Deskriptive Statistik

Lehr- und Arbeitsbuch

Von
Universitätsprofessor
Dr. Georg Bol

6., überarbeitete Auflage

R. Oldenbourg Verlag München Wien

Bibliografische Information Der Deutschen Bibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über <http://dnb.ddb.de> abrufbar.

Rosenheimer Straße 145, D-81671 München
Telefon: (089) 45051-0
www.oldenbourg-verlag.de

Gedruckt auf säure- und chlorfreiem Papier
Herstellung: Books on Demand GmbH, Norderstedt

ISBN 3-486-57612-7
ISBN 978-3-486-57612-2

Vorwort

Die vorliegende Einführung in die Methoden der deskriptiven Statistik entstand aus Aufzeichnungen für das erste Drittel einer zweisemestrigen Vorlesung „Statistik für Wirtschaftswissenschaftler" aus den Jahren 1985 und 1988 an der Universität Karlsruhe. Sie richtet sich damit in erster Linie an Studienanfänger wirtschaftswissenschaftlicher Studiengänge als Ergänzung zu Vorlesungen entsprechenden Inhalts, aber auch zum Selbststudium.

Besonderer Wert wurde dabei auf eine systematische und übersichtliche Darstellung gelegt, wobei Beispiele die Vorgehensweise verdeutlichen sollen. Demgegenüber haben wir einer Einführung entsprechend auf Vollständigkeit bei den dargestellten Verfahren verzichtet; durch Literaturhinweise sind dem interessierten Leser Möglichkeiten zu weiterem Studium aufgezeigt, wegen der Fülle an Büchern und Zeitschriftenartikeln allerdings nur exemplarisch. Vorkenntnisse, die über den Bereich der Schulmathematik hinausgehen, werden nicht benötigt. Durch Übungsaufgaben und ihre Lösungen sind dem Leser Möglichkeiten der Selbstkontrolle gegeben.

Der Autor ist seinen Kollegen am Institut für Statistik und Mathematische Wirtschaftstheorie für ständige Gesprächsbereitschaft und viele Hinweise zu Dank verpflichtet. Ferner dankt er cand. Wi.-Ing. Steffen Hasse und cand. Inf. Karin Münzer für die Durchführung der Schreibarbeit. Herrn Diplom-Volkswirt Martin Weigert und dem Oldenbourg-Verlag danke ich für die Aufnahme des Buches in sein Programm und die reibungslose Zusammenarbeit.

Vorwort zur 2. Auflage

Das Erscheinen einer 2. Auflage hat in erster Linie Gelegenheit gegeben, die vielen Schreib- und Rechenfehler zu beseitigen[1]. Außerdem konnten auch einige Ergänzungen, Erläuterungen und Aktualisierungen der Beispiele vorgenommen werden. Daneben haben wir zu jedem Themengebiet noch zusätzliche Übungsaufgaben angefügt, so daß jetzt ein Lehr- und Arbeitsbuch in einem vorliegt. Leider stand das Textsystem, auf dem die erste Auflage geschrieben wurde, nicht mehr zur Verfügung. Daher mußte der Text gescannt und mit LaTeX neu aufbereitet werden. Diese mühevolle Arbeit haben Frau cand. Wi.-Ing. Monika Kansy und Herr Dipl.Ing. Jörn Basaczek in denkbar kurzer Zeit durchgeführt. Herr cand. Wi.-Ing. Edgar Hotz hat bei den Abbildungen, die teils für die neue Umgebung überarbeitet, teils - da neu hinzugekommen- neu erstellt werden mußten, einen Großteil der Arbeit übernommen. Ferner hat er einen Teil der Berechnungen kontrolliert und beim Register mitgeholfen. Herr Dipl.Wi.-Ing. Johannes Wallacher hat das Manuskript noch ein-

[1] Für Hinweise dazu bin ich vielen meiner Studenten sehr dankbar.

mal durchgelesen, weitere Beispiele eingefügt und den letzten Schliff bei der Formatierung vorgenommen. Auch Frau Rita Frank und meine Tochter Jutta haben beim Korrekturlesen noch etliche Fehler gefunden. Ihnen allen gilt mein herzlicher Dank. Ohne ihre Mitarbeit wäre die zweite Auflage nicht in dieser Form zustande gekommen. Die Zusammenarbeit mit Herrn Weigert und dem Oldenbourg-Verlag war - wie immer - problemlos und angenehm, wofür ich - wie immer - zu danken habe.

Vorwort zur 6. Auflage

15 Jahre nach Erscheinen der ersten Auflage war eine Überarbeitung des Buches und eine Aktualisierung der Beispiele mehr als angebracht. Auch musste der Wechsel von der Deutschen Mark zum Euro berücksichtigt werden. Die Gelegenheit einer Neuauflage wurde ferner dazu benutzt, um die Zeichnungen teilweise neu zu erstellen. Auch waren zwischenzeitlich wieder Hinweise auf Fehler eingegangen, für die jetzt die Korrekturen vorgenommen werden konnten. Dennoch sind sicherlich auch weiterhin Fehler vorhanden, für Hinweise bin ich selbstverständlich immer dankbar. Dem Verlag und insbesondere Herrn Weigert danke ich für die jederzeit erfreuliche Zusammenarbeit.

Georg Bol

Inhaltsverzeichnis

1 Einführung

Ohne sich dessen immer bewusst zu werden, kommt man mit Statistik im Leben nahezu ununterbrochen in Berührung. Dies beginnt bei der Geburt durch die Aufnahme in die Einwohnerkartei der Gemeinde oder Stadt, in der die Eltern amtlich gemeldet sind, wobei Daten wie Geburtsdatum, männlich/weiblich, Name der Eltern, Religionszugehörigkeit etc. für die sogenannte „Bevölkerungsstatistik" erfasst werden, und endet mit dem Tod. Beide Ereignisse finden ihren Niederschlag insbesondere in der Statistik über den Bevölkerungsstand der Gemeinde, die in regelmäßigen Abständen auch veröffentlicht wird. Als Beispiel für eine solche „Bevölkerungsstatistik" sei hier die **Fortschreibung** des Bevölkerungsstandes der Gemeinde Stutensee für den Monat Februar 1988, veröffentlicht im Mitteilungsblatt der Gemeinde, wiedergegeben (s. Abb. 1.1 nächste Seite).[1]

Zwischen diesen beiden extremen Ereignissen Geburt und Tod werden zum Beispiel Ereignisse wie Schuleintritt, Wechsel auf Gymnasium oder Realschule, Studienbeginn, Eintritt in das Berufsleben, etc. statistisch erfasst. Aber auch bei anderen Gelegenheiten kommt man –bewusst oder unbewusst– mit Statistik in Berührung. So z.B. bei Meinungsumfragen, Verkehrszählungen, Zulassung von Kraftfahrzeugen, Eintragungen in die Verkehrssünderkartei in Flensburg etc., nicht zuletzt natürlich auch bei Volkszählungen. Historisch gesehen können diese vielleicht als älteste dokumentarisch nachgewiesene statistische Betätigungen betrachtet werden. So verweisen Hartung et al. (2002) auf die im alten Testament (viertes Buch Moses, zweites Buch Samuel) erwähnten Volkszählungen, allgemein bekannt ist sicher die Volkszählung unter Kaiser Augustus.

Ein anderer Bereich, in dem in vielfältiger Weise Statistiken erstellt werden, ist der betriebswirtschaftliche Bereich. Jedes gut geführte Unternehmen wird neben einer Verkaufsstatistik, in der die Verkaufszahlen der einzelnen Artikel gegliedert nach Zeiträumen (z.B. Monaten) und möglicherweise Verkaufsgebieten, Vertretern etc. aufgeführt sind, eine Personalstatistik, Unfallstatistik etc. erstellen.

Nachdem nun der Begriff Statistik mehrfach verwendet wurde, ist es angebracht, genauer anzugeben, was Statistik ist und welche Aufgaben sie hat. Entstanden ist der Begriff Statistik vermutlich durch die Staatsbeschreibungen G. Achenwalls[2], die dieser „Statistik" nennt, was möglicherweise auf das lateinische Wort „status" (1. Zustand, 2. Staat) zurückzuführen ist.

In dem bisher benutzten Sinn ist eine Statistik eine **Zusammenfassung von**

[1] Auf eine Aktualisierung wurde verzichtet, da die Daten in dieser Form zur Zeit nicht mehr veröffentlicht werden.

[2] Gottfried Achenwall, 1719-1772, begründete durch seine Staatsbeschreibungen den Begriff Statistik.

Bevölkerungsstatistik der Gemeinde Stutensee

Bevölkerungsstand(Fortschreibung)

Monat Februar 1988

	Gesamt	davon					
		männlich			weiblich		
		ev.	kath.	sonst	ev.	kath.	sonst
I. Stand der Bevölkerung am 31. Januar 1988	19 321	5 521	3 001	1 059	5 938	3 090	712
II. Zugang							
a) durch Geburt	17	5	1	2	3	3	3
b) durch Zuzug	88	25	14	11	16	14	8
Summe Zugang	105	30	15	13	19	17	11
III. Abgang							
a) durch Tod	11	4	1	-	5	1	-
b) durch Wegzug	91	24	14	9	25	16	3
Summe Abgang	102	28	15	9	30	17	3
IV. Bevölkerungsstand: Ende Februar 1988	19 324	5 523	3 001	1 063	5 927	3 090	720

Abbildung 1.1 Bevölkerungsstatistik der Gemeinde Stutensee, Februar 1988 (Quelle: Amtsblatt der Gemeinde Stutensee).

Zahlen oder Daten, die gewisse Erscheinungen der Realität beschreiben. Eine derartige Statistik kann je nach ihrem Verwendungszweck unterschiedlich dargestellt sein. Zur Verdeutlichung seien noch einmal einige Beispiele aufgeführt.

- In der Bevölkerungsstatistik der Bundesrepublik bzw. der Länder und Kommunen werden alle dort gemeldeten lebenden Personen aufgegliedert nach Alter, Geschlecht, Religionszugehörigkeit und anderen Kriterien erfasst.
- Die Zulassungsstatistik von Kraftfahrzeugen enthält die Anzahl der zugelassenen Kraftfahrzeuge aufgeschlüsselt nach Typen.
- Die Verkaufsstatistik einer Unternehmung gibt die Verkaufszahlen der einzelnen Artikel aufgegliedert nach Monaten und evtl. weiteren Kriterien wieder.
- Durch Unfallstatistiken wird versucht, Unfallschwerpunkte und -ursachen zu erkennen.
- In den Naturwissenschaften werden Statistiken aufgestellt, um Gesetzmäßigkeiten über den Ablauf von Vorgängen herauszuarbeiten und nachprüfen zu können.

Zum anderen wird der Begriff Statistik aber auch als Bezeichnung für eine Wissenschaftsdisziplin benutzt. Häufig wird der Begriff „Statistik" als die **Gesamtheit aller Methoden (Lehre) zur Untersuchung und Beschreibung von Massenerscheinungen** umschrieben. Dabei sollte man die Statistik als eine **Hilfswissenschaft** auffassen, mit der die **Verbindung zwischen Empirie und Theorie** hergestellt oder zumindest reflektiert wird (vgl. Ferschl (1978), S.13). So gewinnt man z.B. in der Experimentalphysik die allgemeine Gesetzmäßigkeit aus experimentellen Untersuchungen mittels statistischer Methoden. Weitere Bereiche sind unter anderem die Chemie, Biologie, Astronomie, Medizin und insbesondere die Wirtschaftswissenschaft. Es geht also einmal um die **Erhebung, Aufbereitung** und **Betrachtung** der Daten, also die Verarbeitung von empirischem Datenmaterial, und zum anderen darum, aus diesem Datenmaterial **Schlussfolgerungen** zu ziehen, die für die **Entscheidungsfindung** von Bedeutung sind. Man unterscheidet dementsprechend zwischen **deskriptiver** (beschreibender) und **induktiver** (schließender) Statistik. Demnach sind also Statistiken in der ersten Bedeutung des Wortes, wie sie etwa in den Beispielen angegeben waren und wie sie auch in den statistischen Jahrbüchern der statistischen Bundes- und Landesämter herausgegeben werden, das Ergebnis einer Beschäftigung im **Bereich der deskriptiven Statistik**.

Der Aufgabenbereich der induktiven Statistik wird vielleicht am ehesten an zwei Beispielen deutlich.

1.1 Beispiel

Ein Hersteller von Blitzlichtbirnen möchte vor Auslieferung einer Partie von 10.000 Stück prüfen, wie hoch der Anteil der fehlerhaften (nicht funktionierenden) Stücke ist. Eine exakte Feststellung ist nicht möglich, da ja durch das Ausprobieren, also die Prüfung auf Funktionsfähigkeit, die Blitzlichtbirne unbrauchbar wird („zerstörende Kontrolle"). Einzige Möglichkeit ist also, eine Auswahl aus den Blitzbirnen der Partie zu treffen und diese zu prüfen. Dies kann etwa so erfolgen, dass 150 Stück ausgewählt und untersucht werden. Das Ergebnis sehe etwa wie folgt aus:

untersucht	150
davon:	
korrekt	144
fehlerhaft	6
Ausschussanteil der Auswahl	4 %

So weit handelt es sich um deskriptive Statistik. Es wird festgestellt, dass der Ausschussanteil bei den untersuchten 150 Stück 4% beträgt. Es stellt sich nun die Frage, inwieweit es berechtigt ist, davon auszugehen, dass auch der Ausschussanteil der Gesamtpartie 4% oder zumindest nahe bei 4% liegt. Eine solche Aussage kann ja offensichtlich falsch sein, da eine sichere Aussage über den Ausschussanteil nicht möglich ist, wenn nicht alle Stücke geprüft sind. Die induktive Statistik hilft hier weiter, sie liefert eine Aussage darüber, wie zuverlässig eine Übertragung der Ergebnisse der ausgewählten Teile auf die gesamte Partie ist.[3] Die Entscheidung - nämlich ob die Partie ausgeliefert wird oder nicht - wird dann aufgrund des Ergebnisses unter Berücksichtigung der Unsicherheitssituation und einer Bewertung der Konsequenzen erfolgen müssen.

1.2 Beispiel

Aus den Veröffentlichungen des statistischen Bundesamtes sind die Arbeitslosenzahlen sowie die Zahl der Erwerbstätigen seit der Gründung der Bundesrepublik Deutschland bekannt. Jeweils zu Monatsanfang werden die neuesten Zahlen für den letzten Monat vom Bundesamt (jetzt Bundesagentur) für Arbeit in Nürnberg bekanntgegeben. Daraus ergeben sich unmittelbar die Aufgaben:

1. die gegenwärtige Situation zu beurteilen,

[3] vgl. z.B. Uhlmann (1982), Statistische Qualitätskontrolle.

2. die zukünftige Entwicklung unter den gegebenen wirtschaftlichen und sonstigen Rahmenbedingungen zu prognostizieren.

Dabei muss bei 1. versucht werden, die Wirkungen der verschiedenen Einflüsse wie saisonal, witterungsbedingt, konjunkturell etc. zu separieren, um diese dann bei 2. entsprechend berücksichtigen zu können. Es sind also aus dem historischen Datenmaterial Schlussfolgerungen verschiedener Art zu ziehen.

Zusammenfassend sei noch einmal festgestellt:

1. **Deskriptive Statistik** befasst sich mit der Erhebung, Aufbereitung und Auswertung von Daten als solchen. Die Daten werden als historisches Faktum angesehen.

2. **Induktive Statistik** versucht, aus den in der deskriptiven Statistik erhobenen Daten Schlüsse auf die Ursachen und Gesetzmäßigkeiten zu ziehen, die diesen Daten zugrundeliegen. Die induktive Statistik ist hierdurch im Dienste der Entscheidungstheorie und damit Bestandteil der sogenannten statistischen Entscheidungstheorie.

Bindeglied zwischen deskriptiver und induktiver Statistik ist die **Wahrscheinlichkeitstheorie**, die sich systematisch mit dem Phänomen „Zufall" beschäftigt. Die Wahrscheinlichkeitstheorie kann also als Theorie der Gesetzmäßigkeiten des Zufalls bezeichnet werden. Was beim Zufall gesetzmäßig sein kann, wird am leichtesten bei Glücksspielen deutlich. So wird jedem Roulettespieler einleuchten, dass er, wenn er häufig spielt und immer den gleichen Einsatz auf „schwarz" setzt, langfristig keinen Vermögenszuwachs erwarten kann (da ja „schwarz" und „rot" langfristig betrachtet etwa gleichhäufig auftreten).

Ähnlich werden auch, wenn man häufig „mit Stichproben" (also einer zufällig, d.h. ohne jede Systematik ausgewählten Teilgesamtheit) kontrolliert, nur selten große Abweichungen zwischen dem Ausschussanteil der Partie und dem der Teilgesamtheit vorliegen. Mit Hilfe der Wahrscheinlichkeitstheorie können Phänomene dieser Art nicht nur genauer beschrieben, sondern auch quantifiziert, also zahlenmäßig erfasst werden.

Die Verhältnisse können etwa durch Abb. 1.2 verdeutlicht werden.

Ein weiteres Gebiet der Statistik ist seit rund fünfundzwanzig Jahren die sogenannte Explorative-Daten-Analyse, kurz EDA nach dem Titel des 1977 von J.W. Tukey veröffentlichten Buches genannt. Dabei handelt es sich um (neue) Methoden, die eigentlich meist dem Bereich der deskriptiven Statistik zuzuordnen sind und vielfach graphisch sind oder graphikartigen Charakter besitzen, aber ganz gezielt unter dem Gesichtspunkt eingesetzt werden, dass sie zu Ver-

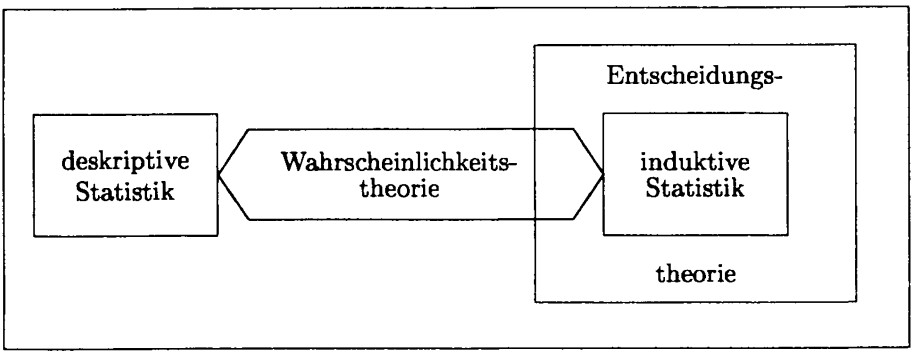

Abbildung 1.2 Beziehung zwischen deskriptiver und induktiver Statistik mittels der Wahrscheinlichkeitstheorie (nach Ferschl (1978), S.16).

mutungen führen, die dann mit Verfahren der induktiven Statistik überprüft werden können.

Die Aussagefähigkeit statistischer Methoden ist vielfach umstritten. Wenn mit sogenannten statistisch gesicherten Fakten argumentiert wird, wird als Gegenargument die Behauptung aufgestellt, dass mit Statistik alles bewiesen werden kann. In diese Richtung zielt auch die Frage nach den drei Formen der Lüge, die mit -die gemeine Lüge, -die Notlüge und -die Statistik beantwortet wird. Damit soll natürlich die Glaubwürdigkeit statistischer Aussagen in Zweifel gezogen werden. Der Grund für die Skepsis gegenüber statistisch begründeten Resultaten liegt darin, dass sich Aussagen scheinbar als falsch erwiesen haben bzw. dass Aussagen, die naturwissenschaftlich offensichtlich unsinnig sind, als statistisch beweisbar dargestellt werden. Dies kann grundsätzlich zwei Ursachen haben. Jede Aussage beruht neben dem vorhandenen Datenmaterial auch auf zusätzlichen Annahmen, die aber oft nicht explizit erwähnt werden und deren Korrektheit nicht immer gewährleistet ist.

So kann z.B. eine Verkaufsprognose auf der Annahme beruhen, dass die Marktbedingungen gleichbleibend sind. Ändert nun ein Konkurrent sein Angebot in Preis und/oder Ausstattung, so wird sich auch die Verkaufsprognose möglicherweise als falsch erweisen. Damit hat sich aber nicht die statistische Aussage als falsch gezeigt; diese lautete ja:

> Wenn die Marktbedingungen gleichbleibend sind, wird der Verkauf im Rahmen der statistischen Unsicherheit soundsohoch sein.

Der zweite Grund liegt - wie auch in anderen Bereichen - in einer unkorrekten Anwendung der Methoden. Dies lässt sich dann häufig dadurch verdeutlichen, dass man dieselbe (unkorrekte) Methode benutzt und zu offensichtlich

unsinnigen Ergebnissen kommt. Ein schönes - und dementsprechend häufig verwendetes - Beispiel dafür ist der angebliche Zusammenhang zwischen der Häufigkeit des Auftretens von Störchen und der Anzahl der Geburten:

In der zweiten Hälfte des vorigen Jahrhunderts ist über einen längeren Zeitraum für Südschweden eine gute Übereinstimmung zwischen der Entwicklung der Storchenbrütungen und der Geburtenrate festgestellt worden. Eine solche Übereinstimmung kann mit Hilfe des sogenannten Korrelationskoeffizienten festgestellt werden. Das Beispiel zeigt damit, dass dieser Korrelationskoeffizient nicht dazu benutzt werden kann - wie es häufig getan wird -, einen direkten kausalen Zusammenhang zwischen der Entwicklung zweier Größen nachzuweisen. Meist wird es nur eine gemeinsame Ursache geben, die möglicherweise auch nicht offen zu Tage tritt. Eine solche Ursache in diesem Beispiel zu finden, sei dem Leser überlassen.

Allerdings erlaubt die Statistik in besonders einfacher und vielfältiger Weise die **Manipulation** von Daten und damit die bewusste Täuschung. Dies wird häufig im Zusammenhang mit politischen Entscheidungen benutzt.[4]

Möglichkeiten hierzu bieten vor allem

- die Auswahl des Bezugspunktes,
- die Auswahl von Vergleichsgrößen,
- die Auswahl der Daten,
- die Art und Weise einer graphischen Darstellung.

Z.B. kann man für das Wachstum einer Branche durch die Auswahl eines geeigneten Vergleichsverfahrens möglicherweise völlig unterschiedliche Eindrücke erwecken. Betrachtet man weiter etwa die Entwicklung der Sozialausgaben eines Staates unabhängig von der Entwicklung anderer Größen wie Steuereinnahmen, sonstigen Ausgaben etc., so entsteht in vielen Fällen ein völlig falscher Eindruck. Auch bei der Auswahl der Daten insbesondere bei Zeitreihen ergeben sich vielfältige Möglichkeiten zur Täuschung, etwa indem man den Beginn der Berichterstattung geeignet wählt, oder durch die Auswahl der Zeitabstände geeignete Wirkungen erzielt (vgl. Abb. 1.3 und 1.4).

Lässt man eine Zeitreihe (s. Abb. 1.3), deren Daten etwa seit dem Zeitpunkt t_1 vorliegen, erst in t_2 oder t_3 beginnen, so erhält der Betrachter intuitiv eine völlig falsche Vorstellung von der Entwicklung der dargestellten Größe. Insbesondere wird der zu vermutende periodisch schwankende Verlauf bei einem Beginn in t_3 nicht deutlich. Gegenüber einem Beginn in t_2 wird dabei auch ein relativ

[4] vgl. hierzu insbesondere Huff (1973). Eine amüsante Lektüre dazu ist auch Krämer, 1998a.

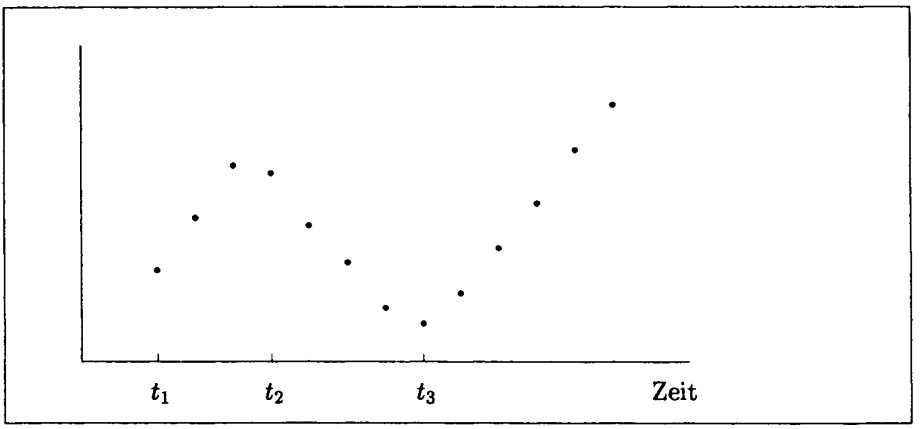

Abbildung 1.3 Manipulation durch Auswahl des Anfangszeitpunktes einer Zeitreihe.

gleichmäßiges Wachstum vorgetäuscht und kaschiert, dass bereits früher ein hohes Niveau dieser Größe erreicht war.

Auch durch die Auswahl von Daten ist eine Verfälschung möglich, wie dies z.B. in der Darstellung 1.4 bei Auswahl der eingekreisten Punkte augenfällig wird. Besonders vielfältig sind die Manipulationsmöglichkeiten, die man unter Ausnutzung optischer Täuschungen bei graphischen Darstellungen erhält. Auf Grund der Darstellung werden beim Betrachter Assoziationen geweckt, die durch das Datenmaterial nicht gerechtfertigt sind. Eine Zusammenstellung solcher Methoden mit Beispielen findet man bei Abels und Degen (1981).

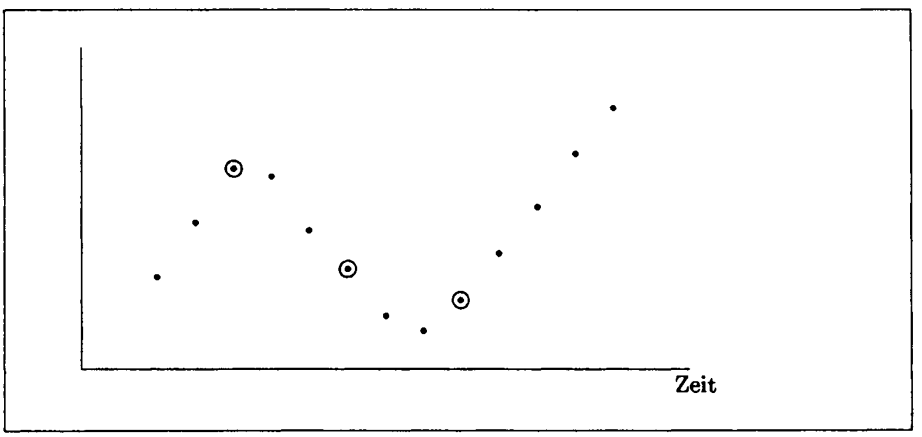

Abbildung 1.4 Manipulation durch Auswahl von Beobachtungszeitpunkten.

2 Grundbegriffe

Bei jeder statistischen Erhebung sind eine Reihe grundsätzlicher Überlegungen durchzuführen, bevor die Einzelheiten technischer und organisatorischer Art festgelegt werden können. Betrachtet man als typisches Beispiel eine Volkszählung [1], bei der ja nicht nur das Volk gezählt, d.h. die Zahl der in der Bundesrepublik lebenden Bürger festgestellt werden soll, sondern bei der auch Informationen über eine Vielzahl von Eigenschaften, Gewohnheiten, äußeren Umständen etc. dieser Bürger ermittelt werden sollen. Ganz offensichtlich ergeben sich unmittelbar folgende Fragen, die vorrangig der Klärung bedürfen:

- Wer soll befragt werden?
- Welche Fragen sollen gestellt werden?
- Welche Antwortmöglichkeiten sind zugelassen?

Durch die Beantwortung von Frage 1 werden die Objekte der statistischen Untersuchung festgelegt, man spricht hierbei von den statistischen Einheiten. Allgemein ist also eine **statistische Einheit** das Einzelobjekt einer statistischen Untersuchung.

In der Regel ist man jedoch nicht an den spezifischen Eigenschaften der Einzelobjekte interessiert, sondern an Informationen über das Gesamtbild. Zum Beispiel

- besteht bei einer Volkszählung nicht die Absicht zu erfahren, ob der Bürger Karl-Heinz Müller ledig, verheiratet, verwitwet oder geschieden ist, welche Einkünfte und Nebeneinkünfte er hat, welche Freizeitbeschäftigungen er bevorzugt usw. Vielmehr möchte man z.B. wissen, wieviel Prozent der befragten Bevölkerung ledig, verheiratet, verwitwet bzw. geschieden sind, wie die Einkommensstruktur der Gesamtbevölkerung ist, usw.
- ist bei der Partie Blitzlichtbirnchen nicht von Bedeutung, ob das Birnchen, das in der rechten unteren Ecke der Schachtel liegt, funktioniert oder nicht. Von Bedeutung ist, wieviel Prozent der Birnchen nicht funktionieren, also wie groß der Ausschussanteil ist.

Die Gesamtheit der statistischen Einheiten bzw. der Einzelobjekte der statistischen Untersuchung nennt man **statistische Masse, Grundgesamtheit**

[1] Die letzte Volkszählung in der Bundesrepublik Deutschland fand zum Stichtag 27. 05. 1987 statt. Wegen der niedrigen Akzeptanz in der Bevölkerung dürfte auf weitere Totalerhebungen in der Zukunft verzichtet werden.

oder **Population**. Man hat also bei jeder statistischen Untersuchung die statistische Masse genau festzulegen. Diese „Konstruktion" der Grundgesamtheit kann in zwei Schritten erfolgen (vgl. Ferschl (1978), S.17):

a) Die Abgrenzung der Grundgesamtheit.
   Von jedem in Betracht kommenden Objekt (Gegenstand der Umwelt) wird festgestellt, ob es zur Grundgesamtheit gehört oder nicht.

b) Die Festlegung der Auswahleinheit („Identifikation").

Die einzelnen Schritte werden vielleicht am ehesten an einem Beispiel deutlich. Bei einer Volkszählung sollen beispielsweise alle an einem bestimmten Tag x mit erstem Wohnsitz in der Bundesrepublik Deutschland gemeldeten Personen über 18 Jahre erfasst werden. Damit ist die exakte Abgrenzung des Personenkreises erfolgt. Die Erfassung kann nun in der Art erfolgen, dass pro Haushalt ein Fragebogen ausgefüllt wird oder aber jede Einzelperson einen Fragebogen erhält. In beiden Fällen werden die Daten über den gewünschten Personenkreis erhoben. Im ersten Fall besteht jedoch die Grundgesamtheit aus den einzelnen Haushalten, im zweiten Fall ist jede Person ein Einzelobjekt, d.h. die Grundgesamtheit besteht aus der Menge aller am Tag x mit erstem Wohnsitz in der Bundesrepublik Deutschland gemeldeten Personen.

Die Abgrenzung von Grundgesamtheiten oder statistischen Massen muss nach den folgenden drei Kriterien eindeutig vollziehbar sein:

- sachlich
- räumlich
- zeitlich

Im Falle der in der Fußnote erwähnten Volkszählung erfolgt dies durch die Angabe des „Stichtages" 27.05.1987 (zeitlich), die Beschränkung erster Wohnsitz in der BRD (räumlich) und Personen über 18 Jahre (sachlich). Auch bei der Identifikation können diese Kriterien herangezogen werden, z.B. durch die sachliche Identifikation: Einzelobjekte sind Personen (nicht Haushalte).

Die beiden Schritte Abgrenzung und Identifikation werden häufig in der Literatur nicht getrennt und in der Praxis auch nicht getrennt vollzogen. Beide Schritte können jedoch nicht unproblematisch sein, da ja das Ergebnis der Untersuchung stark von ihnen abhängen kann. Sollen z.B. bei der Untersuchung des Freizeitverhaltens von Studenten Doktoranden (also Studenten mit Abschluss) miterfasst werden oder nicht (Abgrenzungsproblem)? Oder z.B. bei einer Betriebsstatistik örtlich getrennte Arbeitsstätten einzeln aufgeführt werden, auch wenn sie einem einzigen Unternehmen angehören und/oder eine gemeinsame Betriebsorganisation besitzen (Identifikationsproblem)?

2.1 Beispiele

a) Untersuchung des Wählerverhaltens in Baden-Württemberg für die Bundestagswahl 2006.

Statistische Masse sind alle wahlberechtigten Bürger des Landes, d.h. alle Bürger, die am Tag der Wahl (also nicht am Tag der Untersuchung) die Voraussetzungen dafür erfüllen, in Baden-Württemberg ihre Stimme abgeben zu dürfen (deutsche Staatsbürgerschaft, Mindestalter 18 Jahre, Eintrag ins Wählerverzeichnis etc.).

b) Untersuchung der Verkehrsdichte einer bestimmten Straße.

Die Verkehrsdichte einer Straße kann z.B. dadurch festgestellt werden, dass an bestimmten ausgewählten Beobachtungspunkten die passierenden Fahrzeuge – eingeteilt in Gruppen wie LKW, PKW, Krafträder etc. – in vorgegebenen Zeiträumen gezählt werden. Als statistische Einheiten können hier die Beobachtungspunkte zu einem gegebenen Zeitraum betrachtet werden; diese Beobachtungspunkte werden ja auf den zu diesem Zeitraum durchfließenden Verkehr untersucht. Derselbe Beobachtungspunkt und ein anderer Zeitraum sind damit eine andere statistische Einheit.

c) Untersuchung der Abfüllmenge einer automatischen Abfüllanlage.

Statistische Einheiten sind hier die abgefüllten Flaschen. Grundgesamtheit oder statistische Masse ist die Menge aller abgefüllten Flaschen. Dabei ist natürlich der Untersuchungszeitraum eindeutig festzulegen.

Betrachtet man nochmals die Bevölkerungsstatistik aus Abbildung 1.1, so sieht man, dass in dieser Tabelle zunächst vier verschiedene statistische Massen erfasst sind: Einmal enthält sie in Zeile 1 den Stand der Bevölkerung zum Zeitpunkt 31.1.1988, $24.\underline{00}$. Hier ist also die Anzahl der Einwohner zu diesem Zeitpunkt insgesamt und aufgegliedert nach den angeführten Merkmalen angegeben. Statistische Masse ist die genannte Einwohnerschaft. Die zeitliche Abgrenzung wird durch den Stichtag (genauer „Stichzeitpunkt") gegeben. Unter II. sind die Zugänge erfasst, statistische Einheiten sind hier alle Zugänge im Zeitraum Februar, wobei hier eine weitere Differenzierung nach Geburt und Zuzug vorgenommen ist. Zeitliche Abgrenzung ist hier der Zeitraum Februar. III. erfasst die Abgänge, statistische Einheiten sind alle Abgänge. Zeitliche Abgrenzung ist wieder der Zeitraum Februar. Schließlich gibt IV. den Bevölkerungsstand zum Zeitpunkt 29.2., $24.\underline{00}$ an. Statistische Masse ist hier analog zu I. die gesamte Einwohnerschaft zum Stichtag. Als zeitliche Abgrenzung liegt wieder ein Zeitpunkt vor.

Die zeitliche Abgrenzung einer statistischen Masse kann also einerseits ein Zeitpunkt, andererseits ein Zeitraum sein.

Statistische Massen, deren zeitliche Abgrenzung ein Zeitpunkt ist, heißen **Bestandsmassen**; statistische Massen, deren zeitliche Abgrenzung ein Zeitraum ist, heißen **Ereignismassen**. Die jeweiligen statistischen Einheiten werden **Bestands-** bzw. **Ereigniseinheiten** genannt.

Betrachtet man das Beispiel der Bevölkerungsstatistik, so sieht man, dass jeder der dort erfassten Personen eine „Verweildauer" zugeordnet werden kann, nämlich der Zeitraum, in dem die Person in der Gemeinde wohnhaft (und auch gemeldet) ist. Dabei ist für die Zugehörigkeit zu den Bestandsmassen (I, IV) entscheidend, ob der Zeitpunkt der zeitlichen Abgrenzung („Stichtag") in die Verweildauer fällt oder nicht, während für die Ereignismassen ausschlaggebend ist, ob der Beginn (bei II) bzw. das Ende (bei III) der Verweildauer in den Zeitraum der zeitlichen Abgrenzung fällt. Beginn bzw. Ende der Verweildauer sind Zeitpunkte, in denen etwas geschieht, in denen sich etwas verändert („Ereignisse").

Jeder Bestandseinheit ist also eindeutig ein Zeitraum (die Verweildauer) zugeordnet, während jeder Ereigniseinheit ein Zeitpunkt (nämlich der des Eintretens des Ereignisses) zugeordnet ist. Dabei kann man noch unterscheiden, ob es sich bei dem Zeitpunkt um den Beginn oder das Ende der Verweildauer handelt. Im ersten Fall spricht man von Zugangseinheit (bzw. entsprechend von Zugangsmasse) im zweiten Fall von Abgangseinheit (bzw. Abgangsmasse).

2.2 Beispiele:

Bestandsmassen	Ereignismassen	
	Zugangsmasse	Abgangsmasse
Bevölkerung	Geburten	Todesfälle
zugelassene Kfz	Anmeldungen	Abmeldungen
Lagerbestand	Anlieferungen	Auslieferungen
Kassenbestand	Einzahlungen	Auszahlungen
Touristen	Ankunft	Abreise

Die Beispiele zeigen, dass über die Verweildauer der statistischen Einheiten, die ja durch die Ereignisse „Beginn" und „Ende" festgelegt und begrenzt ist, ein Zusammenhang zwischen Bestands- und Ereignismassen besteht. Dies kann graphisch folgendermaßen verdeutlicht werden:

Die Einheiten werden durchnumeriert und ihre Verweildauer durch einen Streckenzug oberhalb einer Zeitachse abgetragen (s. Abb. 2.1).

Numeriert man die Einheiten nach ihrem Zugangszeitpunkt und lässt die „Verweillinien" auf einer (virtuellen) Achse in spitzem Winkel (z.B. mit 45° Neigung) zur Zeitachse beginnen, so kann man am Abstand zweier Linien die Zeit-

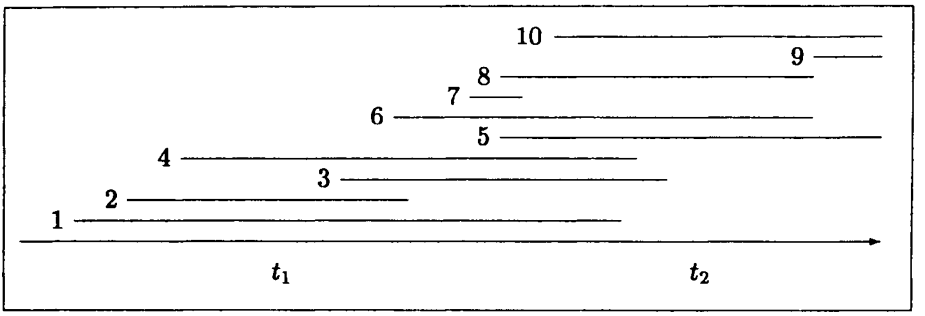

Abbildung 2.1 Graphische Darstellung von Verweildauern.

differenz des Zugangs erkennen (s. Abb. 2.2). Nachteile entstehen bei dieser Darstellung, wenn statistische Einheiten denselben Zugangszeitpunkt haben, da sich dann die Verweillinien überdecken. (Die ursprüngliche Numerierung ist jeweils in Klammer angegeben.)

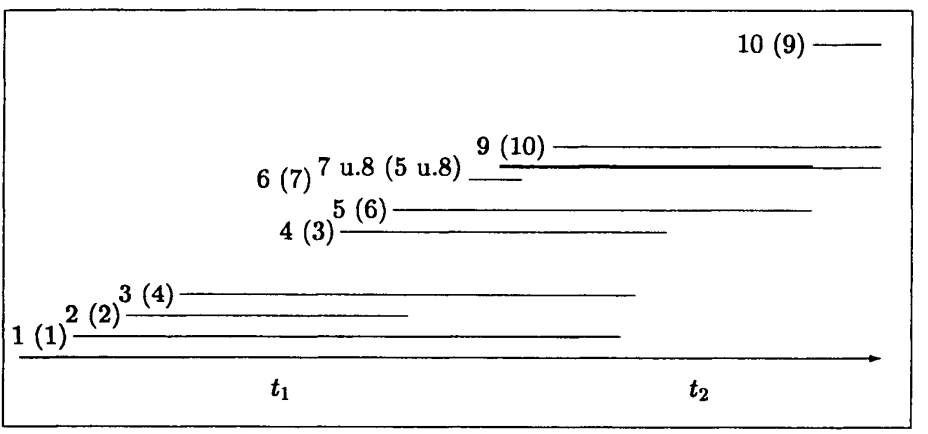

Abbildung 2.2 Graphische Darstellung von Verweildauern (2. Methode).

Zum Zeitpunkt t_1 ergibt sich aus der Graphik als Bestandsmasse die Masse mit den Einheiten 1, 2 und 4 (in der Ausgangsnumerierung). Für den Zeitraum von t_1 bis t_2 erhält man die Zugangsmasse aus den Einheiten 3, 5, 6, 7, 8 und 10 und die Abgangsmasse aus den Einheiten 1, 2, 3, 4 und 7. Die Bestandsmasse zum Zeitpunkt t_2 besteht ersichtlich aus den Einheiten 5, 6, 8 und 10.

Die Bestandsmasse zum Zeitpunkt t_2 erhält man auch, indem man zunächst zur Bestandsmasse zum Zeitpunkt t_1 ($\{1, 2, 4\}$) die Zugangseinheiten 3, 5, 6, 7, 8, 9 und 10 hinzufügt ($\{1, 2, 3, 4, 5, 6, 7, 8, 10\}$) und die Abgangseinheiten 1, 2, 3, 4 und 7 entfernt. Man erhält damit die Bestandsmasse zum Zeitpunkt t_2, die statistische Masse $\{5, 6, 8, 10\}$ wie oben. Dieses Verfahren wird Fortschreibung

genannt.

Man erhält damit als **Fortschreibungsformel**:

a) für die statistischen Massen:
Endbestandsmasse =
Anfangsbestandsmasse $\cup$ Zugangsmasse $\setminus$ Abgangsmasse.

b) für die Anzahl der statistischen Einheiten (zahlenmäßige Fortschreibungsformel):

$$\text{Endbestand} = \text{Anfangsbestand} + \text{Zugang} - \text{Abgang},$$

wobei also mit Endbestand usw. die Anzahl der statistischen Einheiten der Endbestandsmasse usw. bezeichnet sei.

Die Fortschreibungsformel erleichtert die zahlenmäßige Erfassung von Bestandsmassen enorm, da i.a. die Bestandsmassen gegebenüber den Ereignismassen groß sind (vgl. z.B. die Bevölkerungsstatistik) und die Ereignismassen ohnedies erfasst werden. Andererseits birgt sie insofern auch eine Gefahr in sich, da auch jeder Fehler (Rechenfehler, nicht erfolgte oder falsche Erfassung, übersehene Löschung etc.) fortgeschrieben wird. Deshalb wird von Zeit zu Zeit eine Neuerfassung (Volkszählung, körperliche Inventur bei Lagerbeständen, etc.) unumgänglich, wenn man sicher sein will, mit exakten Werten zu arbeiten.[2]

Im allgemeinen interessiert man sich – wie auch im Beispiel der Volkszählung oben – bei einer statistischen Untersuchung nicht nur für die Anzahl der statistischen Einheiten, sondern auch für Eigenschaften dieser statistischen Einheiten. So ist ja auch in der Bevölkerungsstatistik die Grundgesamtheit weiter aufgegliedert worden nach Geschlecht und Religionszugehörigkeit. Auch bei der Volkszählung von 1987 wurde eine Vielzahl von Eigenschaften abgefragt.[3] Mit Hilfe dieser Eigenschaften können die statistischen Einheiten beschrieben werden. Man verwendet hierfür die Bezeichnung **Merkmal**.

2.3 Beispiele für Merkmale

- Statistische Einheit: Student;
  Merkmale: Alter, Geschlecht, Haarfarbe, Studienfach, Größe, Gewicht, ...
- Statistische Einheit: Industriebetrieb;

[2] Dies setzt natürlich voraus, dass bei der Neuerfassung keine Fehler unterlaufen, was häufig auch nicht erreicht werden kann.

[3] Auf die Problematik, die damit verbunden ist, soll hier nicht eingegangen werden, da es sich nicht um ein Problem der statistischen Theorie handelt.

Merkmale: Zahl der Beschäftigten, Rechtsform, Umsatz, Größe des Betriebsgeländes, Standort, ...

- Statistische Einheit: Familienhaushalt;

  Merkmale: Kinderzahl, Einkommen, Anzahl der Berufstätigen, Anzahl der PKW, Ausgaben für Lebensunterhalt, ...

Merkmal ist entsprechend diesen Beispielen nicht eine individuelle Eigenschaft oder Ausprägung, wie blond als Haarfarbe, 19 Jahre als Alter, Wirtschaftswissenschaften als Studienfach etc., sondern die Möglichkeit mit Hilfe dieser Eigenschaften die statistischen Einheiten zu beschreiben.

Ein **Merkmal** ist also eine **Beschreibungsmöglichkeit** für die statistischen Einheiten der betrachteten statistischen Masse.[4] Ein Merkmal muss dabei nicht auf die statistische Masse beschränkt sein, die gerade untersucht wird. So kann das Merkmal Alter oder Geschlecht beispielsweise nicht nur als Beschreibungsmöglichkeit für die Studenten einer Universität oder einer anderen Hochschule herangezogen werden, sondern auch für weit größere Personenkreise.

Die spezielle Eigenschaft, die eine statistische Einheit bzgl. eines Merkmals annimmt („trägt"), nennt man **Merkmalsausprägung**, die statistische Einheit dementsprechend auch **Merkmalsträger**. Merkmalsausprägungen sind spezielle Eigenschaften oder Werte, mit denen die Beschreibung der statistischen Einheit erfolgt.

Vor einer statistischen Untersuchung muss also festgelegt werden, welche Merkmale untersucht („erhoben") werden. Dies hängt natürlich vom Thema oder der Aufgabe der Untersuchung ab. Die Güte der Untersuchungsergebnisse wird offensichtlich ganz wesentlich davon beeinflusst, ob die adäquaten Merkmale erhoben werden. Welche dies sind, ist in vielen Fällen nicht offenkundig (Welche Merkmale sollen beispielsweise erhoben werden, wenn es um die gesundheitlichen Folgen des Rauchens geht?).

Die Menge der theoretisch möglichen Merkmalsausprägungen für ein bestimmtes Merkmal – die bei der Untersuchung tatsächlich auftretenden sind vorher häufig nicht bekannt – ist einerseits bei verschiedenen statistischen Massen unterschiedlich, hängt aber andererseits auch von der Art und der Genauigkeit der Erhebung ab. So kann das Alter von Personen auf den Tag genau, nach vollendeten Lebensjahren oder – seltener – auf Jahre auf- bzw. abgerundet erhoben werden. Durch die Festlegung der Menge der zugelassenen Merkmalsausprägungen wird die oben angesprochene Frage nach den zugelassenen Antwortmöglichkeiten beantwortet.

[4] Auch die Abgrenzung der statistischen Masse erfolgt mit Hilfe von Merkmalen, man spricht dann von Identifikationsmerkmalen.

2.4 Beispiele

Merkmal	Merkmalsausprägungen
Geschlecht	männlich, weiblich
Religionszugeh.	evangelisch, katholisch, sonstige
Note	sehr gut, gut, befriedigend, ausreichend, nicht ausreichend oder 1,0 bis 5,0
Alter	Natürliche Zahlen (Anzahl von Jahren bzw. Tagen)
Größe	Positive reelle Zahlen[5] oder Positive Dezimalzahlen mit maximal zwei Kommastellen[6] oder Natürliche Zahlen[7]

Bei der Festlegung der Merkmalsausprägungen ist auch zu beachten, ob jeder statistischen Einheit genau eine Merkmalsausprägung zugeordnet ist (nicht häufbares Merkmal), oder ob bei einem oder mehreren Merkmalsträgern auch zwei oder mehr Ausprägungen möglich sind (häufbares Merkmal).

Nicht häufbares Merkmal: Eindeutige Zuordnung der Merkmalsausprägungen zu den Merkmalsträgern.

Häufbares Merkmal: Mehrdeutigkeit bei der Zuordnung von Merkmalsausprägung zu Merkmalsträger ist möglich (Mehrere Antworten auf eine Frage sind möglich.). Eine statistische Einheit kann mehrere Merkmalsausprägungen tragen.

2.5 Beispiele für häufbare Merkmale

a) Erlernter Beruf (z.B. Maschinenschlosser und Diplomingenieur).

b) Unfallursache (z.B. überhöhte Geschwindigkeit und Glatteis).

c) Hobby.

Häufbare Merkmale können dadurch auf nicht häufbare Merkmale zurückgeführt werden, dass man als neue Merkmalsausprägungen alle möglichen Kombinationen der ursprünglichen Merkmalsausprägungen einführt. Aus diesem

[4] ohne vorher festgelegte Genauigkeit.
[5] in Meter und Zentimeter.
[6] in Zentimeter.