1
Datentypen

Dieses Kapitel ...
- legt ein System zur Beschreibung verschiedener Datentypen dar;
- erklärt, warum wir die Datentypen genau bestimmen müssen, mit denen wir zu tun haben.

1.1
Kommt es wirklich darauf an?

Ein Statistikbuch aufzuschlagen und gleich mit einer Diskussion anzufangen, in welcher Weise man Daten in verschiedene Typen einteilen kann, das klingt entsetzlich abgehoben. Dennoch besteht der erste Schritt für die Behandlung von Daten im Allgemeinen in der Bestimmung, mit welchen Datentypen wir überhaupt zu tun haben. Das mag trocken sein, aber es hat reale Auswirkungen.

Wir schauen uns drei Arten von Daten an. Alle laufen in der Fachliteratur unter verschiedenen Bezeichnungen. Ich habe hier Namen ausgewählt, die ich am ehesten für selbsterklärend halte, und strebe kein einheitliches Benennungssystem an. Ich werde folgende drei Begriffe verwenden:

- Intervallskala – Daten aus stetigen Messwerten,
- Ordinalskala – Daten aus verschiedenen geordneten Kategorien,
- Nominalskala – Daten aus verschiedenen Kategorien.

1.2
Daten auf einer Intervallskala

Die ersten zwei Datentypen, die wir hier diskutieren, hängen beide mit der Messung eines bestimmten Merkmals zusammen. Daten auf einer „Intervallskala“ (man findet auch die Bezeichnung „Proportionalskala“, obwohl es da streng genommen noch einen kleinen Unterschied gibt, auf den wir hier aber nicht eingehen müssen) werden durch eine stetige Messung gewonnen. Sie sind die aussage-

Statistik für Mediziner und Pharmazeuten, 1. Auflage. P. Rowe.

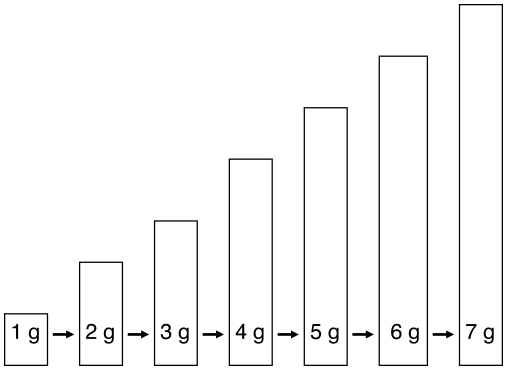

Abbildung 1.1 Daten auf einer Intervallskala – eine Reihe von Massen (1–7 g)

kräftigsten Daten, die man im Labor erzeugen kann, beispielsweise Gewichte, Längen, Zeiten, Konzentrationen, Drücke usw. Stellen Sie sich beispielsweise vor, wir hätten eine Reihe von Objekten mit einer Masse von 1 g, 2 g, 3 g usw. bis 7 g (Abbildung 1.1).

Nun stellen Sie sich die Massendifferenzen vor, die beim Übergang von einem Objekt zum nächsten auftreten. Diese Schritte betragen jeweils eine Einheit auf der Skala und haben folgende Eigenschaften:

- *Die Schritte haben jeweils eine exakt definierte Größe.* Wenn Sie jemandem davon erzählen, dass Sie eine Reihe von Objekten, wie oben beschrieben, haben, weiß Ihr Gegenüber genau, wie groß die Massendifferenzen sind, wenn wir die ganze Messreihe durchgehen.
- *Alle Schritte sind genau gleich groß.* Die Massendifferenz zwischen den Objekten mit 1 g und 2 g ist genauso groß wie der Schritt von 2 zu 3 g oder von 6 zu 7 g usw.

Weil die Messungen Schritte von konstanter Größe zeigen (nämlich Intervalle), spricht man hier von einer Intervallskala. Obwohl die Messwerte in Abbildung 1.1 exakt ganzzahlig sind, könnten sie natürlich genauso gut auch beliebige Werte dazwischen (wie 1,5 g oder 3,175 g) annehmen. Daher nennt man die Maßskala auch „stetig“ oder „kontinuierlich“.

1.3 Daten auf einer Ordinalskala

Auch hier geht es um Messungen, aber die erhobenen Kennzahlen sind meist etwas subjektiver als im vorigen Fall. Es ist schön, wenn man objektive Werte messen kann, wie den Blutdruck oder die Körpertemperatur. Es ist aber ebenso legitim, beispielsweise eine Vorstellung davon zu gewinnen, wie ein Patient seinen Zustand nach einer Behandlung einschätzt. Besonders naheliegend ist es, ein Punkte- oder Notenschema zu verwenden, beispielsweise von –2 bis +2 mit den folgenden Einschätzungen:

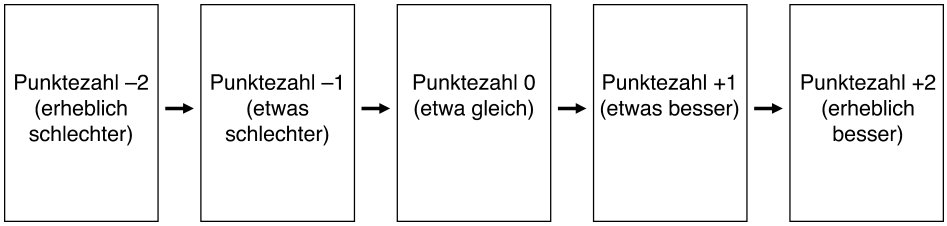

Abbildung 1.2 Daten auf einer Ordinalskala – Punkteschema für die Selbsteinschätzung des Gesundheitszustands von Patienten nach der Behandlung

–2	merklich schlechter
–1	etwas schlechter
0	etwa gleich
+1	etwas besser
+2	erheblich besser

In diesem Fall (Abbildung 1.2) wissen wir nur, dass ein Patient mit einem höheren Wert zufriedener ist mit seiner Behandlung. Wir wissen aber nicht, *um wie viel* zufriedener er ist.

Wir wissen nicht, wie groß die Abstände zwischen den einzelnen Einschätzungen sind; also können wir auch nicht behaupten, sie seien alle gleich groß. Es muss noch nicht einmal so sein, dass die Differenz zwischen den Einschätzungen –2 und 0 größer ist als der zwischen +1 und +2. Keine der Eigenschaften aus einer Intervallskala lässt sich also auf diese Daten übertragen.

Der Begriff „Ordinal“ spiegelt wider, dass die verschiedenen Ergebnisse sich in einer Rangfolge ordnen lassen, von einem Extremwert zum anderen. Daten auf einer Ordinalskala werden daher manchmal auch als „kategorial geordnet“ bezeichnet. In diesem Fall sind die Werte nicht stetig, d. h., die einzelnen Kategorien werden mit –1, +2 usw. bezeichnet, Zwischenwerte gibt es nicht.

1.4 Daten auf einer Nominalskala

In diesem Fall geht es in keinem Fall um die Messung eines Merkmals. Bei diesen Daten verwenden wir eine Einteilung ohne natürliche Rangfolge. Beispielsweise könnte einer der Faktoren, der die Effektivität einer Behandlung beeinflusst, der Hersteller des entsprechenden medizinischen Geräts sein. Die Patienten würde man dann nach den Herstellern „Müller“, „Meyer“ und „Schmidt“ einteilen. Hier gibt es keine natürliche Reihenfolge, es handelt sich nur um drei verschiedene Bezeichnungen.

Bei Ordinaldaten konnten wir wenigstens sagen, dass beispielsweise ein mit +2 bewerteter Fall eher dem Fall mit +1 als einem mit 0 oder –1 ähnelt. Bei Nominaldaten können wir aber nicht davon ausgehen, dass die „Müller“- und die

„Meyer"-Patienten eine irgendwie geartete Ähnlichkeit aufweisen. Die Reihenfolge, in der man sie aufführt, ist völlig beliebig.

Sehr verbreitet sind Einteilungen mit genau zwei Kategorien, etwa männlich/weiblich, lebt/tot oder Erfolg/Misserfolg. In solchen Fällen spricht man von „dichotomen Werten".

Datentypen

- *Intervallskala* – Messungen mit definierten und stets gleich bleibenden Abständen (Intervallen) zwischen aufeinanderfolgenden Werten. Die Werte sind stetig.
- *Ordinalskala* – Messungen mithilfe einer Klassifikation und einer natürlichen Reihenfolge der Werte (vom niedrigsten bis zum höchsten), aber ohne bestimmte Abstände. Die Werte sind nicht stetig.
- *Nominalskala* – Einteilungen ohne natürliche Reihenfolge

1.5 Aufbau dieses Buchs

Das Buch ist so aufgebaut, dass nacheinander die verschiedenen Datentypen durchgenommen werden. Die Kapitel 2–14 behandeln Daten auf Intervallskalen, die bei stetigen Messungen gewonnen werden. Die Kapitel 15 und 16 befassen sich mit kategorialen Daten (auf Nominalskalen), geordnete Daten (auf Ordinalskalen) werden in Kapitel 17 besprochen.

1.6 Kapitelzusammenfassung

Der unerlässliche erste Schritt bei der Auswahl der passenden statistischen Verfahren ist es, die zu behandelnden Datentypen zu erkennen.

Folgende Fälle können auftreten:

- *Intervallskala* – Messungen auf einer Skala mit definierten und stets gleich bleibenden Abständen. Die Werte sind stetig.
- *Ordinalskala* – Messungen mithilfe einer Klassifikation ohne bestimmte Abstände. Die Werte sind nicht stetig.
- *Nominalskala* – Einteilungen ohne natürliche Reihenfolge.