1 Am Anfang stehen die Daten

Ob sie aus einer Umfrage stammen oder aus einem Experiment, ob sie durch Simulationen erzeugt oder aus externen Quellen heruntergeladen wurden – oft bestehen Daten zunächst nur aus einer großen, ungeordneten Sammlung von Werten. Meist sind es Zahlen. Viele Zahlen. Ohne erkennbare Struktur.

Diese Situation ist heute allgegenwärtig. Messgeräte liefern fortlaufend neue Werte, Simulationen produzieren in Sekunden ganze Zahlenkolonnen, digitale Systeme protokollieren jede Interaktion. Daten entstehen schneller, als sie verstanden werden können.

Denn zwischen einer Liste von Zahlen und einem Verständnis dessen, was sie bedeuten, liegt ein weiter Weg.

1.1 Daten als Machtfaktor

Gut ausgewertete Daten verschaffen einen Informationsvorsprung – und damit oft auch einen Handlungsvorteil. Das gilt in nahezu allen gesellschaftlichen Bereichen.

In der Wissenschaft entscheiden sorgfältig analysierte Messreihen darüber, ob Hypothesen bestätigt oder verworfen werden. In der Wirtschaft ermöglichen Datenanalysen Prognosen, Optimierungen und strategische Entscheidungen. In der Politik beeinflussen Umfragen, Statistiken und Modellrechnungen öffentliche Debatten und konkrete Maßnahmen.

Dabei geht es selten um einzelne Zahlen. In all diesen Bereichen beschäftigen sich ganze Abteilungen mit der Analyse und Interpretation gewonnener Daten. Fehlerhaft erfasste Werte müssen erkannt, Zufälliges von Systematischem unterschieden, Muster vorsichtig gedeutet werden. Erst dann lassen sich Entwicklungen einordnen und Entscheidungen begründen.

1.2 *Intelligence*: Erkenntnis aus Daten

Im Englischen bezeichnet der Begriff *intelligence* genau diesen Zwischenschritt: nicht primär eine persönliche Begabung, sondern verdichtete Erkenntnis aus systematisch gewonnenen und ausgewerteten Informationen. Gemeint sind Lagebilder, Entscheidungsgrundlagen, belastbare Einschätzungen unter Unsicherheit.

Der deutsche Begriff *Intelligenz* ist breiter. Er meint meist eine allgemeine geistige Leistungsfähigkeit: die Fähigkeit, Informationen einzuordnen, Zusammenhänge zu erkennen, Unsicherheit auszuhalten und dennoch zu urteilsfähigen Entscheidungen zu gelangen.

Gerade in dieser Unterscheidung liegt ein aufschlussreicher Gedanke. *Intelligence* beschreibt eher das Ergebnis gelungener Datenanalyse, *Intelligenz* eher die Fähigkeit, solche Ergebnisse hervorzubringen, zu verstehen und kritisch zu prüfen. Wo die eine ohne die andere bleibt, entstehen Probleme: Datenbasierte *intelligence* ohne menschliche Urteilskraft begünstigt ein blindes Vertrauen in Zahlen; Intelligenz ohne belastbare Daten bleibt anfällig für Vorurteile, Eindruckseffekte und Bauchentscheidungen.

Der Ausdruck *künstliche Intelligenz* macht diese Spannung besonders deutlich. KI-Systeme können Muster in großen Datenmengen erkennen, Ergebnisse verdichten und Vorhersagen erzeugen. Sie produzieren damit *intelligence* im engeren Sinn. Aber sie verstehen nicht, was diese Muster bedeuten, wo ihre Grenzen liegen oder wann Daten in die Irre führen. Gerade deshalb bleibt menschliche Intelligenz unverzichtbar: nicht um Daten zu verarbeiten, sondern um ihre Bedeutung einzuschätzen.

1.3 Daten sehen heißt noch nicht verstehen

Eine Tabelle mit vielen Zahlen – etwa eine Liste mit Würfelergebnissen – ist zunächst fast bedeutungslos. Erst durch geeignete Darstellungen, durch Vergleiche und durch wiederholte Beobachtung entstehen Hinweise auf Strukturen. Und selbst dann bleibt Vorsicht geboten: Nicht jedes Muster ist stabil, nicht jede Auffälligkeit ist relevant.

Gerade bei zufallsbehafteten Prozessen ist diese Unterscheidung schwierig. Zufall erzeugt Schwankungen, die überzeugend aussehen können. Gleichzeitig verbergen sich hinter scheinbar chaotischen Daten oft erstaunlich stabile Gesetzmäßigkeiten.

Genau an dieser Stelle beginnt die Stochastik.

1.4 Stochastik als Brücke zwischen Daten und Bedeutung

Stochastik beschäftigt sich nicht primär mit exakten Einzelwerten, sondern mit Verhalten bei vielen Wiederholungen, mit typischen Mustern, mit Streuung und mit der Frage, welche Aussagen belastbar sind – und welche nicht.

Dabei spielen Visualisierung und Simulation eine besondere Rolle. Visualisierungen machen Strukturen sichtbar, die in einer bloßen Zahlenliste verborgen bleiben. Simulationen liefern keine „wahren“ Daten, aber sie machen Zufall beobachtbar. Sie erlauben es, Prozesse zu wiederholen, zu variieren und kontrolliert zu untersuchen.

Für den Unterricht ist dieses Zusammenspiel besonders fruchtbar. Daten werden nicht sofort berechnet, sondern zunächst betrachtet, beschrieben, verglichen und hinterfragt. Genau so entsteht schrittweise ein Verständnis dafür, was stochastische Begriffe überhaupt leisten.

1.5 Der Erstkontakt mit Desmos

Folgender Datensatz ist gegeben:

14, 8, 9, 10, 10, 10, 6, 6, 9, 9, 11, 17, 13, 9, 10, 7, 7, 10, 16, 13, 17, 9, 10, 12, 11, 9, 9, 9, 8, 8, 9, 12, 11, 9, 11, 10, 8, 8, 8, 13, 6, 16, 6, 8, 7, 10, 11, 11, 7, 10, 8, 10, 12, 8, 13, 13, 5, 6, 9, 8, 9, 11, 10, 12, 11, 7, 10, 5, 12, 11, 9, 13, 14, 15, 12, 15, 14, 8, 11, 8, 15, 16, 13, 13, 15, 10, 9, 10, 12, 10, 9, 13, 14, 8, 15, 10, 7, 11, 4, 10, 9, 11, 9, 8, 8, 10, 12, 16, 8, 8, 5, 14, 11, 14, 12, 11, 8, 8, 8, 12, 8, 11, 7, 9, 11, 13, 5, 16, 12, 7, 9, 11, 15, 9, 10, 9, 11, 11, 9, 11, 7, 10, 11, 9, 13, 9, 4, 9, 7, 14, 14, 11, 11, 13, 11, 11, 11, 8, 15, 10, 8, 13, 12, 12, 6, 9, 15, 16, 14, 8, 13, 8, 10, 9, 7, 4, 8, 9, 11, 8, 11, 7, 12, 11, 6, 7, 13, 15, 15, 11, 6, 12, 7, 12, 14, 10, 13, 8, 14, 6

Was geht Ihnen beim Betrachten dieser Zahlen durch den Kopf?
Wie ließen sich diese Daten strukturieren?
Worauf würden Sie zuerst achten, welche Fragen würden sich stellen?

Nehmen Sie sich einen Moment Zeit, bevor Sie weiterlesen.

Für die folgenden Schritte wird der frei zugängliche Grafikrechner Desmos verwendet, auf den das gesamte Buch aufbaut. Er erlaubt es, Daten ohne nennenswerte Einstiegshürden zu visualisieren und aus unterschiedlichen Perspektiven zu betrachten. Fachliche Vorkenntnisse sind dafür nicht erforderlich; ein internetfähiges Endgerät genügt, ein größerer Bildschirm und eine Tastatur erleichtern lediglich den Überblick.

Um den Einstieg zu erleichtern, steht ein vorbereitetes Arbeitsblatt mit diesen Daten zur Verfügung. Es kann direkt im Browser geöffnet werden unter Worksheet [1a]: https://ogy.de/dly8. Sie können hier die angegebenen Befehle einfach Schritt für Schritt eingeben. In späteren Kapiteln mit komplexeren Befehlen finden Sie jeweils nach der Erklärung des Codes auch das komplette Worksheet verlinkt. Abtippen ist dann oft zu fehleranfällig und auch das Kopieren funktioniert oft nicht zuverlässig.

1.6 Erste Überlegungen: mögliche Zugänge

Nachdem die Daten im Desmos-Arbeitsblatt vorliegen, stellt sich noch keine Auswertungsfrage. Zunächst geht es darum, überhaupt Ansatzpunkte zu finden. Welche Aspekte dieser Daten könnten interessant sein? Welche Eigenschaften lassen sich untersuchen?

Typische erste Überlegungen sind etwa:

1. Welche Werte kommen häufig vor, welche selten?
2. Gibt es besonders große oder besonders kleine Werte?
3. Wie stark schwanken die Daten insgesamt?
4. Liegt der „typische“ Wert eher im unteren oder im oberen Bereich?
5. Wirken die Daten gleichmäßig verteilt oder eher gebündelt?

Diese Fragen sind bewusst unspezifisch. Sie zielen nicht auf ein Ergebnis, sondern auf eine Haltung: Daten werden nicht sofort berechnet, sondern zunächst beobachtet und befragt.

Auch wenn man durch bloßes Durchsehen vielleicht noch Extremwerte oder die Spannweite erahnen kann, bleiben Häufigkeiten, Verteilungsformen und typische Bereiche ohne Werkzeugunterstützung schwer zugänglich. Genau hier hilft Desmos: Mit wenigen Befehlen wird sichtbar, was in der Zahlenliste verborgen bleibt.

Wir starten mit einem *Dotplot*, einer der einfachsten Formen der Datenvisualisierung. Geben Sie in die zweite Zeile ein:

```
dotplot(D)
```

Klicken Sie anschließend auf die kleine Lupe links neben der Eingabezeile, um die Ansicht im rechten Fenster optimal auszurichten.

Betrachten Sie den entstandenen Dotplot zunächst ohne weitere Auswertung.

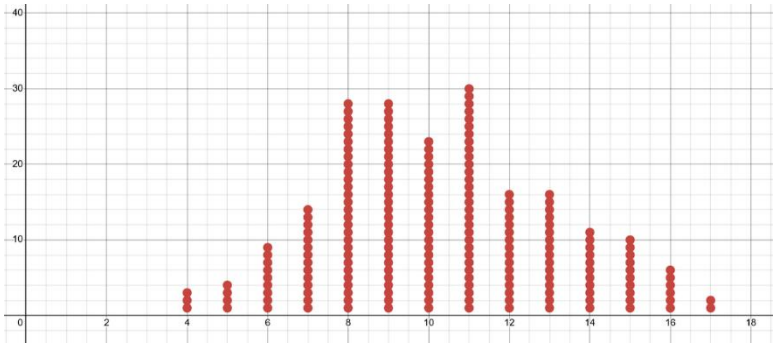

Abb. 1.1: Dotplot

Was fällt Ihnen auf?

Einige Werte treten deutlich häufiger auf als andere. Bei manchen Zahlen stapeln sich die Punkte, andere erscheinen nur vereinzelt. Sehr kleine und sehr große Werte kommen vor, aber offenbar seltener. Die Daten wirken also nicht gleichmäßig verteilt, sondern konzentrieren sich in einem bestimmten Bereich.

Gerade das war in der bloßen Zahlenliste kaum zu erkennen. Der Dotplot macht aus isolierten Einzelwerten erstmals eine beobachtbare Struktur.

Damit lassen sich die ersten Fragen präzisieren:

1. Wo liegt der Bereich der häufigsten Werte?
2. Welche Werte treten nur vereinzelt auf?
3. Gibt es eine erkennbare Mitte, um die sich die Daten sammeln?

Noch wurden keine Kennzahlen berechnet und keine Modelle verwendet. Die Einsicht entsteht allein durch die veränderte Darstellung derselben Daten.

1.7 Weiter verdichten: vom Dotplot zum Boxplot

Der Dotplot erlaubt einen ersten Blick auf Häufungen und Ausreißer. Gleichzeitig bleibt die Darstellung noch stark an den einzelnen Datenpunkten orientiert. Je größer der Datensatz wird, desto unübersichtlicher kann diese Form werden.

Um den Blick weiter zu fokussieren, wird die Darstellung nun erneut verdichtet. Anstelle einzelner Punkte treten zusammenfassende Kenngrößen in den Vordergrund: Lage und Streuung der Daten.

Erstellen Sie dazu einen Boxplot der Daten. Blenden Sie zunächst die Anzeige des Dotplots aus, indem Sie links neben der Definition auf den Kreis mit den farbigen Punkten klicken.

```
boxplot(D)
```

Benutzen Sie wieder die Lupe für eine optimale Darstellung.

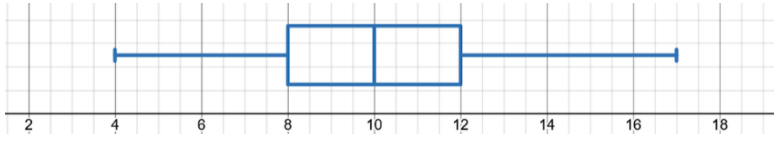

Abb. 1.2: Boxplot

Ohne die zeitliche Abfolge und ohne einzelne Häufungen sichtbar zu machen, zeigt der Boxplot auf einen Blick, wo sich der zentrale Bereich der Daten befindet und wie weit die Werte insgesamt streuen. Extreme Werte treten nun deutlicher als solche hervor, während die große Masse der Daten kompakt dargestellt wird.

Die Perspektive hat sich erneut geändert: Nicht mehr einzelne Häufigkeiten stehen im Vordergrund, sondern typische Bereiche und Abweichungen.

Aus dem Graphen kann man die wichtigsten statistischen Größen ablesen, genauer geht das aber durch einen weiteren Befehl:

```
stats(D)
```

Sofort erhalten Sie die Werte für Minimum und Maximum, das erste und dritte Quartil und den Median. Auch den Mittelwert können Sie durch `mean(D)` bequem berechnen lassen.

Diese Verdichtung hat jedoch ihren Preis. Der Boxplot zeigt Lage und Streuung sehr kompakt, verliert dafür aber weitgehend die Form der Verteilung. Gerade die Häufung im mittleren Bereich, die im Dotplot bereits sichtbar wurde, ist hier kaum noch erkennbar.

Um diese Form wieder sichtbar zu machen, braucht es eine andere Darstellung.

1.8 Die Form der Verteilung: das Histogramm

Die Form der Datenverteilung liefert oft auf einen Blick wertvolle Informationen: Sind die Daten symmetrisch? Gibt es eine klare Häufung um einen Mittelwert? Oder mehrere Häufungszentren?

Um die Verteilungsform sichtbar zu machen, eignet sich ein **Histogramm**. Es fasst die Daten in Klassen zusammen und stellt die Häufigkeiten als aneinandergereihte Rechtecke dar:

```
histogram(D)
```

Die Form ist nun deutlich erkennbar: Die Werte häufen sich in der Mitte – etwa zwischen 8 und 12 –, während sehr kleine und sehr große Werte selten auftreten. Die Verteilung wirkt annähernd symmetrisch und glockenförmig.

Diese Form ist charakteristisch für viele zufallsabhängige Prozesse. Sie lässt sich durch ein theoretisches Modell beschreiben.

1.9 Vergleich mit der Normalverteilung

Die beobachtete Verteilung ähnelt einer **Normalverteilung**. Um diese Ähnlichkeit zu prüfen, lässt sich die theoretische Kurve über das Histogramm legen.

Desmos bietet dafür die Funktion `normaldist`, die eine Normalverteilung mit dem Mittelwert und der Standardabweichung der Daten erzeugt. Damit aber die Höhennormierung der beiden Darstellungen zusammenpasst, müssen Sie beim Histogramm zuerst die *Höhe der Leisten* von *Anzahl* auf *Dichte* umstellen. Geben Sie dann in die nächste Zeile ein:

```
normaldist(mean(D), stdev(D))
```

Hierbei werden Mittelwert und empirische Standardabweichung von Desmos aus den Daten berechnet und als Schätzer der Parameter der Normalverteilung verwendet.

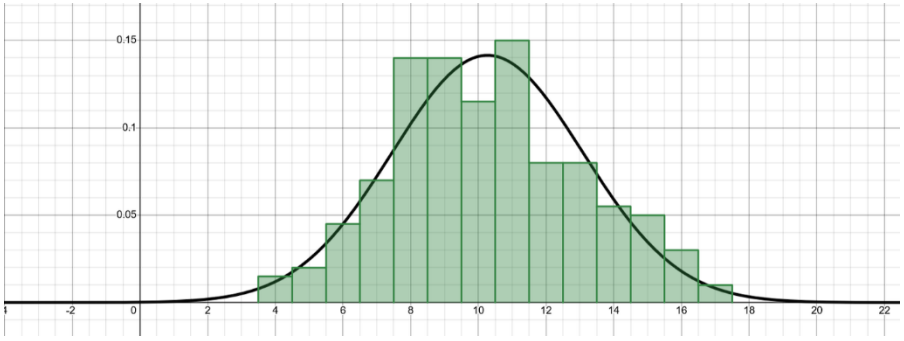

Abb. 1.3: Histogramm/Normalverteilung

Zu beachten ist allerdings, dass schon die Klassenbildung das Aussehen eines Histogramms stark verändern kann. Solche Darstellungen können uns daher auch auf eine falsche Fährte locken.

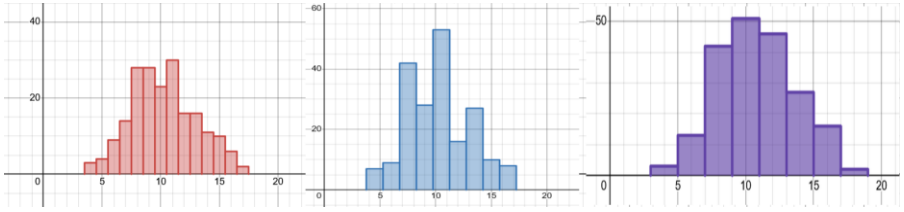

Abb. 1.4: Klassenbreiten

Bei einer Klassenbreite von 1 erhält jeder ganzzahlige Wert eine eigene Klasse. Die Glockenform ist klar erkennbar, und die Normalverteilungskurve passt gut.

Bei einer Klassenbreite von 1,5 liegen die Klassengrenzen zwischen den natürlichen Werten. Dadurch wird die Struktur gestört: Einzelne Werte werden auf verschiedene Klassen verteilt, die Verteilung wirkt unregelmäßiger und die Glockenform ist deutlich schwerer zu erkennen.

Bei einer Klassenbreite von 2 werden die Klassen wieder breiter. Die Verteilung erscheint glatter, die Glockenform wird erneut sichtbar, allerdings auf Kosten von Detailgenauigkeit.

Dieser Vergleich zeigt: Bei diskreten Daten ist die Wahl der Klassenbreite besonders kritisch. Eine Klassenbreite, die nicht zur natürlichen Struktur der Daten passt, kann die Verteilung erheblich verzerren – und dabei den falschen Eindruck erwecken, die Daten seien chaotisch oder würden keiner erkennbaren Form folgen.

Visualisierungen sind mächtige Werkzeuge, aber nicht neutral. Gerade deshalb ist es wichtig, ihre Konstruktionsprinzipien zu verstehen und verschiedene Darstellungen kritisch zu vergleichen.

1.10 Eine letzte Perspektive: die Zeitreihe

Alle bisherigen Darstellungen – Dotplot, Boxplot, Histogramm – ignorierten die *Reihenfolge* der Daten. Sie zeigten, wie oft welche Werte vorkommen, aber nicht, *wann* sie aufgetreten sind.

Eine Zeitreihe macht genau das sichtbar. Sie stellt die Werte in der Reihenfolge dar, in der sie beobachtet oder simuliert wurden.

Wie funktioniert das in Desmos?

1. Sie erstellen eine **Tabelle** über das Plus-Symbol links oben.
2. In die erste Spalte tragen Sie als Index statt `x_1` den Ausdruck `[1...length(D)]` ein (inklusive der Klammern).
3. In die zweite Spalte schreiben Sie statt `x_2` einfach `D`.

Desmos plottet die Werte automatisch. Über das Lupen-Symbol können Sie die Ansicht erneut optimal anpassen.

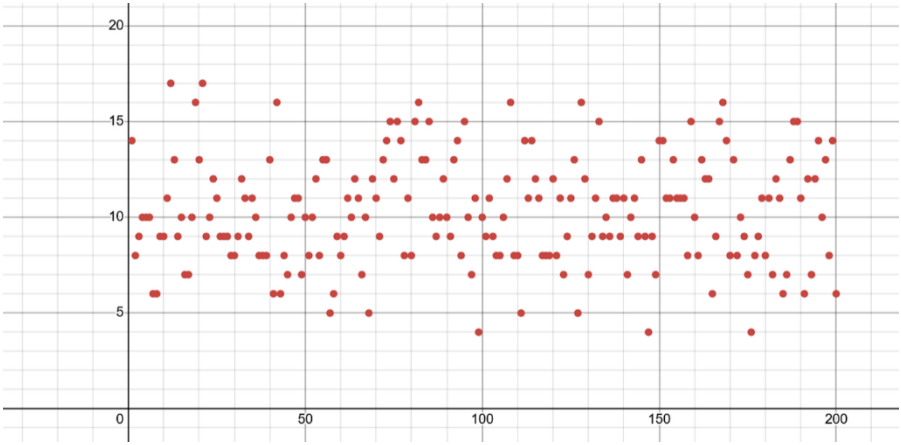

Abb. 1.4: Zeitreihe

Was fällt auf? Die Werte schwanken scheinbar regellos um einen Mittelwert herum. Es gibt keine erkennbaren Trends, keine Zyklen, keine systematischen Muster. Manche Abschnitte wirken unruhiger, andere ruhiger – doch genau das ist typisch für Zufall.

Warum ist das wichtig?

Die Zeitreihe zeigt: Die Reihenfolge spielt hier keine inhaltliche Rolle. Die Werte haben sich rein zufällig ergeben. Für die Verteilung ist der Zeitpunkt ihres Auftretens irrelevant.

Anders wäre es bei *Zeitreihen mit Struktur* – etwa Aktienkursen, Temperaturen oder Herzfrequenzen. Dort ist die Reihenfolge entscheidend, weil frühere Werte Einfluss auf spätere Entwicklungen haben können.

Auch hier zeigt sich: Welche Darstellung man wählt, beeinflusst, was man sieht. Die Zeitreihe macht Reihenfolge sichtbar – und in diesem Fall gerade deren fehlende Bedeutung.

Und jetzt ist es an der Zeit, das Geheimnis zu lüften: Die Daten entstanden als Zufallsergebnisse durch das Würfeln mit drei Laplace-Würfeln und das Aufaddieren der Augensummen.

Vielleicht vermissen Sie jetzt die bei dieser Konstellation möglichen Ergebnisse 3 oder 18. Sollten die bei so vielen Versuchen nicht eigentlich auftreten?

Nun, die Wahrscheinlichkeit, dass beide nicht dabei sind, beträgt

$$\left(\frac{214}{216}\right)^{200} \approx 15{,}6\%$$

– also keineswegs etwas Ungewöhnliches.

1.11 Zum Ausprobieren: echte Messdaten aus Köln

Die Stadt Köln veröffentlicht im Rahmen ihrer Open-Data-Initiative Rohdaten aus der Geschwindigkeitsüberwachung, abrufbar unter [Q8] auf der Internetseite von Köln. Ein aufbereiteter Datensatz mit rund 10.000 Messwerten steht als Desmos-Worksheet [1b]: https://www.desmos.com/calculator/rxvdloz0jb. Der vollständige Datensatz ist dort unter dem Namen `v_mess` gespeichert. Um mit einer überschaubaren Teilmenge zu beginnen, genügt:

```
v = v_mess[1...500]
histogram(v)
m = mean(v)
```

Die erste Zeile sorgt dafür, dass nur die ersten 500 Werte übernommen werden. Sie können das leicht ändern auf `1...100` oder auch `3000...7000`. Wenden Sie nun dieselben Werkzeuge an, die Sie in diesem Kapitel kennengelernt haben: Dotplot, Boxplot, Histogramm, Normalverteilungskurve. Beobachten Sie, was Sie sehen – bevor Sie weiterlesen.

Die Verteilung zeigt einen abrupten Beginn auf der linken Seite und einen langen, langsam abklingenden Schwanz nach rechts – kein sanfter Anstieg, sondern ein fast senkrechter Einstieg direkt beim am Schwellenwert der Erfassung. Wenn Sie versuchen, eine Normalverteilungskurve mit `normaldist(mean(v), stdev(v))` anzupassen, werden Sie feststellen, dass die Übereinstimmung deutlich schlechter ist als bei den Würfelsummen.

Variieren Sie nun den Ausschnitt. Ersetzen Sie die 500 durch 200, durch 1000, durch 5000. Ändert sich das Bild grundlegend? Oder bleibt die Form stabil?

Die Form bleibt stabil – sie ist keine Zufälligkeit des Ausschnitts. Was steckt dahinter? Was dort veröffentlicht wird, sind ausschließlich Messungen, die zu einem Bußgeld geführt haben – also nur Fahrten, bei denen eine Geschwindigkeitsschwelle überschritten wurde. Alle langsameren Fahrzeuge sind in den Daten nicht enthalten. Was wie eine schiefe Verteilung aussieht, ist in Wirklichkeit das rechte Stück einer möglicherweise symmetrischen Verteilung – der die linke Hälfte durch die Messbedingung entzogen wurde. Statistiker sprechen von einer *links abgeschnittenen* Verteilung.

Das ist keine Besonderheit dieses Datensatzes. Immer dann, wenn nur Extremwerte erfasst werden – Überschreitungen, Ausfälle, Rekorde –, zeigt die resultierende Verteilung eine Form, die nicht die eigentliche Grundgesamtheit widerspiegelt, sondern deren selektiv sichtbaren Aus-

schnitt. Die Frage „Woher kommen diese Daten?“ ist damit nicht nur eine historische, sondern eine methodische: Sie entscheidet darüber, was man aus den Daten schließen darf und was nicht.