

# 1 Was ist Statistik?

- 1.1 Aufgaben der Statistik
- 1.2 Zum Begriff der Statistik
- 1.3 Statistische Anwendungen
- 1.4 Statistische Methoden
- 1.5 Wichtige Grundbegriffe

## 1.1 Aufgaben der Statistik

Mit den Ergebnissen statistischer Auswertungen werden wir fast täglich konfrontiert: Umfragen zum Wahlverhalten der wahlberechtigten Deutschen, zum Beliebtheitsgrad von Politikern, zum Freizeitverhalten, zum Zigaretten-, Bier- und Fleischkonsum, aber auch Angaben im Wetterbericht, Meldungen zur aktuellen Konjunkturlage, neueste Daten zur Entwicklung der Arbeitslosigkeit in Deutschland ..., das alles sind Beispiele für die Nutzung der Statistik im Alltag.

Darüber hinaus finden die Statistiker breite Einsatzfelder in allen möglichen wissenschaftlichen Bereichen: Überall da, wo wissenschaftliche Erkenntnisse auf der Basis empirischer Informationen gewonnen werden, sind Statistiker mit ihren Erhebungs-, Auswertungs- und Analysemethoden an der Arbeit. Das gilt in fast allen Bereichen der Sozial- und der Wirtschaftswissenschaften, in der Psychologie und in der Medizin, in der Geographie und in den Sprachwissenschaften und beispielsweise in allen naturwissenschaftlichen Disziplinen, in denen experimentell gearbeitet wird.

Man kann die Ergebnisse der statistischen Arbeit nur dann richtig verstehen und bewerten, wenn man weiß, wie die einzelnen statistischen Methoden funktionieren, und was die Begriffe, mit denen gearbeitet wird (Mittelwert, Streuung, statistische Signifikanz, Prognose, Korrelation usw.), überhaupt bedeuten. Vor allem muss man auch eine Vorstellung davon gewinnen, welche Probleme bei der Ermittlung derartiger Statistiken auftreten und wie fragwürdig deshalb manche statistischen Angaben sind –, glücklicherweise nicht alle! Es besteht die Gefahr, dass man als Nichtfachmann von der Flut unprüfbarer statistischer Einzelinformationen überrollt wird, wenn man sich nicht darum bemüht, einen Überblick darüber zu gewinnen, wie diese Informationen zustande kommen, wo „gemogelt“ werden könnte, und was die präsentierten Ergebnisse deshalb aussagen können (und was nicht). Insbesondere muss man wissen, welche statistischen Auswertungsmethoden bereitstehen, was man damit machen kann, wie sie eingesetzt werden, und welche Befunde erzielt werden können. Deshalb werden in diesem Buch die wichtigsten dieser Methoden vorgestellt.

Der Statistiker hat es fast immer mit umfangreichen Datenbeständen zu tun, die ausgewertet werden müssen. Deshalb empfiehlt sich der Rechnereinsatz. Während in früheren Jahren die rechnergestützte Datenauswertung eine Angelegenheit für Spezialisten war, haben die rasche Verbreitung preiswerter und leistungsfähiger Rechner einerseits und die Möglichkeiten der Nutzung komfortabler Software andererseits dazu beigetragen, dass auch der gelegentliche Statistikanwender und der Anfänger den Rechner- und Softwareeinsatz bevorzugen wird. Auf diese Weise nämlich ist es möglich, sich von der Last eventuell aufwendiger, komplizierter und umfangreicher Rechenarbeiten zu befreien. Mehr noch: Leicht können Alternativberechnungen durchgeführt werden, Änderungen im Ausgangsdatenbestand führen nicht automatisch zu der Notwendigkeit, alles neu berechnen zu müssen, und Rechenfehler werden zuverlässig vermieden. Aber das ist noch immer nicht alles. Was besonders für den statistischen Laien von Bedeutung ist: Er muss nicht mehr die theoretischen Hintergründe statistischer Verfahren beherrschen, wenn er bestimmte Analysemethoden einsetzen will – die Computersoftware bietet ihm die wichtigsten Verfahren an, und was sich an mathematischen Geheimnissen dahinter verbirgt, braucht ihn nicht mehr zu interessieren. Allerdings, das soll in diesem Zusammenhang nicht verschwiegen werden, wer die Hintergründe kennt, auch wenn diese Kenntnisse nicht mehr zwingend erforderlich sind, der versteht natürlich mehr von den Methoden, die er einsetzt, und kann deshalb auch die Ergebnisse leichter und angemessener interpretieren.

Als leistungsfähiges Programm zur Auswertung statistischer Massendaten bietet sich SPSS an. Die Präsentation statistischer Methoden und die Erörterung ihrer Einsatzmöglichkeiten anhand passender Praxisbeispiele geht also in diesem Buch Hand in Hand mit der Besprechung der jeweils geeigneten SPSS-Prozeduren. Damit werden Sie in die Lage versetzt – sofern Sie glücklicher Besitzer dieses Softwareprogramms sind -, die hier besprochenen statistischen Methoden auch sofort einsetzen und die gebotenen Anwendungsbeispiele sofort nachvollziehen zu können.

## 1.2 Zum Begriff der Statistik

Lassen Sie mich mit einer kurzen Erörterung des Begriffs Statistik beginnen. Tabellen mit Angaben (Daten) zur Konjunkturlage, grafische Darstellungen etwa der Wählerstruktur bei einer Bundestagswahl, Angaben zur Entwicklung der Arbeitslosenquote in den letzten Jahren, Auswertungen einer Marktumfrage, Messungen von produzierten Werkstücken in der betrieblichen Qualitätskontrolle u.ä. werden im allgemeinen mit dem Begriff Statistik überschrieben. Genaugenommen sind solche Statistiken aber nur die Ergebnisse statistischer Arbeit. Man erhält diese Ergebnisse, wenn man sich *statistischer Methoden* bedient. Anders ausgedrückt: Wenn man die Ergebnisse statistischer Arbeit kritisch beurteilen will, wenn man sie weiter verwenden will oder aber, wenn man selbst Statistiken erstellen will, dann muss man diese Methoden zuvor kennenlernen. Deshalb

wollen wir im folgenden unter „Statistik“ vor allem die „statistischen Methoden“ verstehen.

Für welchen Zweck werden statistische Methoden benötigt? Folgen wir einer Definition von *R. Wagenführ*, kann man diesen Zweck folgendermaßen umreißen:

*Statistische Methoden* werden benötigt, um Massenerscheinungen zu quantifizieren, zu beschreiben, zu beurteilen, Schlüsse aus ihnen zu ziehen und ihre Erklärung vorzubereiten.

In dieser Definition tauchen einige Begriffe auf, die näher beleuchtet werden müssen, damit man sie richtig verstehen kann.

*Massenerscheinungen* sind Sachverhalte, die man beobachten oder anders erfassen kann, und die in großer Zahl auftreten. Das soll heißen, dass sich der Statistiker in der Regel nicht für Einzelfälle, für einzelne Daten oder für singuläre Angaben interessiert.

*Quantifizierung von Massenerscheinungen* bedeutet, dass der Statistiker die interessierenden Tatbestände oder Sachverhalte mit Zahlen zu belegen versucht. In der Praxis gibt es sehr viele Phänomene, die quantifizierbar sind, also in Zahlen ausgedrückt werden können, auch wenn nicht übersehen werden darf, dass es auch Informationen gibt, die sich einer Quantifizierung entziehen. Nicht umsonst haben deshalb in den letzten Jahren die qualitativen Verfahren deutlich an Bedeutung gewonnen.

*Beschreibung von Massenerscheinungen* soll heißen, dass es nach der Quantifizierung (oder gleichzeitig) darum geht, das statistische Material zu beschreiben. Dabei gehen Sie so vor, als wollten Sie einem Gesprächspartner mitteilen, was Sie festgestellt haben. Sicherlich ist es nicht sinnvoll, jede einzelne Einkommensangabe von 40 Millionen abhängig Beschäftigten in diesem Gespräch mitzuteilen, sondern Sie werden zur mitteilenden Beschreibung sinnvoll zusammenfassen müssen; zum Beispiel können Sie den Datenbestand durch Durchschnittswerte beschreiben oder durch Angaben darüber, wie groß die Anzahl der Beschäftigten in den einzelnen Einkommensklassen ist. Man spricht in diesem Zusammenhang auch von *beschreibender Charakterisierung*.

Mit dem Stichwort *Beurteilung von Massenerscheinungen* wird ein Arbeitsschritt gekennzeichnet, der sich im allgemeinen der Beschreibung anschließt. Der Statistiker beschäftigt sich nämlich nicht nur deshalb mit Massenerscheinungen, um möglichst viele Tabellen anlegen zu können, sondern um Licht in vorher unbekannte Bereiche zu bringen. So kann zum Beispiel die Aussage, dass 73 % aller Haushaltsnettoeinkommen unter 3500 DM monatlich liegen, als Vermutung, als *Hypothese* angesehen werden, die durch statistische Daten, also zum Beispiel durch eine entsprechende Umfrage, überprüft und beurteilt werden soll. Man spricht dann davon, dass der Hypothese die statistischen Daten (Befunde) gegenübergestellt werden. Es soll auf diese Weise eine Entscheidung darüber herbeigeführt werden, ob die Hypothese als bestätigt oder als widerlegt gelten kann – die Aussage der Hypothese wird also überprüft bzw. beurteilt oder, wie man auch sagt, einem Test unterzogen.

In der Definition der statistischen Methoden tauchte noch das Stichwort *Schlussfolgerungen* auf. Auf der Grundlage der vorher geschilderten Aufgaben statistischer Methoden können Schlüsse gezogen werden. Im allgemeinen werden dabei aus den Ergebnissen einer statistischen Auswertung, die sich ja meist auf die Daten einer Stichprobe (Teilerhebung) stützt, Schlüsse auf die Gesamtheit gezogen, aus der diese Stichprobe stammt.

Ergebnis einer solchen Schlussfolgerung könnte zum Beispiel die folgende Aussage sein: Weil sich in einer Stichprobenuntersuchung eine durchschnittliche Körpergröße zufällig ausgewählter Erwachsener von 172 cm ergeben hat, kann mit großer Wahrscheinlichkeit davon ausgegangen werden, dass auch in der Grundgesamtheit, aus der die Stichprobe stammt, die durchschnittliche Körpergröße bei 172 cm liegt.

Eine solche Aussage ist natürlich in dieser Form nicht sonderlich aufregend. Da der Statistiker aber die Wahrscheinlichkeit dafür, dass die Aussage tatsächlich zutrifft, berechnen kann, wird sie interessanter. Er kann feststellen, dass obige Schlussfolgerung von der Stichprobe auf die Gesamtheit z.B. mit einer Wahrscheinlichkeit von 95 % zutrifft. Damit wird eine solche Aussage sehr informativ und brauchbar.

Das letzte Stichwort in der obigen Definition bezog sich auf die *Vorbereitung der Erklärung* von Massenerscheinungen: Die oben beschriebenen Überprüfungsverfahren dienen dieser Vorbereitung insbesondere dann, wenn man mehrere statistische Tatbestände gleichzeitig im Auge hat und sich nicht nur mit einem beschäftigt. Wird in einer statistischen Untersuchung beispielsweise der Gesundheitszustand der Bevölkerung thematisiert, und werden zugleich Informationen über den tagesdurchschnittlichen Zigarettenkonsum gesammelt, dann kann mit geeigneten statistischen Methoden untersucht werden, ob es einen Zusammenhang zwischen beiden Bereichen gibt (Gesundheit und Rauchen), bzw. wie stark dieser eventuelle statistische Zusammenhang ist.

Mit den *beurteilenden Methoden* kann der Statistiker prüfen, ob der eventuelle Zusammenhang deutlich genug ist oder nicht. Auf diese Weise kann der Weg zur Erklärung derartiger Zusammenhänge vorbereitet werden.

Es ist klar, dass diese Erläuterungen zur Definition des Begriffs der statistischen Methoden für Sie noch recht unverbindlich sind. Wenn aber in den späteren Kapiteln dieses Buches die einzelnen Methoden vorgestellt werden, dann werden Sie sich an diese vorläufigen Anmerkungen erinnern.

### 1.3 Statistische Anwendungen

Für welche Anwendungsbereiche taugen die gerade angesprochenen Methoden, und in welchen Bereichen wird dieses Instrumentarium eingesetzt? Auch dem Nicht-Statistiker ist bekannt, soweit er ab und an in die Zeitung schaut, dass ein wichtiger Anwendungsbereich der Statistik der wirtschaftliche Bereich ist. Man kann von *Wirtschaftsstatistik* sprechen, wenn es um Industrieumsätze geht, um die Arbeitslosenquote, um Durchschnittseinkommen, um die Werte für Exporte und Importe, um Lebenshaltungskosten

und Inflationsraten, um Produktivität und Gehälter, um das Brutto sozialprodukt, um Preise, um Produktionsmengen usw. Auch wenn man es nur mit einzelnen Produkten zu tun hat – zum Beispiel Kohleförderung, Entwicklung des Pkw-Bestands oder Erzeugung von Kunststoffen –, sind wirtschaftliche Fragen angesprochen. Offenbar handelt es sich hier um einen sehr breiten Anwendungsbereich statistischer Methoden, der leicht in Teilbereiche untergliedert werden kann. Solche Teilbereiche sind etwa die Preisstatistik, die Produktionsstatistik, die Industriestatistik, die Außenhandelsstatistik oder (ohne dass diese Liste schon vollständig wäre) die *Betriebsstatistik*.

Die Wirtschaftsstatistik ist vielleicht der am ersten ins Auge fallende Anwendungsbereich statistischer Methoden. Es darf jedoch nicht übersehen werden, dass es noch eine ganze Reihe anderer Anwendungsbereiche gibt. Einige wichtige (beileibe nicht alle denkbaren) sollen hier kurz vorgestellt werden:

*Bevölkerungsstatistik:* Die Bevölkerungsstatistik ist derjenige Anwendungsbereich statistischer Methoden, der sich mit der Aufzeichnung, Beschreibung und Analyse von Bevölkerungsbewegungen beschäftigt. Als *Bevölkerungsbewegung* bezeichnet man die Zunahme der Bevölkerung durch Geburten und Zuwanderungen und ihre Abnahme durch Tod und Abwanderungen. Häufig interessiert man sich für Einzelfragen, wie beispielsweise für die Geburtenhäufigkeit bei In- und Ausländern, Entwicklung dieser Kennziffern im Zeitablauf, Verteilung der Todesfälle nach Todesursachen u.ä. – statistische Angaben, die für Politiker, Bildungsplaner, Mediziner und Soziologen, aber auch für den informierten Bürger von großem Interesse sein können. Unter *Bevölkerungsstruktur* versteht man die Zusammensetzung der Bevölkerung etwa nach Geschlecht, nach Alter, nach der Art der Beschäftigung oder zum Beispiel ihre örtliche (geographische) Verteilung: Es gibt eine sehr große Zahl von Merkmalen, nach denen man hier unterscheiden könnte. Dies leitet wieder in die Bereiche der Wirtschaftsstatistik über, wenn man sich etwa für den Umfang der arbeitsfähigen Bevölkerung interessiert, für die Zahl der abhängig Beschäftigten, für ihre Bildung und Ausbildung usw.

*Sozialstatistik:* Auch der Bereich der Sozialstatistik lehnt sich eng an die Wirtschafts- und die Bevölkerungsstatistik an, weil man es hier mit ähnlichen Bereichen der gesellschaftlichen Entwicklung zu tun hat. Die Sozialstatistik wendet sich insbesondere den quantifizierbaren Aspekten von Verteilungen zu, wie etwa der Einkommensverteilung, der Vermögensverteilung, der Verteilung von Bildungsqualifikationen oder der Einteilung der Bevölkerung in soziale Schichten und Klassen.

*Auslandsstatistik:* Häufig interessiert man sich für Fragestellungen der Wirtschafts-, Bevölkerungs- oder Sozialstatistik, die sich auf andere Nationen beziehen – zum Beispiel bei internationalen statistischen Vergleichen. Dann muss man sich mit ausländischen Statistiken beschäftigen und diejenigen Veröffentlichungen heranziehen, die statistisches Material über das betreffende Ausland enthalten.

*Naturwissenschaftliche Statistik:* Nicht nur im wirtschafts- und sozialwissenschaftlichen Bereich haben statistische Methoden weite Verbreitung gefunden, sondern auch in anderen Wissenschaftsgebieten, vor allem in den Naturwissenschaften. So gibt es zum Bei-

spiel eine Medizinstatistik, die etwa die unterschiedlichen Krankheitsbilder, Krankheitsverläufe, Arznei- und Therapiewirkungen und nicht zuletzt auch die Resultate bei Tierversuchen aufzeichnet und statistisch auswertet. Die Physik-Statistik analysiert physikalische Experimente, sofern sie in größerer Zahl durchgeführt werden, oder hilft mit, die in riesigen Mengen auftretenden Satellitenfotos auszuwerten und anderes mehr. Ähnlich ist es beispielsweise in der Psychologie, wo statistische Methoden benötigt werden, um die Ergebnisse von Untersuchungsexperimenten, die mit Personen oder Personengruppen durchgeführt werden, zu durchleuchten.

Es wird Ihnen nach diesen durchaus unvollständigen Aufzählungen klar sein, wo es weitere Anwendungsgebiete für die Statistik gibt: Geographie, Wetterkunde, Biologie, Linguistik (Sprachwissenschaften) usw.

An dieser Stelle ist allerdings der Hinweis angebracht, dass das Statistikprogramm SPSS, mit dem wir es hier zu tun haben, in erster Linie dazu eingesetzt wird, Umfragedaten auszuwerten. Sein Haupteinsatzgebiet ist also der Bereich der empirischen Sozialforschung. Den Einsatz von SPSS in der Wirtschaftsstatistik, in der Geographie oder in der Medizin beobachtet man hingegen eher selten, obgleich SPSS auch in diesen und ähnlichen Bereichen gute Dienste leisten kann, wie in späteren Kapiteln an entsprechenden Beispielen gezeigt wird.

## 1.4 Statistische Methoden

Bevor ich den Einsatz von SPSS selbst bespreche, ist es angebracht, die große Zahl unterschiedlicher statistischer Methoden und Verfahren zu klassifizieren, um einen ersten Überblick zu gewinnen. Dabei bieten sich verschiedene Unterteilungskriterien an.

### *Deskriptive und induktive Statistik*

Zunächst unterteilt man die statistischen Methoden in die beiden wichtigen Bereiche der deskriptiven und der induktiven Statistik. Von *deskriptiver Statistik*, bzw. von *deskriptiven statistischen Methoden* spricht man, wenn das Ziel der eingesetzten Verfahren die Beschreibung des Ausgangsdatenbestandes ist. Zum Beispiel zählt die Berechnung eines arithmetischen Mittels, eines Durchschnitts also, zu dieser Gruppe, weil Sie mit der Mittelwertberechnung Ihren Datenbestand zusammenfassend beschreiben können. Von *induktiven statistischen Methoden* hingegen spricht man, wenn auf der Grundlage von Stichprobendaten Rückschlüsse auf die Grundgesamtheit angestrebt werden, aus der die jeweilige Stichprobe stammt. Diese Rückschlüsse führen, wie Sie noch erkennen werden, zu wahrscheinlichkeitsbehafteten Aussagen, weshalb die Verfahren dieser Gruppe auch manchmal dem Begriff der *Wahrscheinlichkeitsstatistik* untergeordnet werden. Häufig spricht man auch in diesem Zusammenhang von *schließender Statistik* oder von *beurteilender Statistik*.

### ***Uni-, bi- und multivariate Methoden***

Bei den hier genannten Methodengruppen geht es um folgendes: Bei der statistischen Auswertungsarbeit interessiert sich der Statistiker häufig nur für eine einzige Untersuchungsvariable, zum Beispiel für das monatliche Nettoeinkommen einer großen Zahl abhängig Beschäftigter in der Bundesrepublik Deutschland. Alle Methoden, die er einsetzt, um die Einkommensangaben zu analysieren, zählen zu den Methoden der *univariaten Statistik*. Wenn es aber um die Betrachtung von zwei Variablen gleichzeitig geht – zum Beispiel um den tagesdurchschnittlichen Zigarettenkonsum zufällig ausgewählter Erwachsener einerseits und um Angaben zum oberen Blutdruckwert dieser Personen andererseits –, wenn man sich also dafür interessiert, ob es vielleicht Zusammenhänge zwischen diesen beiden Variablen gibt, dann bedient man sich der Methoden der *bivariaten Statistik* (bi = zwei). Schließlich kommt der Statistiker auf die Idee, dass die Angaben zum Blutdruck nicht nur vom Zigarettenkonsum, sondern vielleicht auch vom Alter, vom Geschlecht und vom Beruf der befragten Personen beeinflusst sein könnten. Will er solchen gemeinsamen Beeinflussungen auf die Spur kommen, benötigt er die Methoden der *multivariaten Statistik* (multi = viele = drei oder mehr).

### ***Skalenabhängige Methoden***

Bei diesem Einteilungskriterium geht es um die Frage der sogenannten *Skalenqualität* von statistischen Untersuchungsvariablen. Mit diesem Begriff ist der *Informationsgehalt* von Daten angesprochen. Was damit gemeint ist, sollen einige illustrative Beispiele erläutern:

Stellen Sie sich vor, es wird der Familienstand zufällig ausgewählter Personen erfasst. Hier lassen sich Angaben finden, wie verheiratet, ledig, geschieden oder verwitwet. Man spricht in diesem Zusammenhang von einer *Nominalskala* (genauer von einer *nominalskalierten Variablen*) und will damit zum Ausdruck bringen, dass die Werte, die ermittelt wurden, nur Etiketten sind, gewissermaßen also Namen. Diese erlauben nur Unterscheidungen zwischen einzelnen Personen – sonst nichts. Der Informationsgehalt beschränkt sich also auf die Feststellung von Unterschieden oder von Identitäten.

Anders ist es zum Beispiel mit der Variablen Zeugnisnote. Befragen Sie 20 zufällig ausgewählte Schüler nach ihrer Mathematikzensur, dann beinhalten die 20 Angaben nicht nur Unterscheidungs-/Identitätsinformationen, sondern zusätzlich auch eine Rangordnungsinformation (Schüler Meier ist besser als Müller, Schüler Schmidt ist schlechter als Weber usw.). Eine solche Variable nennen wir *ordinalskalierte Variable*.

Können zusätzlich auch die Abstände (oder sogar die Quotienten) zwischen je zwei Werten einer Untersuchungsvariablen inhaltlich interpretiert werden, sprechen wir von einer *metrischen Skala*. Beispielsweise ist die Variable Körpergröße metrisch skaliert: Der Abstand des Basketballspielers Kreuzer (190 cm) zu Frau Schneeberger-Lüttgenhorst (160 cm) ist doppelt so groß wie der zwischen Professor Wilms (170 cm) und seinem Assistenten Clauß (185 cm). Hier wird also über Abstandsinformationen gespro-

chen. Das Alter unseres Wirtschaftsministers (60 Jahre) ist dreimal so groß wie das Alter meiner Tochter (20 Jahre). Hier wird über Quotienteninformationen gesprochen.

Natürlich können Sie zum Beispiel auch bei einer Ordinalskala Abstände oder Quotienten zwischen je zwei Werten berechnen – aber die Rechenergebnisse besagen inhaltlich nichts: Wenn Schüler Lorenz die Note 2, Schüler Buchmann die Note 4 hat, dann ist zum Beispiel  $4/2 = 2$ . Das heißt aber nicht, dass Lorenz doppelt (zweimal) so gut ist wie Buchmann.

Aus diesen Überlegungen ergibt sich, dass bestimmte mathematische Operationen bei bestimmten Skalen inhaltlich sinnvoll sind, bei anderen aber nicht. Und dies wiederum bedeutet, dass es statistische Methoden gibt, die bei bestimmten Skalen eingesetzt werden können, bei anderen aber nicht.

Beispielsweise ist die Berechnung eines arithmetischen Mittels bei Körpergrößen sinnvoll, bei Angaben zum Familienstand aber sinnlos. Wenn die durchschnittliche Körpergröße 172,3 cm ist, dann können Sie mit dieser Information etwas anfangen. Wenn aber ausgerechnet wird, dass der durchschnittliche Familienstand 1,17 ist, dann ist dies eine sinnlose Information (Sie sollten sich einmal überlegen, warum es auch sinnlos ist, arithmetische Mittel aus Schulzensuren auszurechnen – auch wenn dies in der Realität der statistischen Praxis immer wieder gemacht wird).

Deshalb ist es zweckmäßig, statistische Methoden danach zu unterscheiden, für welche Skalenqualitäten sie eingesetzt werden können.

## 1.5 Wichtige Grundbegriffe

Vor den weiteren Ausführungen ist es zweckmäßig, einige wichtige Grundbegriffe zu klären, die immer wieder auftauchen werden.

### *Merkmale und Merkmalsträger*

Interessiert sich der Statistiker zum Beispiel für die Körpergröße von erwachsenen Personen, dann wird die Körpergröße als *Merkmal* oder als *Variable* bezeichnet. Variable heißt eine solche Größe deshalb, weil sie von Person zu Person unterschiedliche Werte annimmt – nur zufälligerweise werden zwei Personen die gleiche Körpergröße aufweisen. Die Werte der Variablen, also die einzelnen beobachteten oder gemessenen Körpergrößen, werden *Ausprägungen* (*Variablenausprägungen*) oder *Merkmalswerte* (kurz *Werte*) genannt.

Zweckmäßigerweise unterscheidet man zwei Typen von Variablen:

Variablen, die nur ganz bestimmte Werte annehmen können, die streng voneinander getrennt sind, so dass keine Zwischenwerte möglich sind, werden *diskrete Variablen* genannt. Ein typisches Beispiel ist das Merkmal Geschlecht, das als Werte nur männlich oder weiblich annehmen kann. Weitere Beispiele wären etwa die Merkmale Familienstand, Kinderzahl, gewählte politische Partei usw.

Dagegen hat man es mit einer *stetigen Variablen* (Merkmal) zu tun, wenn die Variable im Prinzip jeden Wert und jeden Zwischenwert als Ausprägung annehmen kann. Beispiele dafür sind Einkommen oder generell Geldgrößen (wenn man davon absieht, dass zwischen benachbarten Pfennigangaben im allgemeinen keine Zwischenwerte angegeben werden), metrische Angaben wie Körpergrößen oder -gewichte, gemessene Zeiten, Temperaturen, Prozentangaben usw.

Die *Merkmalsträger* sind bei diesen Beispielen immer einzelne Personen gewesen. Merkmalsträger können aber auch Nationen sein, vielleicht mit den Merkmalen Bevölkerungszahl, Fläche, Brutto sozialprodukt; oder Gemeinden mit den Merkmalen Steueraufkommen, Grünflächenanteil, Bevölkerungszahl; oder Straßenkreuzungen mit dem Merkmal Zahl der Unfälle pro Monat; Werkstücke mit den Merkmalen Durchmesser, Gewicht, Schadhafteigkeit; Zuchtsauen mit den Merkmalen Gewicht, Zahl der Ferkel usw.; Autos mit den Merkmalen Farbe, Hubraum usw.

### ***Stichprobe und Grundgesamtheit***

Wenn sich der Statistiker für bestimmte Sachverhalte, Tatbestände oder Entwicklungen quantitativer Art interessiert, dann muss er versuchen, die entsprechenden Daten zu finden. Es bieten sich generell zwei Wege an: Entweder er betrachtet die *Grundgesamtheit* aller in Frage kommenden Merkmalsträger, oder er beschränkt sich auf eine *Teilerhebung* aus dieser Grundgesamtheit, die man auch *Stichprobe* nennt. Es leuchtet unmittelbar ein, dass eine Stichprobe rascher zu Ergebnissen führt, und dass die Datenerhebung auf Stichprobenbasis viel preiswerter ist als eine *Totalerhebung* (Auszählung der Grundgesamtheit). Weiterhin ist einleuchtend, dass eine Totalerhebung dann vorzuziehen ist, wenn es – aus welchen Gründen auch immer – unbedingt erforderlich ist, alle in Frage kommenden Merkmalsträger zu untersuchen. Genauso einleuchtend ist es, dass in manchen Fällen Totalerhebungen nicht möglich oder nicht sinnvoll sind – beispielsweise dann, wenn produzierte Autos zum Zweck der Qualitätskontrolle zerstörenden Crash-Tests unterzogen werden.

