Kapitel 2

Deskriptive Statistik

"Definition of Statistics: The science of producing unreliable facts from reliable figures.'"

Evan Esar

Lernziel

- Sie kennen Methoden der deskriptiven Statistik, können sie interpretieren und folgende Grössen ausrechnen: arithmetisches Mittel, Standardabweichung, Varianz, Quantil, Median, Kovarianz und Korrelation .
- Sie verstehen die Grundidee der einfachen linearen Regression: wie die Form des Modells ist, wie man die Koeffizienten interpretiert und wie man man die Koeffizienten schätzt .
- Sie können Daten mit folgenden graphischen Methoden darstellen: Histogramm, Boxplot, empirische kumulative Verteilungsfunktion, Streudiagramm .

2.1 Einleitung

Als Einleitung betrachten wir zwei verschiedene Datensätze. Beim ersten Datensatz werden zwei Methoden zur Bestimmung der latenten Schmelzwärme von Eis verglichen. Wiederholte Messungen der freigesetzten Wärme beim Übergang von Eis bei $-0.7°$ C zu Wasser bei $0°$ C ergaben die folgenden Werte (in cal/g):

Methode A	79.98	80.04	80.02	80.04	80.03	80.03	80.04	79.97	80.05	80.03
Methode A	80.02	80.00	80.02							
Methode B	80.02	79.94	79.98	79.97	79.97	80.03	79.95	79.97		

Obwohl die Messungen mit der grösstmöglichen Sorgfalt durchgeführt und alle Störeinflüsse ausgeschaltet wurden, variieren die Messungen von Fall zu Fall. Wir werden diese Variationen innnerhalb der Messreihen als zufällig modellieren, das heisst wir interpretieren diese Werte als Realisierungen von Zufallsvariablen. Jede Messung entspricht also einer Realisierung der Zufallsvariable "freigesetzte Wärme beim Übergang von Eis bei $-0.7°$ C zu Wasser bei $0°$ C".

Wir werden dann die Frage beantworten, ob die Unterschiede zwischen den Methoden ebenfalls als zufällig angesehen werden können, oder ob ein systematischer Unterschied plausibler ist, der auch in der ganzen Population, d.h. in weiteren Messungen, bestehen bleibt. Im letzteren Fall werden wir dann noch zusätzlich angeben, wie gross der systematische Unterschied etwa ist.

Im zweiten Beispiel wurde bei 11 Individuen die Aggregation von Blutplättchen vor und nach dem Rauchen einer Zigarette gemessen. Die folgenden Daten geben den Anteil aggregierter Blutplättchen (in Prozent) nach einer Stimulation an.

Individuum	1	2	3	4	5	6	7	8	9	10	11
Vorher	25	25	27	44	30	67	53	53	52	60	28
Nachher	27	29	37	56	46	82	57	80	61	59	43

Wieder variieren die Werte in einer nicht vorhersehbaren Art. Diesmal handelt es sich jedoch weniger um Messfehler, sondern um Variation zwischen Individuen (vermutlich gäbe es auch noch eine gewisse Variation beim gleichen Individuum, wenn der Test wiederholt würde). Die Aggregation bei diesen 11 Individuen ist meistens, aber nicht immer nach dem Rauchen höher, und die Fragestellung lautet, ob es sich hier um einen zufälligen Effekt handelt, der auf die spezifische Stichprobe beschränkt ist, oder ob man dieses Resultat auf eine grössere Population verallgemeinern kann. Im letzteren Fall möchte man wieder angeben, wie gross die mittlere Zunahme etwa ist.

2.2 Deskriptive Statistik **(Stahel, Kap. 2 und 3.1, 3.2)**

Bei einer statistischen Analyse ist es wichtig, nicht einfach blind ein Modell anzupassen oder ein statistisches Verfahren anzuwenden. Die Daten sollten immer mit Hilfe von geeigneten graphischen Mitteln dargestellt werden, da man nur auf diese Weise unerwartete Strukturen und Besonderheiten entdecken kann. Kennzahlen können einen Datensatz grob charakterisieren. Im folgenden werden die Daten mit $x_1, \ldots, x_n$ bezeichnet.

2.2.1 Kennzahlen

Häufig will man die Verteilung der Daten numerisch zusammenfassen. Dazu braucht man mindestens zwei Kenngrössen, eine für die Lage und eine für die Streuung. Die bekanntesten

solchen Grössen sind das *arithmetische Mittel* für die Lage,

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

und die *empirische Standardabweichung* für die Streuung,

$$s_x = \sqrt{\text{var}} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2}.$$

(der Nenner $n-1$, anstelle von n, ist mathematisch begründet und hat die Eigenschaft, dass kein "systematischer" Fehler auftritt).

Alternative Kenngrössen sind der *Median* als Lagemass und die *Quartilsdifferenz* als Streuungsmass. Diese werden mit Hilfe von sogenannten Quantilen definiert.

Quantile

Das *empirische α-Quantil* ist anschaulich gesprochen der Wert, bei dem $\alpha \times 100\%$ der Datenpunkte kleiner und $(1-\alpha) \times 100\%$ der Punkte grösser sind.

Zur formalen Definition führen wir die geordneten Werte ein:

$$x_{(1)} \leq x_{(2)} \leq \ldots \leq x_{(n)}.$$

Das empirische α-Quantil ist dann gleich

$\frac{1}{2}(x_{(\alpha n)} + x_{(\alpha n+1)})$ falls $\alpha \cdot n$ eine ganze Zahl ist,

$x_{(k)}$ wobei $k = \alpha n + \frac{1}{2}$ gerundet auf eine ganze Zahl; falls $\alpha \cdot n$ keine ganze Zahl ist.

Der (empirische) Median ist das empirische 50%-Quantil: d.h., es markiert die "mittlere" Beobachtung, wenn die Daten sortiert wurden. Es ist also ein Mass für die Lage der Daten.

Die Quartilsdifferenz ist gleich

$$\text{empirisches } 75\%\text{-Quantil} - \text{empirisches } 25\%\text{-Quantil}$$

und ist ein Streuungsmass für die Daten.

Median und Quartilsdifferenz haben den Vorteil, dass sie robust sind: das heisst, dass sie viel weniger stark durch extreme Beobachtungen beeinflusst werden können als arithmetisches Mittel und Standardabweichung.

Beispiel: Messung der Schmelzwärme von Eis mit Methode A
Das arithmetische Mittel der $n = 13$ Messungen ist $\overline{x} = 80.02$ und die Standardabweichung ist $s_x = 0.024$. Ferner ist für $n = 13$: $0.25n = 3.25$, $0.5n = 6.5$ und $0.75n = 9.75$. Damit ist das 25%-Quantil gleich $x_{(4)} = 80.02$, der Median gleich $x_{(7)} = 80.03$ und das 75%-Quantil gleich $x_{(10)} = 80.04$. Wenn bei der grössten Beobachtung ($x_9 = 80.05$) ein Tippfehler passiert wäre und $x_9 = 800.5$ eingegeben worden wäre, dann wäre $\overline{x} = 135.44$, der Median aber nach wie vor $X_{(7)} = 80.03$. Das arithmetsiche Mittel wird also durch Veränderung einer Beobachtung sehr stark beeinflusst, während der Median gleich bleibt - er ist "robust".

Empirische Kovarianz und empirische Korrelation

Wenn wir bei jeder Versuchseinheit zwei verschiedene Grössen messen, d.h. wenn die Daten von der Form $(x_1, y_1), \ldots, (x_n, y_n)$ sind, interessiert man sich in erster Linie für die Zusammenhänge und Abhängigkeiten zwischen den Variablen. Im Datensatz, wo die Aggregation von Blutplättchen vor und nach dem Rauchen gemessen wurde, entspricht die i-te Versuchseinheit einem der 11 Individuen, wobei man mit x_i die Blutaggregation vor dem Rauchen und mit y_i die Blutaggregation nach dem Rauchen des i-ten Individuums bezeichnet.

Für die numerische Zusammenfassung der Abhängigkeit der beiden Grössen ist die **empirische Korrelation** r als Kennzahl (oder auch mit $\hat{\rho}$ bezeichnet) am gebräuchlichsten:

$$r = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{(\sum_{i=1}^n (x_i - \overline{x})^2) \cdot (\sum_{i=1}^n (y_i - \overline{y})^2)}} ,$$

wobei s_{xy} die **emprische Kovarianz** (auch **Stichprobenkovarianz**)bezeichnet und folgendermassen definiert ist

$$s_{xy} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{n-1} .$$

Die **empirische Varianz** var ergibt sich dann als Spezialfall aus der empirischen Kovarianz:

$$\mathrm{var}(y) = s_{yy} , \qquad \mathrm{var}(x) = s_{xx} .$$

Die empirische Korrelation ist eine dimensionslose Zahl zwischen -1 und +1 und misst Stärke und Richtung der **linearen Abhängigkeit** zwischen den Daten x und y. Es gilt

$$r = +1 \text{ genau dann wenn } y = a + bx \text{ für ein } a \in \mathbb{R} \text{ und ein } b > 0,$$
$$r = -1 \text{ genau dann wenn } y = a + bx \text{ für ein } a \in \mathbb{R} \text{ und ein } b < 0 .$$

Überdies gilt:

$$x \text{ und } y \text{ unabhängig} \quad \Longrightarrow \quad r = 0. \tag{2.1}$$

Die Umkehrung gilt i.A. nicht. Man sollte jedoch nie r berechnen, ohne einen Blick auf das Streudiagramm zu werfen, da ganz verschiedene Strukturen den gleichen Wert von r ergeben können. Siehe dazu Abb. 2.1.

Quiz: Angenommen, es stellt sich heraus, dass Personen mit grossem Einkommen auch grosse Weinkenntnisse haben, wenn die Korrelation der beiden Variablen gross ist. Sollte man also einen Kurs über Wein besuchen, um sein Einkommen zu verbessern?

1. Ja, denn die grosse Korrelation beweist, dass grosse Weinkenntnisse ein grosses Einkommen verursachen.
2. Nein, denn die grosse Korrelation beweist, dass es keinen kausalen Zusammenhang zwischen grossen Weinkenntnissen und grossem Einkommen geben kann.
3. Es ist keine Aussage möglich. Die Weinkenntnisse könnten die Ursache für ein grosses Einkommen sein, aber das kann man mit der Korrelation nicht zweifelsfrei beantworten.

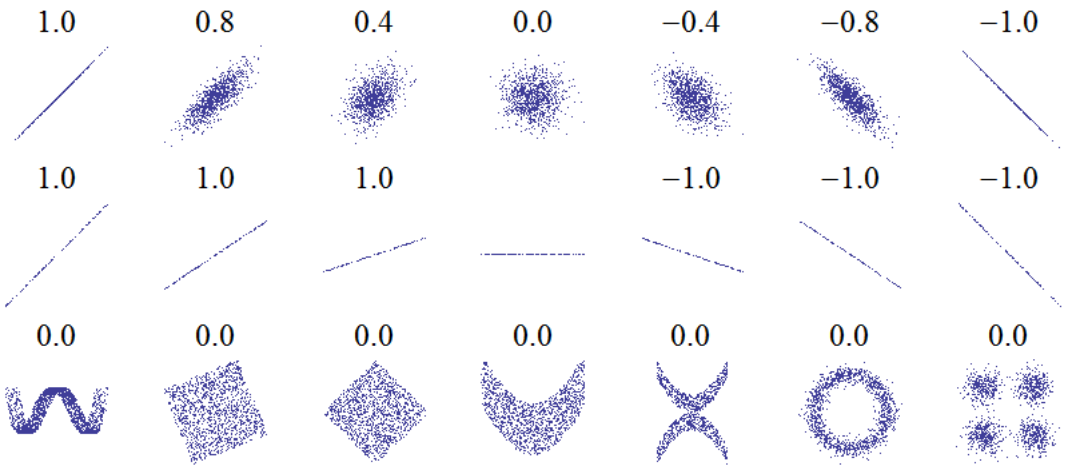

Abbildung 2.1: Es sind 21 verschiedene Datensätze dargestellt, die je aus vielen Beobachtungspaaren (x, y) bestehen. Für jedes Paar wurde ein Punkt gezeichnet. Über jedem Datensatz steht jeweils die zugehörige empirische Korrelation. Bei perfektem linearen Zusammenhang ist die empirische Korrelation +1, -1 oder 0 (je nachdem ob die Steigung positiv, negativ oder null ist; siehe zweite Zeile). Je mehr die Punkte um den linearen Zusammenhang streuen, desto kleiner wird der Betrg der empirischen Korrlation (siehe erste Zeile). Da die empirische Korrelation nur den *linearen* Zusammenhang misst, kann es einen Zusmmenhang zwischen den beiden Variablen x und y geben, auch wenn die empirische Korrelation null ist (siehe unterste Zeile).

Standardisierung

Durch Verschiebung und Skalierung der Werte kann man erreichen, dass zwei oder mehrere Datensätze die gleiche Lage und Streuung haben. Insbesondere kann man einen Datensatz so standardisieren, dass das arithmetische Mittel gleich Null und die Standardabweichung gleich 1 ist. Dies erreicht man mittels der linear transformierten Variablen

$$z_i = \frac{x_i - \overline{x}}{s_x} \ (i = 1, \ldots, n).$$

Alle Aspekte einer Verteilung, die bei einer Verschiebung oder Skalierung unverändert bleiben, machen die Form der Verteilung aus. Dazu gehört insbesondere die Schiefe (Asymmetrie) der Verteilung, für die es auch Kennzahlen gibt.

Übung: Zeigen Sie, dass das arithmetische Mittel von z_i gleich Null und die Standardabweichung von z_i gleich 1 ist.

2.2.2 Einfache lineare Regression

Wir erklären das Modell der einfachen linearen Regression zunächst mit einem fiktiven Beispiel. Je dicker ein Roman (Hardcover) ist, desto teurer ist er in der Regel. Es gibt also einen Zusammenhang zwischen Seitenzahl x und Buchpreis y. Wir gehen in einen Buchladen und suchen zehn Romane verschiedener Dicke aus. Wir nehmen dabei je ein Buch mit der Seitenzahl 50, 100, 150, ..., 450, 500. Von jedem Buch notieren wir die Seitenzahl und den Buchpreis. Damit erhalten wir Tabelle 2.1. :

	Seitenzahl	Buchpreis (SFr)
Buch 1	50	6.4
Buch 2	100	9.5
Buch 3	150	15.6
Buch 4	200	15.1
Buch 5	250	17.8
Buch 6	300	23.4
Buch 7	350	23.4
Buch 8	400	22.5
Buch 9	450	26.1
Buch 10	500	29.1

Tabelle 2.1: Zusammenhang zwischen Buchpreis und Seitenzahl (fiktiv).

Aus der Tabelle sehen wir tatsächlich, dass dickere Bücher tendenziell mehr kosten. Abbildung 2.2(a) zeigt diesen Zusammenhang graphisch.

Übung: Zeigen Sie, dass der Buchpreis positiv mit der Seitenzahl korreliert.

Wenn wir einen formelmässigen Zusammenhang zwischen Buchpreis und Seitenzahl hätten, könnten wir Vorhersagen für Bücher mit Seitenzahlen, die wir nicht beobachtet haben, machen. Oder wir könnten herausfinden, wie teuer ein Buch mit "null" Seiten wäre (das wären die Grundkosten des Verlags, die unabhängig von der Seitenzahl sind: Einband, administrativer Aufwand für jedes Buch, etc.). Wie könnten wir diesen Zusammenhang mit einer Formel beschreiben? Auf den ersten Blick scheint eine Gerade recht gut zu den Daten zu passen. Diese Gerade hätte die Form:

$$y = a + bx \,,$$

wobei y der Buchpreis und x die Seitenzahl sind. a wären dann die Grundkosten des Verlags und b wären die Kosten pro Seite. Versuchen Sie mit einem Lineal eine Gerade durch alle Punkte in Abb. 2.2(a) zu legen. Sie werden feststellen, dass das nicht möglich ist. Die Punkte folgen also nur ungefähr einer Geraden. Wie könnten wir eine Gerade finden, die möglichst gut zu allen Punkten passt? Hier gibt es verschiedene Möglichkeiten. Wir könnten die vertikalen Abstände zwischen Beobachtung und Gerade (siehe Abb. 2.2(b)) zusammenzählen und davon ausgehen, dass eine kleine Summe der Abstände eine gute Anpassung bedeutet. Wir bezeichnen

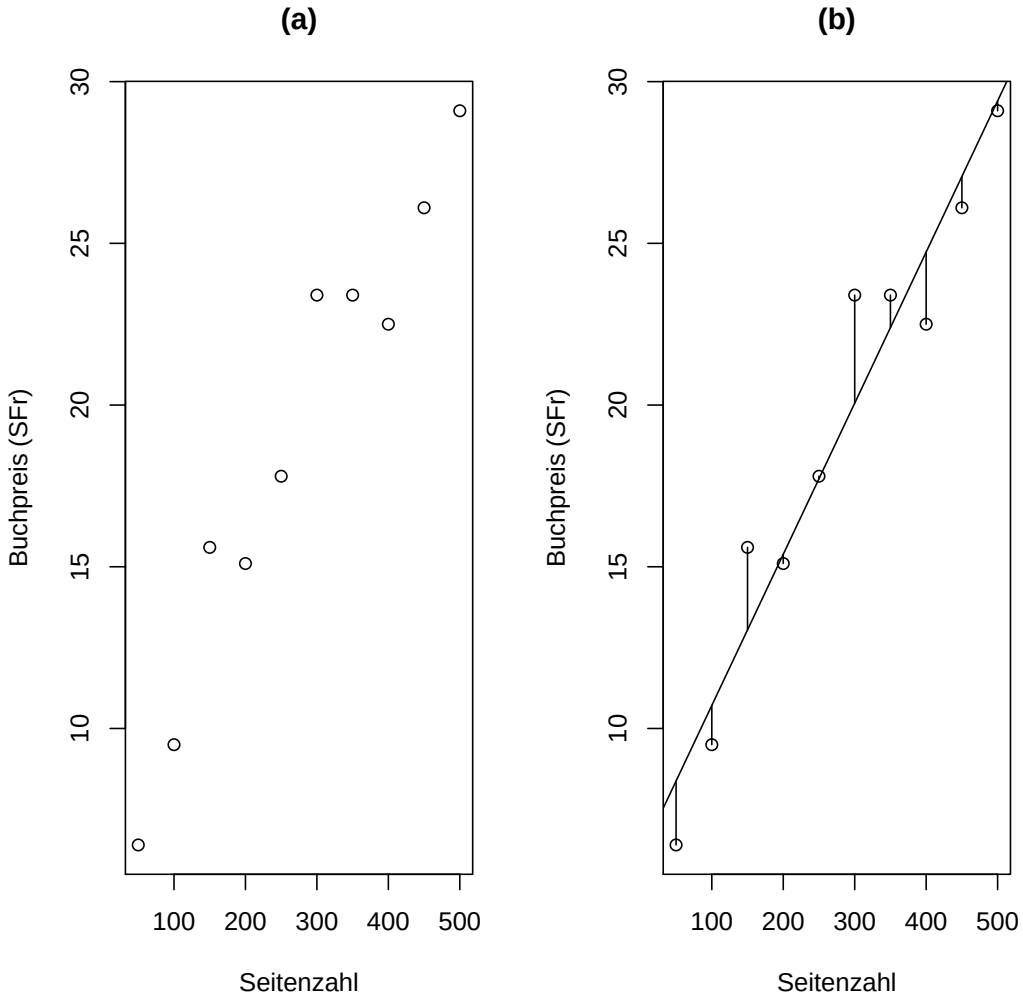

Abbildung 2.2: Zusammenhang zwischen Buchpreis und Seitenzahl (fiktiv).

den vertikalen Abstand zwischen einem Beobachtungspunkt (x_i, y_i) und der Geraden (der Punkt auf der Geraden ist $(x_i, a + bx_i)$) als **Residuum**:

$$r_i = y_i - a - bx_i \, .$$

Wir möchten also $\sum_i r_i$ minimieren. Diese Methode hat aber eine gravierende Schwäche: Wenn die Hälfte der Punkte weit über der Geraden, die andere Hälfte weit unter der Geraden liegen, ist die Summe der Abstände etwa null. Dabei passt die Gerade gar nicht gut zu den Datenpunkten. Die positiven Abweichungen haben sich nur mit den negativen Abweichungen ausgelöscht. Wir müssen also das Vorzeichen der Abweichungen eliminieren, bevor wir zusammenzählen. Eine Möglichkeit besteht darin, den Absolutbetrag der Abweichungen aufzusummieren, also $\sum_i |r_i|$. Eine andere Möglichkeit besteht darin, die Quadrate der Abweichungen aufzusummieren, also $\sum_i r_i^2$. Letztere Methode hat sich durchgesetzt, weil man mit ihr viel leichter rechnen kann, als mit den Absolutbeträgen. Eine Gerade passt (nach unserem Gütekriterium) also dann am besten zu Punkten, wenn die Quadratsumme der vertikalen Abweichungen minimal ist. Dieses Vorgehen ist unter dem Namen **Methode der kleinsten Quadrate** bekannt. In unserem Fall errechnet der Computer die Werte $a = 6.04$ und $b = 0.047$. Die Grundkosten des

Verlags sind also rund 6 SFr. Pro Seite verlangt der Verlag rund 5 Rappen.

Man nennt dieses Modell "einfach", weil nur eine x-Variable vorkommt. Später werden wir die "multiple" Regression kennenlernen, bei der mehrere x-Variablen vorkommen (z.B. Seitenzahl, Genre, Soft-/Hardcover, ...). Man nennt das Modell "linear", weil der Parameter b linear vorkommt.

Parameterschätzungen

Die unbekannten Modell-Parameter in der einfachen linearen Regression sind a und b. Die Methode der Kleinsten-Quadrate liefert die folgenden Schätzungen:

$$\hat{a}, \hat{b} \text{ sind Minimierer von } \sum_{i=1}^{n} (y_i - (a + bx_i))^2.$$

Die Lösung dieses Optimierungsproblem ergibt:

$$\begin{aligned} \hat{b} &= \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{a} &= \bar{y} - \hat{b}\bar{x}. \end{aligned}$$

Übung: Leiten Sie obige Schätzung für die Parameter a und b her.

Der geschätzte Koeffizient $\hat{b}$ lässt sich auch schreiben als:

$$\hat{b} = \frac{s_{xy}}{s_x^2}\,.$$

2.2.3 Graphische Methoden

Einen Überblick über die auftretenden Werte ergibt das *Histogramm*. Es gibt verschiedene Arten von Histogrammen; wir behandeln nur die gebräuchlichste. Um ein Histogramm zu zeichnen, bildet man Klassen konstanter Breite und zählt, wie viele Beobachtungen in jede Klasse fallen. Dann zeichnet man für jede Klasse einen Balken, dessen Höhe proportional zur Anzahl Beobachtungen in dieser Klasse ist (siehe Abbildung 2.3).[1] Wird die vertikale Achse so gewählt, dass die Fläche des Histogramms eins ist, dann ist die Fläche jedes Balkens proportional zur Anzahl Beobachtungen in der zugehörigen Klasse. Auf diese Art kann man eine Wahrscheinlichkeitsdichte (siehe unten) schätzen.

Beim *Boxplot* hat man ein Rechteck, das vom empirischen 25%- und vom 75%-Quantil begrenzt ist, und Linien, die von diesem Rechteck bis zum kleinsten- bzw. grössten "normalen" Wert gehen (per Definition ist ein normaler Wert höchstens 1.5 mal die Quartilsdifferenz von einem der beiden Quartile entfernt). Zusätzlich gibt man noch Ausreisser durch Sterne und den Median durch einen Strich an. Der Boxplot ist vor allem dann geeignet, wenn man die

[1] Eine andere, weit verbreitete Art des Histogramms erlaubt Klassen unterschiedlicher Breite.

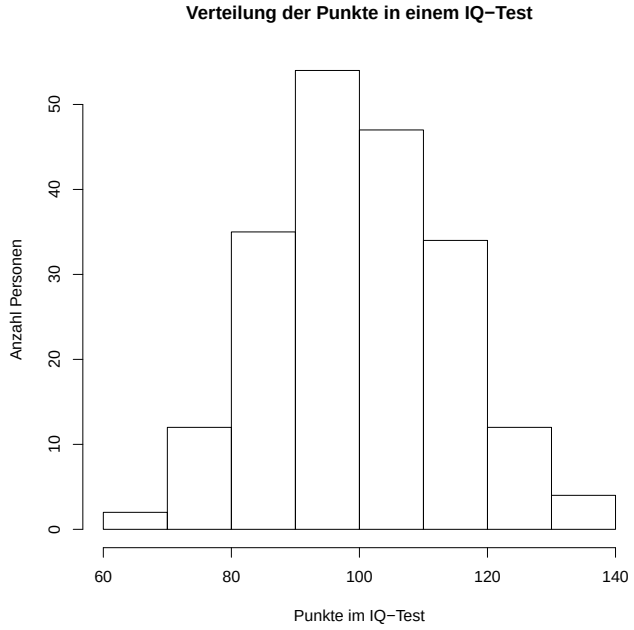

Abbildung 2.3: Histogramm von dem IQ-Test Ergebnis von 200 Personen. Die Breite der Klassen wurde als 10 IQ-Punkte festgelegt und ist für jede Klasse gleich. Die Höhe der Balken gibt die Anzahl Personen an, die in diese Klasse fallen. Z.B. fallen ca. 12 Personen in die Klasse zwischen 120 und 130 IQ-Punkten.

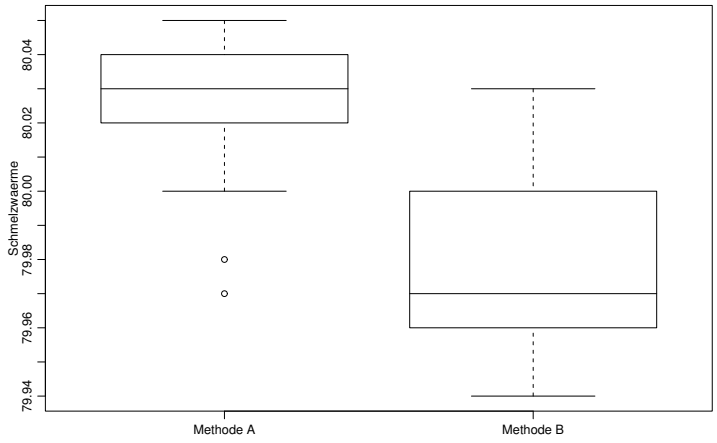

Abbildung 2.4: Boxplots für die zwei Methoden zur Bestimmung der Schmelzwärme von Eis.

Verteilungen einer Variablen in verschiedenen Gruppen (die im allgemeinen verschiedenen Versuchsbedingungen entsprechen) vergleichen will (siehe Abbildung 2.4).

Die *empirische kumulative Verteilungsfunktion* $F_n(\cdot)$ ist eine Treppenfunktion, die links von $x_{(1)}$ gleich null ist und bei jedem $x_{(i)}$ einen Sprung der Höhe $\frac{1}{n}$ hat (falls ein Wert mehrmals vorkommt, ist der Sprung ein Vielfaches von $\frac{1}{n}$). In andern Worten:

$$F_n(x) = \frac{1}{n}\text{Anzahl}\{i \mid x_i \leq x\}.$$

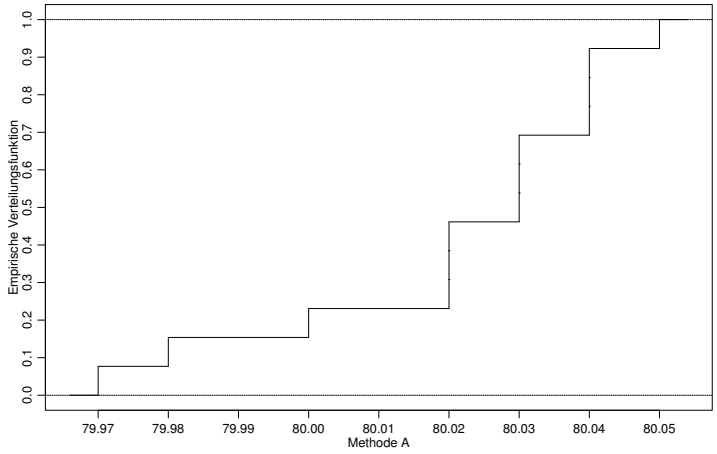

Abbildung 2.5: Empirische kumulative Verteilungsfunktion der Messungen der Schmelzwärme von Eis mit Methode A.

Abbildung 2.5 zeigt die empirische kumulative Verteilungsfunktion für die Messungen der Schmelzwärme von Eis mit Methode A.

Mehrere Variablen

Wenn wir bei jeder Versuchseinheit zwei verschiedene Grössen messen, d.h. wenn die Daten von der Form $(x_1, y_1), \ldots, (x_n, y_n)$ sind, interessiert man sich in erster Linie für die Zusammenhänge und Abhängigkeiten zwischen den Variablen. Diese kann man aus dem *Streudiagramm* ersehen, welches die Daten als Punkte in der Ebene darstellt: Die i-te Beobachtung entspricht dem Punkt mit Koordinaten (x_i, y_i). Die Abbildung 2.6 zeigt das Streudiagramm für die Werte

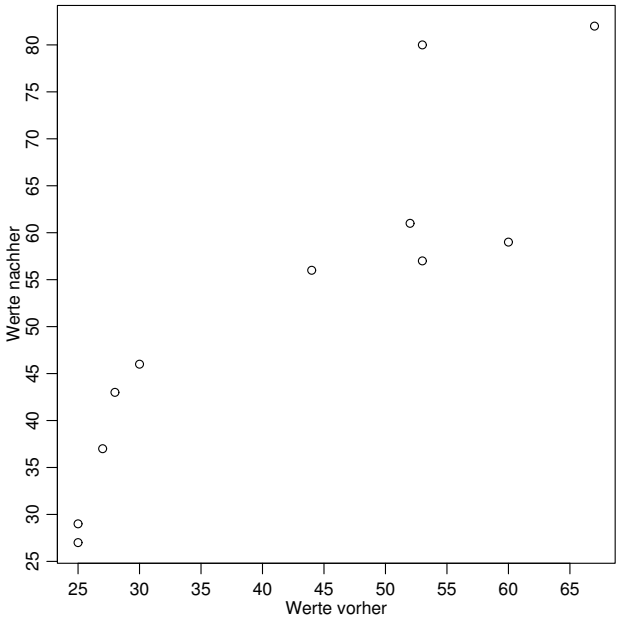

Abbildung 2.6: Streudiagramm der Blutplättchen-Aggregation vor und nach dem Rauchen einer Zigarette.

"vorher" und "nachher" bei der Blutplättchen-Aggregation. Man sieht einen klaren monotonen Zusammenhang, Individuen haben also eine Tendenz zu starker, bzw. schwacher Aggregation, unabhängig vom Rauchen.

2.2.4 Analogien zwischen Modellen und Daten

Zufallsvariablen und Verteilungen beschreiben die Population. Daten $x_1, \ldots, x_n$ interpretieren wir als Realisierungen von Zufallsvariablen $X_1, \ldots, X_n$ (man könnte auch die n Daten als n Realisierungen von einer Zufallsvariablen X interpretieren; die Schreibweise mit mehreren Zufallsvariablen hat jedoch Vorteile, siehe später).

Aus Daten können wir Rückschlüsse auf die zugrunde liegende Verteilung ziehen. Insbesondere haben alle Grössen, die für Zufallsvariablen definiert sind, ein Gegenstück für Datensätze gemäss folgender Tabelle. Die empirischen Grössen sind Schätzungen für die theoretischen Grössen. Diese werden mit wachsendem Stichprobenumfang n immer genauer.

Daten	Population (Modell)
Histogramm	Dichte
empirische kumulative Verteilungsfkt.	theoretische kumulative Verteilungsfkt.
empirische Quantile	theoretische Quantile
Arithmetisches Mittel	Erwartungswert
empirische Standardabweichung	theoretische Standardabweichung.

Es ist jedoch wichtig, die Unterschiede zwischen den empirischen und theoretischen Grössen zu verstehen: Mit Hilfe der Statistik werden wir quantifizieren, wie gross die Unsicherheit ist, wenn wir die theoretischen Grössen mit Hilfe der empirischen Grössen schätzen.