Statistische Beurteilung in der Schule - was und wie

Manfred Borovcnik, Klagenfurt

Kurzfassung: Die Beurteilende Statistik beschäftigt sich mit Methoden der Verallgemeinerung von unvollständiger Information auf größere Gesamtheiten und deren Rechtfertigung. Die Fragestellung wird ausführlich dargestellt. Dabei wird die Bedeutung der zufälligen Auswahl hervorgekehrt. Für die Lösung der Verallgemeinerung gibt es ganz unterschiedliche Ansätze; auf die Schulen der klassischen Statistiker, der Bayesianer und der Exploratorischen Datenanalyse wird näher eingegangen. Es wird gezeigt, worin sich die Ansätze unterscheiden und worin ihre relativen Meriten und Nachteile liegen. Nicht-parametrische Methoden siedeln innerhalb der klassischen Methoden an, versuchen aber den Gehalt an Voraussetzungen zu minimieren und sind deshalb universeller einsetzbar. Insbesondere die in jüngerer Zeit entwickelten Resampling-Verfahren werden die Statistik der Zukunft völlig umgestalten. Die Erörterungen bilden einen breiteren Hintergrund, vor dem curriculare Entscheidungen zu treffen sind.

1 Grundlegendes

In diesem Abschnitt wird das grundsätzliche Problem dargestellt, das in der Beurteilenden Statistik behandelt wird. Es geht um die Verallgemeinerung von eingeschränkter Information. Dabei spielt eine sehr wichtige Rolle, wie man zur Information, i.e. den Daten, gekommen ist. Nur wenn man über den Prozeß, der die Daten "erzeugt", genug weiß, insbesondere, daß er den Regeln des Zufalls entspricht, kann man die Verallgemeinerung von Information durch statistische Schlüsse rechtfertigen. Anschließend wird die Spezifizierung der Fragestellung auf zwei der wichtigsten Kategorien angesprochen, nämlich die Hochrechnung von Anteilen aus einer beobachteten Teilmenge auf die sogenannte Grundgesamtheit sowie die Schätzung von Mittelwerten einer Grundgesamtheit mit Hilfe von Mittelwerten aus Daten.

a) Das Problem

Die Problemstellung in der Beurteilenden Statistik unterscheidet sich grundsätzlich von der in der Wahrscheinlichkeitstheorie, wo aus bekannten Wahrscheinlichkeiten oder Verteilungen neue Wahrscheinlichkeiten berechnet werden. In der Beurteilenden Statistik werden diese Berechnungen wohl zu Hilfe genommen, jedoch wird der Realität nicht nur eine Verteilung als Modell aufgeprägt, sondern i.a. eine Klasse von Verteilungen. Mit Hilfe von Berechnungen aus der Wahrscheinlichkeitstheorie werden Szenarien für die Realität entwickelt. Je glaubwürdiger ein solches Szenarium, umso plausibler das dahinter stehende Modell, ist ein Schlüssel zur Bewertung dieser Modelle.

- *Stichproben und "Grundgesamtheit"*

Man verfügt lediglich über unvollständige Information,

- über eine (endliche) Grundgesamtheit, man hat nur eine Teilmenge davon, eine sogenannte Stichprobe, untersucht,
- über einen "Prozeß des Entstehens von Ereignissen" - man hat ihn nur eine Zeit lang beobachtet, auch hier bezeichnet man die Beobachtungen als Stichprobe.

Wie kann man diese Information auf die "Grundgesamtheit" verallgemeinern, ist die Kernfrage.

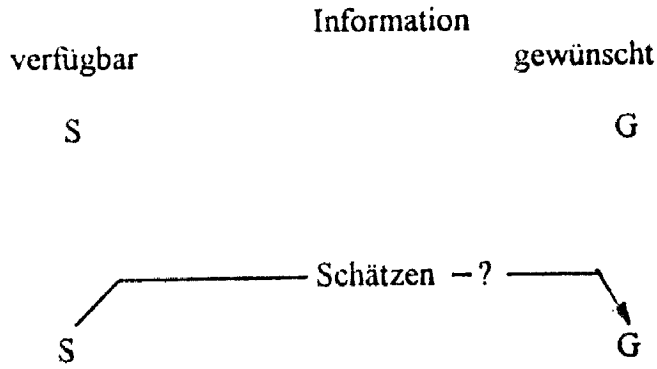

Abb. 1: Schematische Darstellung des statistischen Grundproblems: Verallgemeinern der verfügbaren Information auf eine größere Einheit.

Damit ein Transfer des Wissens über die Teilmenge glaubwürdig und zuverlässig wird, benötigt man fiktive Sichtweisen für die Grundgesamtheit G und Annahmen, wie daraus die Information über S entstehen könnte. Hypothetisches Denken also, was wäre wenn, wie würde sich das und das entwickeln, wenn. Man denkt sich also verschiedene Szenarien aus für G und dafür, wie die Information entstehen könnte. Daraus bezieht man dann einen zahlenmäßigen Transfer der Information von S auf G und dessen Rechtfertigung.

- *Das Umkehrproblem*

Das Szenario für die Grundgesamtheit und dafür, wie die Information auf S entstanden sein könnte, wird mit Hilfe von Wahrscheinlichkeiten beschrieben. Wahrscheinlichkeit ist jedoch nur ein Hilfsbegriff in der Gesamtüberlegung. Am besten vergleicht man den Typ der Fragestellung in der Wahrscheinlichkeitstheorie und in der Statistik direkt miteinander.

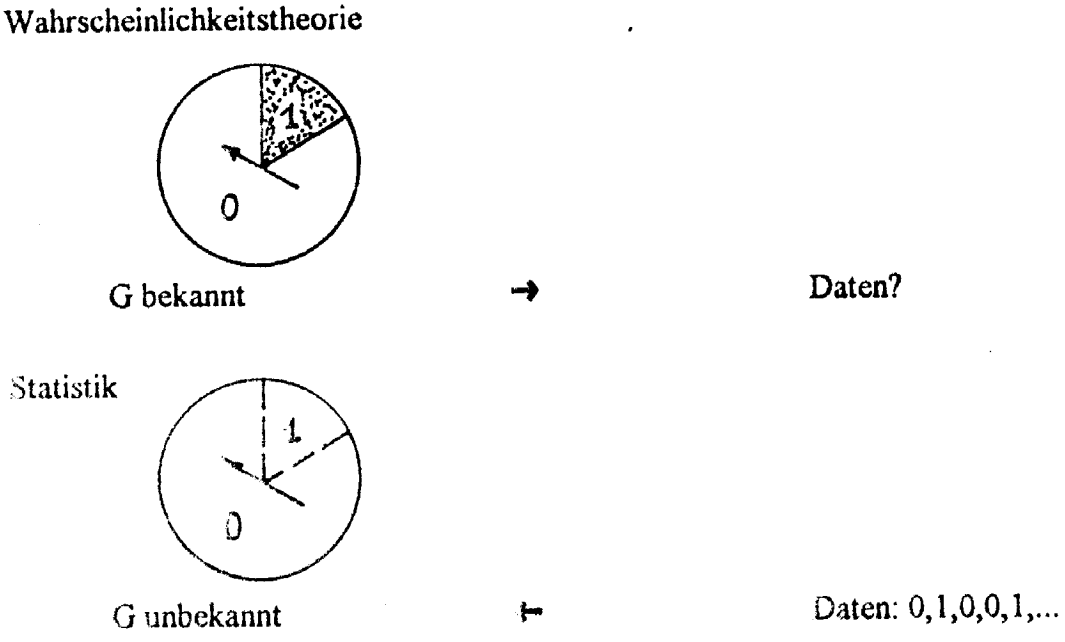

Abb.2: Vergleich der Typen von Fragen in Wahrscheinlichkeitsrechnung und Statistik anhand eines einfachen Glücksrades.

Am einfachsten kann man die unterschiedliche Art der Fragestellung an Glücksrädern erläutern. In der Wahrscheinlichkeitstheorie ist das Glücksrad bekannt (egal zunächst auch, was Wahrscheinlichkeit bedeutet und wie man zu numerischen Werten für Wahrscheinlichkeiten kommt). Welche Daten wird man erhalten? Diese Frage kann man für Einzeldaten nur mit wenig Zuverlässigkeit beantworten, bezieht man sie jedoch auf Anzahl der Ereignisse "1" oder deren Anteil an einer

ganzen Serie, so gibt die Binomialverteilung (mit n als Länge der Serie und p als Anteil des Sektors mit "1") die Antwort. Gleichzeitig wird aus dieser Verteilung ersichtlich: sie ist eingipfelig und (für den Anteil, nicht die Anzahl der Einsen) eng konzentriert um wenige Werte (diese Konzentration nimmt noch mit der Länge der Serie zu). Man kann also einen engen Bereich für den empirischen Anteil angeben, in den bei mehrfacher Drehung des Glücksrades der Anteil der Einsen mit hoher (vorgegebener) Wahrscheinlichkeit hineinfallen wird. Mit einer Metapher aus der Meßtechnik könnte man sagen, daß die Wiederholstreuung der Meßwerte (i.e. die empirischen Anteile) für den (hier bekannten) Anteil p am Glücksrad sehr klein ist. Das bedeutet, es handelt sich um eine zuverlässige "Messung".

In der Statistik dreht man nun das Glücksrad um, der Beobachter sieht nicht mehr die Aufteilung des Rades in Sektoren, ihm wird nur mehr mitgeteilt, daß es sich um die Sektoren mit einer 1 und mit einer 0 handelt. Im folgenden wird das Glücksrad mehrfach gedreht und die Ergebnisse werden mitgeteilt - der Prozeß des Entstehens von Ereignissen wird nur eine Zeit lang beobachtet. Es sind Aussagen über die Grundgesamtheit, also über die Aufteilung des Glücksrades in die beiden Sektoren, zu machen. Jetzt kommt die Metapher aus der Meßtechnik voll zum Tragen, denn jetzt "mißt" man einen unbekannten Anteil p und vom Meßverfahren ist (durch Studium der Wahrscheinlichkeitstheorie) schon bekannt, daß es eine kleine Wiederholstreuung hat, die zudem noch mit größerem Stichprobenumfang kleiner wird.

b) Der Prozeß, der die Daten erzeugt

Das Glücksrad zeigt, wie man zu den Daten gelangt, nämlich mit Hilfe des Zufalls. Im folgenden wird ein Beispiel gegeben, wie wenig eine Übertragung von Information von einer Teilmenge sinnvoll und gerechtfertigt ist, wenn die Datengewinnung nicht "kontrolliert" wird, wenn also dabei andere Mechanismen als der Zufall am Werk sind, die dann mit der Fragestellung interferieren und eine Verzerrung des Antwortverhalten nach sich ziehen. Sodann wird für die Glücksräder die Konsequenz aus dem Zufall, wie dadurch bedingt Daten entstehen, untersucht. Dabei wird auch die typische Denkweise erläutert, die da lautet: Eines der möglichen Modelle (eine konkrete Verteilung) für die Erzeugung der Daten wird danach beurteilt, ob die Daten, die man schon hat, unter diesem Modell eine kleine oder große Wahrscheinlichkeit besitzen. Schließlich wird mit Hilfe der Glücksräder ein Bild entworfen, wie man sich den Prozeß, der die Daten erzeugt, zu denken hat. Insbesondere wird dabei klar, welche Rolle der Zufall spielt, wenn man von einer Stichprobe spricht. Stichproben sind insbesondere keine beliebigen Teilmengen der Grundmenge; nur im Fall von kontrollierten Stichproben sind Transfers von Wissen über die Stichprobe auf die Grundgesamtheit durch Methoden der Beurteilenden Statistik gerechtfertigt, in anderen Fällen sind solche Transfers oft reine Spekulation.

- *Datengewinnung nicht kontrolliert*

Hierzu soll ein einfaches Beispiel genügen, das aufzeigt, wie stark verzerrt die Verhältnisse in der untersuchten Teilmenge sein können, wenn man den Mechanismus zur Auswahl der Objekte (der befragten Personen z.B.) nicht von deren Eigenschaften (dem Antwortverhalten z.B.) sauber trennt. Der Zufall soll dies und nichts anderes leisten.

Beispiel: In der Kolumne von Lynn Anders in einer US-Zeitschrift wurde die Frage erörtert, ob jemand seinen Partner wieder heiraten würde. "Würden Sie heute Ihren Partner wieder heiraten?" war die an die Leser gestellte Frage und die Antworten sollten zu den Erörterungen einen empirischen Beleg liefern, wie die Dinge wirklich liegen. Von den 15 000 (!) Antworten lauteten 70% (!) auf "nein". Klar war hier ein Selektionsbias (eine Verzerrung) wirksam; jene, die sich von dem Thema nicht besonders angesprochen fühlten, hatten gar keinen Grund zu antworten, während wieder andere, die etwa mit ihrem Partner sehr unzufrieden waren, hier eine Möglichkeit sahen, dies zu artikulieren. Die Grundgesamtheit läßt sich also in zwei Schichten einteilen, jene die antworten (A) und jene die nicht antworten auf diese Frage ($\overline{A}$). Die folgende Abbildung zeigt schematisch das unterschiedliche Antwortverhalten.

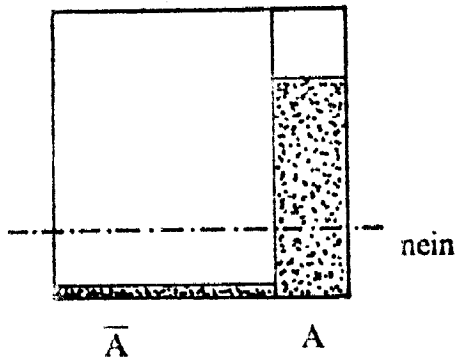

Abb. 3: Verteilung der Antwort "nein" in den zwei Schichten von Personen, die antworten und die nicht antworten - Mittlerer Anteil von "nein" als gewogener Mittelwert.

Eine kontrollierte Studie später, in welcher die befragten Personen durch das Zufallsprinzip ausgewählt wurden, in welcher also die Schichten A und $\overline{A}$ im richtigen Verhältnis vertreten waren, ergab dann auch tatsächlich ein Quote von 10% für die Antwort "nein". Das Beispiel macht deutlich, daß noch so viele Daten keinen Anspruch auf Verallgemeinerung haben, hier sind es z.B. 15 000. Es ist viel wichtiger, daß die Daten durch den Zufall (wie schwierig das auch sein mag) gewonnen werden; dieser Zufall trennt i.a. die Aufnahme in die Stichprobe und das Antwortverhalten der Aufgenommenen und Nicht-Aufgenommenen.

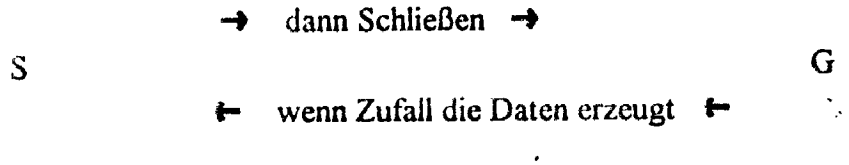

Abb.4: Schematische Darstellung des statistischen Schlusses von einer Stichprobe S auf die Grundgesamtheit G - der Schluß ist nur gerechtfertigt, wenn der Zufall die Daten erzeugt.

Die meisten Fehlurteile liegen gar nicht in der Schwierigkeit der statistischen Verfahren, sondern weit vorher, nämlich in der schlechten Kontrolle der Gewinnung der Daten. Aus verzerrten Daten kann man alles "beweisen" und dann Laien noch mit großen Stichprobenumfängen beeindrucken - "15 000 Personen lügen nicht!" Hier zeigt sich, daß noch so viele Daten wertlos sind, wenn ihre Gewinnung nicht kontrolliert ist. Die meisten statistischen Flops sind gar nicht durch falsche Methoden in der Analyse verursacht, sondern durch mangelnde Kontrolle der Datengewinnung. Man hat einfach irgendwelche "confounders" übersehen, diese üben aber einen starken Einfluß auf die Gewinnung der Daten aus. Nur durch eine wohlüberlegte Systemanalyse kann man frühzeitig auf solche "confounders" kommen und sie im Stichprobenplan geeignet ausschalten.

- *Datengewinnung durch Zufall kontrolliert*

Wie man die Gewinnung von Daten durch Zufall kontrolliert und welche Auswirkung das auf das Entstehen der Daten hat, ist im Fall stetiger Verteilungen für ein Merkmal mathematisch recht anspruchsvoll. Man kann die Dinge aber im Spezialfall von Anteilen sehr gut illustrieren. Dabei benötigt man nur das Hilfsmittel der Binomialverteilung. Wieder lassen sich die Dinge mit dem Glücksrad schön veranschaulichen. Man geht davon aus, daß eine bekannte Grundgesamtheit vorhanden ist, man studiert dann unter den Regeln des Zufalls, wie sich Stichproben entwickeln. In der Metapher der Meßtechnik, man studiert die Wiederholstreuung des "Meßverfahrens" im Labor, wo man die "wahren" Größen kennt.

Für die Grundgesamtheit sei der Anteil eines Merkmals mit 1/6 vorausgesetzt. Man kann den Prozeß der Auswahl von Personen und damit der Gewinnung der Daten durch die n-fache Dre-

hung des folgenden Glücksrades darstellen; i.f. seien n = 100 Daten unterstellt.

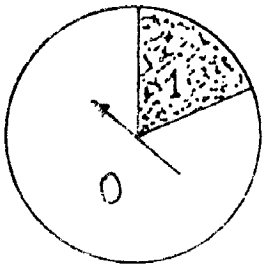

Abb. 5: Glücksrad mit Sektoren 0 und 1 - 1 steht für das Vorhandensein des Merkmals, 0 für dessen Fehlen - eine Stichprobe vom Umfang 100 entsteht durch 100faches Drehen des Glücksrades.

Neben den konkreten Daten $x_1, x_2, \ldots, x_{100}$ hat man sich Gedanken zu machen über die Entstehung dieser x_i. Dies wird üblicherweise in Form von Wahrscheinlichkeiten bzw. einer Zufallsvariablen gemacht; im konkreten Fall ist diese eine Alternativverteilung mit dem Parameter p = 1/6. Für die Entstehung der Summe $x_1 + x_2 + \ldots + x_{100}$ hat man dann nach Sätzen aus der Wahrscheinlichkeitstheorie eine Binomialverteilung mit den Parametern n=100 und p=1/6. Die Entstehung der Anteile $h = \frac{x_1 + x_2 + \ldots + x_{100}}{100}$ hat man nur die entsprechende Binomialverteilung auf das Intervall [0, 1] zu stauchen. Die folgende Abbildung zeigt die Binomialverteilung mit den Paramtern n=100 und p=1/6, gestaucht auf dieses Intervall.

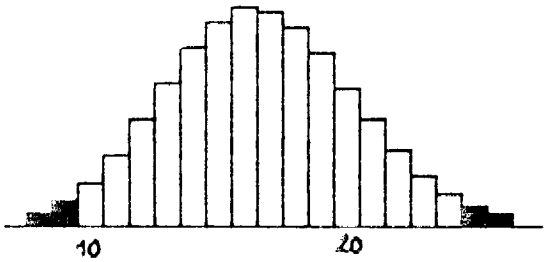

Abb. 6: Auf das Intervall [0, 1] gestauchte Binomialverteilung mit n=100 und p=1/6. So modelliert man den Prozeß des Entstehens von Daten unter Zufall und der Annahme p=1/6 für die Grundgesamtheit - Extreme Bereiche sind markiert.

Ein statistischer Schluß basiert nun auf der Annahme mehrerer Verteilungen für die Grundgesamtheit; hier sind es alle Alternativverteilungen mit p aus [0, 1]. Für ein konkretes p, z.B. p=1/6, studiert man die (theoretische) Verteilung des Prozesses der Entstehung der Daten, das sind die Stichprobenanteile h. Für diese Verteilung markiert man sich einen Extrembereich, der unter p=1/6 sehr unwahrscheinlich ist, wobei andere Werte für p diesen Extrembereich mit entsprechend größerer Wahrscheinlichkeit ausstatten würden. Der Schluß liegt nahe, für Daten, die in den Extrembereich für p=1/6 fallen, diesen Wert für die Grundgesamtheit auszuschließen, weil die Daten mit p=1/6 "nicht verträglich" sind. Der Schluß birgt ein statistisches Risiko mit sich (es könnte ja auch etwas Ungewöhnliches passiert sein) und basiert auf dem Zufall, mit dem die Daten entstanden sind. Der Normalbereich ist in dem Beispiel (0,092, 0,241); alle Stichprobenanteile darunter oder darüber würden daher in den Extrembereich und somit zur Ablehnung des möglichen Wertes 1/6 für p führen. Die Überlegung kann man mit speziellen Werten von p oder mit allen Werten von p durchführen. Ersteres führt zum statistischen Test, letzteres zu statistischen Vertrauensintervallen.

- *Stichprobe*

Der Begriff Stichprobe ist für die Rechtfertigung des Transfers von Daten auf die zugrunde liegende Grundgesamtheit grundlegend, jedoch in seiner Allgemeinheit von technisch-mathematischen Feinheiten gekennzeichnet. Dennoch ist wichtig, eine Idee davon zu bekommen. Diese Idee wird hier an Glücksrädern entwickelt.

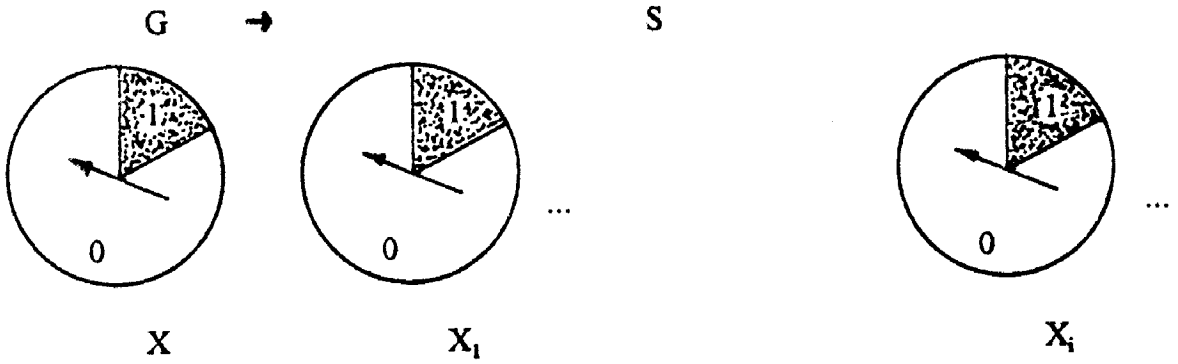

Abb. 7: Grundgesamtheit als Glücksrad, die Entstehung einzelner Daten als Kopien desselben Glücksrades, das "unabhängig" voneinander wiederholt gedreht wird.

Die Grundgesamtheit, symbolisch X wird durch eine Verteilung modelliert, die vorübergehend (fiktiv) als bekannt vorausgesetzt wird. Im Fall der Glücksräder ist dies eine Alternativverteilung, allgemeiner könnte es eine Normalverteilung oder irgendeine stetige Verteilung sein. Die Entstehung des i-ten Datums X_i wird ebenso durch dieselbe Verteilung modelliert: $_iX \sim X$. Man spricht plakativer von einer "Kopie" von X. (Die Großbuchstaben deuten an, daß es nicht um konkrete Daten sondern um die Entstehung dieser Daten geht, also um Zufallsvariable.) Diese Kopien sind "unabhängig", im Fall der Glücksräder heißt das, sie werden eigenständig gedreht, die Zwischenergebnisse sind ohne Einfluß (im kausalen Sinne) und ohne Informationswert (im Sinne einer Änderung der Verteilung, weil andere Ergebnisse nun bekannt sind). Wenn die Entstehung der Daten diesem Bild genügt, so spricht der Statistiker von einer Stichprobe. Die Bezeichnung Zufallsstichprobe ist redundant, denn Stichprobe bedingt bereits, daß der Zufall im Spiel war. Anders geartete Teilmengen, über die man Information hat, sollten nicht mit Stichprobe tituliert werden.

$$S: \quad \overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n} \qquad G: \quad \mu$$

Abb. 8: Schematischer Transfer von der Stichprobe auf die Grundgesamtheit.

Mit Hilfe der Wahrscheinlichkeitstheorie kann man dann aus den bekannten Verteilungen für die i-ten Daten X_i und der Unabhängigkeit eine weitere Verteilung für die Summe der Daten sowie für den Mittelwert der Daten ableiten. Mit Hilfe dieser Verteilung kann man für jeden konkreten Wert des theoretischen Mittelwerts μ der Grundgesamtheit solche Extrembereiche wie oben berechnen und prüfen, ob ein vorhandener empirischer Mittelwerte mit einem speziellen Wert für die Grundgesamtheit verträglich ist oder nicht. Die Berechnung der Verteilung ist wieder nur dann einfach, wenn man Anteile schätzen will, für stetige Verteilungen muß man die Ergebnisse einfach mitteilen oder durch Simulation illustrieren. Wieder ist festzuhalten, daß die Daten nur dann für die Verträglichkeitsprüfung geeignet sind, wenn sie aus einer Stichprobe stammen, d.h. wenn sie durch Zufall erzeugt wurden.

c) Problemtyp: Anteile

Im folgenden wird das Problem behandelt, wie man einen unbekannten Anteil in einer Grundgesamtheit durch den Anteil in einer Stichprobe schätzt. Die Grundgesamtheit kann wieder durch ein Glücksrad modelliert werden, der Prozeß des Entstehens von Daten durch die wiederholte Drehung desselben. Einsen stehen dabei für "Merkmal vorhanden", Nullen für "Merkmal nicht vorhanden", die Summe der Daten ist dann die Anzahl der Objekte mit dem Merkmal, der Anteil der Daten entspricht dem Anteil an Objekten in der Stichprobe mit dem Merkmal. Dieser Anteil wird als Schätzwert für den i.a. unbekannten Anteil p des Merkmals in der Grundgesamtheit genommen.

Die folgende Abbildung zeigt nun, wie der Prozeß der Entstehung der Daten sich auswirkt, wenn drei spezielle Werte für den unbekannten Anteil p unterstellt werden. Die Entstehung der Summe der Daten wird durch eine entsprechende Binomialverteilung modelliert, für die Anteile wird diese Verteilung auf das Intervall [0, 1] gestaucht. Unabhängig von dem speziellen Wert für p gilt: Die Verteilungen für die Entstehung des Anteils in Stichproben hat einen Erwartungswert p und eine Varianz $p(1-p)/n$. Dies bedeutet in der Metapher des Messens einer unbekannten Größe p, daß der Meßprozeß (wiederholte unabhängige und zufällige Daten und Anteilsbildung) richtig zentriert ist (man mißt tatsächlich, was zu messen ist) und daß die Wiederholstreuung mit zunehmender Serie von Daten kleiner wird. Es ist der Zufall, der diese Aussagen, insbesondere über die Wiederholgenauigkeit berechenbar macht. In diesem Sinn spricht man von statistischer Information über p, welche dann in Form statistischer Tests oder Vertrauensintervalle nutzbar gemacht werden kann.

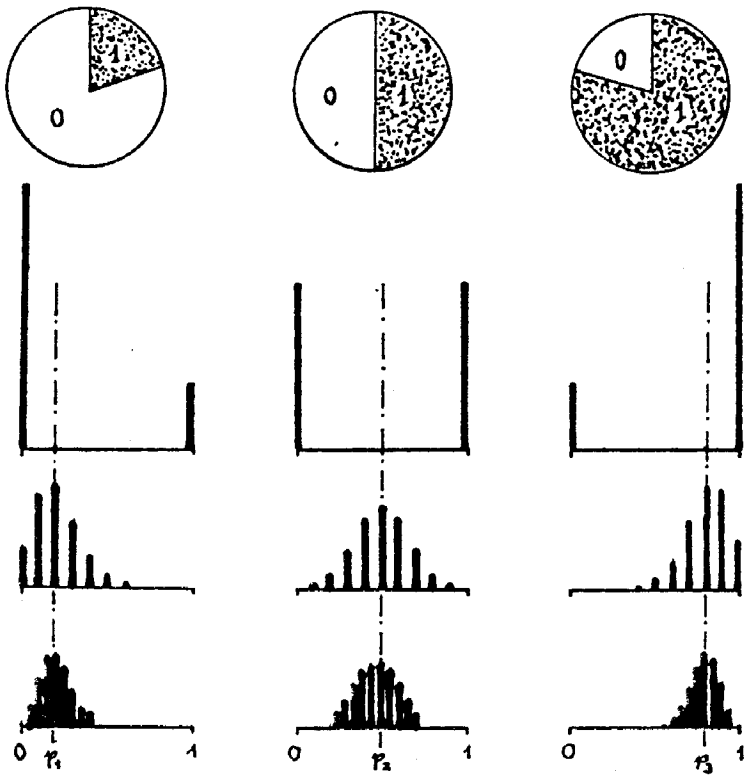

Abb. 9: In den "Zeilen" der Fig. stehen die Grundgesamtheiten sowie die entsprechende Verteilung der Entwicklung der Anteile in Stichproben für n=1 (entspricht der Grundgesamtheit) sowie für einen kleinen und einen großen Stichprobenumfang. In den Spalten steht diese Entwicklung für einen speziellen Wert für den Anteil p. Unabhängig von p zeigt sich: die statistische Information ist richtig zentriert und wird mit größerer Serie von Daten besser, weil die Wiederholstreuung abnimmt.

d) Problemtyp: Mittelwerte

Im folgenden wird der zweite Grundtyp von Problemen statistischer Schlußfolgerungen behandelt, es geht um die Übertragung von Mittelwerten aus Stichproben auf Grundgesamtheiten. Die Grundgesamtheit X wird durch eine Normalverteilung modelliert, üblicherweise wird die Varianz σ^2 dieser als bekannt vorausgesetzt, nur um den Mittelwert geht es, der ist unbekannt und heißt nun als Parameter einer Wahrscheinlichkeitsverteilung Erwartungswert; er sei mit μ bezeichnet. Der Prozeß des Entstehens von Daten führt hier zu einer Folge von Normalverteilungen und schließlich (über Sätze aus der Wahrscheinlichkeitstheorie) zu einer Normalverteilung für die Entstehung der Mittelwerte $\overline{X}$. Die letztere hat genau denselben Erwartungswert μ und die viel kleinere Stichprobenvarianz σ^2/n (die mit dem Faktor 1/n abnimmt, n die Zahl der Daten).

Folgende, schon mehrfach verwendete Metapher aus der Meßtechnik ist hier hilfreich: Eine bestimmte physikalische Größe μ ist zu messen. Die Verteilung der möglichen Meßwerte mit einem bestimmten Meßgerät ist eine Normalverteilung mit eben diesem Erwartungswert μ und einer gerätespezifischen Streuung σ, je größer , umso mehr werden sich aufeinanderfolgende Meßwerte im Durchschnitt unterscheiden, man spricht von kleiner oder großer Wiederholstreuung oder umgekehrt von großer oder kleiner Wiederholgenauigkeit. Man kann nun ein teureres Gerät anschaffen, das eine größere Wiederholgenauigkeit hat, oder man kann sorgfältige, unabhängige (!) Messungen mit dem einen Gerät durchführen und den Mittelwert $\overline{x}$ der Meßserie $x_1, x_2, \ldots, x_n$ als neuen Meßwert angeben. Schreibt man Großbuchstaben, um anzudeuten, daß es sich um den Prozeß des Entstehens von Meßwerten handelt, so erhält man die Aussage, daß dieser Meßprozeß richtig zentriert ist und daß sich die Präzision im Sinne der Wiederholgenauigkeit mit $\sigma/\sqrt{n}$ verringert. Man kann also ein teureres, präziseres Meßinstrument durch wiederholte, unabhängige Messungen simulieren.

Die folgende Abbildung 10 zeigt nun, wie der Prozeß des Entstehens der Daten sich auswirkt, wenn drei spezielle Werte für den unbekannten Erwartungswert μ unterstellt werden. Die Entstehung der Mittelwerte der Daten wird zu einer entsprechenden Normalverteilung (nach Sätzen der Wahrscheinlichkeitstheorie), wenn die Grundgesamtheit (i.e. die Einzelmessung) als normalverteilt modelliert wird. Unabhängig vom speziellen Wert für μ gilt: Die Verteilung des Entstehens der Mittelwerte hat einen Erwartungswert μ und eine Varianz σ^2/n. Dies bedeutet, das Meßverfahren ist richtig zentriert (ohne systematischen Fehler) und die Wiederholstreuung wird geringer, weshalb es Sinn macht, mehrere Messungen durchzuführen. Es ist wieder der Zufall und die Unabhängigkeit, welche diese Aussagen, insbesondere über die Wiederholgenauigkeit berechenbar machen. In diesem Sinn spricht man von statistischer Information über μ. Für statistische Methoden muß man noch Extrembereiche in der Verteilung bestimmen und den Vergleich von Mittelwerten aus Daten mit den theoretischen Werten μ über die Grundgesamtheit im Sinne der Verträglichkeit durchführen. Neben der Genauigkeit muß man also auch noch ein Risiko von statistischen Aussagen beachten.

Ganz besonders wichtig ist nun der Umstand, daß die Voraussetzung der Normalverteilung für die Grundgesamtheit, i.e. die Einzelmessung fallen gelassen werden kann. Die Bedingungen an die Entstehung der Daten sind, wie immer für Stichproben, dieselbe Verteilung für jedes Datum und die Unabhängigkeit der Entstehung der Daten, insbesondere muß die Varianz σ^2 für die Grundgesamtheit existieren (eine mathematische Feinheit). Dann ist die Verteilung der Entstehung der Mittelwerte immer ähnlicher einer Normalverteilung (genauer: ihre Standardisierung konvergiert gegen die (0, 1)-Normalverteilung), das ist Inhalt des Zentralen Grenzverteilungssatzes. Die Verteilung der Grundgesamtheit kann also, wie im Fall der Glücksräder, auch eine Alternativverteilung sein, und dennoch werden Mittelwerte (Anteile sind spezielle Mittelwerte) annähernd normalverteilt. Das erleichtert die Berechnung extremer Bereiche und die daran geknüpften statistischen Methoden wesentlich und verbreitet die Anwendungsmöglichkeiten dieser Methoden enorm.

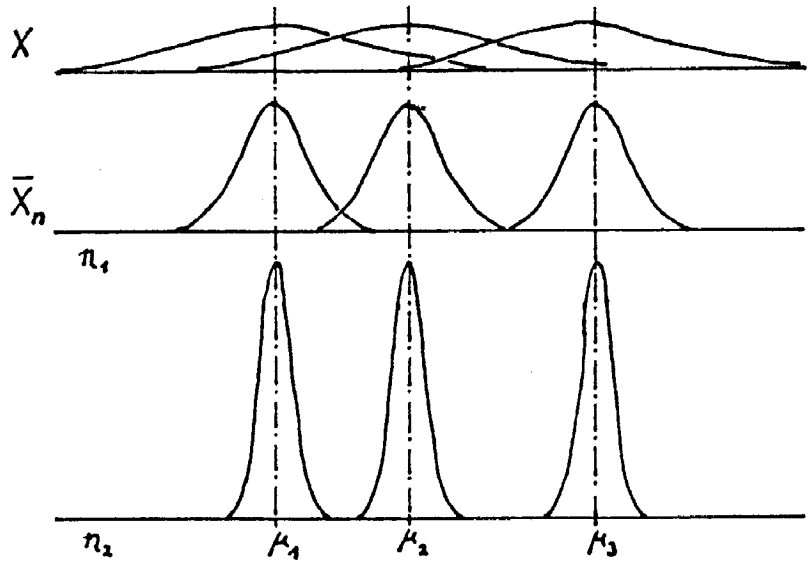

Abb. 10: In den Zeilen der Fig. stehen die Grundgesamtheiten sowie die entsprechenden Verteilungen der Entwicklung der Mittelwerte in Stichproben für n=1 identisch mit der Grundgesamtheit sowie für einen kleineren und einen größeren Umfang der Daten. In den Spalten steht diese Entwicklung für einen speziellen Wert von μ. Unabhängig von diesem μ zeigt sich: die statistische Information ist richtig zentriert und wird mit größerer Serie von Daten besser, weil die Wiederholstreuung abnimmt.

2 Die großen Alternativen

Es gibt gänzlich unterschiedliche Ansätze, Wissen aus einer Stichprobe auf eine Grundgesamtheit zu verallgemeinern. Die Schulen dahinter haben je ihre Anhänger, die sich gegenseitig vorwerfen, das Problem nicht sauber zu lösen. In diesem Abschnitt wird ein kurzer Überblick über diese Schulen gegeben, es wird auf ihre Voraussetzungen und Ziele eingegangen. Danach wird in den einzelnen Ansätzen eine je andere Fragestellung mit einer je anderen Antwortmöglichkeit behandelt. Schließlich werden die Ansätze der Explorativen Datenanalyse und der klassische Ansatz bzw. der Bayes-Ansatz mit dem klassischen verglichen. Es geht darum, herauszustreichen, was die spezifischen Vor- und Nachteile der aus den Schulen folgenden Methoden sind.

a) Ein Überblick über die Schulen zum statistischen Schluß

Im folgenden werden die klassische Statistik, die Bayes-Statistik und die Exploratorische Datenanalyse (EDA) miteinander verglichen. Den einzelnen Schulen liegt eine unterschiedliche Auffassung von Wahrscheinlichkeit zugrunde. Hier soll zur Einbegleitung eine Übersicht gegeben werden (Abb. 11). Die klassische Schule begründet ihre Methoden auf einen mathematischen Wahrscheinlichkeitsbegriff, der rein frequentistisch gedeutet wird. Der Bayes-Ansatz ist in einer gewissen Hinsicht umfassender, insofern als die Deutung von Wahrscheinlichkeit als relative Häufigkeit auf lange Sicht in ihr als Spezialfall vorkommt, geht jedoch grundsätzlich von einem qualitativen, subjektbezogenen Wahrscheinlichkeitsbegriff aus. Eine Person hat aufgrund aller ihr zur Verfügung stehenden Information die Wahrscheinlichkeiten für eine unsichere Sache zu bemessen. Der Bayes-Ansatz wirft den klassischen Statistikern vor, daß deren Methoden irrationale Lücken in der Begründung aufweisen und in konkreten Fällen zu unsinnigen Schlüssen führen. Die Klassiker wiederum werfen den Bayesianern vor, ihre Methoden auf einen qualitativen, nicht-objektivierbaren Wahrscheinlichkeitsbegriff aufzubauen. Die EDA ihrerseits wirft den Klassikern vor, von komplizierten Modellen Gebrauch zu machen, die in der Praxis einfach nicht auf ihre Passung geprüft werden können; die EDA weist einen Weg "back to basics", zu einfachen

Modellen und Begriffen, die direkt (d.h. ohne Bezug auf eine komplizierte Theorie und komplizierte Deutungen etwa des Wahrscheinlichkeitsbegriffs) verstanden werden können. Die Klassiker werfen den Anhängern der EDA wiederum vor, Daten ohne Modelle und gezielte Vorüberlegungen zu (detektivisch) nach etwaigen Mustern zu untersuchen; wenn man nur lange genug hinsieht, würde man in allen Daten ein Muster sehen, eine Absicherung gegen Artefakte etwa mit Sicherheitswahrscheinlichkeiten gäbe es aber nicht.

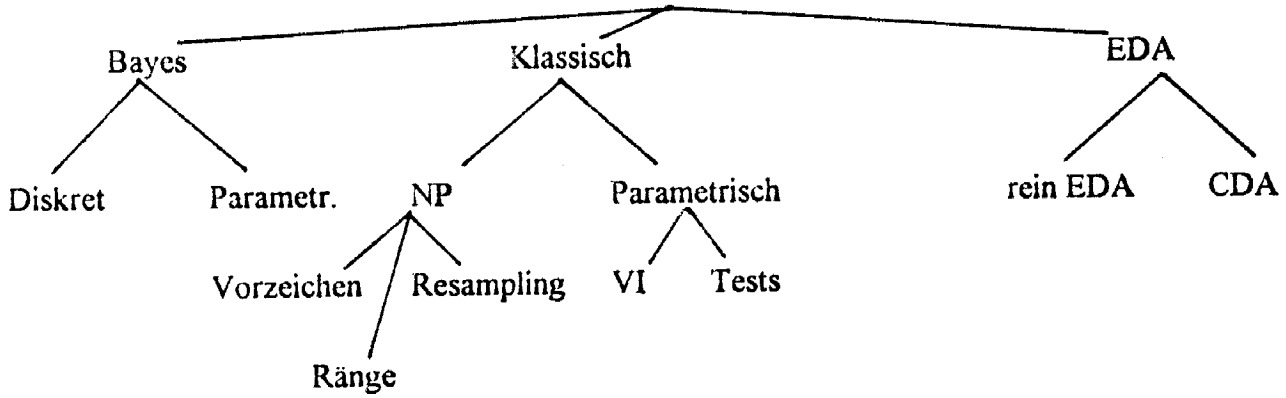

Abb. 11: Eine Übersicht und Untergliederung statistischer Schulen - Parametr. = Parametrisch; VI Vertrauensintervalle; NP Nicht-Parametrische Verfahren; CDA confirmatory data analysis.

Statt den Grundlagenstreitigkeiten zu folgen ist man heute in den Anwendungen und bei theoretischen Statistikern, die im anwendungsnahen Feld forschen zur Einsicht gelangt, daß man aus den verschiedensten Ansätzen das jeweils beste herausnehmen soll. Die darauf aufbauenden Methoden haben je ihre Meriten, ihre besonderen Einsatzgebiete, wo sie einfach besser als die konkurrierenden Schulen sind. Mehr dazu kann man in Borovcnik (1992) bzw. (1990) nachlesen.

b) Voraussetzungen und Ziele der statistischen Schulen

Es wird eine grobe Übersicht über die Unterschiede in den Schulen gegeben, dabei wird auf die Voraussetzungen an die Daten, den Charakter der verwendeten Modelle, die Art der Analyse sowie die Ziele der Analyse eingegangen. Insgesamt wird qualitativ geklärt, daß die Methoden der diversen Schulen andere Fragen mit anderen Antwortmöglichkeiten behandeln. Für Details sei auf die angegebene Literatur verwiesen. Die folgende Tabelle wird weiter unter etwas näher erläutert werden.

Für den Bayes-Ansatz ist zentral, daß alles, worüber man unsicher ist, eine Wahrscheinlichkeit besitzt, die der Analyst durch "Introspektion" aus sich herauszufinden hat. Wenn er also einen unbekannten Anteil p untersucht, so "hat" er darüber eine Verteilung zu haben, z.B. eine Gleichverteilung, welche einer Situation mit sehr wenig Information über den Parameter entspricht. Erhält er jetzt in 10 Daten 7 Fälle mit dem in Frage stehenden Merkmal, so spitzt sich sein aktueller Informationsstand bei 0,7 auf, er wird eine neue Wahrscheinlichkeitsverteilung haben, die über das ganze Intervall [0, 1] Wahrscheinlichkeiten hat, jedoch in der Nähe von 0,7 einen Gipfel. Hat er 100 Daten mit 70 Fällen, so wird seine Verteilung die gleiche Qualität aufweisen, jedoch wird der Gipfel bei 0,7 wesentlich stärker ausgeprägt sein. Diese Verteilung ist der jeweils aktuelle Informationsstand (Information wird bei Bayesianern immer durch Wahrscheinlichkeiten wiedergegeben), der die a priori-Information (hier die Gleichverteilung) und die Information aus den Daten zusammenfaßt. Das Bayes-Theorem bildet die Grundlage der Neuberechnung der Verteilungen; Details kann man z.B. in Borovcnik (1992) oder in Kleiter (1980) nachlesen.

Der klassische Statistiker dagegen kann eine a priori-Verteilung nicht angeben, weil sie keine Prüfung durch ein Experiment mit Daten zuläßt (sie stellt eine überwiegend qualitative Zusammenfassung des bisherigen Kenntnisstandes zusammen, der natürlich auch Daten aus vergangenen Experimenten beinhaltet). Für den Klassiker ist der unbekannte Anteil p eine Konstante, über den er indirekt aus den Daten eine Beurteilung erhält. Entweder hat er eine Vermutung, eine Hypo-

these über einen konkreten Wert von p, dann kann er untersuchen, ob die Daten einer Zufallsstichprobe mit diesem Vorgabe-Wert verträglich sind, das führt zu einem statistischen Test. Oder er hat keine solche Vorgabe, dann prüft er alle Werte von p durch und prüft, welche mit den Daten verträglich sind, das führt zu Vertrauensintervallen.

	Bayes	Klassisch	EDA
Daten	Zufall	Zufall	egal
Modell	ja	jein	nein
Parameter	Zufallsvariable	Konstante	allgemein
Analyse	eine	eine	viele
Ziele	Ws. für Parameter	Beurteilung indirekt	Muster & Besonderheiten
Beispiel: Unbekannter Anteil p	aus einer a priori-Verteilung durch Daten eine neue Verteilung	Test oder Vertrauensintervall für p	Häufung und Lücken - Zerlegen der Daten und Erklären der Besonderheiten

Der EDA-Analyst hat gar keine Vorgabe hinsichtlich des Modells. Er prüft etwa, ob sich zeitlich einige Merkwürdigkeiten hinsichtlich des Auftretens des Merkmals ergeben, vielleicht gibt es einen Zusammenhang mit der Örtlichkeit der Objekte? Vielleicht gibt es einen Zusammenhang des Auftretens dieses Merkmals mit anderen Eigenschaften der Objekte, ist etwa ein geschlechtsspezifisches Häufen des Merkmals vorhanden? Diese Untersuchungen setzen keine zufälligen Daten voraus, das Argument zur Verallgemeinerung von gefundenen Ergebnissen muß daher von anderer Natur sein, etwa die spontan im Ergebnis inneliegende Einsicht von der Substanzwissenschaft her. Dazu ist es mindestens notwendig, die Zwischenschritte der Analyse immer mit dem Kontext, aus dem die Daten stammen, rückzukoppeln. Die eigenartige, interaktive Analyse verwendet zwar einfachere Methoden aus der Statistik, verlangt vom Analysten jedoch eine Menge Kenntnisse aus der Bezugswissenschaft, auf die sich die Daten beziehen.

Wenn man qualitatives Vorwissen über Parameter einer Verteilung für die Daten hat, so wird man mit Vorteil den Bayes-Ansatz verfolgen können und das Vorwissen geeignet einbringen. Lohn wird etwa sein, daß man aus weniger Daten noch immer scharfe Schlüsse ziehen kann. Ist dieses Vorwissen allzu vage, oder kostet es zu viel Zeit und oder Geld, es zu quantifizieren, so werden klassische Verfahren vielleicht passender sein. Hat man von den untersuchten Phänomenen noch zu wenig analysiert, befindet man sich etwa in der Pilotphase eines Projekts, wo man grundsätzliche Richtungen der weiteren Forschung abklären will, so werden gezielte Fragen nicht zu stellen sein und man wird mit Vorteil sich der interaktiven Art der Analyse der EDA zuwenden, um einen Überblick zu bekommen. CDA, die confirmatory data analysis, würde dann erst sich in der zweiten Phase wieder verstärkt klassischer Verfahren bedienen.

Details können aufgrund von Platzmangel nicht hervorgekehrt werden. Die Ideen hinter dem Bayes-Ansatz sind ausführlich in Borovcnik (1992) dargestellt, zur EDA und zum Vergleich mit der klassischen Statistik findet man weitere Information in Borovcnik und Ossimitz (1987).

c) EDA versus klassische Statistik

Tabellarisch und an einem Beispiel soll der Vergleich zwischen diesen beiden Schulen weiter geführt werden. Wesentliche Unterschiede liegen im Modellverständnis der beiden Ansätze. Während klassische Statistik ein mathematisch-naturwissenschaftliches Modellverständnis zur Basis hat (oft in einem naiven Sinn der Deskription von einer real existierenden Situation), hat EDA durch ihre merkwürdige Interaktion zwischen den Zwischenergebnissen und dem Sachbezug keine vollständige Trennung zwischen Realität und Modell vollzogen; Modell und Wirklichkeit ändern sich fortwährend mit der Analyse. Zu diesem unterschiedlichen Modellverständnis findet man mehr in Borovcnik (1996).

	Klassisch	EDA
Daten	Zufall	egal
Analyse	numerisch eindeutig modellgebunden, geplant im voraus	visuell & mehrfach frei und interaktiv
Ziel	Prüfung von Annahmen	Trennung: D = F + R
Realität	getrennt von Modell	Motor der Analyse

Die Gleichung D = F + R meint die Zerlegung der Daten in Fit plus Residual, in ein Muster, das allgemein gültig ist (und das nach klassischem Stil untersucht werden kann und soll) und die Residuen, die Besonderheiten der Daten, die aus dem Kontext der Daten heraus interpretiert und erklärt werden sollen. Gerade diese Residuen sind eine Quelle von Einsicht und der Motor der Analyse nach Zwischenergebnissen. Insgesamt kann man der EDA die Rolle, Modelle zu erzeugen, zuschreiben, und der klassischen Statistik die Rolle, Modelle zu prüfen. Klar, daß man dazu schon mehr Vorwissen haben muß, um solcherart gezielte Fragen zu stellen, woraus sich eine natürliche Abgrenzung ergibt, wo sich die Ansätze je mit Vorteil einsetzen lassen.

Ein kleines Beispiel soll diese Trennung in Fit plus Residual illustrieren. Merkwürdig mag anmuten, daß hier der Kontext der Daten gar nicht erklärt wird, wo doch weiter oben behauptet wurde, daß sich gerade aus der Interpretation der Zwischenergebnisse vom Kontext her die weiteren Schritte der Analyse ergeben sollen. Hier soll lediglich demonstriert werden, wie die Methode der EDA die Residuen technisch vergrößert, damit eine solche Interpretation überhaupt erst ansetzen kann, weil nun die Residuen durch die Darstellungstechnik erst augenfällig geworden sind.

Die folgende Verteilung zeigt potentielle Ausreißer nach oben. Die Differenzen vom Mittelwert, i.e. $|x_i - \bar{x}|$, sind für die merkwürdigen oberen Punkte nicht allzu groß. Erstens wird der Mittelwert gerade durch diese merkwürdigen Punkte mit nach oben gezogen. Zweitens werden dadurch für ansonsten normale untere Punkte die Abweichungen vom Mittelwert größer. Insgesamt ergibt sich, daß sich die Abweichungen für diese oberen Ausreißerkandidaten nicht allzu groß darstellen. Durch den Median und die Bildung der Differenzen der Einzeldaten zum Median, i.e. $|x_i - \tilde{x}|$, werden die Residuen vergrößert. Erstens reagiert der Median gar nicht auf die merkwürdigen Punkte. Zweitens sind die Residuen der unteren Punkte nicht durch ein allfälliges Hochziehen des Bezugspunktes (des Fits) schon vergrößert. Insgesamt ergibt sich, daß der Median die Residuen wie durch eine Lupe vergrößert. Daran kann sich dann die sachliche Interpretation des durch die Darstellung geschickt vergrößerten Residuums anschließen. In der EDA werden also Modelle und Begriffe besonders wichtig, die erlauben, den Fit so zusammenzufassen, daß er überhaupt nicht auf potentiell merkwürdige Punkte reagiert. Die Residuen zu diesem Modell sind dann automatisch augenfälliger. Neben diesem technischen Punkt der Darstellung ist dann aber die sachgemäße Aufklärung des einmal gefundenen Residuums die Hauptaufgabe der EDA.

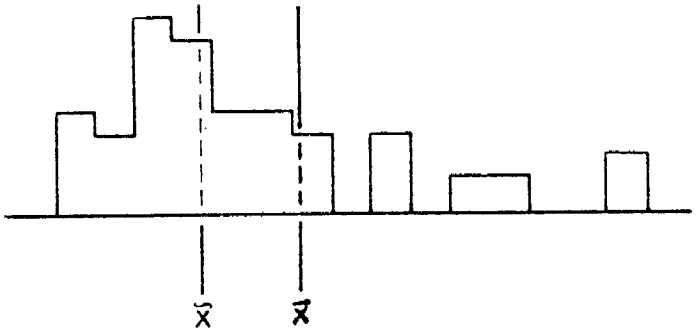

Abb. 12: Daten mit merkwürdigen Punkten nach oben - der Median faßt das Gewöhnliche in den Daten zusammen, ohne von diesen Punkten beeinflußt zu werden. Ganz anders der hingegen Mittelwert.

d) Bayes versus klassische Statistik

Tabellarisch und an einem Beispiel soll der Vergleich zwischen klassischer Statistik und dem Bayes-Ansatz weitergeführt werden. Ein wesentlicher Unterschied liegt darin, daß der Bayesianer qualitative Information in sein Modell integriert. Er macht damit von nicht objektivierbarer Information Gebrauch, weil seine Wahrscheinlichkeitsangaben nicht durch ein Zufallsexperiment überprüft werden können. Der Einwurf gilt allerdings nicht ganz, denn der klassische Statistiker macht in seinen Modellen zwar theoretisch von Wahrscheinlichkeit lediglich als relativer Häufigkeit Gebrauch, er verwendet jedoch so komplexe Modelle, deren Passung auf eine spezielle Situation ganz einfach nicht, jedenfalls nicht in einem objektivierbaren Sinn, überprüft werden kann.

	Klassische Statistik	**Bayes-Statistik**
Analyse	numerisch	numerisch
Daten	zufällig	zufällig
Ziel	indirekte Prüfung von Hypothesen	Bewertung von Hypothesen durch Wahrscheinlichkeiten
Realität	getrennt vom Modell	in der a priori-Bewertung starke Vernetzung zum Modell in der Person des Analysten

An einem Beispiel aus dem Kontext der medizinischen Diagnose soll verdeutlicht werden, wie der klassische Statistiker indirekte Aussagen über Hypothesen anstrebt und wie der Bayesianer seine Information a priori durch Wahrscheinlichkeiten ausdrückt und die Daten zu einer neuen Wahrscheinlichkeitsbewertung der Hypothesen heranzieht. Das Beispiel ist viel einfacher als das schon angesprochene Beispiel mit dem unbekannten Anteil p eines Merkmals in einer Grundgesamtheit, weil es von diskreten (insbesondere von zwei) Zuständen ausgeht.

Der Einfachheit halber sei angenommen, bei der Diagnose ginge es lediglich darum, zu prüfen, ob eine Person Virus V hat oder nicht. Über die Zustände V und $\overline{V}$ hat man die (subjektiven) Wahrscheinlichkeiten 0,005 bzw. 0,995 - diese Werte könnten etwa der frequentistischen Prävalenz von Virusträgern in der gesamten Bevölkerung entsprechen. Abb. 13 gibt Auskunft über die Sicherheiten der Diagnose, etwa wird ein Virusträger mit Wahrscheinlichkeit 0,99 als solcher erkannt. Mit Hilfe der Bayes-Formel berechnet man nun "neue" Wahrscheinlichkeiten für Virusträger, wenn die untersuchte Person einen positiven Befund hatte.

Die direkte Wahrscheinlichkeitsaussage für Virusträger nach positivem Befund bedarf der Wahrscheinlichkeit, Virusträger zu sein vor allen Befunden, also a priori, i.e. W(V). Kann man diese Wahrscheinlichkeit nicht spezifizieren, so kann man das Ergebnis der Bayes-Formel nicht übernehmen. Es bleibt dann von klassischer Sicht nur ein statistischer Test über, etwa als Test von

$$H_0: \overline{V} \qquad \text{gegen} \qquad H_1: V \qquad \text{mit} \quad \alpha = \beta = 0{,}01$$

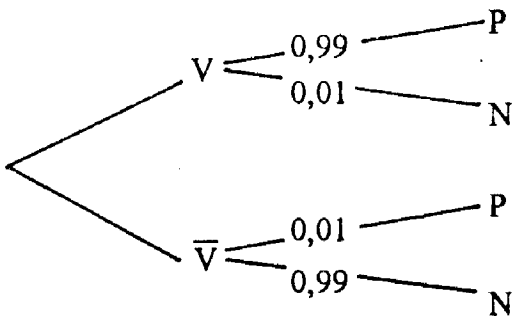

Abb. 13: Schematische Darstellung der Prävalenzen von Virusträgern und den Sicherheiten und Fehlerwahrscheinlichkeiten der Diagnosen.

Für Schwierigkeiten, die sich aus dem Bayes-Ansatz ergeben, etwa weil der behandelnde Arzt eine andere a priori-Wahrscheinlichkeit hat als der Patient und daher noch unangenehme und gefährliche Untersuchungen anschließen muß, während der Patient schon längst Klarheit hat, sowie für Schwierigkeiten mit der Interpretation des klassischen Tests sei auf Borovcnik (1992) verwiesen. Zusammenfassend kann man sagen, daß die sehr wichtige Frage nach der Wahrscheinlichkeit einer Fehldiagnose (der Patient ist positiv, aber dennoch virusfrei, z.B.) im Rahmen der klassischen Statistik nicht behandelt werden kann. Die sogenannten Irrtumswahrscheinlichkeiten α und β beziehen sich eigentlich auf die Untersuchungsprozedur und schließen den Patienten als einen (gleichwertigen) Fall von vielen ein und ergeben von daher auch recht magere Antworten den Patienten betreffend. Auch hierzu sei auf die angegebene Literatur verwiesen.

Der Disput zwischen klassischen Statistikern und Bayesianern kann so charakterisiert werden: Der Bayesianer verwendet auch qualitative Information und kann damit wichtige Fragen, die im klassischen Rahmen nicht behandelt werden können, beantworten.

3 Nichtparametrische Statistik

In diesem Abschnitt werden nicht-parametrische Methoden behandelt. Sie dienen zur Befreiung von Annahmen, die den klassischen Modellen zugrunde liegen und die oft nicht erfüllt sind oder deren Zutreffen gar nicht überprüft werden kann. Die Unterschiede zu klassischen Verfahren werden herausgestrichen, typische Methoden werden kurz dargestellt. Der Verlust an ursprünglich in den Daten vorhandene Information durch Übergang zu Vorzeichen oder Rängen ist dem Gewinn durch geringere Voraussetzungen gegenüberzustellen. Neuerdings bieten Resampling-Methoden einen vernünftigen Kompromiß.

a) Nicht-parametrische versus parametrische Verfahren

Tabellarisch soll ein kurzer Abriß über die unterschiedlichen Voraussetzungen und Ziele der beiden Ansätze gegeben werden. Illustrierende Beispiele folgen dann im nächsten Abschnitt.

Vorteil des nicht-parametrischen Ansatzes (NP) ist, das Modell braucht nicht spezifiziert zu werden, es braucht daher auch nicht geprüft zu werden, ob z.B. Normalverteilung für das untersuchte Merkmal zutrifft (wenn man dies als Modell gewählt hat) oder nicht. Die Ergebnisse passen daher auch, wenn die dem Modell zugrunde liegenden Annahmen nicht erfüllt sind. Die Methoden können auch dann angewendet werden, wenn das Meßniveau lediglich ordinal ist, d.h. die Daten können nur angeordnet werden, ihre Differenzen haben aber sachlich keine Interpretation. Dies ist etwa, ernst genommen bei Schulnoten der Fall, dies tritt aber häufig ein, wenn Daten nur durch Vergleich gewonnen werden, wenn also eine Rangfolge bestimmt wird. Diesen Vorteilen von weniger Input steht auch der Nachteil von weniger Output gegenüber. Die Fragen, die gestellt werden können, sind weniger präzise, man kann z.B. nicht nach der Größe von Unterschieden zwischen zwei zu vergleichenden Gruppen fragen, die Analyse muß sich lediglich darauf zurückziehen, ob man statistisch nachweisen kann, daß Unterschiede bestehen.

	NP	parametrisch
Daten	zufällig	zufällig
Modell	beliebig	bestimmte Familie von Verteilungen, etwa Normalvert.
Metrik	ordinal	intervallskaliert
Ziel	allgemein	auf Parameter fixiert

b) Nicht-parametrische Methoden

- *Vorzeichen-Test*

Das folgende Beispiel ist aus Lorenz (1988), wo man Details findet, hier sei aus Platzgründen der

Kontext und die Methode nur angedeutet. In einem chemotaktischen Versuch werden Käfer durch ein Gangsystem mit einer Verzweigung geschickt. In einem Seitenzweig befindet sich eine duftende chemische Substanz. Die Frage ist, ob der Käfer sich an der Verzweigung ohne besonderen Grund (d.h. zufällig) "entscheidet", welchen Zweig er geht, oder ob der eine Präferenz für einen der beiden Gänge hat. Damit man "confounder" ausschaltet, muß man die Substanz durch Zufall in einem der beiden Gänge placieren. Es könnte sich sonst ein Lerneffekt ergeben, oder der Käfer geht mit größerer Wahrscheinlichkeit nach rechts oder ... Zu testen ist also

H_0: Keine Seitenstetigkeit $\equiv p=0,5$ gegen H_1: Seitenstetigkeit vorhanden $\equiv p \neq 0,5$

Der datenerzeugende Prozeß wird so modelliert: $X_i = 1$ (+) mit Wahrscheinlichkeit p bzw. 0 (-) mit 1-p. Die Summe der Daten $X = X_1 + X_2 + ... + X_n$ hat dann eine Binomialverteilung mit n und p als Parameter, falls das Experiment ordentlich durchgeführt wird und damit die möglichen Confounder ausgeschaltet werden. Praktisch ist das sehr einfach und erfordert kaum Voraussetzungen. Über den Parameter der Binomialverteilung kann man dann gewöhnliche statistische Tests durchführen.

- *Ränge*

Das folgende Beispiel ist ebenfalls aus Lorenz. Es geht um einen Mastversuch von Kälbern und um die Untersuchung der mittleren täglichen Gewichtszunahme (in kg) bei unterschiedlicher Fütterung, insbesondere bei Fütterung mit Trockenmagermilch (TMM) bzw. frischer Magermilch (FMM) nach einem bestimmten Zeitraum; die Daten sind:

x TMM: 970, 1010, 1150, 1050, 1280, 1030; y FMM: 1000, 870, 970, 1020, 1130

Statt jetzt wie im parametrischen Verfahren die Differenzen der Mittelwerte geeignet zu normierten und den t-Test für unverbundene Stichproben durchzuführen (der auf die Normalverteilung und die gleichen Varianzen in beiden Fütterungsgruppen zurückgreifen muß), kann man auch die Daten der kombinierten Stichprobe der Größe nach anordnen und in dieser Folge den Daten Ränge zuordnen. Daraus ergibt sich für die x-Daten wie für die y-Daten eine Rangsumme R_x bzw. R_y. Ein Unterschied in den beiden Erwartungswerten μ_x und μ_y wird sich in den Rängen bemerkbar machen. Ist der Erwartungswert bei TMM sehr viel kleiner als bei FMM, so werden die Ränge für die x-Daten sehr klein ausfallen. Zu testen ist H_0: $\mu_x = \mu_y$ gegen H_1: $\mu_x \neq \mu_y$. Unter der Nullhypothese, daß keine Unterschiede in den beiden Erwartungswerten bestehen (und der zusätzlichen Oberhypothese, daß die Varianzen gleich sind), werden alle Zuordnungen aus der kombinierten Rangfolge zu den x- bzw. y-Daten gleich wahrscheinlich. Man kann daher diese gleichwahrscheinlichen Fälle der Größe der Rangsumme R_x nach anordnen und die z.B. 2,5% kleinsten und 2.5% größten Rangsummen zum Extrembereich zusammenfassen; Daten aus diesem Bereich gelten dann (auf dem 5%-Niveau) als nicht verträglich mit der Nullhypothese und führen zu deren Ablehnung.

Bei der Suche nach den kleinsten und größten x-Rangsummen hilft die Kombinatorik. Die Ergebnisse sind in ausführlichen Tabellen zum sogenannten Wilcoxon- bzw. Mann-Whitney-Test etwa in Lorenz (1988) enthalten. Der Wilcoxon-Test bedarf wie der t-Test auch der gleichen Varianzen, weil bei Verletzung dieser Voraussetzung die Auswirkung auf die Rangsummen nicht mehr erfaßt werden kann. Der Test hat ferner den Nachteil, daß durch die Reduktion des Meßniveaus auf Ränge manchmal ganz kleine und ganz große Differenzen einander gleich gemacht werden. Er hat aber den Vorteil, von der Normalverteilung, hier für die Gewichtszunahme der Tiere, absehen zu können. Dies könnte man bei 5 oder 6 Daten ohnehin nicht überprüfen.

- *Resampling-Methoden*

Das Beispiel mit dem Mastversuch soll nun wie folgt analysiert werden: In der kombinierten Stichprobe werden jetzt nicht die Ränge sondern die Summe der Originaldaten bestimmt. Unter der Nullhypothese haben wieder alle möglichen Zuordnungen aus der kombinierten Stichprobe zu den x- bzw. y-Daten dieselbe Wahrscheinlichkeit, Es geht also um die Suche nach den z.B. 2,5% kleinsten und 2.5% größten möglichen x-Summen. Dabei hilft keine Kombinatorik mehr, da ja nicht Ränge und damit fortlaufende Zahlen von 1 bis $n_x + n_y$ vorhanden sind, sondern eben die

Originaldaten. Man kann jedoch die Verteilung aller möglichen x-Summen durch Simulation approximativ bestimmen, man nimmt also eine Stichprobe von n_x Daten aus allen und bestimmt ihre Summe; man "resamplet" die vorhandene Stichprobe, daher der Name.

Die Vorteile der Resampling-Methode liegen auf der Hand. Man benötigt kaum Voraussetzungen für das Verfahren. Es ist begrifflich sehr einfach, man benötigt keine speziellen Verteilungen, die Normalverteilung wird eigentlich überflüssig, die eigentliche Lösung wird auf dem Computer durch Simulation herbeigeführt. An Nachteilen ergeben sich: Der Verlust an Information, noch immer; manche Probleme werden so nicht lösbar, insbesondere macht auch hier die Voraussetzung ungleicher Varianzen eine Schwierigkeit (man beachte, daß der t-Test für diesen Fall durch die sogenannte Welch-Korrektur der Freiheitsgrade adaptiert werden kann, siehe z. B. Lorenz, 1988); die Lösung erfordert den Computer und wird rechenintensiv und zeitaufwendig, wenn die Zahl der Daten steigt; der Verlust an Theorienotwendigkeit läßt sich in seinen Auswirkungen noch gar nicht abschätzen. Für weitere Beispiele und eine Erläuterung der Methode sowie deren Auswirkungen auf die Statistik siehe Borovcnik (1994). Die Resampling-Methoden sind didaktisch reizvoll - manches ist nicht (noch nicht) möglich.

4 Zusammenfassung

Die Diskussion der verschiedenen Ansätze und deren Vor- und Nachteile zeigt:

- Es gibt viele Wege, Informationen von Stichproben zu verallgemeinern
- Die Wege führen zu unterschiedlichen Zielen
- Die Begriffsvielfalt läßt die einzelnen Ansätze und deren Methoden viel besser verstehen
- Anwendungen drängen auf einen pluralistischen Ansatz
- Software zur Unterstützung ist nicht so einfach wie nötig

Für die Planung eines zukünftigen Curriculums kann der Autor folgende Empfehlung abgeben:

- Etwas Bayes, vor allem im diskreten Fall der medizinischen Diagnose
- Dann Nicht-parametrische Methoden - viel davon
- Schließlich klassische statistische Methoden - wie die Zeit vorhanden ist
- EDA ist interessant, jedoch soll der Umfang nicht übertrieben werden

Der Hintergrund zu den hier entwickelten Ideen ist vor allem in Borovcnik (1992), die Konkretisierung ist teilweise, insbesondere was die Anbindung an einige Bayes-Gedanken betrifft, in Laub e.a. (1988) enthalten. Durch die neueren Entwicklungen, insbesondere der Resampling-Methoden sind "mathematikfreiere" Ansätze zur statistischen Beurteilung in die Nähe gerückt. Nicht aufgehoben dadurch ist die indirekte Denkweise der Statistik, wonach man Daten aus der Sicht von mehreren möglichen Modellen bzw. Szenarien her beurteilt und eine Passung oder Ablehnung dieser Szenarien nach Wahrscheinlichkeiten vornimmt.

Literatur

Borovcnik, M.: Explorative Datenanalyse - Techniken und Leitideen. In: Didaktik der Mathematik 18 (1990), 61-80.

Borovcnik, M.: Stochastik im Wechselspiel von Intuitionen und Mathematik, Mannheim: Bibliographisches Institut 1992.

Borovcnik, M.: Der Einfluß des Computers auf die Statistik-Ausbildung. In: H. Friedl (Hrsg.): Was ist Angewandte Statistik. Grazer Mathematische Berichte Nr. 34, Graz: Technische Universität 1994.

Borovcnik, M.: Exploratory Data Analysis - A new approach to modelling. In: K. Houston e.a. (Hrsg.): Mathematical Modelling, Chichester: Alias Horwood 1996.

Borovcnik, M. und G. Ossimitz: Materialien zur Beschreibenden Statistik und Explorativen Datenanalyse, Wien: Hölder-Pichler-Tempsky 1987.

Kleiter, G. D.: Bayes-Statistik. Grundlagen und Anwendungen, Berlin: de Gruyter 1980.

Laub, J., M. Bernhard und M. Borovcnik: Mathematik für Bildungsanstalten, Bd.4, Wien: Hölder-Pichler-Tempsky 1988.

Lorenz, R.: Grundbegriffe der Biometrie, Stuttgart: G. Fischer 1988.