WWW.PEARSON-STUDIUM.DE
CWS
COMPANION WEBSITE

bio
biologie

Matthias Rudolf
Wiltrud Kuhlisch

Biostatistik

Eine Einführung für Biowissenschaftler

Studentengetestet!

PEARSON
Studium

Biostatistik

Matthias Rudolf
Wiltrud Kuhlisch

Biostatistik

Eine Einführung für Biowissenschaftler

PEARSON
Studium

ein Imprint von Pearson Education
München · Boston · San Francisco · Harlow, England
Don Mills, Ontario · Sydney · Mexico City
Madrid · Amsterdam

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.

10 9 8 7 6 5 4 3 2 1

10 09 08

ISBN 978-3-8273-7269-7

ein Imprint der Pearson Education Deutschland GmbH,
Martin-Kollar-Straße 10-12, D-81829 München

www.pearson-studium.de

Lektorat: Dr. Stephan Dietrich, sdietrich@pearson.de;
Christian Schneider, cschneider@pearson.de
Korrektorat: Dunja Reulein, München
Herstellung: Martha Kürzl-Harrison, mkuerzl@pearson.de
Satz: Reemers Publishing Services GmbH, Krefeld, www.reemers.de
Einbandgestaltung: Thomas Arlt, tarlt@adesso21.net
Druck und Verarbeitung: Kösel Druck, Krugzell (www.KoeselBuch.de)
Printed in Germany

Inhaltsübersicht

ÜBERBLICK

Inhaltsverzeichnis

Vorwort

Dieses Buch richtet sich an Anwender statistischer Methoden in den Biowissenschaften, speziell in der Biologie oder der Biotechnologie. Es wendet sich an Studierende biowissenschaftlicher Studiengänge, die das Buch gleichermaßen als Begleitbuch zu einer Vorlesung oder als ergänzende Lektüre zum Selbststudium nutzen können. Daneben ist es auch für Biowissenschaftler geeignet, die ihr biostatistisches Wissen auffrischen möchten oder einen Einstieg in die Thematik suchen. Der Leser benötigt lediglich die üblichen Grundkenntnisse der Elementarmathematik, die an Gymnasien unterrichtet werden.

Die grundlegenden Verfahren der Biostatistik sollen in diesem Buch anwendungsorientiert und gut nachvollziehbar dargestellt werden. Der Inhalt umfasst die klassischen Teilbereiche der Biostatistik, die in den entsprechenden Lehrveranstaltungen in Bachelor- oder Diplomstudiengängen angeboten werden. Daneben werden moderne Verfahren behandelt, deren Bedeutung in den letzten Jahren zugenommen hat. Da eine gute Versuchsplanung die Grundlage für verwertbare Versuchsergebnisse bildet, haben wir ein eigenständiges Kapitel zu den Grundlagen der biostatistischen Versuchsplanung aufgenommen.

Bei der Auswertung biowissenschaftlicher Untersuchungen ist heute die Anwendung leistungsfähiger Computer selbstverständlich. Auch die statistischen Berechnungen werden nahezu ausschließlich mit Hilfe von Statistikprogrammen durchgeführt. Deshalb verfolgen wir bei der Behandlung der statistischen Verfahren stets zwei Ziele: Einerseits sollen die jeweils behandelten Methoden so ausführlich und detailliert dargestellt werden, dass der Leser die grundlegenden Überlegungen und das konkrete Vorgehen gut nachvollziehen kann. Andererseits bereiten wir den Anwender darauf vor, die Methoden mit Statistikprogrammen anzuwenden und die Ergebnisse sachgerecht zu interpretieren.

Inhaltliche und didaktische Gestaltung

Das wichtigste Anliegen dieses Buches besteht darin, dem Leser ein Grundverständnis biostatistischer Denkweisen und Methoden zu vermitteln. Deshalb beschreiben wir die Grundlagen der jeweiligen Verfahren ausführlich, wobei wir auf unnötige mathematische Herleitungen verzichten. Wichtige Tests und Berechnungen werden detailliert dargestellt, damit die jeweilige Vorgehensweise konkret nachvollzogen werden kann. In allen Darstellungen gehen wir davon aus, dass die spätere Anwendung der Verfahren vorwiegend unter Benutzung von Statistikprogrammen erfolgen

wird. Deshalb verzichten wir zum Beispiel weitgehend auf die Angabe von Schnellrechenformeln und beschreiben das Prinzip von Testentscheidungen vorrangig auf der Grundlage der p-Werte.

Beispiele

Der unmittelbare Bezug der Darstellung zur Praxis wird durch Anwendungsbeispiele gewährleistet, die am Anfang der jeweiligen Kapitel vorgestellt werden. Der inhaltliche Rahmen dieser Beispiele kommt aus der Biologie oder der Biotechnologie und berührt oft Fragestellungen, mit denen Studierende biowissenschaftlicher Disziplinen im Rahmen ihrer Praktika beschäftigt sind. Die in den Beispielen verwendeten Daten haben wir künstlich erzeugt, um die zu beschreibenden Verfahren bestmöglich illustrieren zu können. Die Umsetzung aller im entsprechenden Kapitel behandelten Verfahren wird auf der Grundlage der Daten des Anwendungsbeispiels beschrieben.

Aufgaben und Lösungen

Am Ende der einzelnen Kapitel sind Beispielaufgaben angegeben, die das Spektrum der im jeweiligen Kapitel behandelten Methoden weitgehend abdecken. Wir empfehlen, die Aufgaben, soweit es sich um Rechenaufgaben handelt, zunächst per Hand und Taschenrechner zu rechnen, da damit das Verständnis der behandelten Verfahren weiter vertieft werden kann. Selbstverständlich können die Aufgaben auch unter Verwendung von Statistik-Software bearbeitet werden.

Lösungshinweise zu allen Aufgaben sowie weitere Aufgaben mit Lösungen sind auf der Companion Website zum Buch zu finden.

CD: Anwendung von Statistik-Software

Neben dem Verständnis der statistischen Verfahren bereitet vielen Anwendern erfahrungsgemäß der Einstieg in den Umgang mit Statistik-Software erhebliche Schwierigkeiten. Aus diesem Grund bieten wir auf der dem Buch beiliegenden CD eine Einführung in drei gebräuchliche Programme zur statistischen Datenanalyse an, die sehr unterschiedliche Vor- und Nachteile haben: R, SPSS und Excel.

- R ist ein kostenfreies Programm, das im Internet zur Verfügung steht. Viele Fachleute entwickeln das Programm kontinuierlich weiter, so dass mittlerweile sehr viele Verfahren realisiert sind. Neben den grundlegenden Methoden der Biostatistik sind Module für sehr komplexe Verfahren verfügbar. Allerdings müssen die notwendigen Befehle vom Anwender selbst eingegeben werden, entsprechende Auswahlfenster stehen nicht zur Verfügung.
- SPSS ist ein besonders im Bereich der Human- und Sozialwissenschaften weitverbreitetes Programmsystem. Neben dem großen Umfang verfügbarer Verfahren kann es auf eine sehr komfortable Bedienung verweisen. Der Anwender kann die notwendigen Methoden aus entsprechenden Fenstern auswählen, was die Einarbeitung in das Programm und die Arbeit mit dem Programm sehr erleich-

tert. Es ist in vielen Rechnerkabinetten der Universitäten und Fachhochschulen installiert. Die Kosten für das Programm sind relativ hoch, allerdings bietet SPSS kostengünstige Studentenversionen an.

- MS-Excel ist kein Statistikprogramm, enthält aber verschiedene Möglichkeiten zur Realisierung statistischer Verfahren. Unterschiedliche Grafiken und einfache statistische Verfahren lassen sich in Excel realisieren, die Durchführung komplexerer Analysen ist jedoch nur eingeschränkt möglich. Wegen seiner weiten Verbreitung im Rahmen der MS-Office-Produkte wird Excel häufig für einfache statistische Berechnungen verwendet.

Auf der beiliegenden CD werden die im Buch angegebenen Beispielrechnungen mit dem jeweiligen Programm nachvollzogen. Dabei beschreiben wir sowohl die erforderlichen Eingaben als auch die Ergebnisse. Zu ausgewählten Kapiteln wird zusätzlich die Analyse von Praxisdaten aus biowissenschaftlichen Forschungsvorhaben demonstriert.

Die Beschreibungen zu den Programmen werden auf der Companion Website zum Buch aktualisiert, sobald neue Versionen der Programme erscheinen.

Danksagung

Unser Dank gilt allen, die uns bei der Anfertigung dieses Lehrbuches unterstützt haben. An erster Stelle bedanken wir uns bei den Studentinnen Claudia Huth, Viktoria Decker und Bianca Kranzusch, die uns bei der Erstellung der Anwendungsbeispiele, durch Korrekturlesen und Nachrechnen der Beispiele geholfen haben. Wir danken allen Kolleginnen und Kollegen, die uns ihre Daten zur Verfügung gestellt haben. Viele Kolleginnen und Kollegen der Fachrichtungen Biologie, Hydrobiologie, Mathematik und Psychologie sowie des Biotechnologischen Zentrums (BIOTEC) der Technischen Universität Dresden haben uns bei unserem Vorhaben unterstützt. Ihnen allen sei herzlich gedankt. Frau Helga Mettke danken wir für die sorgfältige Bearbeitung der Grafiken. Sehr herzlich danken wir unseren Familien für ihre Unterstützung. Wir bedanken uns bei SPSS München, die uns Version 15 für die Arbeit an diesem Buch zur Verfügung gestellt haben.

Unser besonderer Dank gilt Herrn Dr. Stephan Dietrich und Herrn Christian Schneider, den Lektoren des Verlags Pearson Studium, für die stets angenehme und konstruktive Zusammenarbeit bei der Verwirklichung dieses Buchprojekts.

Matthias Rudolf und Wiltrud Kuhlisch

Einführung

1

ÜBERBLICK

In diesem einführenden Kapitel soll die Bedeutung der Biostatistik im Prozess biowissenschaftlicher Forschung veranschaulicht werden. Dabei wird auf die beschreibende Statistik und die Inferenzstatistik als grundlegende Teilbereiche der Biostatistik eingegangen. Die Unterschiede von Hypothesen erzeugenden und Hypothesen prüfenden Datenanalysen sollen verdeutlicht werden.

In weiteren Abschnitten werden einführend Aspekte der Versuchsplanung behandelt, deren Kenntnis für das Verständnis der in den folgenden Kapiteln behandelten statistischen Methoden unentbehrlich ist. Das betrifft einerseits die Unterscheidung von Populationen und von aus diesen Populationen gewonnenen Stichproben. Andererseits werden die unterschiedlichen Skalenarten behandelt, mit denen biowissenschaftliche Größen erfasst werden können.

1.1 Biostatistik als Bestandteil biowissenschaftlicher Forschung

Unter dem Begriff Biostatistik werden die Anwendungen der Methoden der mathematischen Statistik in den Biowissenschaften zusammengefasst. Neben Disziplinen wie der Biologie oder der Biotechnologie bietet vor allem die Medizin vielfältige Anwendungsbereiche für die Biostatistik. Für die Anwendung statistischer Methoden in der Medizin wird oft die Bezeichnung medizinische Statistik verwendet. Im Unterschied dazu wird in diesem Lehrbuch ausschließlich auf Beispiele aus der Biologie und der Biotechnologie zurückgegriffen. Das Buch richtet sich demnach primär an Studierende und Wissenschaftler dieser und benachbarter Disziplinen.

Die problemangepasste Anwendung biostatistischer Methoden ist ein integraler Bestandteil biowissenschaftlicher Forschung. Allgemein lässt sich die Durchführung von biowissenschaftlichen Forschungsvorhaben grob in drei Phasen unterteilen:

- Versuchsplanung,
- Versuchsdurchführung,
- Versuchsauswertung.

In diesem Ablauf hat die Anwendung biostatistischer Methoden besonders in den Phasen der Versuchplanung und der Versuchsauswertung große Bedeutung.

Ausgehend von der fachwissenschaftlichen Fragestellung und von den damit verbundenen inhaltlichen Hypothesen müssen bereits in der *Planungsphase* biostatistische Überlegungen einbezogen werden, um eine sachgerechte Versuchsdurchführung für die Beantwortung der gestellten Fragen zu gewährleisten. Schon unter Berücksichtigung der später durchzuführenden Datenauswertung sind die zu untersuchenden Merkmale und deren Skalenniveau festzulegen (siehe Abschnitt 1.3). Es

ist zu entscheiden, welcher konkrete Versuchsplan unter Berücksichtigung aller Rahmenbedingungen für die gegebene Fragestellung am besten geeignet ist (siehe Kapitel 9). Dabei muss bereits in der Planungsphase gesichert werden, dass die geplante Methode der Datenauswertung mit dem ausgewählten Versuchsplan überhaupt möglich ist. Integraler Bestandteil der Versuchsplanung sind Festlegungen zum notwendigen Stichprobenumfang. In allen Anträgen auf Bewilligung von Forschungsgeldern ist der Nachweis zu führen, dass einerseits nur die notwendige Anzahl von Untersuchungseinheiten für die geplante Untersuchung vorgesehen ist, um unnötige Kosten zu vermeiden. Andererseits muss der Antragsteller ebenfalls belegen, dass der geplante Stichprobenumfang groß genug ist, um das angestrebte Ergebnis überhaupt mit hinreichender statistischer Sicherheit erzielen zu können.

Im Rahmen der Versuchsauswertung müssen die adäquaten biostatistischen Methoden zur Datenanalyse eingesetzt werden. Dabei lassen sich die statistischen Verfahren in zwei grundlegende Klassen einteilen: in die Verfahren der deskriptiven (beschreibenden) Statistik und in die Methoden der Inferenzstatistik (der schließenden Statistik).

Die Verfahren der deskriptiven Statistik haben das Ziel, erhobene Daten so darzustellen, dass ihre bezüglich der aktuellen Fragestellung wesentlichen Eigenschaften veranschaulicht werden können. Zu diesem Zweck werden die Daten in Tabellen, in grafischen Darstellungen und mit Hilfe statistischer Maßzahlen zusammengefasst (siehe Kapitel 2). Methoden der beschreibenden Statistik sind wichtige Werkzeuge im Rahmen explorativer (Hypothesen generierender) Datenanalysen.

Definition

Explorative Datenanalysen dienen der Beschreibung gegebener Daten oder der Suche nach unbekannten Strukturen in komplexen Datenmengen. Mit ihrer Hilfe können Hypothesen über die untersuchten Merkmale gewonnen werden.

Die explorative Datenanalyse geht über die reine beschreibende Statistik hinaus. Mit Hilfe moderner leistungsfähiger Computeralgorithmen ist es in explorativen Untersuchungen zusätzlich möglich, nach unbekannten Strukturen in komplexen Datenmengen zu suchen und auf diesem Wege Hypothesen zu finden, wenn die eigentliche Forschungsfrage noch nicht genau definiert ist oder noch kein geeignetes statistisches Modell bestimmt werden konnte.

Inferenzstatistische Verfahren der Datenanalyse gehen von statistischen Modellen und Hypothesen aus. Sie basieren auf der Wahrscheinlichkeitstheorie (Kapitel 3). Auf dieser Grundlage können Hypothesen über Eigenschaften der untersuchten Populationen bestätigt oder abgelehnt werden, wobei alle Aussagen nur mit vorgegebenen Wahrscheinlichkeiten getroffen werden können (Kapitel 4 und 5). Inferenzstatistische Methoden sind die Grundlage konfirmatorischer (Hypothesen prüfender) Datenanalysen.

Definition

Konfirmatorische Datenanalysen dienen zur Entscheidung über *vor* der Untersuchung aufgestellte Hypothesen auf der Grundlage von inferenzstatistischen Methoden.

Der grundsätzliche Unterschied, aber auch der oft fließende Übergang zwischen explorativen und konfirmatorischen Datenanalysen soll an folgendem Beispiel veranschaulicht werden:

In einem industriell wenig erschlossenen Gebiet wurde ein großes Zuliefererwerk der Autoindustrie errichtet, dessen Abwässer in die benachbarten Flüsse gelangen. Es gibt keine inhaltlich begründeten Vermutungen, wie sich die Abwässer auf den Nitratgehalt der Flüsse auswirken. Mit Methoden der beschreibenden Statistik wird im Rahmen einer explorativen Datenanalyse der mittlere Nitratgehalt an unterschiedlichen Messstellen ermittelt. Dabei wird festgestellt, dass sich der Nitratgehalt nach der Errichtung des Werks mehr als verdreifacht hat. Ergebnis der explorativen Vorgehensweise ist damit die Hypothese, dass sich der durchschnittliche Nitratgehalt im Ergebnis der Veränderungen der Umwelt verdreifacht hat. Diese Hypothese kann nun im Rahmen einer konfirmatorischen Untersuchung geprüft werden. Dazu müssen neue Daten erhoben werden, zum Beispiel an anderen Flüssen, an anderen Messpunkten oder in angemessen großem zeitlichem Abstand zur ersten Messung. Im Ergebnis dieser Untersuchung kann die vor dieser Messung aufgestellte Hypothese bestätigt oder verworfen werden.

Dabei ist streng zu beachten, dass nur eine *vor* der Untersuchung aufgestellte Hypothese beurteilt werden kann. So wäre es denkbar, dass im Rahmen der zweiten Untersuchung festgestellt wird, dass sich der Nitratgehalt der Flüsse sogar verzehnfacht hat. Da diese deutlich höhere Nitratbelastung aber nicht vor der Untersuchung angenommen wurde, kann die erhöhte Belastung mit dieser Untersuchung nicht nachgewiesen werden. Gewissermaßen als wichtiges „Nebenprodukt“ der konfirmatorischen Datenanalyse hat sich eine neue Hypothese ergeben, die nun erneut unter Verwendung neu erhobener Daten bestätigt werden muss. Dieses Beispiel macht deutlich, dass explorative und konfirmatorische Datenanalysen oft keine starren Grenzen aufweisen, sondern ineinander übergehen können.

1.2 Population und Stichprobe

Daten werden in den Biowissenschaften immer an einzelnen Untersuchungseinheiten gewonnen. Solche Untersuchungseinheiten können sehr unterschiedlich sein, zum Beispiel Mikroorganismen, Säugetiere, Menschen oder Landschaftsschutzgebiete.

Im Anwendungsbeispiel in Kapitel 7 werden Daten an 24 Flüssen erhoben. Jeder Fluss ist eine Untersuchungseinheit. Bei diesen 24 Flüssen handelt es sich um eine Auswahl aller für die Untersuchung relevanten Flüsse. Die Gesamtheit aller ver-

gleichbaren Flüsse bildet die Population (oft auch als Grundgesamtheit bezeichnet). Sie enthält alle Untersuchungseinheiten, über die man Aussagen gewinnen will.

Definition

Unter einer Population (Grundgesamtheit) versteht man die Menge aller potentiellen Untersuchungseinheiten für eine bestimmte Fragestellung.

Die Population muss sehr genau abgegrenzt werden. Im Beispiel kann sie aus allen Flüssen eines bestimmten Gebietes bestehen. Die untersuchte Fragestellung kann sich aber ebenso auf alle Flüsse Deutschlands beziehen, aus denen sich in diesem Fall die Population zusammensetzen würde.

Bei der Betrachtung von Populationen kann man zwischen endlichen, unendlichen und hypothetischen Populationen unterscheiden. Wenn in einem Aquarium eine bekannte Anzahl an Fischen lebt, handelt es sich um eine endliche Population. Als ein Beispiel einer unendlichen Population kann die Menge aller Fische in den Ozeanen angesehen werden. Ein Beispiel für eine hypothetische Population sind alle Fische, die jemals gelebt haben.

Unter Teilpopulationen versteht man eine nach einem oder mehreren Gesichtspunkten eingegrenzte Population. Im Beispiel kann eine Teilpopulation durch alle Flüsse des untersuchten Gebietes beschrieben werden, deren Quelle im Gebirge zu finden ist.

Da es in biowissenschaftlichen Untersuchungen in den meisten Fällen nicht möglich ist, die interessierenden Populationen komplett zu erfassen und an allen Untersuchungseinheiten Messungen vorzunehmen, stehen in den Untersuchungen typischerweise nur ausgewählte Untersuchungseinheiten aus der Population zur Verfügung. Diese tatsächlich untersuchten Einheiten bilden die Stichprobe, deren Daten für die statistischen Analysen verwendet werden können.

Definition

Als Stichprobe bezeichnet man eine Teilmenge einer Population, die zufällig oder nach bestimmten Kriterien ausgewählt wurde. Sie enthält alle für die statistische Analyse verwendeten Untersuchungseinheiten.

Aus der Stichprobe sollen Daten gewonnen werden, mit denen unter Verwendung der in den folgenden Kapiteln behandelten Verfahren die Eigenschaften der Population möglichst genau beschrieben werden können. Auf die Bildung von Zufallsstichproben und auf weitere Typen von Stichproben wird in Kapitel 9 ausführlich eingegangen.

An den Untersuchungseinheiten der Stichprobe werden die interessierenden Größen erhoben. Im folgenden Abschnitt soll auf die unterschiedlichen Eigenschaften der zu untersuchenden Merkmale eingegangen werden, da deren Kenntnis für die Auswahl und Anwendung der in den folgenden Kapiteln behandelten statistischen Verfahren notwendig ist.

1.3 Merkmale und Skalenarten

Die an den Untersuchungseinheiten erhobenen interessierenden Eigenschaften werden als Merkmale oder Variablen bezeichnet. Dabei kann jedes Merkmal unterschiedliche Werte (Merkmalsausprägungen) annehmen. So kann zum Beispiel das Merkmal Geschlecht beim Menschen die Werte männlich oder weiblich annehmen. Das Merkmal Nitratkonzentration kann in Flüssen bei theoretisch unbegrenzter Messgenauigkeit unendlich viele Merkmalsausprägungen haben, praktisch ist die Anzahl möglicher Werte in Folge der eingeschränkten Messgenauigkeit natürlich begrenzt.

Für die spätere statistische Datenanalyse ist es notwendig, Merkmale hinsichtlich ihrer Eigenschaften zu klassifizieren.

Die einfachste Einteilung von Merkmalen kann nach der Anzahl der möglichen Werte erfolgen.

Definition

Ein Merkmal wird als diskret bezeichnet, wenn es nur endlich viele Werte annehmen kann. Ein stetiges Merkmal kann alle Werte eines Intervalls annehmen.

Beispiele für diskrete Merkmale sind das Geschlecht mit den möglichen Ausprägungen männlich und weiblich, die Variable Schädlingsbefall mit den Ausprägungen vorhanden oder nicht vorhanden sowie das Merkmal Schulnote mit den möglichen Ausprägungen 1, 2, 3, 4, 5 und 6. Stetige Merkmale sind zum Beispiel Nitratkonzentration, Größe oder Geschwindigkeit.

Die Einteilung in die beiden Merkmalstypen ist nicht immer völlig eindeutig. So bezeichnet man Merkmale als quasi-stetig, bei denen durch Begrenzung der Messgenauigkeit nicht jeder beliebige Wert in einem Intervall, sondern nur eine endliche Zahl von Merkmalsausprägungen angenommen werden kann. Wenn zum Beispiel die Größe von Menschen untersucht wird und hier davon ausgegangen werden soll, dass nur Werte zwischen 100 cm und 250 cm realistisch sind, so können bei einer Messgenauigkeit von 1 cm nur 151 mögliche Werte ermittelt werden.

Andererseits ist es manchmal auch sinnvoll, die Ausprägungen eines stetigen Merkmals in Gruppen zusammenzufassen. Die so erzeugten gruppierten Daten können als diskret angesehen werden. Beispielsweise kann es bei Untersuchungen am Menschen aus Gründen des Datenschutzes notwendig sein, das Alter nicht genau, son-

dern in Altersgruppen zu erfassen (unter 20 Jahre, 20–30 Jahre usw.). Ein anderes Beispiel für gruppierte Daten sind Klausurergebnisse, bei denen nicht die konkreten Fehlerzahlen, sondern die erreichten Noten festgehalten werden.

Für die statistische Datenanalyse, besonders für die Auswahl des adäquaten statistischen Verfahrens, ist das Skalenniveau (Skalentyp, Skalenart) des betrachteten Merkmals von besonderer Bedeutung.

Merksatz

Die Auswahl des geeigneten statistischen Verfahrens für die Datenanalyse ist unmittelbar vom Skalenniveau der untersuchten Merkmale abhängig.

Grundsätzlich können vier Skalentypen unterschieden werden, die für die Auswahl statistischer Verfahren relevant sind: Nominalskala, Ordinalskala, Intervallskala und Verhältnisskala.

Definition

Ein Merkmal wird als nominalskaliert bezeichnet, wenn die Merkmalsausprägungen diskrete Kategorien sind. Zwischen den Kategorien besteht keine Ordnungsrelation.

Ein typisches Beispiel für ein nominalskaliertes Merkmal ist das Geschlecht. Jeder der Untersuchungseinheiten kann eine der Kategorien weiblich oder männlich zugewiesen werden. Die Kategorien sind gleichwertig, eine Ordnungsrelation besteht nicht. Ein nominalskaliertes Merkmal mit zwei möglichen Merkmalsausprägungen wird auch als dichotomes (alternatives) Merkmal bezeichnet.

Ein anderes Beispiel eines nominalskalierten Merkmals ist die Farbe von Bakterienkolonien (siehe Kapitel 2) mit den Ausprägungen gelb, weißlich, braun, orange, farblos, rosa und grün. Auch hier sind die Kategorien gleichwertig ohne eine Ordnungsstruktur. Bei der Datenspeicherung werden den Kategorien üblicherweise Zahlen zugeordnet (1: gelb, 2: weißlich, 3: braun und so weiter). Diese Zahlen sind jedoch nur als Kennzeichnungen der Kategorien zu verstehen. Es ist nicht sinnvoll, numerische Rechenoperationen mit diesen Zahlen durchzuführen. Möglich ist lediglich die Feststellung der Häufigkeiten des Auftretens der einzelnen Kategorien in der gegebenen Stichprobe. Die später für nominalskalierte Merkmale beschriebenen statistischen Verfahren benutzen lediglich diese Häufigkeitsinformationen.

Die Nominalskala ist die Skala mit dem niedrigsten Informationsgehalt. Zusätzliche Informationen beinhaltet eine Ordinalskala.

Definition Ein Merkmal wird als ordinalskaliert bezeichnet, wenn die Merkmalsausprägungen in eine Rangfolge gebracht werden können, ihre Abstände aber nicht interpretierbar sind.

Die Ordinalskala soll an einem Beispiel erläutert werden, das in Kapitel 2 verwendet wird. Das Merkmal Antibiotikaresistenz von Bakterienkolonien hat die möglichen Merkmalsausprägungen sehr sensitiv, sensitiv, intermediär, resistent und sehr resistent. Die Merkmalsausprägungen weisen eine Ordnung auf und können in die Reihenfolge 1: sehr sensitiv, 2: sensitiv, 3: intermediär, 4: resistent und 5: sehr resistent gebracht werden (hier in einer Reihenfolge nach dem Grad der Sensitivität). Zur Häufigkeitsinformation, wie bei den nominalskalierten Merkmalen, kommt also noch die Rangfolgeinformation hinzu. Es gibt aber keine Informationen, ob der Unterschied zwischen einer Kolonie mit dem Wert sehr sensitiv und einer zweiten Kolonie mit der Merkmalsausprägung sensitiv kleiner, ebenso groß oder größer ist als der Abstand zwischen zwei Kolonien mit den Ausprägungen intermediär und resistent. Verfahren zur Analyse ordinalskalierter Merkmale benutzen die Häufigkeits- und die Ranginformationen der Daten.

Die Skalenarten mit dem höchsten Informationsgehalt sind die Intervall- und die Verhältnisskala.

Definition Ein Merkmal wird als intervallskaliert bezeichnet, wenn die Abstände der Merkmalsausprägungen durch eine Skala erfasst werden. Intervallskalen besitzen keinen absoluten Nullpunkt. Im Gegensatz dazu weisen Verhältnisskalen zusätzlich einen absoluten Nullpunkt auf. Intervall- und Verhältnisskala können unter dem Begriff metrische Skala (Kardinalskala) zusammengefasst werden.

Im Vergleich zu einer Ordinalskala erlaubt die Intervallskala, zusätzlich zur Anordnung der Merkmalsausprägungen deren Abstände zu interpretieren. Als typisches Beispiel einer Intervallskala wird in der Literatur die Temperatur angeführt, die in Grad Celsius (°C) gemessen wird. Ein Temperaturunterschied zwischen 1 °C und 9 °C und eine Differenz zwischen 21 °C und 29 °C sind mit jeweils 8 °C gleich groß. Allerdings gibt es bei dieser Skala keinen absoluten Nullpunkt, da der Nullpunkt dieser Skala als Gefrierpunkt des Wassers willkürlich festgelegt wurde. Deshalb ist es physikalisch nicht sinnvoll, davon zu sprechen, dass 20 °C doppelt so warm wie 10 °C seien.

Im Unterschied zur Celsius-Skala weist die Kelvin-Skala der Temperatur einen absoluten Nullpunkt auf. Deshalb ist bei dieser Skala eine Quotientenbildung möglich. Es ist physikalisch sinnvoll, bei 200 K von der doppelt so hohen Temperatur gegen-

über 100 K zu sprechen. Die meisten Merkmale, die in biowissenschaftlichen Untersuchungen betrachtet werden, weisen das Niveau einer Verhältnisskala auf. Beispiele sind Längen-, Größen- und Gewichtsmaße.

Verfahren zur Analyse von Merkmalen mit dem Skalenniveau einer Verhältnisskala können neben den Häufigkeits- und den Ranginformationen der Daten auch arithmetische Operationen mit den Daten (Addition, Subtraktion, Produkt- und Quotientenbildung) durchführen und deren Ergebnisse benutzen.

Für die Mehrzahl biostatistischer Verfahren gibt es kaum praktische Unterschiede zwischen einer Intervall- und einer Verhältnisskala, die deshalb oft unter dem Oberbegriff metrische Skala (Kardinalskala) zusammengefasst werden. Metrisch skalierte Merkmale werden als metrische Merkmale bezeichnet.

Die dargestellten Skalenarten weisen einen unterschiedlichen Informationsgehalt auf, der für ein- und dasselbe Merkmal von der Nominalskala über die Ordinalskala zur metrischen Skala ansteigt. Statistische Analyseverfahren, die für ein bestimmtes Skalenniveau geeignet sind, können unter entsprechendem Informationsverlust auch auf Merkmale eines niedrigeren Skalenniveaus angewendet werden. Diese Möglichkeit kann in der statistischen Datenanalyse genutzt werden, wenn zum Beispiel die für die Analyse metrischer Merkmale zur Verfügung stehenden Verfahren wegen verletzter Voraussetzungen (zum Beispiel bei nicht normalverteilten Daten) nicht angewendet werden können. In einem solchen Fall kann ein Ausweg darin bestehen, nichtparametrische Verfahren anzuwenden, die lediglich die Informationen ordinalskalierter Daten benutzen (siehe Kapitel 6).

Zusammenfassung

Die Biostatistik ist ein integraler Bestandteil biowissenschaftlicher Forschung. Dabei kann sie im Rahmen explorativer Datenanalysen dazu beitragen, Hypothesen über die untersuchten Merkmale oder über unbekannte Datenstrukturen zu gewinnen. In konfirmatorischen Analysen können mit Hilfe biostatistischer Methoden Entscheidungen über Hypothesen getroffen werden, die vor der Untersuchung aufgestellt wurden.

Ziel biowissenschaftlicher Untersuchungen sind typischerweise Aussagen über die Eigenschaften untersuchter Populationen. Da in den Untersuchungen oft nur die Daten aus Stichproben ausgewertet werden können, muss bei der Bildung der Stichproben sichergestellt werden, dass sie die Verhältnisse in der Population möglichst genau widerspiegeln. Besonders günstig sind dafür Zufallsstichproben.

Voraussetzung für die Auswahl der geeigneten biostatistischen Methoden zur Datenanalyse ist die Kenntnis des Skalenniveaus der untersuchten Merkmale. Vier wichtige Skalentypen sind die Nominalskala, die Ordinalskala, die Intervall- und die Verhältnisskala. Die Skalenarten weisen bezüglich ihres Informationsgehaltes eine hierarchische Struktur auf.

Übungsaufgaben

Aufgabe 1.1

Welches Skalenniveau weisen folgende Merkmale auf, die an Ratten erhoben werden sollen:

a. Alter (in Monaten),

b. Fellfarbe (fünf Farben),

c. Anzahl der Zähne,

d. Geschlecht,

e. Gewicht?

Aufgabe 1.2

Um welchen Skalentyp handelt es sich bei

a. Schulnoten,

b. Windstärken,

c. Waldschadensklassen?

Ausführliche Lösungen sowie weitere Aufgaben finden Sie auf der Companion Website zum Buch unter **http://www.pearson-studium.de**

Beschreibende Statistik eines Merkmals

2

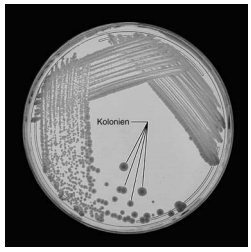
ÜBERBLICK

Nach der Erhebung biologischer Daten und ihrer Erfassung in entsprechenden Datenbanken oder Statistikprogrammen besteht die erste Aufgabe der Auswertung in der Regel darin, einen Überblick über die Daten und ihre wichtigsten Eigenschaften zu gewinnen. Diesem Ziel der reinen Beschreibung von Daten dienen Methoden der beschreibenden (deskriptiven) Statistik. Dabei ist für alle Auswertungen das Skalenniveau des untersuchten Merkmals für die Auswahl der jeweils geeigneten Methode zu beachten. In den folgenden Abschnitten werden die drei grundsätzlichen Vorgehensweisen beschrieben:

- Darstellung der Daten in Tabellen.
- Grafische Darstellung der Daten.
- Datenbeschreibung durch charakteristische Maßzahlen.

Dabei sollen in diesem Kapitel ausschließlich Verfahren der beschreibenden Statistik *eines* Merkmals (d.h. eindimensionaler Verteilungen) behandelt werden. Entsprechende Methoden zur Beschreibung des Zusammenhanges von *zwei* Merkmalen (d.h. für zweidimensionale Verteilungen) werden im Kapitel zur Zusammenhangsanalyse (Kapitel 7) behandelt.

Anwendungsbeispiel



Bakterienkolonien

In einer mikrobiologischen Untersuchung sollten die Eigenschaften von Mikroorganismen in der Luft untersucht werden. Dazu wurde ein Nährboden auf einer runden Agarplatte 30 Minuten bei Zimmertemperatur offen im Raum stehen gelassen. Nach Inkubation über drei Tage waren 40 Pilz- bzw. Bakterienkolonien gewachsen. Von diesen 40 Kolonien wurden der Durchmesser, die Farbe sowie die Antibiotikaresistenz (ordinalskaliert auf einer fünfstufigen Skala) bestimmt. Die erfassten Merkmale und die Daten sind in ►Tabelle 2.1 bzw. in ►Tabelle 2.2 dargestellt. Mit Methoden der beschreibenden Statistik soll die Verteilung der Merkmale beschrieben werden. Dabei soll zusätzlich betrachtet werden, ob sich die Verteilungen der Durchmesser zwischen den Kolonien unterschiedlicher Farbe unterscheiden.

Merkmal	Skalenniveau	Erläuterungen
X: Durchmesser	metrisch	in mm
Y: Antibiotika-resistenz	ordinal	5 Ausprägungen (1: sehr sensitiv; 2: sensitiv; 3: intermediär; 4: resistent; 5: sehr resistent)
Z: Farbe	nominal	7 Ausprägungen (1: gelb; 2: weißlich; 3: braun; 4: orange; 5: farblos; 6: rosa; 7: grün)

Tabelle 2.1: Variablen im Anwendungsbeispiel.

Nummer	Durchmesser	Resistenz	Farbe
1	0.5	sehr sensitiv	gelb
2	4.1	sensitiv	gelb
3	4.4	intermediär	gelb
4	5.6	resistent	gelb
5	6.8	sehr resistent	gelb
6	7.2	sehr sensitiv	gelb
7	7.7	resistent	gelb
8	7.8	intermediär	gelb
9	8.2	resistent	gelb
10	9.5	sehr resistent	gelb
11	9.2	sehr sensitiv	gelb
12	9.9	sensitiv	gelb
13	11.9	intermediär	gelb
14	2.1	resistent	weißlich
15	2.2	sehr resistent	weißlich
16	2.2	sehr sensitiv	weißlich
17	4.1	sensitiv	weißlich
18	5.8	sehr sensitiv	weißlich
19	5.8	sensitiv	weißlich
20	5.8	sehr sensitiv	weißlich
21	6.2	sensitiv	weißlich
22	6.4	sehr sensitiv	weißlich
23	6.4	sensitiv	weißlich
24	7.9	sehr sensitiv	weißlich
25	9.8	sensitiv	weißlich
26	9.8	sehr sensitiv	weißlich
27	10.1	sehr sensitiv	weißlich
28	0.2	sehr sensitiv	orange
29	1.5	sensitiv	orange
30	2.8	intermediär	farblos
31	3.2	resistent	farblos
32	2.4	sehr resistent	farblos
33	6.6	sensitiv	farblos
34	4.2	resistent	rosa
35	8.1	intermediär	rosa
36	5.8	intermediär	rosa
37	6.2	sehr sensitiv	rosa
38	10.1	sehr sensitiv	braun
39	3.3	intermediär	grün
40	4.2	intermediär	grün

Tabelle 2.2: Daten im Anwendungsbeispiel (Urliste).

2.1 Darstellung der Daten in Tabellen

Im Anwendungsbeispiel liegen 40 Datensätze vor, in praktischen biologischen Untersuchungen werden häufig noch weit mehr Daten erhoben. Aus den vorliegenden Daten lassen sich unmittelbar keine Schlüsse auf die Verteilungsform und auf weitere Eigenschaften der Daten ziehen. Der erste Schritt zur Beschreibung der Daten besteht deshalb in der Erzeugung von Häufigkeitstabellen, die außerdem die Grundlage für viele grafische Darstellungen bilden. Dabei ist zu unterscheiden, ob die Daten in kategorisierter Form (wie im Beispiel bei den Merkmalen Farbe [sieben Klassen] oder Antibiotikaresistenz [fünf Stufen]) vorliegen oder in metrischer, nicht kategorisierter Form, wie im Beispiel beim Merkmal Durchmesser. Im ersten Fall können die Häufigkeitstabellen unmittelbar auf der Grundlage der vorliegenden Daten erstellt werden. Beim Vorliegen nicht kategorisierter Merkmale müssen die Daten dagegen zunächst in Klassen zusammengefasst werden. Die dazu notwendige Vorgehensweise soll im folgenden Abschnitt am Beispiel des Merkmals Durchmesser beschrieben werden.

2.1.1 Anzahl und Breite der Klassen

In Tabelle 2.2 sind die Daten des Merkmals Durchmesser ungeordnet in der Urliste dargestellt. Der erste Schritt zur Erzeugung einer Klasseneinteilung besteht darin, die Daten der Größe nach geordnet in einer Primärliste darzustellen (►Tabelle 2.3).

0.2	0.5	1.5	2.1	2.2	2.2	2.4	2.8	3.2	3.3
4.1	4.1	4.2	4.2	4.4	5.6	5.8	5.8	5.8	5.8
6.2	6.2	6.4	6.4	6.6	6.8	7.2	7.7	7.8	7.9
8.1	8.2	9.2	9.5	9.8	9.8	9.9	10.1	10.1	11.9

Tabelle 2.3: Primärliste des Merkmals Durchmesser der Kolonien (in mm).

Aus der Primärliste können Minimum (*Min*) und Maximum (*Max*) der gegebenen Werte abgelesen werden. Im **Anwendungsbeispiel** (alle Angaben zum Durchmesser in mm) ergeben sich die Werte $Min = 0.2$ bzw. $Max = 11.9$. Aus der Differenz von Maximum und Minimum ergibt sich die Variationsbreite V.

Formel

$$V = Max - Min \tag{2.1}$$

- V: Variationsbreite
- *Max*: Maximum
- *Min*: Minimum

Im **Beispiel** beträgt die Variationsbreite $V = 11.9 - 0.2 = 11.7$. Sie muss durch die zu bestimmenden Klassen abgedeckt werden. Dazu sind die Anzahl und die Breite der

Klassen festzulegen. Die Anzahl der Klassen sollte mit steigendem Datenumfang und mit größerer Variationsbreite zunehmen. In der Praxis hat sich die Faustregel von Sturges (1926) bewährt, die einen wichtigen Anhaltspunkt für die zu wählende Klassenanzahl m bietet.

Formel

$$m \approx 1 + 3.32 \cdot \lg(n) \tag{2.2}$$

- m: Anzahl der Klassen
- n: Anzahl der Messwerte

Für $n = 10$ ergibt sich der Wert $m \approx 1 + 3.32 \cdot \lg(10) = 4.32$. Naheliegend wäre demnach die Wahl von vier Klassen. Für $n = 100$ erhält man $m \approx 1 + 3.32 \cdot \lg(100) = 7.64$, danach würde vorrangig eine Klassenanzahl von sieben oder acht in Betracht kommen. Im **Anwendungsbeispiel** ergibt sich der Richtwert $m \approx 1 + 3.32 \cdot \lg(40) = 6.32$, wonach sechs oder sieben Klassen empfehlenswert wären.

Die tatsächlich günstigste Klassenzahl und die Breite der Klassen können aber nur festgelegt werden, wenn gleichzeitig die Variationsbreite sowie Minimum und Maximum berücksichtigt werden. Dabei ist zu beachten, dass sowohl die Klassengrenzen als auch die Mitten der Klassen möglichst runde Werte sein sollen. Wenn die Klassengrenzen viele Nachkommastellen aufweisen, wird die Übersichtlichkeit und Lesbarkeit der Tabellen und der darauf basierenden grafischen Darstellungen wesentlich erschwert. Deshalb sollte die Genauigkeit der Klassengrenzen und der Mitten der Klassen die Messgenauigkeit des Merkmals nach Möglichkeit nicht überschreiten. Wenn also die Messwerte ganzzahlig erfasst werden, sollten die Werte zur Beschreibung der Klassen ebenfalls ganzzahlig sein; wenn die Messwerte eine Nachkommastelle haben, sollten die Klassengrenzen nach Möglichkeit ebenfalls höchstens eine Nachkommastelle haben. Um dieses Ziel erreichen zu können, ist es oft hilfreich, die Variationsbreite künstlich etwas zu erweitern.

Das Prinzip soll am **Anwendungsbeispiel** veranschaulicht werden (alle Angaben zum Durchmesser in mm). Hier beträgt die Variationsbreite $V = 11.7$, der kleinste Wert ist 0.2. Der Versuch, diese Variationsbreite auf sechs oder sieben Klassen aufzuteilen, würde in keinem Fall zu annähernd glatten Klassengrenzen führen. Wegen der Messgenauigkeit von 0.1 mm beinhaltet die Variationsbreite von 11.7 genau 118 mögliche Messwerte (0.2, 0.3, 0.4, …, 11.8, 11.9). 118 ist weder durch 6 noch durch 7 teilbar. Eine sinnvolle Erweiterung der Variationsbreite besteht in diesem Beispiel darin, zwei zusätzliche mögliche Messwerte vorzusehen. Damit ergeben sich 120 mögliche Messwerte, die in sechs gleich große Klassen eingeteilt werden können (mit je 20 möglichen Messwerten). Es bietet sich an, den zusätzlichen Wert 0.1 am Anfang des Wertebereichs anzufügen und den Wert 12.0 am Ende. Damit ist einer-

seits der zusätzliche Wertebereich symmetrisch auf das Ende und auf den Anfang der Verteilung aufgeteilt. Andererseits ergeben sich glatte Klassengrenzen und Klassenmitten. Nach diesen Vorüberlegungen ist es möglich, die in ►Tabelle 2.4 angegebene Klasseneinteilung vorzunehmen.

Klasse	Grenzen der Klasse	Mitte der Klasse
1	$0 < x \leq 2$	1
2	$2 < x \leq 4$	3
3	$4 < x \leq 6$	5
4	$6 < x \leq 8$	7
5	$8 < x \leq 10$	9
6	$10 < x \leq 12$	11

Tabelle 2.4: Klassen für die Darstellung der Durchmesser der Kolonien.

Es soll an dieser Stelle ausdrücklich darauf hingewiesen werden, dass es für die Klasseneinteilung eines metrischen Merkmals oft mehrere sinnvolle Lösungen gibt. Zur Plausibilitätsüberprüfung einer gefundenen Klasseneinteilung sollten folgende Fragen beantwortet werden:

- Ist die Klassenanzahl angemessen? Im **Beispiel** ergab sich die Klassenanzahl nach Formel (2.2). Sechs Klassen sind zur Beschreibung von 40 Werten sinnvoll.
- Sind die Klassengrenzen und die Klassenmitten angemessen „rund"? Im **Anwendungsbeispiel** sind die Grenzen und die Mitten der Klassen ganzzahlig. Grenzen und Klassen mit einer Nachkommastelle wären ebenfalls befriedigend.
- Sind die Klassen disjunkt, d.h. sind alle möglichen Messwerte eindeutig zu genau einer Klasse zuzuordnen? Diese notwendige Bedingung für sinnvolle Klassen ist im **Beispiel** erfüllt, die Klassengrenzwerte (2, 4, 6, 8, 10 und 12) sind eindeutig zu jeweils einer der Klassen zugeordnet.
- Sind die Klassen gleich groß? Diese Bedingung ist im **Beispiel** gegeben, alle Klassen haben die Breite 2 bzw. beinhalten 20 mögliche Messwerte. Da die Klasseneinteilung einen objektiven Überblick über die Merkmalsverteilung sichern soll, ist die Forderung nach gleichen Klassenbreiten notwendig. Eine Ausnahmesituation kann dann gegeben sein, wenn in großen Merkmalsmengen einzelne Ausreißer vorkommen, die die Klassenbildung und damit den Informationsgehalt der entstehenden Häufigkeitsverteilung stark beeinflussen würden. In einem solchen Fall kann es sinnvoll sein, mit offenen Randklassen zu arbeiten (zum Beispiel Klasse 6: $x > 10$).
- Überdecken die Klassen den Wertebereich vollständig? Im **Beispiel** überdecken die Klassen den Wertebereich von 0.2 bis 11.9 komplett.
- Ist die künstliche Erweiterung der Variationsbreite angemessen? Im **Beispiel** umfasst die Variationsbreite $V = 11.7$ 118 mögliche Messwerte. Diese Anzahl wurde um 2 erhöht, um ganzzahlige Werte für die Beschreibung der Klassen

angeben zu können. Diese künstliche Erweiterung der Variationsbreite ist sinnvoll. In der Praxis trifft man jedoch oft auf Situationen, in denen die Entscheidung weniger eindeutig ist. In jedem Einzelfall muss die Entscheidung getroffen werden, ob der Gewinn an Übersichtlichkeit der Darstellung die jeweilige künstliche Erweiterung der Variationsbreite rechtfertigt. Das Problem besteht darin, dass bei zu starken Erweiterungen automatisch Randklassen mit geringen Häufigkeiten entstehen. Deshalb sollte angestrebt werden, nur die objektiv notwendigen Erweiterungen vorzunehmen und diese zusätzlichen Wertebereiche weitgehend gleichmäßig auf die Randklassen aufzuteilen.

2.1.2 Merkmalsverteilung

Die Klasseneinteilung des Wertebereiches des zu untersuchenden Merkmals ist die Grundlage für die Bestimmung der Merkmalsverteilung. Ausgehend von der Primärliste (Tabelle 2.3) werden klassifizierte Häufigkeiten und klassifizierte Summenhäufigkeiten bestimmt.

Formel

$$h_i^{\%} = \frac{h_i}{n} \cdot 100\% \tag{2.3}$$

$$H_i = \sum_{j=1}^{i} h_j \tag{2.4}$$

$$H_i^{\%} = \frac{H_i}{n} \cdot 100\% \tag{2.5}$$

- n: Anzahl der Messwerte
- m: Anzahl der Klassen
- h_i: Klassifizierte Häufigkeit der Messwerte, d.h. Anzahl der Messwerte in Klasse i ($i = 1,\ldots,m$)
- $h_i^{\%}$: Prozentuale klassifizierte Häufigkeit der Messwerte in Klasse i ($i = 1,\ldots,m$)
- H_i: Klassifizierte Summenhäufigkeit, d.h. Anzahl der Messwerte bis einschließlich Klasse i ($i = 1,\ldots,m$)
- $H_i^{\%}$: Prozentuale klassifizierte Summenhäufigkeit, d.h. prozentualer Anteil der Messwerte bis einschließlich Klasse i ($i = 1,\ldots,m$)

Für die Durchmesser der Kolonien im **Beispiel** ergibt sich auf der Grundlage der Klasseneinteilung aus Tabelle 2.4 die in ►Tabelle 2.5 dargestellte empirische Häufigkeitsverteilung.

Klasse i	Grenzen	h_i	$h_i^{\%}$	H_i	$H_i^{\%}$
1	$0 < x \leq 2$	3	7.5 %	3	7.5 %
2	$2 < x \leq 4$	7	17.5 %	10	25.0 %

Tabelle 2.5: Klassifizierte Häufigkeitstabelle der Durchmesser der Kolonien.

Klasse i	Grenzen	h_i	$h_i^{\%}$	H_i	$H_i^{\%}$
3	$4 < x \leq 6$	10	25.0 %	20	50.0 %
4	$6 < x \leq 8$	10	25.0 %	30	75.0 %
5	$8 < x \leq 10$	7	17.5 %	37	92.5 %
6	$10 < x \leq 12$	3	7.5 %	40	100.0 %

Tabelle 2.5: Klassifizierte Häufigkeitstabelle der Durchmesser der Kolonien (Fortsetzung).

Die Häufigkeitsangaben können analog auch für die unklassifizierten Messwerte vorgenommen werden. Von einer Statistik-Software kann man sich diese Angaben routinemäßig liefern lassen. Da diese Darstellungen in der Regel aber keine Hilfe für die Bestimmung empirischer Merkmalsverteilungen sind, soll darauf nicht eingegangen werden. Bei nominal- bzw. ordinalskalierten Variablen (wie im **Beispiel** Farbe und Antibiotikaresistenz) sind die Klassen in der Regel durch die Kategorien der Variablen vorgegeben. Sehr schwach besetzte Kategorien können gegebenenfalls zusammengefasst werden. Bei nominalskalierten Merkmalen ist die Berechnung von Summenhäufigkeiten nicht sinnvoll.

2.2 Grafische Darstellung der Daten

Die in Tabellen zusammengefassten klassifizierten Häufigkeiten bilden den Ausgangspunkt für die grafische Veranschaulichung der empirischen Merkmalsverteilung. Grafische Darstellungen bieten im Vergleich zu Tabellen bessere Möglichkeiten, die typischen Eigenschaften von erhobenen Daten anschaulich abzubilden. Aus einer Vielzahl von grafischen Darstellungsmöglichkeiten sollen nachfolgend die wichtigsten Typen vorgestellt werden. Die Anwendung der verschiedenen Grafiktypen ist vom Skalenniveau des untersuchten Merkmals abhängig. Im folgenden Text werden jeweils Standardtypen der verwendeten Grafiken dargestellt. Auf weiterführende Gestaltungsmöglichkeiten der Grafiken wird bei der Behandlung der Vorgehensweisen in den Statistikprogrammen auf der beiliegenden CD eingegangen.

Für nominal- oder ordinalskalierte Merkmale, aber auch bei klassifizierten metrischen Variablen bieten sich Diagramme an, die die klassifizierten Häufigkeiten oder die prozentualen klassifizierten Häufigkeiten darstellen. Typische Beispiele für solche Darstellungen sind Balken- und Kreisdiagramme.

2.2.1 Balkendiagramm

In Balkendiagrammen werden die klassifizierten Häufigkeiten h_i $(i = 1, \ldots, m)$ des betrachteten Merkmals dargestellt. Die Balken sind separat in der Grafik angeordnet, so dass die Klassen bzw. Kategorien keine Ordnung oder Beziehung untereinander aufweisen. In ►Abbildung 2.1 ist das Balkendiagramm der klassifizierten Häufigkeiten der Antibiotikaresistenz dargestellt. Aus dem Diagramm wird deutlich, dass

mehr als 50 % der Kolonien sensitiv oder sehr sensitiv gegenüber Antibiotika waren, die meisten Kolonien erwiesen sich als sehr sensitiv.

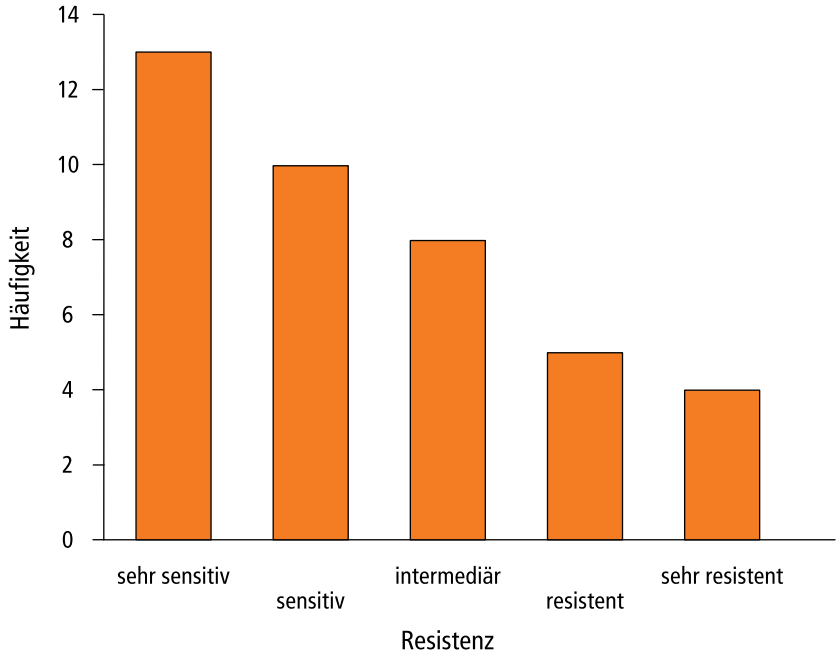


Abbildung 2.1: Balkendiagramm der Häufigkeitsverteilung des Merkmals Antibiotikaresistenz.

2.2.2 Kreisdiagramm

Kreisdiagramme werden sehr häufig zur Darstellung der Häufigkeitsverteilung von nominal- bzw. ordinalskalierten Merkmalen benutzt. Hier werden die prozentualen klassifizierten Häufigkeiten in Kreissegmente transformiert (Winkel $\alpha = h_i^{\%} \cdot 3.6°, 0 \leq h_i^{\%} \leq 100$). Aus ►Abbildung 2.2 wird deutlich, dass im **Beispiel** die Farben gelb und weißlich deutlich häufiger als die restlichen Farben auftreten. Am häufigsten wurden weißliche Kolonien beobachtet. Gelbe, weißliche und Kolonien sonstiger Farbe traten annähernd gleich häufig auf. Bei der Interpretation von Kreisdiagrammen ist zu beachten, dass hier lediglich Aussagen über die relativen Anteile der einzelnen Kategorien möglich sind (falls die absoluten Häufigkeiten nicht zusätzlich angegeben werden). Ähnliche Aussagen liefern Komponentenstabdiagramme, bei denen anstelle des Kreises eine schmale, stabförmige Rechteckfläche im analogen Verhältnis zur Häufigkeit der Merkmalsausprägungen unterteilt wird.

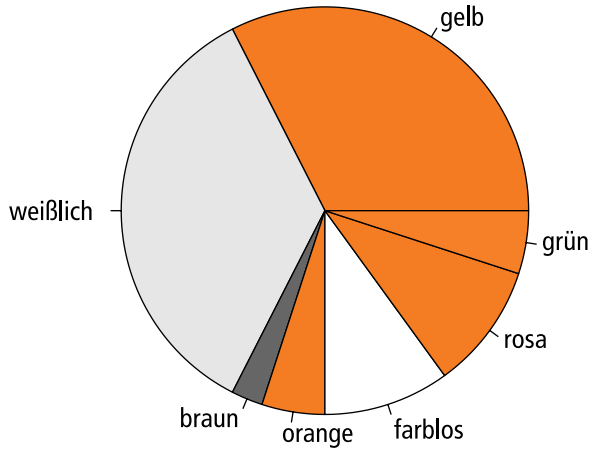


Abbildung 2.2: Kreisdiagramm der Häufigkeitsverteilung des Merkmals Farbe der Kolonien.

Die bisher vorgestellten Diagrammtypen können für klassifizierte metrische Merkmale grundsätzlich ebenfalls verwendet werden. Aussagekräftiger sind in diesem Fall aber Diagrammtypen, die den Informationsgehalt der Intervall- bzw. Verhältnisskala berücksichtigen. Die wichtigsten Grafiktypen für die Darstellung von Häufigkeitsverteilungen metrischer Merkmale sind deshalb Histogramme und Polygone.

2.2.3 Histogramm

Bei der Darstellung in Histogrammen werden auf der Abszisse die Klassengrenzen oder die Klassenmitten der Klasseneinteilung dargestellt, die für das Anwendungsbeispiel in Tabelle 2.4 enthalten sind. Im Unterschied zu der Darstellung in Balkendiagrammen sind die Klassen hier untereinander verbunden. Die klassifizierten Häufigkeiten werden über jeder Klasse abgetragen. Wenn die Breite jeder Klasse auf den Wert 1 standardisiert wird, entspricht die Fläche unter der Histogrammkurve der Anzahl der Messwerte. Damit kann man aus der Grafik unmittelbar Informationen über die Verteilung der Werte entnehmen. Aus ►Abbildung 2.3 wird deutlich, dass die Durchmesserwerte symmetrisch verteilt sind. In ►Abbildung 2.4 sind die Histogramme der Häufigkeitsverteilung der Durchmesser dargestellt, die sich getrennt für die gelben Kolonien ($n = 13$), die weißlichen Kolonien ($n = 14$) sowie für die Kolonien sonstiger Farbe ($n = 13$) ergeben. In den Diagrammen erkennt man unterschiedliche Verteilungsformen, bei deren Interpretation im vorliegenden Beispiel allerdings der sehr geringe Umfang der Teilstichproben zu berücksichtigen ist. Während die Häufigkeitsverteilung bei den weißlichen Kolonien ebenfalls symmetrisch ist, ist sie bei den gelben Kolonien leicht rechtssteil und bei den Kolonien sonstiger Farbe leicht linkssteil.

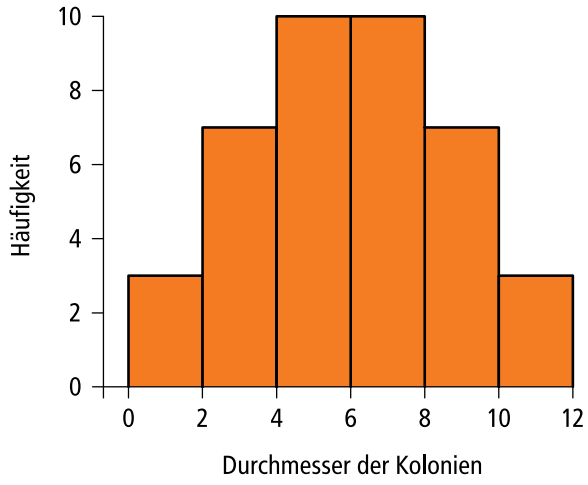


Abbildung 2.3: Histogramm der klassifizierten Häufigkeiten des Merkmals Durchmesser.

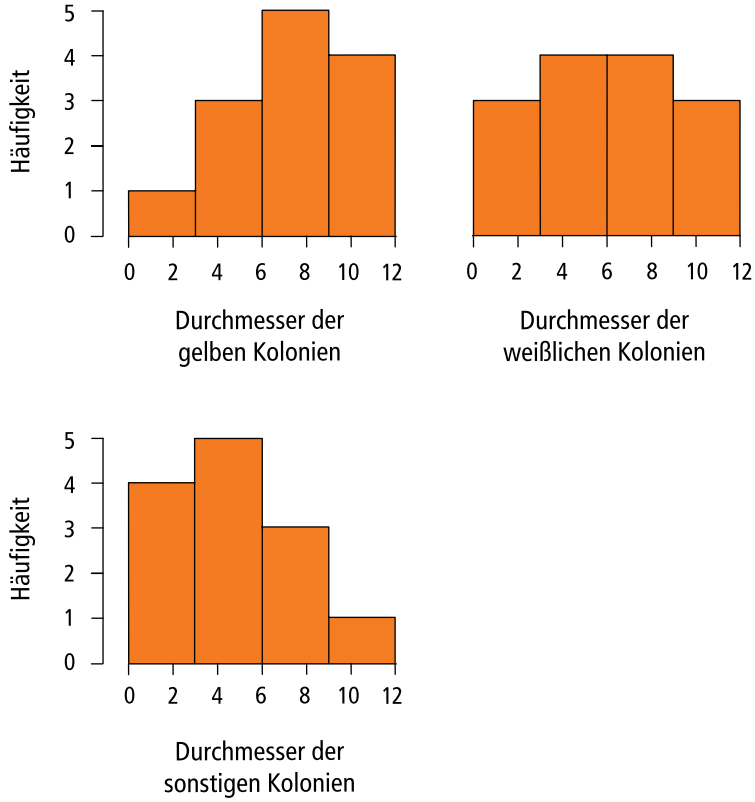


Abbildung 2.4: Histogramme der klassifizierten Häufigkeiten des Merkmals Durchmesser in den Teilstichproben.

2.2.4 Polygon

Oft noch deutlicher als aus Histogrammen lassen sich typische Verteilungseigenschaften aus Polygonzügen ableiten. Grundlage für die Darstellung sind erneut die klassifizierten Häufigkeiten h_i (Tabelle 2.5). Im Unterschied zur Darstellung in Histogrammen werden hier die klassifizierten Häufigkeiten über den Klassenmitten abgetragen und linear verbunden. Damit der Polygonzug bei 0 beginnt und damit – wie noch zu sehen sein wird – die Fläche unter dem Polygon der Fläche unter dem entsprechenden Histogramm entspricht, wird vor der ersten und nach der letzten Klasse jeweils eine zusätzliche leere Klasse mit der Häufigkeit 0 angefügt. Im **Beispiel** wird vor der ersten Klasse (Tabelle 2.4) die zusätzliche Klasse mit den Grenzen -2 und 0 eingefügt und nach der letzten Klasse die leere Klasse mit den Grenzen 12 und 14. An den Mitten dieser zusätzlichen Klassen (-1 bzw. 13) wird jeweils die klassifizierte Häufigkeit 0 eingetragen, der Polygonzug endet in diesen Punkten bei 0. In ►Abbildung 2.5 ist der Polygonzug der klassifizierten Häufigkeitsverteilung aus dem Anwendungsbeispiel dargestellt. Wie schon beim Histogramm (Abbildung 2.3) wird die Symmetrie der Häufigkeitsverteilung deutlich.

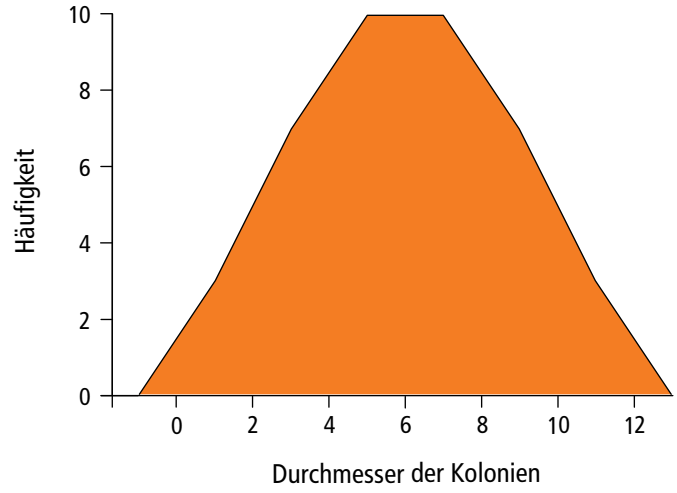


Abbildung 2.5: Polygon der klassifizierten Häufigkeiten des Merkmals Durchmesser.

►Abbildung 2.6 veranschaulicht den Zusammenhang von Histogramm und Polygon. Die Flächen unter beiden Kurven sind gleich. Dieser Sachverhalt wird deutlich, wenn man jeweils paarweise die vertikal schraffierten Dreiecksflächen (hier liegt der Polygonzug über der Histogrammkurve) mit den benachbarten horizontal schraffierten Dreiecksflächen (bei denen das Histogramm über dem Polygon liegt) vergleicht. Wenn – analog zur Interpretation der Histogramme – die Breite einer Klasse auf den Wert 1 standardisiert wird, entspricht die Fläche unter dem Polygon ebenfalls der Anzahl der Messwerte.

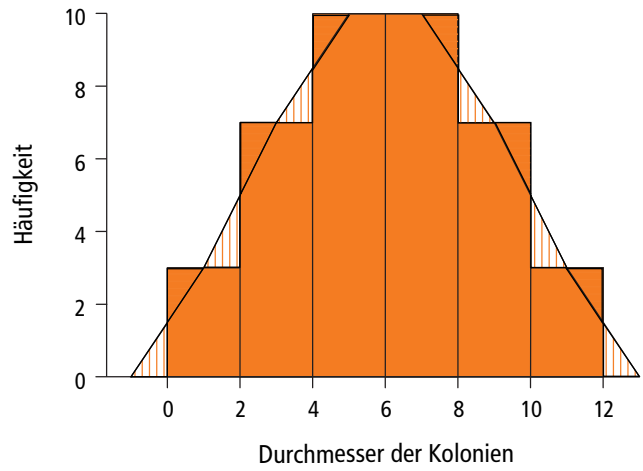


Abbildung 2.6: Zusammenhang von Polygon und Histogramm.

Mit dem Polygon wird die Verteilung der Messwerte über dem Wertebereich beschrieben. Dabei können – gegebenenfalls nach entsprechender Glättung der Polygone – typische Verteilungsmuster identifiziert werden (siehe ►Abbildung 2.7). Neben der Beschreibung der Daten kann die Verteilungsform erste Hinweise für die spätere statistische Analyse liefern, speziell für die Auswahl eines angemessenen statistischen Verfahrens.

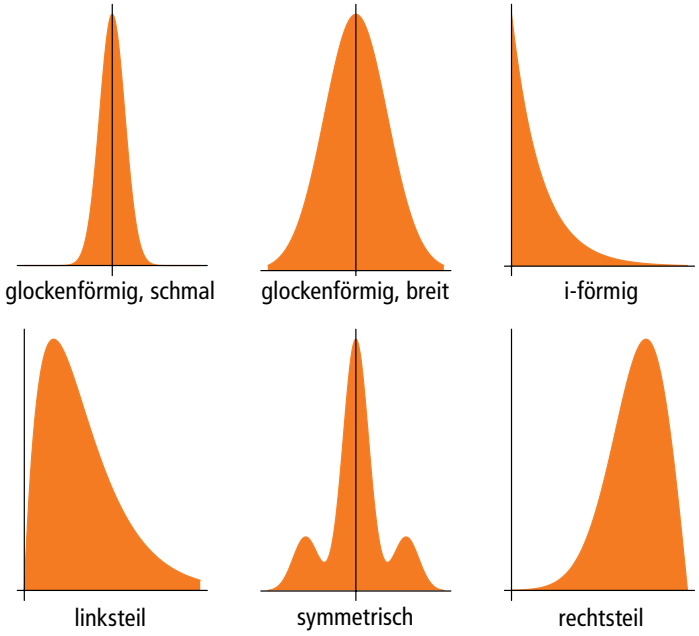


Abbildung 2.7: Typische Verteilungsformen.

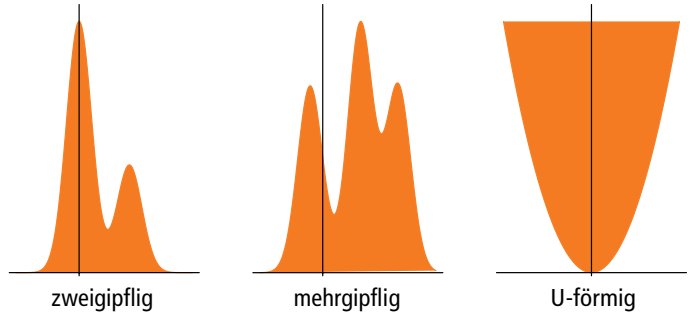


Abbildung 2.7: Typische Verteilungsformen (Fortsetzung).

Histogramm- und Polygondarstellungen liefern einen Eindruck von der Häufigkeitsverteilung der Messwerte über dem Wertebereich. Eine grafische Veranschaulichung der Summenhäufigkeiten liefern Summenhistogramme und Summenpolygone.

2.2.5 Summenhistogramm

Die Darstellung der klassifizierten Summenhäufigkeiten H_i in Summenhistogrammen erfolgt analog zur Histogrammdarstellung der klassifizierten Häufigkeiten h_i. Die Summenhäufigkeiten werden über den Klassen des untersuchten Merkmals abgetragen. Im **Anwendungsbeispiel** ergibt sich mit den in Tabelle 2.5 dargestellten Summenhäufigkeiten das in ►Abbildung 2.8 dargestellte Summenhistogramm. Es beschreibt die Zunahme der Summenhäufigkeiten über dem Wertebereich.

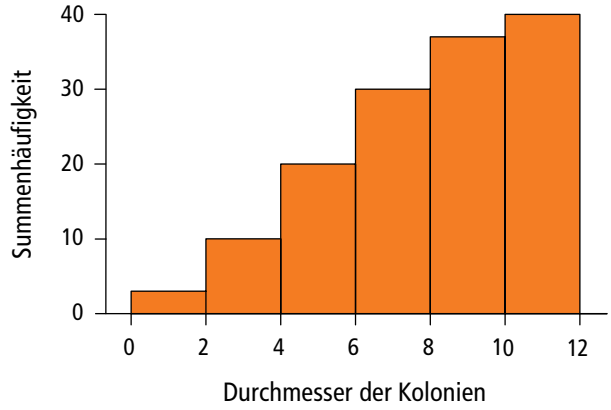


Abbildung 2.8: Summenhistogramm der klassifizierten Summenhäufigkeiten der Durchmesser.

2.2.6 Summenpolygon

In einem Summenpolygon werden die klassifizierten Summenhäufigkeiten H_i über dem jeweiligen Ende der Klassen abgetragen und linear verbunden. Die Darstellung beginnt am Anfang der ersten Klasse mit der Summenhäufigkeit 0 und endet am Ende der letzten Klasse mit der Summenhäufigkeit n (Anzahl der Werte). Das im **Anwendungsbeispiel** resultierende Summenpolygon ist in ►Abbildung 2.9 dargestellt. Aus ►Abbildung 2.10 wird deutlich, dass das Summenpolygon *unter* dem Summenhistogramm verläuft, beide Kurven treffen sich über den Klassengrenzen. Das Summenpolygon beschreibt die Anzahl der Werte, die *bis* zu einem bestimmten Punkt erfasst werden. So liegen bis zum Anfang der vierten Klasse (bis zum Wert 6.0) 20 Messwerte vor, bis zum Ende dieser Klasse sind es 30 Werte. Innerhalb der Klasse erfolgt eine lineare Interpolation.

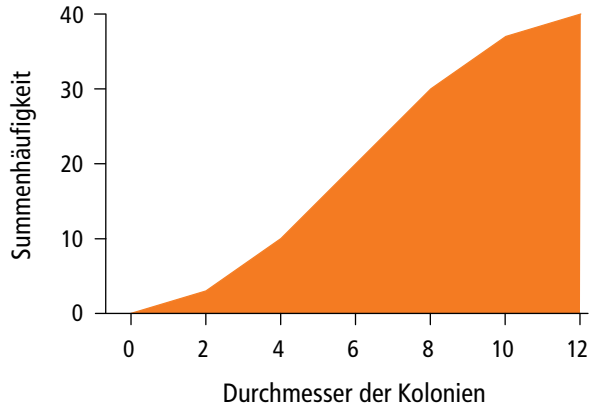


Abbildung 2.9: Summenpolygon der klassifizierten Summenhäufigkeiten der Durchmesser.

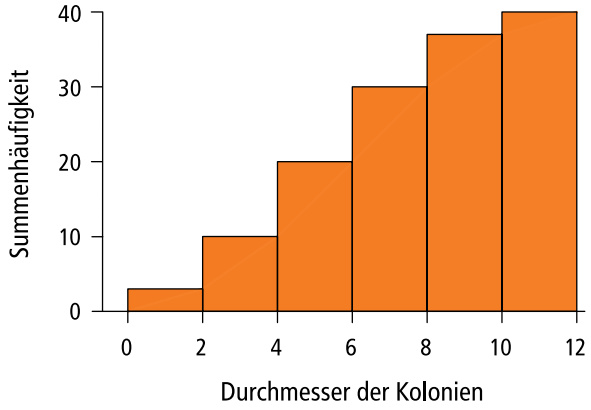


Abbildung 2.10: Zusammenhang von Summenpolygon und Summenhistogramm.

2.3 Statistische Kennwerte

Die grafische Darstellung von Messwerten liefert einen anschaulichen Eindruck von der Verteilung der Daten. Statistische Kennwerte repräsentieren die Daten mit wenigen Kennwerten und erlauben den quantitativen Vergleich unterschiedlicher Häufigkeitsverteilungen. Dabei ist grundsätzlich zwischen Lageparametern und Streuungsparametern zu unterscheiden. Lageparameter beschreiben die Lage (bzw. den Schwerpunkt oder die zentrale Tendenz) der Messwerte im Wertebereich. Streuungsparameter beschreiben demgegenüber die Streuung (bzw. die Unterschiedlichkeit) der Messwerte um den jeweiligen Lageparameter.

2.3.1 Lageparameter

Wichtige Parameter zur Beschreibung der Lage (des Schwerpunkts, der zentralen Tendenz) einer Verteilung von Messwerten sind der Modalwert, der Median und der arithmetische Mittelwert. Die Berechnung dieser Kenngrößen und ihre inhaltliche Interpretation sind sehr unterschiedlich. Die Möglichkeiten ihrer Verwendung hängen vom Datenniveau der Messwerte ab.

Modalwert

Der Modalwert ist besonders zur Beschreibung der Lage nominalskalierter Merkmale geeignet.

Definition

Als Modalwert (Mo) einer Menge von Messwerten wird der am häufigsten auftretende Wert bezeichnet. Wenn zwei oder mehr Werte am häufigsten vorkommen, gibt es mehrere Modalwerte.

Für das nominalskalierte Merkmal Farbe aus dem **Anwendungsbeispiel** (siehe Abbildung 2.2) ergibt sich als Modalwert die am häufigsten ermittelte Farbe (gelb).

Der Modalwert kann ebenfalls zur Beschreibung der Lage ordinalskalierter bzw. metrischer Merkmale eingesetzt werden. Allerdings ist bei unklassifizierten metrischen Messwerten die Angabe des Modalwerts oft wenig sinnvoll, weil die einzelnen Merkmalsausprägungen nur in geringer Anzahl auftreten. Wenn anstelle der Messwerte lediglich bereits klassifizierte Merkmale vorliegen, wird als Modalwert oft der mittlere Wert der am häufigsten auftretenden Klasse bzw. der am häufigsten auftretenden Klassen verwendet.

Median

Der Median kann für ordinalskalierte und für metrische Merkmale ermittelt werden.

Definition

Der Median (Md) ist ein Wert mit der Eigenschaft, dass in der Menge der nach der Größe geordneten Messwerte gleich viele Daten unterhalb und oberhalb des Medians liegen.

Im ersten Schritt müssen die Messwerte der Größe nach geordnet werden. Für das weitere Vorgehen muss unterschieden werden, ob eine gerade oder eine ungerade Anzahl an Messwerten vorliegt.

Für eine ungerade Anzahl ergibt sich der Median als der mittlere Wert. Bei fünf gegebenen Messwerten entspricht der Median also dem dritten Wert der geordneten Messwertreihe. Bei gegebenen, bereits geordneten Werten 4, 6, 9, 12, 14 erhält man $Md = 9$. Oberhalb und unterhalb von 9 liegen gleich viele (jeweils zwei) Messwerte.

Bei einer geraden Anzahl von Messwerten gibt es keinen unmittelbaren mittleren Wert. Der Median wird in diesem Fall als Durchschnittswert der beiden mittleren Werte berechnet. Wenn die geordneten Werte 1, 3, 4, 8, 12, 13 vorliegen, sind die mittleren dieser sechs Werte der dritte Wert (4) und der vierte Wert (8). Der Median ergibt sich danach als $Md = (4 + 8)/2 = 6$. Auch in diesem Fall liegen oberhalb und unterhalb des Medians gleich viele Messwerte, nämlich jeweils drei.

Im **Anwendungsbeispiel** wurden die Durchmesser von $n = 40$ Kolonien ermittelt. Die der Größe nach geordneten Messwerte sind in der Primärliste (Tabelle 2.3) enthalten. Die beiden mittleren Messwerte dieser geordneten Reihe sind der zwanzigste (5.8) und der einundzwanzigste (6.2) Wert. Damit ergibt sich der Median dieser Daten als

$$Md = (5.8 + 6.2)/2 = 6.$$

Oberhalb und unterhalb von 6 befinden sich jeweils 20 Messwerte. Für das ordinalskalierte Merkmal Antibiotikaresistenz ist ebenfalls die Rangfolge der Merkmalsausprägungen zu bestimmen. In geordneter Abfolge wurde dreizehnmal der Wert 1 (sehr sensitiv), zehnmal der Wert 2 (sensitiv), achtmal der Wert 3 (intermediär), fünfmal der Wert 4 (resistent) und viermal der Wert 5 (sehr resistent) ermittelt. Der zwanzigste und einundzwanzigste Wert dieser Rangreihe sind jeweils der Wert 2 (sensitiv). Damit ergibt sich $Md = 2$ (sensitiv). Jeweils 20 Werte liegen bei 2 (sensitiv) und darüber bzw. bei 2 (sensitiv) und darunter.

Arithmetischer Mittelwert

Der arithmetische Mittelwert kann nur für metrische Daten sinnvoll berechnet und interpretiert werden. Er ergibt sich als Durchschnitt der gegebenen Messwerte.

Formel

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (2.6)$$

- $\bar{x}$: arithmetischer Mittelwert
- x_i: Messwerte ($i = 1,\ldots,n$)
- n: Anzahl der Messwerte

Im **Beispiel** erhält man für den arithmetischen Mittelwert der Durchmesser der Kolonien mit den Daten aus Tabelle 2.2

$$\bar{x} = \frac{0.5 + 4.1 + 4.4 + \ldots + 10.1 + 3.3 + 4.2}{40} = 5.9.$$

Eigenschaften der Lageparameter

Der arithmetische Mittelwert ist für metrische Daten der Lageparameter mit dem höchsten Informationsgehalt, da alle Informationen der Daten in die Berechnung eingehen. Allerdings ist er dadurch auch anfällig gegenüber Messfehlern und Ausreißern. Wenn bei der Dateneingabe im **Beispiel** beim 38. Wert das Komma vergessen worden wäre, also anstelle 10.1 der Wert 101 eingetragen wäre (siehe Tabelle 2.2), hätte sich anstelle von $\bar{x} = 5.9$ ein arithmetischer Mittelwert $\bar{x} \approx 8.2$ ergeben, der vom tatsächlichen Mittelwert erheblich abweicht. Demgegenüber nutzt der Median zwar nicht alle Informationen der Daten aus, ist aber sehr robust gegenüber Ausreißern in den Daten. Im Beispiel hätte sich der Median überhaupt nicht verändert, wenn der beschriebene Fehler passiert wäre.

Eine weitere wichtige Eigenschaft der beiden Lageparameter kann veranschaulicht werden, wenn der Median und der arithmetische Mittelwert getrennt für die gelben, weißlichen und Kolonien sonstiger Farbe berechnet werden. Die grafische Darstellung der Verteilungsformen kann Abbildung 2.4 entnommen werden. Dort wird deutlich, dass die Durchmesser der gelben Kolonien leicht rechtssteil verteilt sind, die Durchmesser der weißlichen Kolonien eine symmetrische Verteilung aufweisen und die Durchmesser der sonstigen Kolonien leicht linkssteil verteilt sind. Diesen Verteilungsmustern entsprechen die in ►Tabelle 2.6 angegebenen Beziehungen von arithmetischem Mittelwert und Median.

Teilstichprobe	Verteilungsform	Arithmetischer Mittelwert	Median	Vergleich
gelbe Kolonien	rechtssteil	7.1	7.7	$\bar{x} < Md$
weißliche Kolonien	symmetrisch	6.0	6.0	$\bar{x} \approx Md$
sonstige Kolonien	linkssteil	4.5	4.2	$\bar{x} > Md$

Tabelle 2.6: Vergleich von arithmetischem Mittel und Median.

Wenn bei größeren Stichproben zusätzlich der Modalwert zum Vergleich herangezogen wird, ergeben sich die in ►Abbildung 2.11 dargestellten typischen Beziehungen bei linkssteilen, rechtssteilen und symmetrischen Verteilungsmustern.

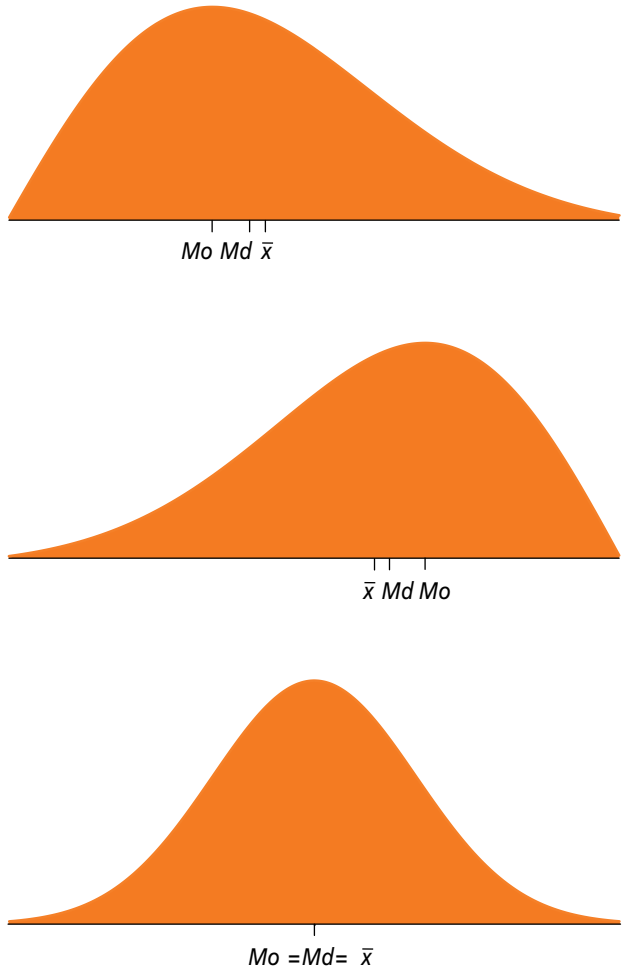


Abbildung 2.11: Vergleich von Mittelwert, Median und Modalwert bei unterschiedlichen Verteilungsformen.

Bei der praktischen Analyse von metrischen Daten sind die Beziehungen zwischen Median und arithmetischem Mittelwert von besonderer Bedeutung, die sich in zwei Punkten zusammenfassen lassen:

- In symmetrischen Verteilungen stimmen der Median und der arithmetische Mittelwert überein. Je schiefer die Häufigkeitsverteilung ist, desto mehr weichen der Median und der arithmetische Mittelwert voneinander ab. Bei linkssteilen Verteilungen ist der Median kleiner als der arithmetische Mittelwert, bei rechtssteilen Verteilungen ist der Median größer.
- Größere Unterschiede zwischen dem Median und dem arithmetischen Mittelwert können auf Ausreißer bzw. Messfehler hindeuten. Der Median wird von Extremwerten wenig beeinflusst, während der arithmetische Mittelwert sensibel reagiert.

Für praktische Auswertungen metrischer Daten ist es deshalb oft sinnvoll, zur Beschreibung der Daten neben dem arithmetischen Mittelwert zusätzlich den Median anzugeben.

Gewogenes arithmetisches Mittel

Manchmal stehen für die Datenauswertung lediglich Zwischenergebnisse verschiedener Teilstichproben zur Verfügung, aus denen auf die Kenngrößen der Gesamtstichprobe geschlossen werden soll. Der Sachverhalt soll am **Anwendungsbeispiel** veranschaulicht werden. In ►Tabelle 2.7 sind die Mittelwerte der Durchmesser und die Anzahlen der gelben, weißlichen und der sonstigen Kolonien zusammengestellt.

Teilstichprobe	Mittelwert (gerundet)	Anzahl
Gelbe Kolonien	7.14	13
Weißliche Kolonien	6.04	14
Sonstige Kolonien	4.51	13

Tabelle 2.7: Mittelwerte des Merkmals Durchmesser in den Teilstichproben.

Gesucht ist der Gesamtmittelwert aller Durchmesser. Die Bildung des Durchschnitts der gegebenen Mittelwerte wäre der falsche Weg, da hierbei die unterschiedlichen Teilstichprobenumfänge nicht berücksichtigt würden. Der entstehende Fehler wäre umso größer, je unterschiedlicher die Stichprobenumfänge in den Untergruppen wären. Die korrekte Ermittlung des Mittelwerts aller Messwerte ist durch den gewogenen arithmetischen Mittelwert möglich, bei dem die gegebenen Mittelwerte der Teilstichproben mit der jeweiligen Anzahl der Messwerte nach Formel (2.7) gewichtet werden.