar*, um im Mathematikunterricht (der Oberstufe) sinnvoll behandelt werden zu können. Tatsächlich beschränken sich die rechentechnischen Anforderungen im Wesentlichen auf das Rechnen mit Prozenten, Brüchen, Quadraten, Wurzeln, Betrags- und Summenzeichen, allenfalls trifft man in der zweidimensionalen Statistik bei der Herleitung der Regressionsgleichung auf „höhere Mathematik".
Andere hingegen sind der Auffassung, dass die *beschreibende Statistik* sehr hohe Anforderungen hinsichtlich Kontextverständnis, Darstellung und Interpretation stelle und aus diesem Grund für den Schulunterricht *eher zu anspruchsvoll* sei. Tatsächlich ist die beschreibende Statistik allein innermathematisch kaum angemessen behandelbar, außermathematische Sachverhalte erfordern aber nicht selten weiterreichendes Kontextwissen und -verständnis (und ein gewisses Maß an Motivation, sich mit solchen Problemstellungen zu befassen) sowie einen kritischen Umgang mit Darstellungen und Interpretationen der kontextbezogenen Daten.
Beide Standpunkte führen zur selben Konsequenz: einer in vielen Fällen eher stiefmütterlichen und oft auch recht unangemessenen unterrichtlichen Behandlung der beschreibenden Statistik.

Folgt man der Sichtweise von R. FISCHER und G. MALLE (1985, S. 221), wonach sich Mathematik „im Wechselspiel zwischen Darstellen, Interpretieren und Operieren" vollzieht, wie auch ihrem (vielfach bestätigten) Befund, dass das Operative den traditionellen Mathematikunterricht so stark beherrscht, dass (zu) wenig Raum für das Darstellen und Interpretieren bleibt, dann kommt zur o. a. Praxisrelevanz ein weiteres Argument für die unterrichtliche Behandlung der beschreibende Statistik hinzu: die Chance, dem Darstellungs- und Interpretationsaspekts im Mathematikunterricht einen angemesseneren Stellenwert einzuräumen.

In diesem Beitrag sollen zentrale Tätigkeiten, lokale Bedeutungen und globale Ideen der beschreibenden Statistik herausgearbeitet und damit ein didaktischer Rahmen („roter Faden") zur Planung, Entwicklung, Durchführung und Bewertung entsprechender curricularer Konzepte angeboten werden.

2. Beschreibende Statistik im Spannungsfeld zweier globaler Ideen

Für die beschreibende Statistik lassen sich mindestens zwei globale („fundamentale") Ideen identifizieren, die die Entwicklung dieses Gebiets nachhaltig beeinflusst haben und ihr heutiges Erscheinungsbild prägen: Die *Idee der Mustererkennung und -darstellung* sowie die *Idee der Raffung*.

Viele statistische Tätigkeiten und Arbeitsweisen (insbesondere diverse Verfahren der grafischen und halbgrafischen Darstellung) zielen darauf ab, in großen Datensätzen explorativ **globale Muster** (Strukturen, Verläufe, Beziehungen) zu **erkennen**, diese **sichtbar** zu **machen** und zu **beschreiben** (darzustellen). In Abb. 1 etwa wird es wohl darum gehen, die starken saisonalen Schwankungen in der Energieaufbringung Klagenfurts (bei leicht steigender Gesamtentwicklung) zu erkennen bzw. deutlich zu machen, Abb. 2 macht die starken Geburtenrückgänge seit 1970 wie jene der (Nach-)Kriegsjahre ebenso erkenn- und sichtbar wie den Frauenüberschuss in den älteren Jahrgängen u. v. a. m.

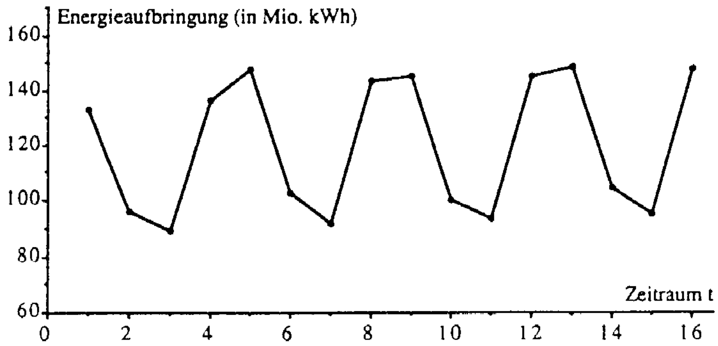


Abb. 1: Energieaufbringung in Klagenfurt 1990 - 1993

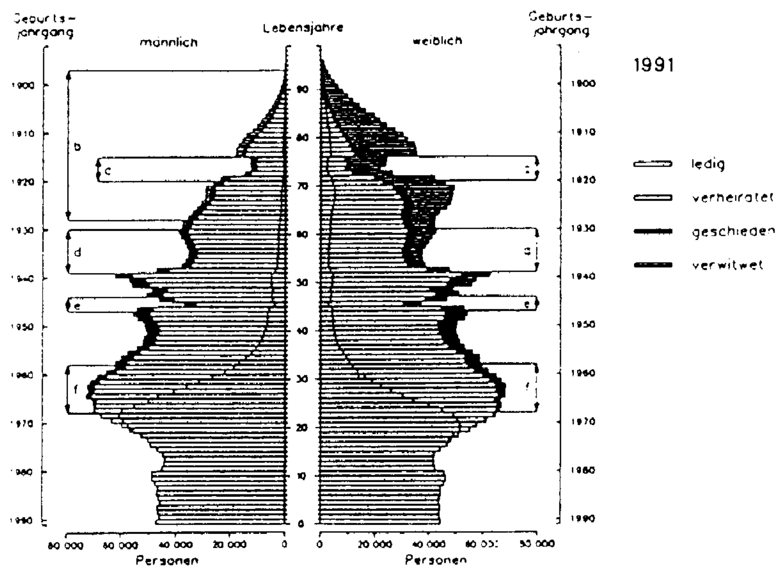


Abb. 2: Alterspyramide für Österreich

Der **Raffung** als weiterer globaler Idee der beschreibenden Statistik liegt eine grundlegende menschliche und gesellschaftlich höchst bedeutsame Denkstrategie zugrunde: Wer umfangreiche, komplexe Informationen (Datensätze) vergleichen will oder muss, wer angesichts unübersichtlicher Informationslage Entscheidungskriterien sucht, kann versuchen, sich durch operative Komprimierung der Daten Überblick, Vergleichsmöglichkeiten oder intersubjektiv nachvollziehbare Entscheidungsgrundlagen zu schaffen: Addition und deren Kommutativität ermöglichen es uns, die in einem Restaurant konsumierten Speisen und Getränke mit einem Gesamtbetrag zu bewerten und zu bezahlen, die verschiedenen Noten im Verlauf eines Schuljahres werden zu einer Endnote (mit entsprechenden Konsequenzen) verdichtet, das quantitative Verhältnis zwischen Budgetdefizit (selbst Produkt einer extremen Raffung) und Bruttonationalprodukt (ebenfalls eine extreme Raffung) macht uns für den Euro (un-)tauglich usw. Erst solche Raffungen machen die Mathematik - in unserem Fall die Statistik - so „einfach, klar und undurchsichtig“, wie dies R. FISCHER (1986, S. 123) einmal formuliert hat. In der beschreibenden Statistik finden wir die Idee der Raffung fast durchgängig, beginnend bei der Datenerhebung (Zusammenfassung in Gruppen) über die Klasseneinteilung und das Histogramm bis hin zu - und ganz besonders bei - den Zentral- und Streuungsmaßen; der damit erreichbaren Vereinfachung und Übersichtlichkeit steht das Problem der Reduktion und geringeren Transparenz gegenüber.

Wesentlichstes Anliegen der Raffung ist *Gewinn an Übersicht*. Damit unterstützt die Raffung einerseits die Mustererkennung (globaler Überblick über Verläufe, Trends, globale Auffälligkeiten), läuft ihr andrerseits aber auch zuwider, denn der Raffung ist im Prozess der Komprimierung ja praktisch keine Grenze gesetzt: die Raffung hat ihr Ziel dann am besten erreicht, wenn es gelingt, alle Daten auf ein zentrales Moment, eine einzige Kennzahl, ein einziges grafisches Symbol zu reduzieren. Eine solche Verdichtung der Information ist jedoch i. A. mit *Verlust an Information* verbunden, d. h., Gewinn an Übersicht bedeutet meist auch Verlust an Information. Abb. 3 soll diesen Zusammenhang veranschaulichen:

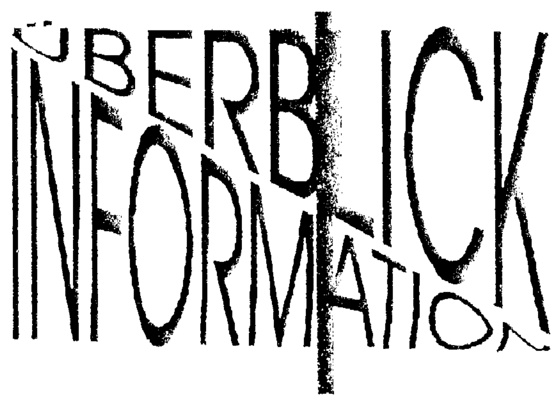


Abb. 3: Überblick vs. Informationsgehalt bei der Raffung

In der Praxis sollte das subjektive und zweckabhängige Bedürfnis nach Mustererkennung dem Prozess der Raffung dort Einhalt gebieten, wo ausreichend Überblick erreicht ist, die wesentlichen (Einzel-)Informationen und Strukturen aber noch erkennbar sind (vgl. dazu Abschnitt 3, Abb. 7 - 9 sowie Abschnitt 4).

3. Zentrale Tätigkeiten und lokale Deutungen in der eindimensionalen beschreibenden Statistik

Neben einer geeigneten *Datenerhebung* (mit ihrer ganzen Problematik), auf die hier nicht näher eingegangen wird, gehören die *Darstellung der Daten* sowie die *Ermittlung von statistischen Kennzahlen* (Zentral- und Streuungsmaße, Schiefe u. a.) zu den zentralen Tätigkeiten der eindimensionalen beschreibenden Statistik.

Die **Darstellung statistischer Daten** lässt sich in Form von *Tabellen* meist besonders einfach und platzsparend realisieren. Andrerseits sind Tabellen manchmal aber auch etwas unübersichtlich, man muss sich oft erst „einlesen" und man kann Größenvergleiche nur über die angegebenen Zahlenwerte durchführen. *Grafische Darstellungen* hingegen zielen darauf ab, einen raschen Überblick zu ermöglichen, Größenverhältnisse und Muster rasch zu erkennen. Im Vergleich zu tabellarischen Darstellungen sind sie meist aufwendiger zu erstellen, benötigen mehr Platz und sind anfälliger für Missverständnisse und gegenüber Fehlinterpretationen.
Bei der Behandlung der Grundtypen grafischer Darstellungen (vor allem Stab-, Kreis- und Streifendiagramm, Piktogramm, Liniendiagramm) sind deren Vor- und Nachteile (lokale Bedeutungen) mitzubedenken: Stabdiagramme etwa ermöglichen sehr viel einfacher als Kreisdiagramme sichere Größenvergleiche zwischen zwei Werten, letztere wiederum stellen den Anteil der einzelnen Werte an der Gesamtheit dar und begünstigen so u. a. den Vergleich zwischen zwei Gruppen von Werten:

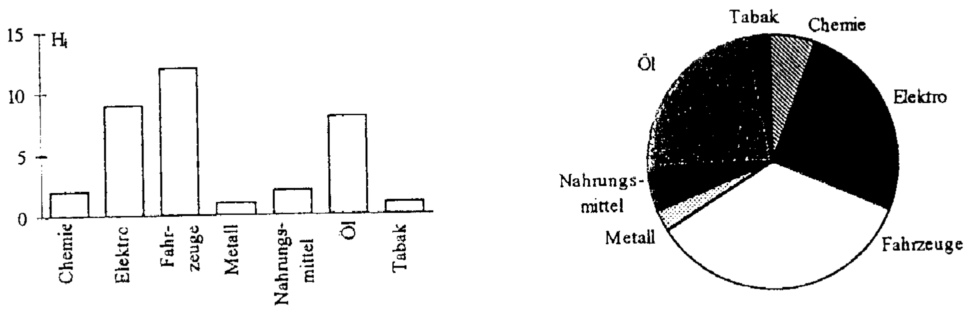


Abb 4: Häufigkeitsverteilung der Variablen „Branche" für 35 Industrieunternehmen als Stab- und als Kreisdiagramm

Aus dem Stabdiadramm in Abb. 4 erkennt man schneller als aus dem Kreisdiagramm, dass die Anzahl der Elektrounternehmen größer ist als die Anzahl der Ölunternehmen, aus dem Kreisdiagramm wiederum ist sehr viel einfacher zu erkennen, dass der Anteil der Elektro-, Öl- und Nahrungsmittelunternehmen zusammen größer ist als der aller anderen Unternehmen.
Die Liniendiagramme in Abb. 5 und Abb. 6 stellen thematisch sehr ähnliche Zeitreihen dar - und sind doch sehr unterschiedlich zu interpretieren: Während das Liniendiagramm in Abb. 5 Bestandsdaten darstellt, die zu bestimmten Zeitpunkten erhoben wurden, handelt es sich bei den in Abb. 6 dargestellten Daten um Bewegungsdaten, die jeweils für einen bestimmten Zeitraum ermittelt wurden. Das hat Konsequenzen: Während die Verbindungslinien in Abb. 5 einigermaßen sinnvoll interpretiert werden können (sofern man eine gleichmäßige Entwicklung zwischen den gemessenen Werten unterstellt), ist eine entsprechende Interpretation der Verbindungslinien in Abb. 6 nicht möglich (die Darstellung dieser Zeitreihe in einem Stabdiagramm würde allfälligen Fehlinterpretationen besser vorbeugen).

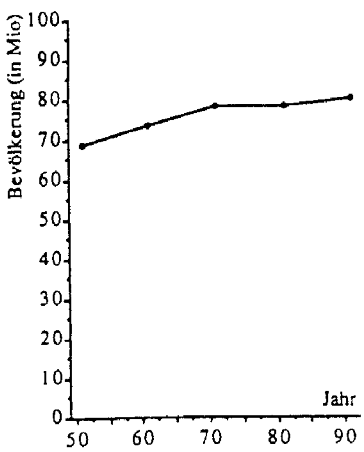


Abb. 5: Bevölkerungsentwicklung in Deutschland

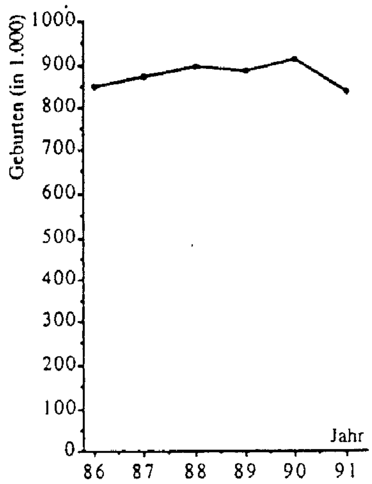


Abb. 6: Entwicklung der Geburtenzahlen in Deutschland

Das *Wechselspiel von Raffung und Mustererkennung* wird gerade auch bei grafischen Darstellungen recht deutlich: Die Darstellung der Beschäftigtenzahlen in Abb. 7 ist unübersichtlich und lässt für die Häufigkeitsverteilung wenig „Struktur" erkennen. Verzichtet man auf die Unterscheidung nahe beieinander liegender Werte und fasst benachbarte Werte zu einer Klasse zusammen (Raffung!), so erhält man z. B. die in Abb. 8 gezeigte, in mancher Hinsicht weit aufschlussreichere Darstellung.

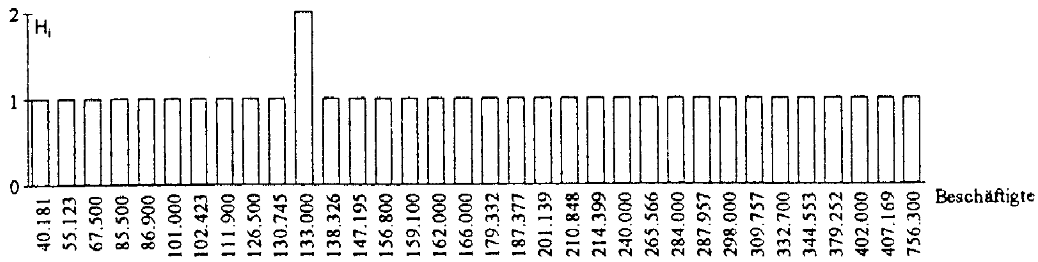


Abb. 7: Beschäftigtenzahlen in 35 Industrieunternehmen

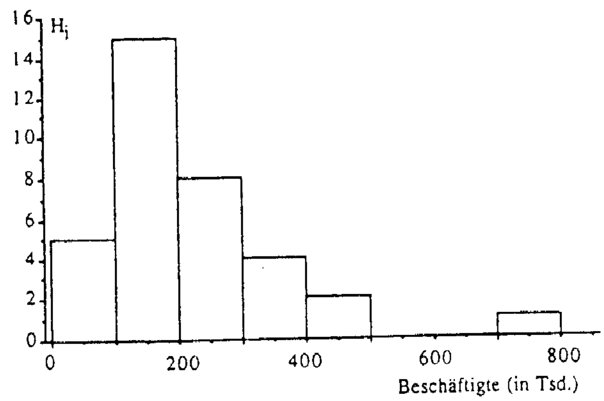


Abb. 8: Beschäftigtenzahlen in 35 Industrieunternehmen

Natürlich hätte man die Raffung auch sehr viel weiter treiben, die Klassenbreiten auf 200.000, 300.000 oder gar 800.000 vergrößern können. Das Ergebnis wäre dann zwar noch übersichtlicher als in Abb. 8, allerdings wäre der Informationsgehalt nur noch sehr gering. Insbesondere würde man durch eine so extreme Raffung fast alles an (Einzel-)Information verlieren und die „Struktur" der Verteilung völlig zerstören (vgl. Abb. 9).

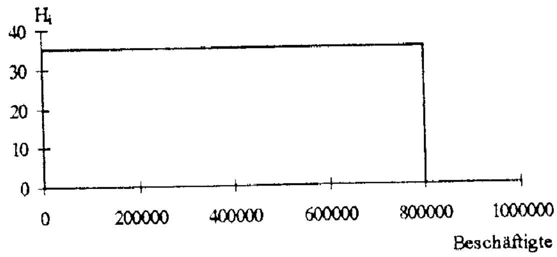


Abb. 9: Beschäftigtenzahlen in 35 Industrieunternehmen

Fast täglich stößt man auf Grafiken, die nicht sosehr statistische Daten als vielmehr ein grobes Unverständnis für die Zielsetzungen und die Bedeutung grafischer Darstellungen erkennen lassen; Abb. 10 und Abb. 11 mögen hier als - nicht weiter kommentierte - Beispiele unzweckmäßiger bis unsinniger grafischer Darstellungen statistischer Daten dienen.

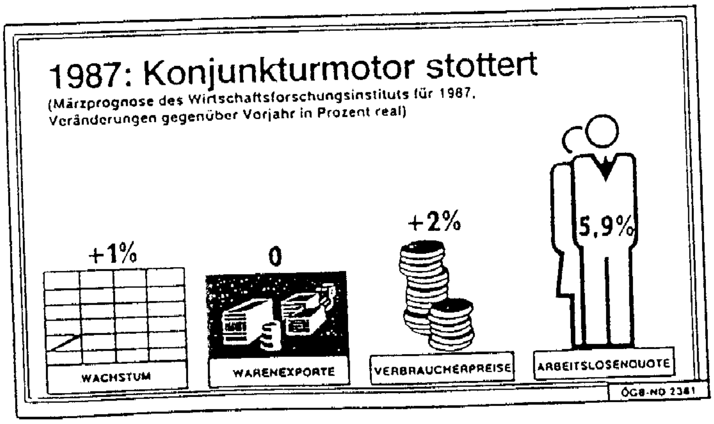


Abb. 10

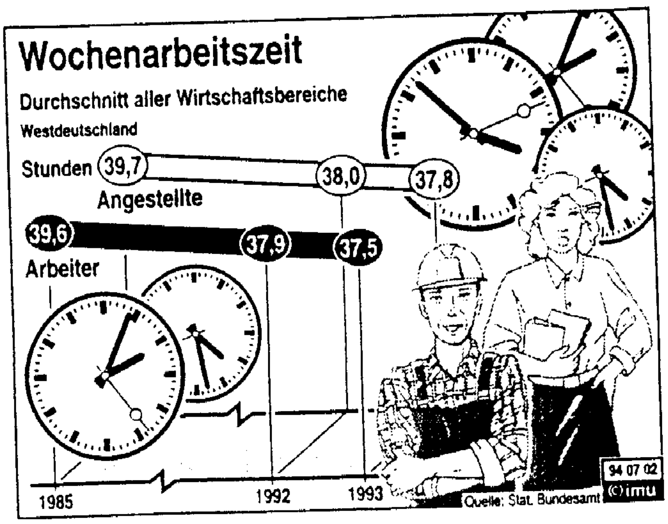


Abb. 11

Schon bei einfachen Grafiken ist man mit grundlegenden Problemen der grafischen Darstellung von statistischen Daten und verschiedenen Manipulationsmöglichkeiten konfrontiert. In den beiden folgenden Beispielen soll dies zumindest exemplarisch gezeigt werden, wobei in Beispiel 1 allein durch verschiedene Festlegungen der Achseneinheiten (gibt es eine „richtige“?), in Beispiel 2 mit verschiedenen Klasseneinteilungen und Bezugsbasen „manipuliert“ wird.

Beispiel 1: Besucher/innenzahlen im Strandbad Klagenfurt:

Jahr	1970	1972	1976	1978	1981
Besucher/innen	755.000	815.000	845.000	840.000	945.000

Opposition: *Es ist ein Skandal: Die Besucher/innenzahlen im Strandbad explodieren und die Stadtregierung weigert sich nach wie vor, finanzielle Mittel für den Ausbau des Strandbades bereitzustellen!*

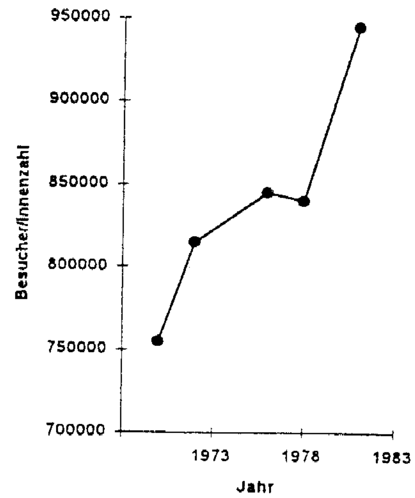


Abb. 12a

Stadtregierung: *Die Entwicklung der Besucher/innenzahlen im Strandbad ist uns wohlbekannt, wir können daraus aber keinen derartigen Zuwachs erkennen, der die Finanzierung einer Erweiterung des Strandbades rechtfertigen würde!*

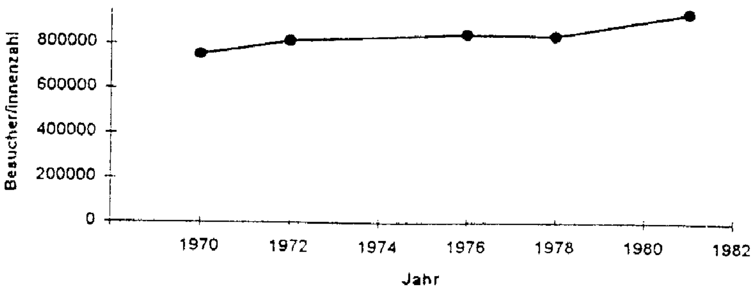


Abb. 12b

Beispiel 2: An einer politischen Diskussionsveranstaltung nahmen 140 Personen teil. Die Altersverteilung der Teilnehmer/innen an dieser Veranstaltung ist aus Abb. 13a - d ersichtlich:

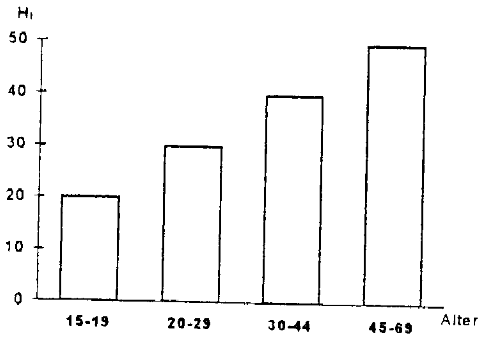


Abb. 13a: *Das Interesse an der Politik steigt mit zunehmendem Alter!*

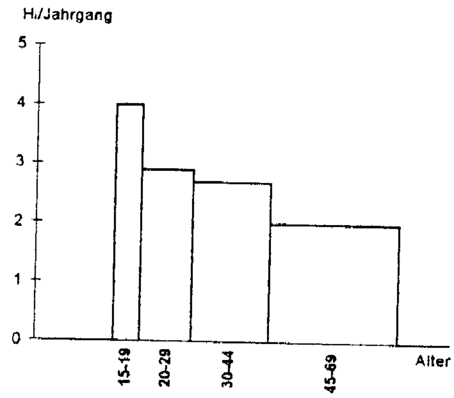
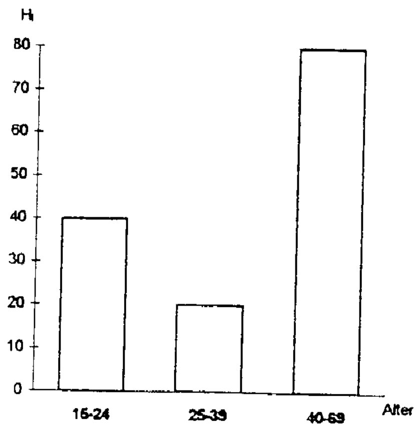


Abb. 13b: *Das Interesse an der Politik sinkt mit zunehmendem Alter!*



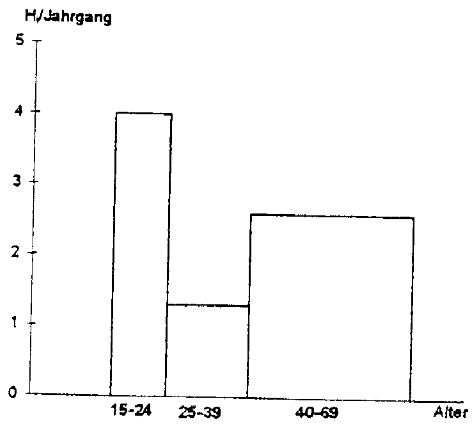


Abb. 13c: *Gleichgültig, ob man nach der in Abb. 13a oder nach der in Abb. 13b gewählten Methode vorgeht, in jedem Fall ist es die mittlere Generation, die sich am wenigsten für Politik interessiert! Das Interesse an der Politik sinkt mit zunehmendem Alter!*

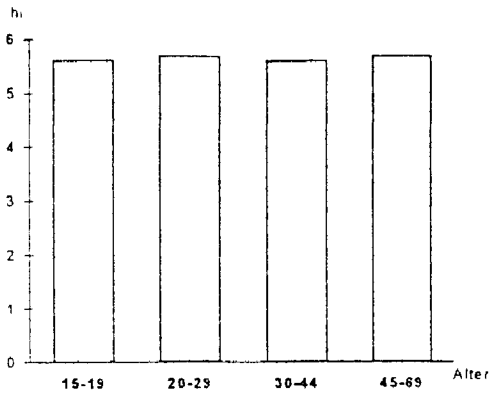


Abb. 13d: *Bezogen auf die Altersverteilung in dieser Kleinstadt zeigten alle Altersgruppen ungefähr gleich großes Interesse bzw. Desinteresse an dieser Veranstaltung!*

Mit **statistischen Kennzahlen** versucht man, bestimmte Aspekte statistischer Daten durch ein einziges Datum zu beschreiben.
So etwa ist das wesentliche Anliegen der *Zentralmaße*, ihre lokale Bedeutung, darin zu sehen, dass eine ganze (oft unübersichtliche) Liste von Daten durch einen einzigen, „repräsentativen", „durchschnittlichen" oder „typischen" Wert ersetzt wird. Dieser Wert soll die gesamte Liste der Daten möglichst gut charakterisieren und so auch verschiedene Listen von Daten vergleichbar machen. Natürlich stellt eine derartige Komprimierung vieler Daten in ein einziges Zentralmaß eine sehr *extreme Raffung* dar, bei der viele (oft sehr wichtige) Einzelinformationen verlorengehen; die beschreibende Statistik kennt für diese Art der Raffung verschiedene Modelle (diverse Mittelwerte, Median, Modus). Didaktisch und im Hinblick auf Anwendungen sind dabei weniger die jeweiligen Berechnungsmethoden von Interesse als vielmehr die lokalen Bedeutungen der einzelnen Modelle, ihre Aussagekraft und Anwendbarkeit. Wenn ein Statistiker etwa sagt, dass die 100 Mitarbeiter/innen eines Unternehmens „durchschnittlich" ATS 30.000,-- pro Monat verdienen, so kann er damit meinen, dass bei gleichmäßiger Verteilung der Summe aller Gehälter auf alle Mitarbeiter/innen ein/e Mitarbeiter/in ungefähr ATS 30.000,-- monatlich bekommen würde (*arithmetisches Mittel*), er kann damit aber auch meinen, dass 50 Mitarbeiter/innen *höchstens* ATS 30.000,--, 50 Mitarbeiter/innen hingegen *mindestens* ATS 30.000,-- pro Monat verdienen (*Median*). Denkbar wäre weiters auch die Interpretation, dass das am häufigsten ausbezahlte Entgelt ATS 30.000,-- beträgt (*Modus*). Das Konzept des arithmetischen Mittels ist das der Summenbildung (daher muss diese möglich und sinnvoll sein!) und der gleichmäßigen Verteilung - in diesem Sinne ist es „egalitär", es reagiert stark auf extreme Werte („Ausreißer") und setzt metrische Daten voraus. Das Konzept des Medians besteht darin, zwischen einer „besseren" und einer „schlechteren" Hälfte zu unterscheiden, es ist in diesem Sinn „elitär", darüber hinaus ausreißerunempfindlich (robust) und verlangt nur Ordinaldaten. Das Konzept des Modus schließlich sucht nach einem häufigsten Wert, der aber wohl nur dann als „typischer" Wert akzeptabel erscheint, wenn er sehr viel häufiger auftritt als alle anderen Werte; der Modus ist auch auf nominalskalierte Daten anwendbar.
So simpel die Zentralmaße auf den ersten Blick erscheinen mögen, wenn man bedenkt, dass selbst gesetzliche Bestimmungen (etwa Stipendiengesetze, Studienzulassungen) die Verwendung des arithmetischen Mittels für die Mittelung von (üblicherweise ordinalskalierten) Noten

zwingend vorschreiben und auch stoffdidaktische Arbeiten mit diesem Thema gewisse Schwierigkeiten haben (siehe etwa Reichel, Hanisch und Müller 1987, S. 31), dann kann man erahnen, dass die Interpretation der lokalen Bedeutungen mancher Zentralmaße schwieriger, zugleich aber auch interessanter ist, als man zunächst angenommen hatte.
Durch die extreme Raffung, die zu den Zentralmaßen führt, gehen viele Informationen verloren, unter anderem auch Informationen darüber, wie gut das Zentralmaß die vorliegenden Daten repräsentiert bzw. wie stark die Einzeldaten um das Zentralmaß *streuen*. (Würden Sie etwa als Nichtschwimmer einen Teich durchwaten, von dem Sie lediglich wissen, dass er „durchschnittlich" 80 cm tief ist?) Im Hinblick auf solche Beurteilungen ist die Angabe eines Zentralmaßes alleine völlig unzureichend, die Raffung zu extrem. Die beschreibende Statistik versucht, diesen Informationsverlust durch Angabe eines *Streuungsmaßes* etwas zu mildern. *Spannweite* und *Interquartilspannweite* (Abstand zwischen *oberer* und *unterer Quartile*) sind sehr einfache, anschauliche Streuungsmaße, die für die Anliegen der beschreibenden Statistik meist völlig genügen, die *mittlere absolute Abweichung* (vom Median oder vom arithmetischen Mittel) als arithmetisches Mittel der Abstände („durchschnittlicher Abstand") modelliert recht gut die intuitive Vorstellung von „mittlerer Streuung". Das in der beschreibenden Statistik am häufigsten verwendete Streuungsmaß hingegen, die *empirische Standardabweichung*, erweist sich als unanschaulich und schwer interpretierbar (Seitenlänge des „durchschnittlichen Abstandsquadrats", genauer: Seitenlänge jenes Quadrats, dessen Flächeninhalt das arithmetische Mittel der Flächeninhalte all jener Quadrate repräsentiert, die man erhält, wenn man den Abstand jedes einzelnen Werts vom arithmetischen Mittel als Seite eines Quadrats versteht). Mit dieser, von intuitiven Vorstellungen sehr abgehobenen Modellierung der Streuung läuft die empirische Standardabweichung den Anliegen der beschreibenden Statistik (Gewinn an Einsicht und Überblick) geradezu zuwider und erscheint daher für die beschreibende Statistik weniger geeignet; ihre eigentliche Bedeutung kommt erst in der Wahrscheinlichkeitsrechnung und in der schließenden Statistik zum Tragen.

4. Zusammenspiel von Raffung und Mustererkennung - ein Beispiel

Die folgende Tabelle zeigt die Bruttoeinkommen (in ATS) von 50 Beschäftigten in 25 verschiedenen Wirtschaftsbereichen:

11.211	15.357	21.099	32.608	18.004	26.365	13.757	21.896	13.677	20.118
12.338	17.571	11.841	15.590	14.040	18.082	15.408	25.653	17.399	28.316
17.064	24.891	16.277	23.393	16.936	23.539	15.500	21.067	13.608	20.560
12.560	14.637	15.106	18.910	20.693	31.063	15.535	20.700	10.679	14.964
14.429	19.499	15.715	22.006	13.570	19.654	17.167	20.934	7.221	6.017

Das *Kastenschaubild* in Abb. 14 liefert für diese 50 Einkommen eine grafische Veranschaulichung von Median, unterer und oberer Quartile (und damit auch des Interquartilabstands) sowie des höchsten und des niedrigstens Wertes (und damit der Spannweite): Das höchste Einkommen liegt bei ca. ATS 32.500,--, das niedrigste bei ca. ATS 6.000,--, die (mittlere) Hälfte der Beschäftigten verdient monatlich zwischen ca. ATS 14.500,-- und knapp ATS 20.500,--, der „Durchschnittsverdiener" (im Sinne des Medians) kommt auf monatlich ca. ATS 17.000,--. Es sind zwei extrem niedrige und drei vergleichsweise sehr hohe Einkommen erkennbar, zwischen ca. ATS 10.500,-- und ca. ATS 26.500,-- darf man annehmen, dass die Daten „dicht" liegen (keine großen Lücken auftreten), wobei die Verteilung etwas „rechtsschief" ist (d. h., die Daten streuen bei den oberen Einkommen etwas stärker als in der unteren Hälfte).

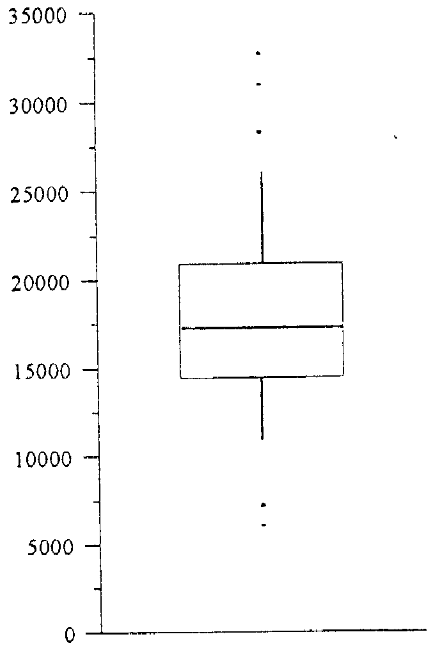


Abb. 14: Bruttoeinkommen von 50 Beschäftigten (in ATS)

Aus diesen stark gerafften Daten (Zentral- und Streuungsmaße) lässt sich also schon sehr viel über die Verteilung der Einkommen aussagen, allerdings ist aus den dargestellten Kennzahlen (Raffung!) die Struktur der Verteilung nicht mehr erkennbar. Diese oft wichtige Information kann z. B. recht gut einem *Stengel-Blatt-Diagramm* (halbgrafische Darstellung) entnommen werden; im vorliegenden Beispiel könnte anhand des Stengel-Blatt-Diagramms in Abb. 15a etwa auffallen, dass sich die Einkommen um ATS 15.000,-- wie auch bei ca. ATS 20.000,-- häufen, somit eher eine zweigipfelige Verteilung vorliegt als eine Gauß'sche Glockenkurve. Mehrgipfelige Verteilungen sind aber immer alamierend, sie verweisen auf eine mögliche Inhomogenität der zugrundeliegenden Daten. So könnten etwa die zwei Gipfeln im Stengel-Blatt-Diagramm von Abb. 15a durch Einkommensunterschiede zwischen Arbeitern/innen und Angestellten bedingt sein, oder auch durch Einkommensunterschiede zwischen Frauen und Männern. Im Stengel-Blatt-Diagramm von Abb. 15b wurden die Einkommen von Frauen (fett, kursiv) und Männern verschieden gekennzeichnet, das Ergebnis ist recht aufschlussreich!
Eine Konsequenz aus dem in Abb. 15a erkannten Muster (und dem in Abb. 15b erkannten, möglichen Ursachen) könnte eine für die Einkommen von Frauen und Männern getrennte (aber vergleichende) Darstellung relevanter Kennzahlen in zwei Kastenschaubildern sein (Abb. 16).

Tausend	Hundert
6	0
7	2
8	
9	
10	6
11	28
12	35
13	5667
14	0469
15	1345557
16	29
17	0135
18	009
19	46
20	05679
21	008
22	0
23	35
24	8
25	6
26	3
27	
28	3
29	
30	
31	0
32	6

Abb. 15a: Bruttoeinkommen (in ATS)

Tausend	Hundert
6	0
7	*2*
8	
9	
10	*6*
11	*28*
12	*35*
13	*5667*
14	*04*69
15	*1345557*
16	*29*
17	*013*5
18	*0*09
19	46
20	05*6*79
21	*0*08
22	0
23	35
24	8
25	6
26	3
27	
28	3
29	
30	
31	0
32	6

kursiv: Frauen

Abb. 15b: Bruttoeinkommen (in ATS)

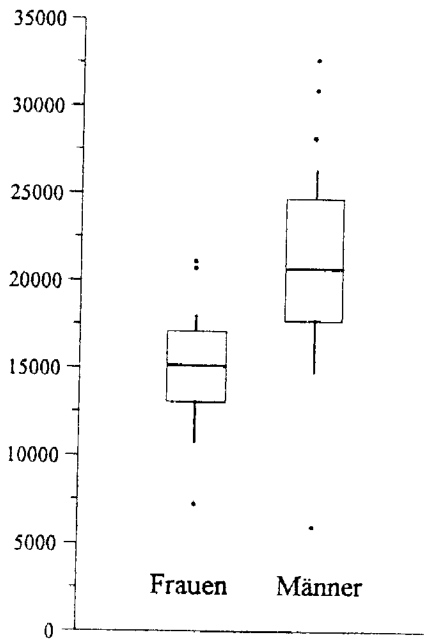


Abb. 16: Bruttoeinkommen von 25 Frauen und 25 Männern

Bei den in diesem Beispiel verwendeten Daten handelt es sich im übrigen um die Durchschnittseinkommen (im Sinne des Medians) aller weiblichen und männlichen Beschäftigten in 25 Wirtschaftsbereichen in Österreich im Jahre 1994.

5. Zweidimensionale beschreibende Statistik

Bei der eindimensionalen Betrachtung von Merkmalen gehen Zusammenhänge zwischen zwei (oder mehreren) Merkmalen verloren, zum Beispiel mögliche Zusammenhänge zwischen den jährlichen Umsätzen und den Beschäftigtenzahlen in verschiedenen Unternehmen. Methoden zur Beschreibung von Zusammenhängen zwischen zwei Merkmalen liefert die zweidimensionale beschreibende Statistik. Dabei werden dieselben globalen Ideen und ähnliche lokale Bedeutungen bzw. zentrale Tätigkeiten wie in der eindimensionalen Statistik sichtbar:
Zunächst wird man die Ausprägungen der beiden Merkmale wohl tabellieren. Falls für beide Merkmale metrische Daten vorliegen, wird man meist auch versuchen, diese in einem *Streudiagramm* grafisch darzustellen (Abb. 17).

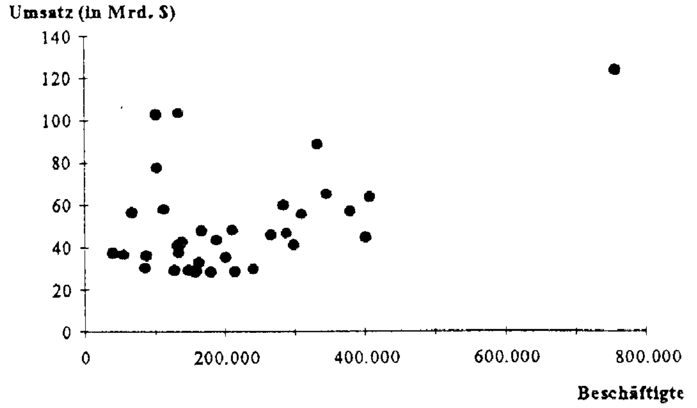


Abb. 17: Beschäftigtenzahlen und Umsätze in 35 Industrieunternehmen

Anhand des Streudiagramms bzw. anhand der durch die eingezeichneten Punkte gegebenen *Punktwolke* lässt sich ein gewisser (hier nicht sonderlich stark ausgeprägter) Trend dahingehend erkennen, dass Unternehmen mit höheren Beschäftigtenzahlen *tendenziell* auch höhere Umsätze aufweisen. Dieser Trend in Abb. 17 lässt sich verdeutlichen und in den Vordergrund rücken, wenn man ihn als Graph einer (im einfachsten Fall linearen) Funktion modelliert und diesen Graphen im Streudiagramm einzeichnet. Die in Abb. 18 dargestellte *Trendgerade* wurde „nach Gefühl" eingezeichnet; sie ist für die vorliegenden zweidimensionalen Daten in ähnlichem Sinn repräsentativ wie ein Zentralmaß für eine eindimensionale Liste von Daten.

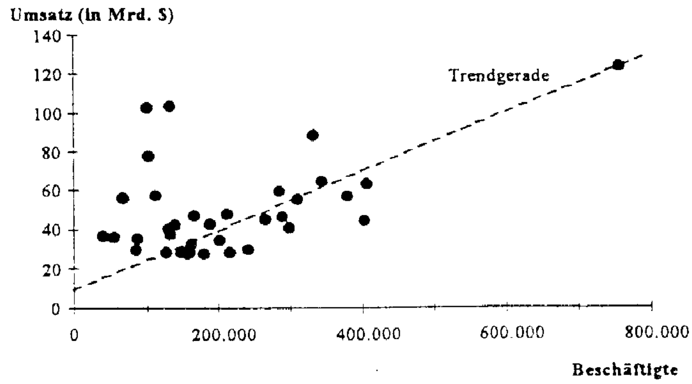


Abb. 18: Beschäftigtenzahlen und Umsätze in 35 Industrieunternehmen

Aus der Grafik lässt sich leicht eine Gleichung der Trendgeraden ermitteln, aus der dann entsprechende quantitative Aussagen ablesbar bzw. berechenbar werden.

Gefühle sind subjektiv und nicht normiert, daher kann es für die nach Gefühl gezeichnete Trendgerade kein eindeutiges Ergebnis geben. Wer ein intersubjektiv eindeutiges Ergebnis möchte oder braucht, wird den Trend daher nach einem *genormten Verfahren*, etwa der *Methode der kleinsten Quadrate*, ermitteln und zum Beispiel durch eine *Regressionsgerade* beschreiben. Je nachdem, ob man dabei die Abstandsquadrate der Punkte von der Regressionsgeraden in vertikaler Richtung (*1. Regressionsgerade*) oder in horizontaler Richtung (*2. Regressionsgerade*) minimiert, erhält man eine Gerade, die gegenüber der nach Gefühl gezeichneten Trendgeraden zu flach oder zu steil erscheint - in jedem Fall also eine Gerade, die der intuitiven (auf Normalabstände ausgerichteten) Vorstellung von einer gut passenden Geraden nicht optimal entspricht (Abb. 19). Der Grund für diese Unzulänglichkeit liegt im Verfahren - darauf soll hier jedoch nicht näher eingegangen werden.

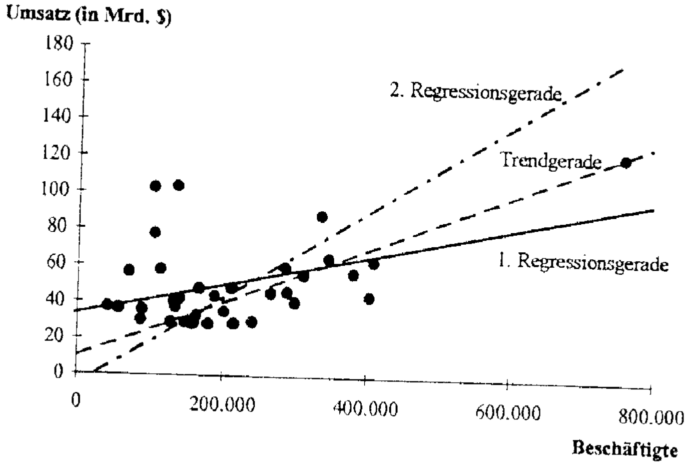


Abb. 19: Beschäftigtenzahlen und Umsätze in 35 Industrieunternehmen

Die beiden Regressionsgeraden lassen sich für jede Punktwolke ermitteln, also auch für solche, bei denen kaum ein linearer Trend erkennbar ist. Ähnlich wie bei den Zentralmaßen gilt auch hier: Bei entsprechend großer Streuung um die Trendgerade ist diese wenig aussagekräftig und kaum geeignet, die vorliegenden Daten in geeigneter Weise zu repräsentieren. Ein (auf Werte zwischen -1 und 1 genormtes) Maß für diese Streuung ist der *Pearsonsche Korrelationskoeffizient*, dessen Formel sehr komplex und undurchsichtig erscheint, der sich aber als Quadratwurzel aus dem Verhältnis der Steigungen der beiden Regressionsgeraden (oder auch als geometrisches Mittel der Anstiege) doch recht anschaulich erklären und anhand von Streudiagramm und Regressionsgeraden zumindest intuitiv auch recht gut nachvollziehen lässt. Der Korrelationskoeffizient ist als ein Maß für die Repräsentativität der Regressionsgeraden zu verstehen; es kommt ihm damit in der zweidimensionalen Statistik eine ähnliche Funktion zu, wie den Streuungsmaßen in der eindimensionalen beschreibenden Statistik: Die Repräsentativität und Aussagekraft der Regressionsfunktion ist umso geringer, je näher der Wert des Korrelationskoeffizienten bei null liegt (d. h., je stärker die Daten um die Regressionsgeraden streuen bzw. je weiter die Regressionsgeraden auseinanderklaffen).

Auch Trend bzw. Regression und Korrelation basieren wesentlich auf der Idee der (extremen) Raffung: Ein großer, zweidimensionaler Datensatz wird durch eine einzige (meist lineare) Funktion repräsentiert und ersetzt, die Streuung (Repräsentativität), wird durch einen einzigen Koeffizienten beschrieben. Dabei gehen natürlich Informationen verloren, Strukturen und Muster, die im Streudiagramm vielleicht noch erkennbar sind, verschwinden in der Regressions- und Korrelationsanalyse. So etwa könnte der nur mäßige Zusammenhang, die stark streuenden Werte in Abb. 17 - 19 darauf zurückzuführen sein, dass die Zusammenhänge zwischen Beschäftigtenzahlen und Umsätzen in unterschiedlichen Branchen sehr unterschiedlichen Trends folgen (vgl. die *Clusterung* nach Branchen in Abb. 20), die dann - wie die Einkommen von Männern und Frauen in Abb. 16 - besser getrennt untersucht werden sollten.

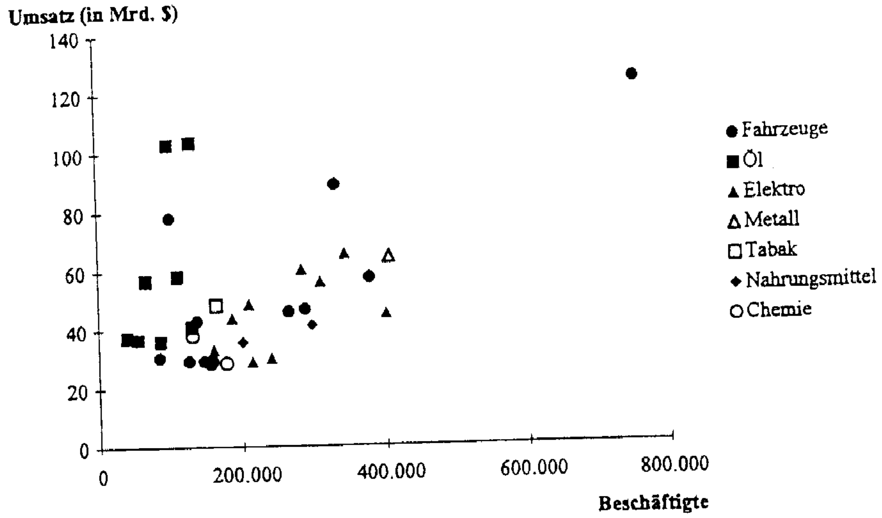


Abb. 20: Beschäftigtenzahlen und Umsätze in 35 Industrieunternehmen

Streudiagramm, Regressions- und Korrelationsanalyse verlangen metrische Daten. Bei der Untersuchung von Zusammenhängen zwischen qualitativen Merkmalen wird man sich auf die Darstellung in Tabellen und die Ermittlung eines *Kontingenzkoeffizienten* (der ähnlich wie der Pearsonsche Korrelationskoeffizient zu interpretieren ist) beschränken müssen.

Bei der Interpretation von Trend, Regression, Korrelation bzw. Kontingenz ist Vorsicht geboten: Zum einen sind Extrapolationen über den vorliegenden Datensatz hinaus jedenfalls problematisch (z. B. längerfristige Prognosen bei Zeitreihen), weiters sollten Verallgemeinerungen, die über die vorliegenden Daten hinausgehen („Hochrechnungen"), der schließenden Statistik vorbehalten bleiben und schließlich sind eventuell konstatierte Zusammenhänge zunächst nur formaler Art und ohne tiefere, kontextbezogene Einsichten in mögliche Ursachen und Wirkung nicht kausal interpretierbar: Ob die häufigen Absenzen mancher Schüler die Ursache für schlechte Noten sind oder die schlechten Noten Ursache für häufige Absenzen (oder ob der Zusammenhang doch viel komplexer ist), lässt sich anhand von Regression und Korrelation allein nicht beurteilen, eine kausale Interpretation der hohen Korrelation zwischen dem Verschwinden der Störche und dem drastischen Geburtenrückgang in einer schwedischen Kleinstadt dürfte bei schulpflichtigen Kindern doch berechtigte Zweifel auslösen.

6. Resumé

In der folgenden Tabelle sind die globalen Ideen, lokalen Bedeutungen und zentralen Tätigkeiten der beschreibenden Statistik noch einmal übersichtlich zusammengestellt:

BESCHREIBENDE STATISTIK Ordnen und Darstellen von vorhandenen (erhobenen) Daten, Charakterisierung der verfügbaren Daten und Aussagen darüber.		
Globale Ideen	**Lokale Bedeutungen**	**Zentrale Tätigkeiten**
	Möglichkeiten, Grenzen und Probleme der Datenerhebung; Datentypen	Konkretisieren des Problems; Erheben von Daten
MUSTER ERKENNEN und SICHTBAR MACHEN	Aussagekraft tabellarischer, halbgrafischer oder grafischer Darstellungen	Daten ordnen und (tabellarisch oder grafisch) übersichtlich darstellen
RAFFUNG	Deutung diverser Zentral- und Streuungsmaße, von Trend (Regression) und Korrelation bzw. Kontingenz	Kennzahlen ermitteln

Literatur

FISCHER, R. (1986): Zum Verhältnis von Mathematik und Kommunikation. In: mathematica didactica 9, S. 119-131.

FISCHER, R. und MALLE, G. (1989): Mensch und Mathematik. Eine Einführung in didaktisches Denken und Handeln. B. I. Wissenschaftsverlag, Mannheim-Wien-Zürich.

KRÖPFL, B., PESCHEK, W., SCHNEIDER, E. und SCHÖNLIEB, A. (1994): Angewandte Statistik. Eine Einführung für Wirtschaftswissenschaftler und Informatiker. Hanser Verlag, München-Wien.

KRONFELLNER, M. und PESCHEK, W. (1998): Angewandte Mathematik 4. Verlag Hölder-Pichler-Tempsky, Wien.

REICHEL, H.-C., HANISCH, G. und MÜLLER, R. (1987): Wahrscheinlichkeitsrechnung und Statistik. Verlag Hölder-Pichler-Tempsky, Wien.