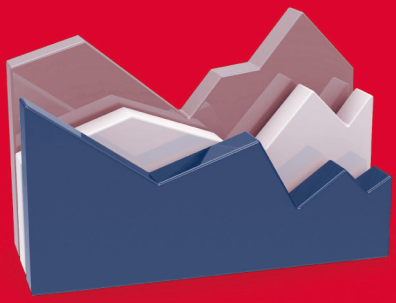
utb.

Christian FG Schendera

Deskriptive Statistik



utb 3969

Eine Arbeitsgemeinschaft der Verlage

Böhlau Verlag · Wien · Köln · Weimar
Verlag Barbara Budrich · Opladen · Toronto
facultas · Wien
Wilhelm Fink · Paderborn
A. Francke Verlag · Tübingen
Haupt Verlag · Bern
Verlag Julius Klinkhardt · Bad Heilbrunn
Mohr Siebeck · Tübingen
Nomos Verlagsgesellschaft · Baden-Baden
Ernst Reinhardt Verlag · München · Basel
Ferdinand Schöningh · Paderborn
Eugen Ulmer Verlag · Stuttgart
UVK Verlagsgesellschaft · Konstanz, mit UVK/Lucius · München
Vandenhoeck & Ruprecht · Göttingen · Bristol
Waxmann · Münster · New York

Christian FG Schendera

Deskriptive Statistik verstehen

UVK Verlagsgesellschaft mbH · Konstanz
mit UVK/Lucius · München

Dr. Christian FG Schendera ist SAS / SPSS Experte, Statistical Analyst und Scientific Consultant. Mehr Informationen zum Autor finden Sie auf Seite 382.

Abbildungen

Die Abbildungen des Buches finden Sie in Teilen auch online unter **www.uvk-lucius.de/schendera**.

Online-Angebote oder elektronische Ausgaben sind erhältlich unter www.utb-shop.de.

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.

Lektorat: Rainer Berger
Umschlagmotiv: © sukruengshop – fotolia.com
Einbandgestaltung: Atelier Reichert, Stuttgart
Druck und Bindung: fgb · freiburger graphische betriebe, Freiburg

UVK Verlagsgesellschaft mbH
Schützenstr. 24 · 78462 Konstanz
Tel. 07531/9053-0 · Fax 07531/9053-98
www.uvk.de

UTB-Nr. 3969
ISBN 978-3-8252-3969-5

Vorwort

„Wenn man mir die Freude am Fußball nimmt,
hört der Spaß bei mir auf."
Thomas Häßler

Was für ein Sommer!

Deutschland ist Fußballweltmeister, Miro Klose ist nun alleiniger Rekordtorschütze bei Fußballweltmeisterschaften, und Manuel Neuer erhielt den Goldenen Handschuh als bester Torhüter des Turniers. Deutschland überholte außerdem mit 223 Treffern bei Weltmeisterschaften den bisherigen Rekordhalter Brasilien, und führt wegen der Siege v.a. in der WM-Endrunde seitdem auch die Weltrangliste an.

Man darf mit einiger Berechtigung annehmen, dass Fußball, mindestens jedes Wochenende, umso mehr an internationalen Wettbewerben wie z.B. Champions League, Europa- oder Weltmeisterschaft, deutlich beliebter als Mathematik und Statistik sein könnte. Was liegt da näher, als die Faszination am Fußball auch ein wenig auf die deskriptive Statistik scheinen zu lassen? Umso mehr, da das DFB-Team während der WM Big-Data-Analysen einsetzte, die eben auch auf deskriptiver Statistik basiert (vgl. SAP News, 2014; Stier, 2014). Die deskriptive Statistik ist ebenfalls ein Teamsport: Sie funktioniert nach Regeln, nach Erfolgen (Titeln, Renommee, Punkten oder Toren), erfordert Koordination und Zusammenspiel, die Leistungen Einzelner tragen zum Ganzen bei, und sie kann auch eine breite Öffentlichkeit haben, z.B. in der Gestalt eines anspruchsvollen Publikums oder des Teams selbst. Also, los geht's…

Dieses Schema gibt den Aufbau des Buches wieder:

	Inhalt	Ziel
1	**Deskriptive Statistik**	■ Überblick ■ Disziplin
2	*„Heimspiel“* **Grundlagen** innerhalb einer Datentabelle	■ Beispiel: Bundesligatabelle ■ Zahlen, Ziffern und Werte ■ Messniveaus ■ Konsequenzen des Messniveaus
3	*„Vor dem Anpfiff“* ***Vor* dem Beschreiben** außerhalb einer Datentabelle	■ Datenerhebung ■ Verborgene Strukturen ■ Datenqualität ■ Strukturierung und Verarbeitung ■ Werte und Missings
4	*„Das Herz“* **Maßzahlen**	■ Mengen / Anteile ■ Lage-, Streu-, Formmaße ■ Grenzen und Bereiche ■ ROC ■ Zeit ■ Prozesse
5	*„Für das Auge“* **Tabellen und Grafiken**	■ Tabellenkonstruktion: 0×– bis höher klassierte Tabellen ■ Grafiken: je nach Daten, Zweck (Aussage) und Skalenniveau
6	*„Dream-Team“* **Datenqualität**	■ Vollständigkeit ■ Einheitlichkeit ■ Doppelte ■ Fehlende Werte ■ Ausreißer ■ Plausibilität
7	*„Jonglieren“*	■ Gewichte ■ Zahlen als Text
8	*„Werkzeuge“* **Einführungen**	■ SAS Enterprise Guide (kurz: EG) ■ IBM SPSS Statistics (kurz: SPSS)
9	**Literatur**	

Kapitel 1 geht in Abschnitt 1.1 zunächst der Frage nach: Was ist deskriptive Statistik? Deskriptive Statistik ist ein Teilbereich der Statistik und darin die regelgeleitete Anwendung eines Methodenkanons auf u.a. numerische oder Textdaten. Das Beherrschen der deskriptiven Statistik ist *auch* Kompetenz. Anschließend geht Abschnitt 1.2 darauf ein, was deskriptive Statistik *nicht* ist: Deskriptive Statistik ist *keine* explorative Analyse, konfirmatorische Analyse oder Inferenzstatistik. Deskriptive Statistik kommt auch nicht ohne Qualität und Hintergrundinformation über die Daten aus. Auch ist sie keine Projektionsfläche willkürlicher Auslegungen oder Spielball hemmungslosen Verallgemeinerns.

Kapitel 2 stellt die Grundlagen der deskriptiven Statistik als ein „Heimspiel" vor. Mit einem Heimspiel ist gemeint: Man spielt mit dem eigenen Team im eigenen Stadion vor eigenem Publikum. Man kennt sich bestens aus. Die Grundlagen der deskriptiven Statistik sind bekannt, man ist bestens vorbereitet. Abschnitt 2.1 beginnt daher mit einer der am häufigsten betrachteten Tabellen in Deutschland, nämlich einer Bundesligatabelle. Das Ziel ist, anhand dieser Tabelle die wichtigsten Grundbegriffe der deskriptiven Statistik zu erläutern. Fußball erklärt also die deskriptive Statistik. Abschnitt 2.2 beginnt mit der Erläuterung des Inhalts von Datentabellen und geht auf Begriffe wie z.B. Zahlen, Ziffern und Werte anhand von Beispielen aus dem Fußball ein. Abschnitt 2.3 geht anschließend mit der Frage: „Was hat Messen mit meinen Daten zu tun?" auf das sog. Messniveau einer Variablen über. Anhand der Bundesligatabelle werden Messniveaus und ihre grundlegende Bedeutung für jede (nicht nur deskriptive) Statistik erläutert. Abschnitt 2.4 hebt die Konsequenzen des Messniveaus für die praktische Arbeit mit Daten hervor. Begriffe wie z.B. Genauigkeit, Reliabilität und Validität sowie Objektivität werden z.B. mittels Torjägern veranschaulicht. Heimspiel bedeutet auch, dass man es durch eine gute Vorbereitung selbst in der Hand hat, auch ein anspruchsvolles Auswärtsspiel in die Kontrollierbarkeit und Niveau eines Heimspiels zu wandeln. Der Fokus von Kapitel 2 ist *daten-nahe*, und beschränkt sich daher auf Information *in* einer Datentabelle. Kapitel 3 beschreibt dagegen den *Kontext* von Daten, also Information, die man nicht notwendigerweise durch das Analysieren einer Datentabelle erfährt.

Kapitel 3 stellt grundlegende Fragen zusammen, die *vor* der Durchführung einer deskriptiven Statistik geklärt sein sollten. Den Anfang macht Abschnitt 3.1, der fragt: Wie wurden die Daten erhoben?

und stellt damit z.B. Fragen nach dem Messvorgang. Abschnitt 3.2 stellt Fragen nach verborgenen Strukturen, wie z.B. Ziehung und Auswahlwahrscheinlichkeit. Anhand von Entdeckungsreisenden in Sachen Fußball wird erläutert, was eine naive von einer systematischen Ziehung und Gewichtung von Daten unterscheidet. Aber selbst wenn diese Frage zufriedenstellend geklärt ist, ist damit noch nicht selbstverständlich, dass eine deskriptive Statistik erstellt werden kann. Abschnitt 3.3 fragt nach der Fitness der Daten (Darf eine deskriptive Statistik überhaupt erstellt werden?) und stellt mehrere mögliche Spielverderber vor. Abschnitt 3.4 ist eine Art Exkurs („Auszeit") und stellt Strukturen von Datentabellen vor, welche technische Eigenschaften (Attribute) sie haben und wie sie u.a. von Software verarbeitet werden. Abschnitt 3.5 widmet sich abschließend der womöglich spannendsten Frage: Was kann ich an meinen Daten beschreiben? Die Antwort darauf *muss* lauten: „Es kommt darauf an…"

Kapitel 4 beschreibt (*endlich!*) die Reise ins Herz der deskriptiven Statistik. Abschnitt 4.1 erläutert Maße für das Beschreiben von Mengen und Anteilen: Summe ($\sum$), Anzahl (N, n) und Häufigkeit (h, f, H, F). Abschnitt 4.2 erläutert die gebräuchlichsten Maße für das Beschreiben des Zentrums einer Verteilung (Lagemaße): Modus (D), Median (Z), Mittelwert ($\bar{x}$). Zur Illustration des Effekts von Missings sind die Beispiele für Lagemaße *ohne* und *mit* Missings berechnet. Abschnitt 4.3 erläutert die gebräuchlichsten Maße für das Beschreiben der Abweichung vom Zentrum einer Verteilung (Streuungsmaße): Spannweite R, Interquartilsabstand, Varianz, Standardabweichung, und Variationskoeffizient. Auch die Beispiele für Streuungsmaße sind *ohne* und *mit* Missings berechnet. Abschnitt 4.4 erläutert die gebräuchlichsten Maße für das Beschreiben der Abweichung von der Form einer Normalverteilung (Formmaße): Schiefe und Exzess. Abschnitt 4.5 erläutert das Beschreiben von Grenzen und Bereichen anhand von Quantilen (u.a. Median, Quartile, Dezentile) als eine Art Kombination aus Lage- und Streumaß. Abschnitt 4.6 erläutert das Beschreiben von Treffern, z.B. bei Wetten mit *zwei* Ausgängen („hopp oder topp"). Für einen „Wettkönig" werden für Wetten mit *vier* Ausgängen Sensitivität, Spezifität, ROC/AUC sowie Gewinn-Verlust-Matrix ermittelt. Abschnitt 4.7 stellt drei Möglichkeiten für das Beschreiben von Zeit vor: das geometrische Mittel (4.7.1), die Regressionsanalyse (4.7.2) sowie die Methode der exponentiellen Glättung als Trend bzw. Prognose (4.7.3). Bevor es an die praktische deskriptive Statistik geht, veran-

schaulicht Abschnitt 4.8, dass wer sich in der deskriptiven Statistik auskennt, auch andere als die „üblichen" Visualisierungen „lesen" kann. Deskriptive Statistik eben als Kompetenz. Abschnitt 4.8 stellt das Beschreiben von Prozessen vor, z.B. Funnel Charts (Trichterdiagramme usw.) für z.B. Pipelines. Abschnitt 4.9 verschafft einen schnellen Überblick, wo die meisten dieser Maße im SAS Enterprise Guide (4.9.1) und in IBM SPSS Statistics zu finden sind (4.9.2).

Kapitel 5 beschreibt die Grundlagen der Struktur und Interpretation von Tabellen und Grafiken zur Visualisierung von Daten. Abschnitt 5.1 beginnt beim Grundsätzlichen und erläutert die Konstruktion von 0- bis n×klassierten Tabellen; darunter Ausrichtung, Verschachtelung, die Vor- und Nachteile von Tabellen und wie mit SAS und SPSS 0- bis n×klassierte Tabellen erzeugt werden können. Abschließend wird eine einfache 0×(gesprochen: „nullfach") klassierte Tabelle vorgestellt. Eine solche Tabelle ist *nicht* nach einer Klassifikationsvariablen strukturiert. Abschnitt 5.2 beginnt mit den Grundlagen einer 1×klassierten Tabelle und geht dann zu spezielleren Themen über. Anhand *einer* Klassifikationsvariablen auf *Nominalniveau* werden die Grundlagen 1×klassierter Tabellen erläutert (5.2.1); an einer Klassifikationsvariablen auf *Ordinalniveau* werden Besonderheiten wie z.B. Ranginformation (5.2.2) oder Missings (5.2.3) vertieft. Unterabschnitt 5.2.4 erläutert eine 1×klassierte Tabelle für Variablen auf *Intervallniveau*, z.B. eine Mittelwerttabelle. Abschnitt 5.3 geht auf 2×klassierte Tabellen über, darin definieren *zwei* Kategorialvariablen eine Tabelle. Trotz komplexerer Tabellenstrukturen kommen mathematisch gesehen dieselben Rechenoperationen zum Einsatz. 5.3.1 beschreibt detailliert die Anforderung und Interpretation einer Kreuztabelle, u.a. Zellhäufigkeit und -prozente sowie Spalten- und Zeilenhäufigkeit und -prozente. Unterabschnitt 5.3.2 erläutert eine Tabelle, die wie eine Kreuztabelle strukturiert ist, jedoch die Werte einer dritten Variablen auf Intervallskalenniveau als Mittelwerte wiedergibt. Abschnitt 5.4 behandelt die Kommunikation von Werten und Daten mittels Diagrammen. Die Unterabschnitte sind anwendungsorientiert auf bestimmte Aussagen ausgerichtet: Wiedergabe von Datenpunkten (einzelne Werten einer Variablen, z.B. univariates Dot-Plot; vgl. 5.4.2), Wiedergabe von zusammengefassten Werten einer Variablen (vgl. 5.4.3, z.B. Balkendiagramm; ggf. gruppiert nach einer zweiten Variablen), Wiedergabe von bivariaten Messwertpaaren (z.B. eines Streudiagramms; vgl. 5.4.4) sowie Aggregierung und Gruppierung zweier Variablen und andere Fälle (z.B. Butterfly-Plot, vgl. 5.4.5). Allem

voran geht ein Crashkurs (Übersicht) mit Tipps (Dos), was man tun sollte und was besser nicht (Don'ts; vgl. 5.4.1).

Kapitel 6 vertieft das Thema der Datenqualität. Letztlich sind Datenqualität *und* deskriptive Statistik ein *Dream-Team*. Nur mit geprüfter Datenqualität macht eine deskriptive Statistik Sinn. Für jeden „Spielverderber" werden Sie seine besondere Bedeutung (um nicht zu sagen: *Gefahr*) und meist mehrere unkomplizierte Maßnahmen zur Prüfung kennenlernen. Der Umgang mit einem gefundenen Fehler hängt dabei von Art und Ursache des Fehlers ab. Die Systematik des Vorgehens orientiert sich an Schendera (2007). Abschnitt 6.1 beginnt, wenig überraschend, mit der Vollständigkeit. Abschnitt 6.2 geht zur Einheitlichkeit über. Abschnitt 6.3 behandelt doppelte (Doubletten) und Abschnitt 6.4 fehlende Werte (Missings). Abschnitt 6.5 stellt das Überprüfen auf Ausreißer vor; genau betrachtet wird bei Ausreißern auch die Gültigkeit eines Erwartungshorizonts geprüft. All dieses Prüfen von Datenqualität strebt (zunächst) das Ziel der Plausibilität an. Abschnitt 6.6 schließt mit Maßnahmen zur Prüfung der Plausibilität (Daten sollten unbedingt auf Plausibilität geprüft werden!). Abschnitt 6.7 schließt mit konkreten Trainingseinheiten zur Prüfung von Datenqualität.

Kapitel 7 schließt die Einführung in die deskriptive Statistik mit zwei spezielleren Anwendungen des Umgangs mit deskriptiven Statistiken: dem praktischen Umgang mit Gewichten (vgl. 7.1) und dem Umgang mit Zahlen beim Abfassen von Texten (vgl. 7.2). Abschnitt 7.1 führt in das Erstellen einer deskriptiven Statistik unter Einbeziehung von *Gewichten* ein. Gewichte haben einen großen Einfluss bei der Ermittlung deskriptiver Statistiken. Unterabschnitt 7.1.1 wird zuerst den Effekt von Gewichten an Beispielen aus dem Fußball, der Politik, und der Wirtschaft veranschaulichen. Gewichtete Ergebnisse sind nur mit Kenntnis der dahinterstehenden Annahmen und Interessen nachvollziehbar. Unterabschnitt 7.1.2 wird den Effekt von Gewichten an zahlreichen Streu- und Lagemaßen veranschaulichen. Unterabschnitt 7.1.3 wird als „Hintergrundbericht" die Frage klären: Was sind eigentlich Gewichte? Dabei wird auf die Funktion und Varianten von Gewichten eingegangen, von selbstgewichteten Daten über Designgewichte (disproportionale Ansätze) bis hin zur Poststratifizierung. Abschnitt 7.2 führt in das Verfassen einer deskriptiven Statistik als Text ein, und stellt u.a. Empfehlungen zusammen, wann eine Zahl als Ziffer („Zahl") und wann als Zahlwort („Text") geschrieben werden sollte. Unterabschnitt 7.2.1 stellt den Umgang mit allgemein ge-

bräuchlichen Zahlen vor. Unterabschnitt 7.2.2 behandelt den Umgang mit präzisen Maßen bzw. Messungen. Unterabschnitt 7.2.3 schließt mit Symbolen und Statistiken.

Kapitel 8 bietet zwei Kurzeinführungen in zwei der bekanntesten Werkzeuge für das Erstellen einer deskriptiven Statistik, den Enterprise Guide von SAS und SPSS Statistics von IBM. Die Berechnungen und Visualisierungen erfolgten mit dem Enterprise Guide 6.1, SAS v9.4, sowie SPSS v22. Die Zitate am Anfang eines jeden Kapitels sind überwiegend Michael Schaffraths (2013²) „Fußball ist Fußball" entnommen.

Zu Dank verpflichtet bin ich für Freundschaft, fachlichen Rat und/oder auch einen Beitrag in Form von Syntax, Daten und/oder auch Dokumentation unter anderem: Prof. William Greene (NYU Stern), Prof. em. Gerd Antos (Martin-Luther-Universität Halle-Wittenberg), Prof. Mark Galliker (Universität Bern, Schweiz), Roland Donalies (SAS Heidelberg), Ralph Wenzl (Zürich). Bei Sigur Ros, Jónsi und Alex sowie auch bei Walter Moers (Zamonien) bedanke ich mich für die langjährige künstlerische Inspiration. Meiner Frau Yun danke ich für ihre Geduld, Weitsicht und für ihr Verständnis.

Mein Dank gilt Patric Märki und Markus Grau von SAS Switzerland (Wallisellen) für die großzügige Bereitstellung von SAS Software und technischer Dokumentation. Herrn Rainer Berger vom UVK Verlag danke ich für das Vertrauen, dieses Buch zu veröffentlichen, sowie die immer großzügige Unterstützung. Stephan Lindow (Hamburg) entwarf diverse Grafiken. Falls in diesem Buch noch irgend etwas unklar oder fehlerhaft sein sollte, so liegt die Verantwortung alleine beim Autor.
An dieser Stelle möchte ich mich auch für die positiven Rückmeldungen und Vorschläge zu meinen weiteren Veröffentlichungen bedanken, u.a. zu SQL (2012, 2011), zur Clusteranalyse (2010), Regressionsanalyse (2014²), zur Datenqualität (2007), zu Syntaxprogrammierung mit SPSS (2005) sowie einführend in die Datenanalyse und Datenmanagement mit dem SAS System (2004). Die wichtigsten Rückmeldungen, Programme und Beispieldaten stehen auf der Webseite des Autors *www.method-consult.ch* zum kostenlosen Download bereit.

Hergiswil/Haikou, Februar 2015

Dr. CFG Schendera

Inhalt

1 Deskriptive Statistik: Was ist deskriptive Statistik?

„Entscheidend is auf'm Platz."
Adi Preißler

Dieses Kapitel geht in Abschnitt 1.1 der Frage nach: Was ist deskriptive Statistik? Deskriptive Statistik ist ein Teilbereich der Statistik und darin die regelgeleitete Anwendung eines Methodenkanons auf u.a. numerische oder Textdaten. Das Beherrschen der deskriptiven Statistik ist auch *Kompetenz*. Anschließend geht Abschnitt 1.2 darauf ein, was deskriptive Statistik *nicht* ist: Deskriptive Statistik ist keine explorative Analyse, konfirmatorische Analyse oder Inferenzstatistik. Deskriptive Statistik kommt auch nicht ohne Qualität und Hintergrundinformation über die Daten aus. Auch ist sie keine Projektionsfläche willkürlicher Auslegungen oder Spielball hemmungslosen Verallgemeinerns.

Die deskriptive Statistik ist ein Teilbereich der Statistik (vgl. Schulze, 2007; von der Lippe, 2006). Als eine allgemeine Definition könnte man die Statistik als die wissenschaftliche Anwendung mathematischer Prinzipien auf die Sammlung, Analyse und Präsentation (alpha)numerischer Daten verstehen. Teilbereiche der *Statistik* sind u.a. die Theoretische und Mathematische Statistik, darin eingebettet als Unterbereich die *Angewandte Statistik* (darin die Deskriptive Statistik und Inferenzstatistik) und darin wiederum als Unterbereich eingebettet der Bereich der *Datenanalyse* mit der explorativen und der konfirmatorischen Analyse.

In der folgenden Abbildung sind Bezüge zur Nachbarin der Statistik, der *Wahrscheinlichkeit* ausgeschlossen, z.B. bei der Inferenzstatistik (vgl. Mosler & Schmid, 2003), um die Hinführung zur deskriptiven Statistik stromlinienförmig zu gestalten. Anmerkungen zur Wahrscheinlichkeit und der damit verbundenen Unsicherheit (als wahrscheinlichkeitstheoretisches Konzept) sind bei der Deskriptiven Statistik nicht nötig (und aus diesem Grund auch in der eingangs allgemeinen Definition von Statistik nicht erwähnt). Was ist nun eine deskriptive Statistik? Eine erste Antwort ist: ein Methodeninstrumentarium, das auf Daten unabhängig von Erhebung

(online, POS, Fragebogen, Interview, Beobachtung, Experiment, Simulation), Studiendesign (Querschnitt, Längsschnitt, Panel usw.), Ziehungsart oder Umfang (Stichprobe, Vollerhebung) angewandt wird. Als weitere Antwort verdeutlicht diese Grafik den Stellenwert der deskriptiven Statistik: Wer die deskriptive Statistik als Teilbereich der angewandten Statistik beherrscht, hat damit auch das Werkzeug für die explorative Datenanalyse (klassisch: Tukey, z.B. 1980, 1977) *und* auch eine der zentralen Voraussetzungen vor der Durchführung einer inferenzstatistischen Analyse. Die Übergänge zwischen deskriptiver Statistik, explorativer und konfirmatorischer Datenanalyse sowie Inferenzstatistik werden sich dabei (wie so oft) als fließend herausstellen (vgl. Behrens, 1997; Cochran, 1972, 19). Gigerenzer (1999, 606ff.) zählt deskriptive Statistiken zu den wichtigsten Methoden aus der „Werkzeugkiste" für das Prüfen von Hypothesen. Während Tukey (1977) eine explorative Analyse als „attitude", als Einstellung, bezeichnet, werden wir hier sagen: Eine deskriptive Statistik ist *auch* Kompetenz.

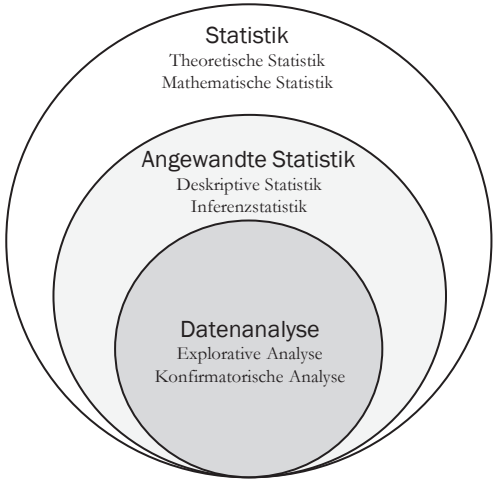


Abb. 1: Die Deskiptive Statistik als Teilbereich der Statistik

1.1 Was ist deskriptive Statistik?

Was ist der Sinn von deskriptiver Statistik? Die deskriptive (auch: darstellende, beschreibende) Statistik ist die Vorstufe und das Fundament jeder professionellen Analyse von Daten. Die deskriptive Statistik ist dabei keineswegs ignorierbar oder trivial. Im Gegenteil, ihre Funktionen sind vielfältig, ihre Maßzahlen sind allgegenwärtig, und ihre Bedeutung kann nicht hoch genug eingeschätzt werden. Die deskriptive Statistik ist die Grundlage und in vielen Fällen die Voraussetzung für den sinnvollen Einsatz der Inferenzstatistik. Je nach Datenart kann sie diese ggf. sogar ersetzen. Eine deskriptive Analyse geht einer professionellen Datenanalyse, sei sie nun inferenzstatistisch oder nicht, immer voraus. Im ersteren Falle gilt: Keine Inferenzstatistik ohne deskriptive Statistik!

Die deskriptive Statistik besitzt zahlreiche wichtige Funktionen:

- **Methoden und Kennziffern**: Die grundlegende Funktion der deskriptiven Statistik als *Disziplin* ist, ein Instrumentarium an Methoden und Kriterien zur statistischen oder visuellen (1) Reduktion von Daten und (2) Beschreibung durch z.B. Kennziffern, Tabellen oder Graphiken bereitzustellen. Die *explorative* Datenanalyse verwendet meist dieselben Methoden und Kriterien, hat jedoch das Ziel, anhand v.a. visueller Analyse der Daten *neue* Annahmen und Hypothesen über Strukturen, Ursachen oder Zusammenhänge aufzustellen (vgl. Behrens, 1997). Die im Weiteren beschriebenen Funktionen beziehen sich auf die deskriptive Statistik als *Methode*.
- **Datenreduktion**: Die grundlegende Funktion der deskriptiven Statistik als Methode ist die *Datenreduktion*, also die Reduktion von unüberschaubaren Mengen an Daten auf wenige, aber überschaubare Kennzahlen, Tabellen oder z.B. Graphiken, und damit auch die *Beschreibung* durch sie (vgl. auch Ehrenberg, 1986). Das Ziel der deskriptiven Statistik ist *nicht* der inferentielle Schluss auf eine nicht-verfügbare, hypothetische Grundgesamtheit.
- **Zusammenfassen**: Zahlreiche Einzelwerte können in einem einzelnen Wert zusammengefasst werden. Die Anzahl aller Einwohner eines Landes kann z.B. in einem einzigen Summenwert ausgedrückt werden. Auf diese Weise kann eine unübersehbare Menge an Daten übersichtlich aufbereitet werden.

- **Beschreiben**: Die Information zahlreicher Einzelwerte kann durch einen einzelnen Wert beschrieben werden. Das durchschnittliche Alter aller Einwohner eines Landes kann z.B. durch einen einzelnen Mittelwert beschrieben werden.
- **Strukturieren**: Für das Strukturieren zahlreicher Einzelwerte gibt es verschiedene Möglichkeiten: z.B. über Häufigkeitstabellen, Streudiagramme oder Maßzahlen, ggf. zusätzlich unterteilt (aggregiert) nach einer sog. Gruppierungsvariablen. All diese Möglichkeiten können Strukturmerkmale von Daten (also ihrer Verteilung) deutlich machen. Je nach Datenmenge und -verteilung können bestimmte Ansätze geeigneter sein als andere. Bei sehr großen Datenmengen sieht man z.B. bei Graphiken u.U. nur noch „schwarz". Häufigkeitstabellen geraten oft unübersichtlich. Letztlich verbleiben oft nur (gruppierte) Maßzahlen in Kombination mit Grafiken.
- **Herausheben**: Die wesentliche Information soll hervorgehoben werden. Gegebenenfalls erforderliche Vereinfachungen sollen den Informationsgehalt der deskriptiven Statistik so wenig als möglich einschränken. Ein klassisches Beispiel ist z.B., dass bei der Angabe eines Mittelwerts immer auch eine Standardabweichung angegeben werden sollte, um anzuzeigen, ob der Mittelwert tatsächlich die einzelnen Daten angemessen repräsentiert oder ob sie substantiell von ihm abweichen (was eben die mit angegebene Standardabweichung zu beurteilen erlaubt).
- **Grundlegen**: Die deskriptive Statistik ist oft die Wirklichkeit hinter innovativ klingenden Verfahren. Googles MapReduce ist z.B. aus der Sicht der deskriptiven Statistik nichts anderes als umfangreiche Freitexte in einzelne Elemente (z.B. Worte) zu zerlegen, diese zu sortieren und abschließend ihre Häufigkeit zu ermitteln. Das Umwandeln des Freitexts in die Wortliste wird als Erzeugen der „Map" bezeichnet, und das Auszählen und Ersetzen vieler gleicher Worte durch einen Repräsentanten und die dazugehörige Häufigkeit als das „Reduce". „MapReduce" mag interessanter klingen als „Auszählen von Zeichenketten" (vgl. z.B. Schendera, 2005, 133–136 zur Analyse von Text mit SPSS v13). Zentral für das verteilte Text Mining auch sehr großer Datenmengen sind jedoch die Prinzipien der deskriptiven Statistik und die erscheint spätestens jetzt so richtig spannend. Wer weiß, welche Geheimnisse andere Data-Mining-Verfahren verbergen…

- **Schließen**: *Im Allgemeinen* ist mittels der deskriptiven Statistik nur der Schluss auf die Stichprobe möglich, an der die Daten erhoben wurden; mittels Inferenzstatistik ist dagegen auch der Schluss von der Stichprobe auf die Grundgesamtheit möglich (u.a. Zufallsziehung vorausgesetzt). Die deskriptive Statistik *kann* die schließende Statistik allerdings ersetzen, und zwar dann *und nur dann(!)*, wenn es sich bei den Daten um eine *Vollerhebung* handelt, z.B. bei Daten einer Volkszählung oder auch um unternehmensinterne Kundendaten in einem DWH. In diesem Falle, *und nur in diesem Falle(!)*, kann auf die Inferenzstatistik verzichtet werden. Stammen die Daten aus einer Vollerhebung, ist jegliche deskriptive Statistik gleichzeitig auch eine Beschreibung einer (verfügbaren!) Grundgesamtheit; Inferenzschlüsse auf diese Grundgesamtheit sind somit nicht mehr erforderlich (dies kann auch Konsequenzen für die Wahl der Formeln haben). *Nur in diesem Fall* ist mittels der deskriptiven Statistik auch die Überprüfung von Hypothesen möglich (jedoch nicht im strikt inferenzstatistischen Sinne). Bei einer Stichprobe beschränkt sich die Aussage also *im Allgemeinen* auf die *beschriebenen* Daten; bei einer Vollerhebung gilt die Aussage auch für die Grundgesamtheit (*weil* die beschriebenen Daten die Grundgesamtheit *sind*). An dieser Stelle eröffnet sich ein fließender Übergang zur konfirmatorischen Analyse, die in Form der Abweichung der Daten von einem Modell zwar einen Modelltest darstellt, jedoch keinen Hypothesentest im inferenzstatistischen Sinne.
- **Screening**: Die deskriptive Statistik beschreibt die Daten, so wie sie sind. „as is" wird in der IT oft dazu gesagt. Dies bedeutet auch, dass die deskriptive Statistik gegebenenfalls auch Fehler in den Daten erkennen lassen kann (vgl. Schendera, 2007). Was also an dieser Stelle hervorgehoben werden sollte: Die Funktionen des Aggregierens, Beschreibens, Heraushebens bzw. Schließens sind dieser Funktion als Priorität und in der Zeit nachgeordnet. Die beste Beschreibung nützt leider nur wenig, wenn sie noch auf fehlerhaften Daten beruht. Das Screening mittels deskriptiver Statistik ist also ein mehrfach durchlaufener Prozess: Am Anfang wird keine Qualität von Daten vorausgesetzt (sie wird jedoch überprüft) („vorläufige deskriptive Statistik"), sie sollte jedoch am Ende des Screenings geprüft und schlussendlich als gegeben vorliegen („finale deskriptive Statistik").

- **Kommunikation von Vertrauen**: Während die Funktion des Screenings ein iterativ durchlaufener *Prozess* ist, ist die resultierende Datenqualität am Ende dieses Prozesses *auch* ein *Wert* mit der Funktion des Kommunizierens von Qualität und Vertrauen in die Daten. Die Funktion dieses Wertes ist, dass sich Leser und Anwender auf Maßzahlen und Aussagen auf Basis der deskriptiven Statistik verlassen können.
- **Unterstützung der Datenanalyse und Inferenzstatistik**: Die („finale") deskriptive Statistik unterstützt die Datenanalyse (v.a. explorative und konfirmatorische Analyse) und die Inferenzstatistik in mehrerer Hinsicht: z.B. um (1) sich einen ersten Eindruck von Voraussetzungen der Daten (z.B. Verteilungsform) zu verschaffen, (2) z.B. deskriptive Statistiken zu erzeugen, die konfirmatorische oder inferenzstatistische Analysen nicht standardmäßig ausgeben, (3) ihre Daten und Analysen besser nachzuvollziehen, und (4) (ggf. unterstützt durch einen eher explorativen Zugang) letzten Endes zusätzliche Hinweise für das weitere Vorgehen aufzudecken.

Die statistische Beschreibung mittels deskriptiver Statistik kann auf unterschiedliche Weise erfolgen:

- **Maßzahlen**: Maßzahlen reduzieren die Information unübersehbarer Datenmengen auf wenige Zahlen, die bestimmte Facetten dieser Datenmenge möglichst gut beschreiben. Man kann sich das so vorstellen, dass eine einzelne Maßzahl nur eine „Perspektive" auf die Daten ist, z.B. ihr Durchschnitt. Um nun die Daten auch aus anderen Blickwinkeln „betrachten" zu können, werden daher mehrere Maßzahlen berechnet, z.B. auch ihre Streuung. Dadurch wird auch einem möglichen Informationsverlust durch die Datenreduktion vorgebeugt. Maßzahlen werden in Lage-, Streu- und Formparameter unterteilt, z.B. Mittelwert (MW) und Standardabweichung (SD).

  > Beispiel
  > Daten a: 2, 2, 2 MW = 2,0, SD = 0,0
  > Daten b: 1, 2, 3 MW = 2,0, SD = 1,0
  > Daten c: 0, 2, 4 MW = 2,0, SD = 2,0

- **Tabellen**: Daten können in Tabellenform nonaggregriert (Rohdaten), aggregiert (z.B. Häufigkeitstabellen), kreuztabelliert oder hochverschachtelt wiedergegeben werden. Ist die gewählte Tabellenstruktur (z.B. uni-/multivariat und/oder ein-/mehrdimen-

sional) der konkreten Datenverteilung angepasst, wird die Information großer Datenmengen überschaubar wiedergegeben, oft z.B. in Kombination mit Grafiken.

- **Grafiken**: Daten können auch in grafischer Form als „fixierte Bilder" wiedergegeben werden. Hier stellt der Forschungsbereich der visuellen Statistik bzw. der statistischen Visualisierung vielfältige Diagrammvarianten zur Verfügung, von nonaggregierten, aggregierten, gruppierten bis hin zu uni-/multivariaten und/oder ein-/mehrdimensionalen Diagrammformen. Angefangen von Balken-, Kreis- und Liniendiagrammen bis hin zu Streu-, Bubble- oder Mosaik-Diagrammen, um nur einige zu nennen (vgl. 5.4).
- **Animationen**: Daten können auch als „bewegte Bilder" wiedergegeben werden. Der Phantasie sind hier keine Grenzen gesetzt: angefangen von animierten Standardgrafiken über Cockpits und Dashboards (v.a. für Unternehmen) bis hin zu (ggf. sogar in Echtzeit aktualisierten) Visualisierungen von Kunden-, Waren- bzw. Nutzungsströmen, die fast schon an Videoclips grenzen.

Empfehlungen, welche Darstellungsform den anderen vorgezogen werden können, lassen sich nicht allgemeingültig aussprechen. Die Übersichtlichkeit und damit auch ihr Informationsgehalt werden letztlich auch von der konkreten empirischen Verteilung und der Relevanz der jeweiligen Kenngrößen mitbestimmt. Die Kombination von Maßzahlen und Grafiken (Visualisierungen) gilt i. Allg. als das aufschlussreichste Vorgehen.

Was sind die Voraussetzungen einer erfolgreichen deskriptiven Statistik?

- **Daten**: So banal das klingen mag, eine deskriptive Statistik ist nicht ohne Daten, also Werte, möglich. Die untere Datenmenge liegt je nach deskriptiver Maßzahl zwischen N=0 (z.B. Summe) und um N=5 (z.B. für bestimmte Verfahren aus der Zeitreihenanalyse). Nach oben gibt es keine Grenze außer der Leistungsfähigkeit des Analysesystems selbst. Metadaten, also Informationen über Daten, erleichtern die Arbeit mit Daten ungemein. Zu den Informationen zum *Erheben bzw. Definieren* von Daten gehören z.B. semantische Definitionen (inkl. Ein- und Ausschlusskriterien), Informationen zur Datenquelle (Ort, Anzahl) oder auch zum Erhebungsmodus (Kunden- bzw. Haushaltsbefragungen) usw. (vgl. Schendera, 2007, 393–395).

- **Vollständigkeit**: Die deskriptive Statistik setzt die Vollständigkeit der zu beschreibenden Daten voraus. Damit ist nicht gemeint, dass Daten aus einer Vollerhebung stammen sollen, sondern dass alle Daten einer zu beschreibenden Stichprobe oder Vollerhebung auch tatsächlich vollständig vorhanden sind. Vollständigkeit ist eines der grundlegenden Kriterien für Datenqualität und damit auch für die deskriptive Statistik – vielleicht mit der Präzisierung, dass es sich dabei um die *richtigen* Daten handeln muss.
- **Datenqualität**: Datenqualität ist *die* zentrale Voraussetzung für die deskriptive Statistik (i.S.e. „finalen deskriptive Statistik"). Deskriptive Statistik auf der Basis fehlerhafter Daten kann nicht hinreichend die gemessenen Entitäten beschreiben und kann einer (Selbst-)Täuschung gleichen. Datenqualität stellt sicher, dass sich Anwender auf Maßzahlen und Aussagen verlassen können. Auf Datenqualität wird einführend in Abschnitt 3.3 und ausführlich in Kapitel 6 eingegangen.
- **Messniveau**: Die deskriptive Statistik setzt die Kenntnis der Messeinheiten der zu beschreibenden Daten voraus. Erst Messeinheiten und das zugrunde liegende Referenzsystem machen aus Zahlen erst Werte, die Zustände, Unterschiede oder auch Veränderungen *korrekt* zu beschreiben und vor allem auch zu interpretieren erlauben. Eine der ersten Fragen, die man sich bei der Beschreibung von Daten stellen sollte, ist: In welcher Einheit sind diese Zahlen und wie sind sie zu interpretieren? Messeinheiten werden in Abschnitt 2.2 vorgestellt.
- **Erhebung**: Die deskriptive Statistik kann auf Daten jeglicher Ziehungsart und jeden Umfangs angewandt werden; es empfiehlt sich jedoch die Klärung der Umstände ihrer Erhebung. „Erhebung" umfasst drei thematisch verschiedene Aspekte, die aber oft zusammen auftreten, nämlich *Art, Umfang* und *Design* einer Erhebung: (1) Vor dem Erzeugen einer deskriptiven Statistik ist es notwendig zu prüfen, ob die Daten aus Vollerhebungen oder Stichproben stammen. (2) Stammen die Daten aus einer Vollerhebung, ist jegliche deskriptive Statistik gleichzeitig auch eine Beschreibung der Grundgesamtheit. Stammen die Daten aus einer Stichprobe, so sind u.a. das Verhältnis Ziehungs- und Erhebungsgesamtheit und die Abhängigkeit der statistischen Signifikanz vom ggf. nicht unerheblichen N zu beachten (vgl. z.B. Schendera, 2007, 395, 406). Bei der „Grauzone", wenn sich

die Größe der Stichprobe einer Vollerhebung, also einer Grundgesamtheit annähert, stehen Anwender letztlich vor der Wahl, ihre Daten als Grundgesamtheit oder Stichprobe zu definieren. Die Merkmale einer (Zufalls-)Stichprobe werden mit zunehmender Größe derjenigen der Grundgesamtheit immer ähnlicher (Gesetz der großen Zahl). (3) Mit dem Design einer Erhebung ist gefordert, dass eine Zufallsziehung vorliegt und dass im Falle ungleicher Auswahrscheinlichkeit der Fälle ihre Gewichte (idealerweise im selben Datensatz) vorliegen und ihre Ermittlung als Erhebungsdesign dokumentiert ist (vgl. 3.2 und 7.1).

- **Gewichte**: Üblicherweise wird jeder Wert in der deskriptiven Statistik mit dem Gewicht 1 in die Analyse einbezogen. Ein Gewicht von 1 bedeutet, dass dieser Wert nur einen Fall repräsentiert, also nur für sich selbst steht. Je nach Analysekontext ist es sehr gut möglich, dass ein Fall jedoch nicht nur für sich selbst alleine steht, sondern für mehrere andere. In diesem Fall wird diesem Fall explizit ein anderes Gewicht zugewiesen, z.B. 10. Ein Wert mit dem Gewicht 10 repräsentiert daher zehn Fälle, und nicht nur einen. Gewichte werden aus diversen Gründen vergeben, z.B. um Auswahlwahrscheinlichkeiten (z.B. Oversampling) anzugleichen. Eine der ersten Fragen, die man sich bei der Beschreibung von Daten stellen sollte, ist: Sind die Daten gewichtet oder nicht? Falls die Daten gewichtet sind, wo sind die Gewichte dokumentiert und abgelegt? Zwei Abschnitte mit zwei völlig unterschiedlichen, aber einander ergänzenden Schwerpunkten führen in die deskriptive Statistik unter Einbeziehen von Gewichten ein. Abschnitt 3.2 richtet zunächst die Aufmerksamkeit auf Designstrukturen, Auswahlwahrscheinlichkeiten und Zufallsziehung. Abschnitt 7.1 befasst sich genauer mit der Herleitung von Gewichten und veranschaulicht das Berechnen deskriptiver Maße unter Zuhilfenahme von Gewichten.

1.2 Was ist deskriptive Statistik *nicht*?

Die deskriptive Statistik wird, eventuell abgesehen von der zugrunde liegenden Mathematik oder Statistik, überwiegend als recht unproblematisch vermittelt. Die Erfahrung zeigt, dass in der praktischen Anwendung der deskriptiven Statistik oft etwas großzügig (meist unbedacht) mit dem Sinn, aber vor allem mit den *Grenzen* der

deskriptiven Statistik umgegangen wird. Was sind erfahrungsgemäß häufige Fallstricke bei der Arbeit mit der deskriptiven Statistik?

- **Kein Plan:** Keinen Plan zu haben, kann manchmal etwas Befreiendes an sich haben; bei der Erstellung einer deskriptiven Statistik könnte dies u.U. zu heiklen Situationen führen. Nach allgemeiner Erfahrung *ist* die deskriptive Statistik ein *unterschätztes* Instrumentarium an Methoden, Kriterien und Voraussetzungen. Keinen Plan zu haben, meint weniger die Anforderung einer deskriptiven Statistik „auf Knopfdruck", sondern, dass dabei wesentliche Hintergrundinformationen (Metadaten) über die Daten nicht bekannt sind oder berücksichtigt werden. Hilfreiche Stichworte für einen Plan können z.B. sein: Vollerhebungen vs. Stichproben; falls Stichproben: Ziehungs-/Erhebungsgesamtheit (inkl. Ausfälle), Ein-/Ausschlusskriterien, Erhebungsdesign (Strukturen, Ziehungsplan, Gewichte, usw.), Variablen (Definitionen, Messniveaus, Einheiten, Maße, usw.), Analysepläne (Designstrukturen, Klassifikationsvariablen), (Grad der) Datenqualität oder auch, *wie* Zahlen im Text dargestellt werden sollen. Abschnitt 7.2 stellt diverse Vorschläge für das Schreiben von „zahlenlastigen" Texten zusammen.
- **Verwechslung:** Explorative Analyse, konfirmatorische Analyse und Inferenzstatistik haben andere Ziele wie die deskriptive Statistik – die deskriptive Statistik reduziert und beschreibt die Daten, *so wie sie sind*. Mit einem Quentchen Salz könnte man vielleicht sagen: Die deskriptive Statistik ist daten-geleitet, die konfirmatorische Analyse ist modell-geleitet, die Inferenzstatistik ist hypothesen-geleitet und die explorative Analyse ist neugierde-geleitet: Die explorative Analyse sucht nach *neuen* Strukturen und Zusammenhängen in den Daten (meist *auch* mit den Methoden der deskriptiven Statistik!). Die konfirmatorische Analyse prüft, ob die Verteilung der Daten *vorgegebenen* Modellen folgt (Modelltests). Die Inferenzstatistik schließt über Hypothesentests *von Stichproben auf Grundgesamtheiten*.
- **Sicherheit:** Die deskriptive Statistik beschreibt die Daten, so wie sie sind. Nicht weniger, aber auch nicht mehr. Dies bedeutet auch, dass die deskriptive Statistik keine „Sicherheit" von Aussagen einzustellen bzw. zu errechnen erlaubt, wie z.B. Alpha, p-Werte, „Fehler" usw. Auf der einen Seite braucht es diese Sicherheit auch gar nicht, weil keine Aussagen über Grundgesamtheiten getroffen werden. Auf der anderen Seite hilft eine

kluge Kombination von Lage- mit Streumaßen abzusichern, dass sie eine Verteilung von Daten ohne substantiellen Informationsverlust repräsentieren.

- **Datenqualität:** Die deskriptive Statistik setzt Datenqualität voraus, z.B. vollständige und geprüfte Daten. Nur weil eine deskriptive Statistik „auf Knopfdruck" abgerufen werden kann, bedeutet dies nicht automatisch, dass die Daten auch in Ordnung sind. Das Resultat ist höchstens eine *vorläufige* deskriptive Statistik. Keine deskriptive Statistik ohne zuvor geprüfte Datenqualität. Dieses Thema ist so wichtig, das ihm eine Einführung (Abschnitt 3.3) und eine Vertiefung (Kapitel 6) gewidmet sind.

Erfahrungsgemäß ist die deskriptive Statistik eine erste Belohnung für die harte Arbeit des Erhebens, Eingebens, Korrigierens und oft auch häufig genug komplizierten Transformierens von Daten. In der IT werden diese oft auch als ETL-Prozesse bzw. -Strecken abgekürzt („Extract", „Transform", „Load"). Entsprechend groß ist die Begeisterung, erste Einblicke in den (wünschenswerten) Erfolg der ganzen Unternehmung haben zu können. Wie die Erfahrung zeigt, treten an dieser Stelle gleich mehrere Fehler bei der Interpretation der deskriptiven Statistik auf. Um sie besser auseinanderhalten zu können, werden sie separat dargestellt; allesamt könnte man sie als Varianten des Über- bzw. Fehlinterpretierens der deskriptiven Statistik zusammenfassen:

- **Projektionsfläche** (Messgegenstand): Eines der häufigsten, größten und unerklärlicherweise immer noch stiefmütterlich behandelten „Fettnäpfchen" ist, den in der deskriptiven Statistik wiedergegebenen Daten Bedeutungen zu unterstellen, die gar nicht Gegenstand der Messung waren. Oft werden z.B. *sozio*demographische Variablen (z.B. Alter, Geschlecht, Einkommen) erhoben, und dann in der Gesamtschau als z.B. *psycho*logische Merkmale (z.B. „extrovertierter Konsumhedonist") überinterpretiert (vgl. Schendera, 2010, 20–21). Diese verkaufsfördernde bzw. arbeitserleichternde, jedoch an (Selbst-)Täuschung grenzende Unsitte ist leider nicht selten anzutreffen und keinesfalls auf eine bestimmte Disziplin beschränkt. Beispiele sind allgegenwärtig. In anderen Forschungsfeldern kann man es durchaus erleben, dass deskriptive Statistiken zu *Einstellungen* zum Lernen erhoben, aber als *Kognitionen* interpretiert werden (was inhaltlich etwas völlig anderes ist).

- **Hemmungsloses Verallgemeinern** (Merkmalsträger): Ein- und Ausschlusskriterien legen die Stichprobe, ggf. auch die Grundgesamtheit fest, auf die die deskriptive Statistik verallgemeinert werden kann. Mit dem „hemmungslosen Verallgemeinern" ist ein Interpretieren über diese Grenzen hinaus gemeint. Häufige Verstöße sind z.B. (1) die deskriptive Statistik einer *Stichprobe* als die einer *Grundgesamtheit* zu überinterpretieren. Die deskriptive Statistik einer Stichprobe kann *nicht* auf eine Grundgesamtheit verallgemeinert werden. Aussagen über die Grundgesamtheit, allein auf der Grundlage von *Stichproben*daten, sind ohne Absicherung nicht zulässig. (2) Zu den Verstößen zählt auch, die deskriptive Statistik einer Teilmenge (z.B. alte Menschen) auch für andere Teilmengen (z.B. junge Menschen) zu verallgemeinern. (3) „Projektion" ist z.B. die nicht seltene Praxis, z.B. bei der Korrelations- oder auch der Trendanalyse, die deskriptive Statistik über den Bereich der erhobenen Werte hinaus zu interpretieren.
- **jumping to conclusions** (Extrapolieren und Schlussfolgerung innerhalb einer Erwartungshaltung, dem „frame"): Der Begriff „jumping to conclusions" drückt, meine ich, schön aus, wie man bei der Interpretation der deskriptiven Statistik aus Begeisterung, und damit fehlender Zurückhaltung, leider vorschnellen Schlüssen über die darin wiedergegebenen Daten verfallen kann. Dieses „jumping to conclusions" ist, meiner Erfahrung mit Statistik-Einsteigern nach, eine Erscheinungsform des *gezielten Suchens* von Zusammenhängen oder Unterschieden innerhalb eines Frames. Dieses Phänomen lässt sich wohl am besten als kognitiver Ersatz eines erwartungsgeleiteten Hypothesentests umschreiben. Bei der Überinterpretation der deskriptiven Statistik (vor allem anhand von Stichproben) werden Unterschiede oder Zusammenhänge „gesehen", die in Wirklichkeit in den beschriebenen Daten gar nicht vorkommen. Das „jumping to conclusions" ist an sich gesehen nichts Schlechtes; allerdings sollte man diese „Schlussfolgerungen" nicht als abgesichertes Ergebnis eines „Hypothesentests" missverstehen, sondern als noch zu prüfende spekulative Annahme, die explizit einem echten Hypothesentest unterzogen werden sollte.
- **Der blinde Fleck** (Schlussfolgerung außerhalb eines Frames): Während ein erwartungsgeleiteter „Hypothesentest" dazu führt, dass „große" Unterschiede (die gar nicht so groß sind) zwischen

deskriptiven Parametern oft überschätzt werden, bezieht sich der „blinde Fleck" auf Phänomene, die außerhalb der eigenen Erwartungshaltung (frame) liegen (Schendera, 2007, 165–169). Hier tritt der gegenteilige Effekt auf: Erwartungswidrige Effekte werden oft erst gar nicht wahrgenommen, geringe Unterschiede dagegen oft leider unterschätzt. Erfahrungsgemäß werden bei der Interpretation oft andere relevante Aspekte übersehen, z.B. die unterschiedliche Größe der miteinander verglichenen Gruppen (vgl. dazu auch die Stichworte Designstruktur, Auswahlwahrscheinlichkeit und Gewichtung).

Die deskriptive Statistik hat ihre Grenze eindeutig dann erreicht, sobald es nicht mehr um das Beschreiben einer Stichprobe, sondern um das Ziehen von Schlüssen über eine Grundgesamtheit geht, z.B. in Gestalt von Hypothesentests, Punkt- oder Intervallschätzungen. Ausgehend von *Stichproben* erlaubt die deskriptive Statistik keine Aussagen zur Grundgesamtheit. Die Inferenzstatistik wird in diesem Buch nicht behandelt; ich erlaube mir für ausgewählte Verfahren z.B. auf Schendera (2014², 2010) zu verweisen.

Diese Einführung in Sinn und Grenzen der deskriptiven Statistik fokussiert grundlegende Konzepte. Abgeschlossen werden soll mit einem Hinweis darauf, dass manche der erwähnten Begriffe, wie z.B. „Grundgesamtheit", „Zufallsstichprobe" und m.E. vor allem „Repräsentativität" deutlich komplexer sind, als sie in dieser notwendigerweise vereinfachenden Darstellung womöglich anmuten (vgl. Prein et al., 1994). Allerdings beziehen sich Diskussion und Konzepte auf die Gültigkeit des Schlusses von einer „repräsentativen" Zufallsstichprobe auf eine unbekannte Grundgesamtheit, was nicht Aufgabe der deskriptiven Statistik und damit auch nicht Gegenstand dieser Einführung ist.

2 Ein Heimspiel: Grundlagen der deskriptiven Statistik

„Fußball ist einfach, deshalb ist es ja so kompliziert."
Berti Vogts

„Der Fußball ist einer der am weitesten verbreiteten religiösen Aberglauben unserer Zeit. Er ist heute das wirkliche Opium des Volkes."
Umberto Eco

„The best thing about being a statistician is that you get to play in everyone else's backyard."
John Tukey, Bell Labs, Princeton University

Mit einem Heimspiel ist gemeint: Man spielt mit dem eigenen Team im eigenen Stadion vor eigenem Publikum. Man kennt sich bestens aus. Die Grundlagen der deskriptiven Statistik sind bekannt, man ist bestens vorbereitet. Heimspiel bedeutet also auch: Durch eine gute Vorbereitung hat man es selbst in der Hand, auch ein anspruchsvolles Auswärtsspiel in die Kontrollierbarkeit und Niveau eines Heimspiels zu wandeln.

Der Fokus von Kapitel 2 beschränkt sich daher auf Informationen *in* einer Datentabelle. Informationen, die man nicht notwendigerweise durch das Analysieren einer Datentabelle erfährt, also den *Kontext* von Daten, beschreibt dagegen Kapitel 3. Abschnitt 2.1 beginnt daher mit einer der an Wochenenden wohl am häufigsten gesehenen Tabellen im deutschen Fernsehen, nämlich einer Bundesligatabelle. Das Ziel ist, anhand dieser Tabelle die wichtigsten Grundbegriffe der deskriptiven Statistik zu erläutern. Fußball erklärt also die deskriptive Statistik. Abschnitt 2.2 beginnt mit dem Erläutern des Inhalts von Datentabellen und erläutert Begriffe wie z.B. Zahlen, Ziffern und Werte an Beispielen aus dem Fußball. Anschließend geht Abschnitt 2.3 mit der Frage: „Was hat Messen mit meinen Daten zu tun?" auf das sog. Messniveau einer Variablen ein. Anhand der Bundesligatabelle werden Messniveaus und ihre grundlegende Bedeutung für jede (nicht nur deskriptive) Statistik erläutert. Abschnitt 2.4 hebt die Konsequenzen des Messniveaus für die praktische Arbeit mit Daten hervor. Begriffe wie z.B. Genauigkeit, Reliabilität und Validität sowie Objektivität werden z.B. mittels Torjägern veranschaulicht.