```
quantile(methodeA, 0.7, type = 2)

##    70%
## 80.04
```

Rund 10 % der Messwerte sind kleiner oder gleich 79.98 . Entsprechend sind rund 70 % der Messwerte kleiner oder gleich 80.04.

Beispiel 2.1.7

In einer Schulklasse mit 24 SchülerInnen gab es an einer Prüfung folgende Noten:

4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9, 6, 4, 3.7, 5, 5.2, 4.5, 3.6, 5, 6, 2.8, 3.3, 5.5, 4.2, 4.9, 5.1

Wir berechnen nun mit R verschiedene Quantile:

```
noten.1 <- c(4.2, 2.3, 5.6, 4.5, 4.8, 3.9, 5.9, 2.4, 5.9,
    6, 4, 3.7, 5, 5.2, 4.5, 3.6, 5, 6, 2.8, 3.3, 5.5, 4.2,
    4.9, 5.1)

quantile(noten.1, seq(0.2, 1, 0.2), type = 2)

##  20%  40%  60%  80% 100%
##  3.6  4.2  5.0  5.6  6.0
```

Rund 20 % der SchülerInnen haben also eine 3.6 oder waren schlechter. Genau 20 % der SchülerInnen ist nicht möglich, da dies 4.8 SchülerInnen entsprechen würde. Das 60 %-Quantil besagt, dass 60 Prozent der SchülerInnen eine 5 haben oder schlechter waren. Folglich haben 40 % eine 5 oder sind besser.

□

2.1.5. Graphische Methoden

Histogramm

Einen graphischen Überblick über die auftretenden Werte erhalten wir mit einem sogenannten *Histogramm*. Histogramme helfen uns bei der Frage, in welchem Wertebereich besonders viele Datenpunkte liegen. Ist die Datenmenge gross, so macht es keinen Sinn, alle Werte einzeln zu betrachten. Wir bilden sogenannte *Klassen*, die

jeweils einen Ausschnitt des Beobachtungsbereiches darstellen. Um ein Histogramm zu zeichnen, bildet man Klassen (einfachheitshalber mit konstanter Breite) und zählt, wie viele Beobachtungen in jede Klasse fallen. Es gibt verschiedene Arten von Histogrammen; wir behandeln hier nur die gebräuchlichste.

Beispiel 2.1.8

In Abbildung 2.1 sehen wir ein Histogramm von dem Ergebnis eines IQ-Testes von 200 Personen.

- Die Breite der Klassen wurde mit 10 IQ-Punkten festgelegt und ist für jede Klasse gleich.
- Die Höhe der Balken gibt die Anzahl Personen an, die in diese Klasse fallen. Zum Beispiel fallen ca. 14 Personen in die Klasse zwischen 120 und 130 IQ-Punkten.

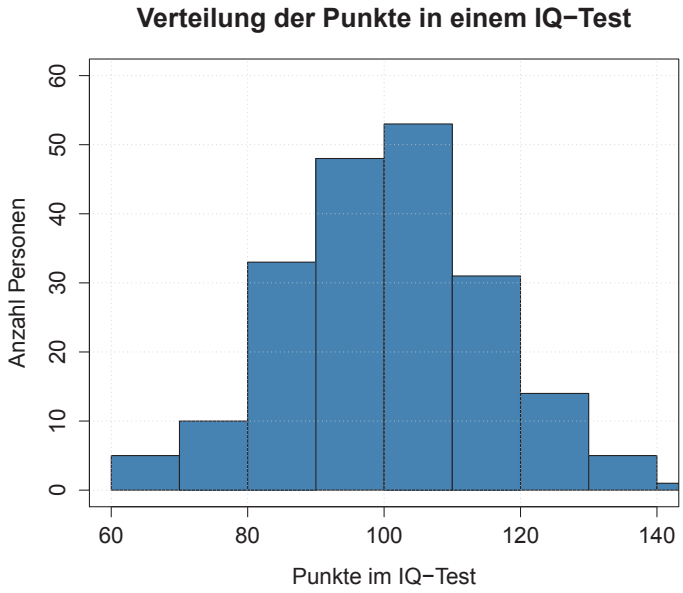


Abbildung 2.1.: Histogramm von dem IQ-Test Ergebnis von 200 Personen.

□

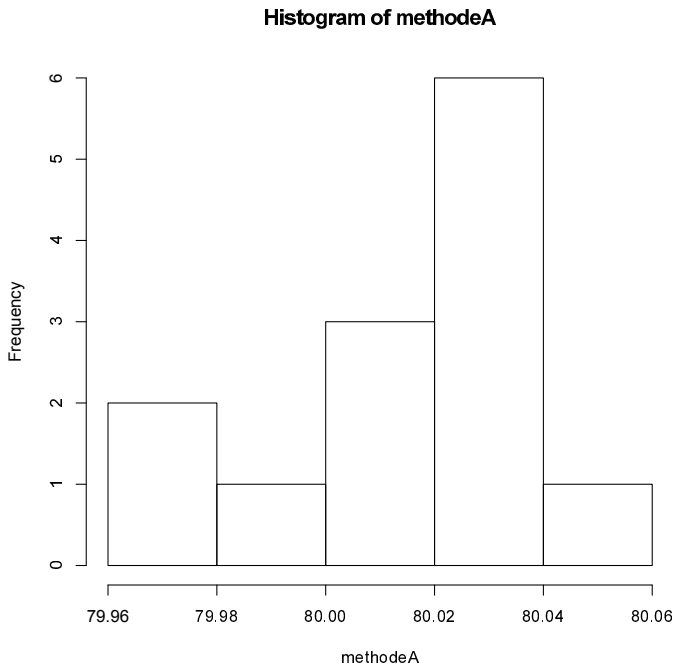
Schrittweise Konstruktion eines Histogramms:

- Wir teilen die Datenmenge in Klassen ein. Für die Festlegung der Anzahl der Klassen bzw. Rechtecke existieren verschiedene Faustregeln: bei weniger als 50 Messungen ist die Klassenzahl 5 bis 7, bei mehr als 250 Messungen wählt man

10 bis 20 Klassen.[2] Im einfachsten Fall wird die gleiche Breite für alle Klassen gewählt, was aber nicht unbedingt der Fall sein muss.

- Dann zeichnen wir für jede Klasse einen Balken, dessen Höhe proportional zur Anzahl Beobachtungen in dieser Klasse ist.
- Dividieren wir die Anzahl Beobachtungen in einer Klasse durch die Gesamtzahl der Beobachtungen, so erhalten wir den prozentualen Anteil einer Klasse zur Gesamtbeobachtung.

```
hist(methodeA)
```



Bemerkungen:

i. Da die Methode A nur aus 13 Messungen besteht, wählt man 5 Balken (Sturges-Regel: $k = 1 + \log_2 13 \approx 5$).

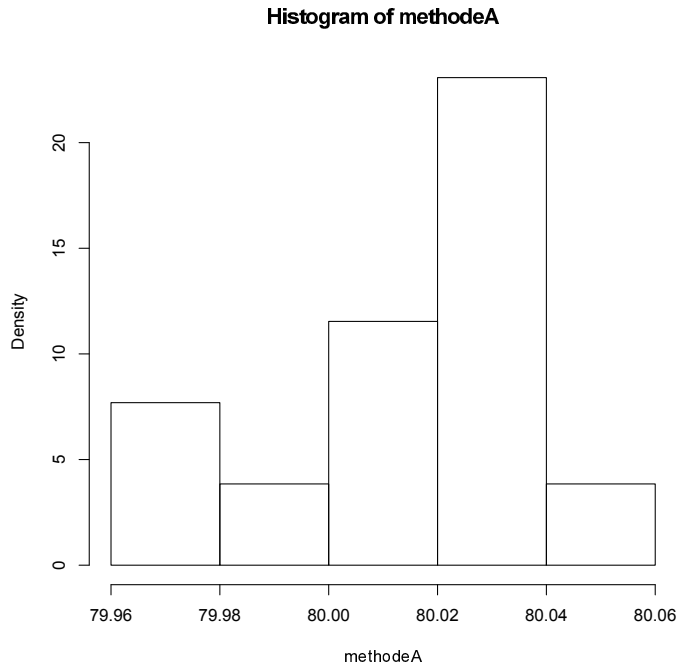
ii. Bedeutung der Anzahlen (Frequency): in der 1. Klasse 79.96-79.98 sind die Beobachtungen mit den Werten 79.97 und 79.98 berücksichtigt; in der 2. Klasse 79.99 und 80.00; usw. Der linke Rand wird also nicht berücksichtigt, der rechte dagegen schon. Man hätte dies auch umgekehrt machen können, und das Histogramm würde etwas anders aussehen. Bei grossen Datensätzen spielen solche Überlegungen kaum eine Rolle.

[2] Gegebenenfalls kann man die Anzahl der Klassen k auch nach der Sturges-Regel berechnen: $k = 1 + \log_2 n = 1 + 3.3 \cdot \log_{10} n$, wobei n die Anzahl Messungen ist.

iii. Mit dem R-Befehl lassen sich auch die Anzahl und die Breiten der Klassen festlegen, Überschriften ändern, usw. (siehe Übungen).

Im Histogramm oben entspricht die Höhe der Balken gerade der Anzahl der Beobachtungen in einer Klasse. Oft ist es besser und übersichtlicher, wenn wir die Balkenhöhe so wählen, dass die Balkenfläche dem prozentualen Anteil der jeweiligen Beobachtungen an der Gesamtanzahl Beobachtungen entspricht. Die Gesamtfläche aller Balken muss dann gleich eins sein.

```
hist(methodeA, freq = F)
```



Auf der vertikalen Achse sind nun die Dichten angegeben. Wir können also herauslesen, dass sich über $(80.04 - 80.02) \cdot 20 = 0.4$, also über 40 % der Daten zwischen 80.02 und 80.04 befinden. Die Balkenhöhe ermittelt sich, indem man die Gesamtanzahl Beobachtungen mit $\frac{1}{n}$ multipliziert und diese Zahl durch die Balkenbreite dividiert. Diese Darstellung hat den Vorteil, dass man Messungen mit unterschiedlichen Umfängen besser miteinander vergleichen kann. Würde man also mit Methode A nun eine Messung mit 30 Beobachtungen durchführen, liessen sich mit Dichten besser die Verteilungen von Messwerten auf die jeweiligen Klassen vergleichen.

Boxplot

Der *Boxplot* (siehe Abbildung 2.2) besteht aus

- einem Rechteck, dessen Höhe vom empirischen 25 %- und vom 75 %-Quantil begrenzt wird,
- Linien, die von diesem Rechteck bis zum kleinsten- bzw. grössten „normalen“ Wert führen (per Definition ist ein „normaler“ Wert höchstens 1.5 mal die Quartilsdifferenz von einem der beiden Quartile entfernt),
- einem horizontalen Strich für den Median,
- kleinen Kreisen, die Ausreisser markieren.

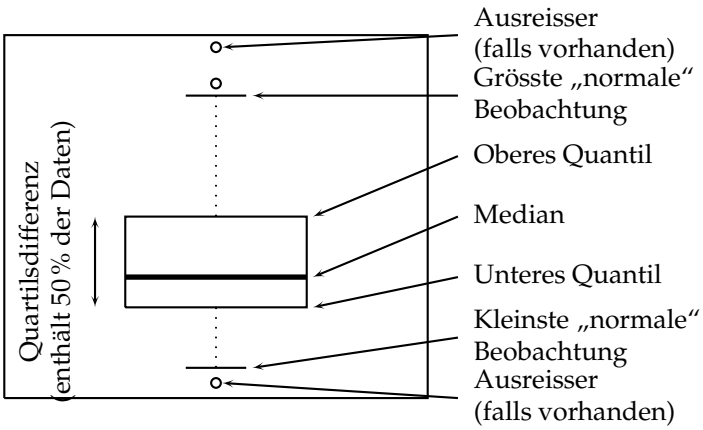
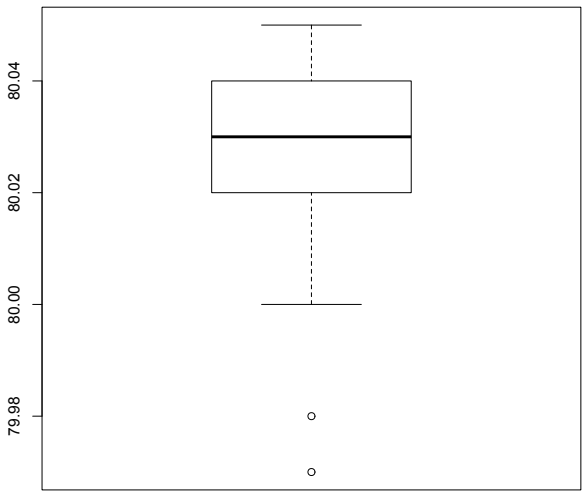


Abbildung 2.2.: Boxplot

```
boxplot(methodeA)
```



Bemerkungen:

i. Die Hälfte der Beobachtungen befindet sich zwischen dem oberen Quartil 80.04 und dem unteren Quartil 80.02, mit Quartilsdifferenz 0.02

ii. Der Median liegt bei 80.03.

iii. Der „normale“ Bereich der Werte liegt zwischen 80.00 und 80.05.

iv. Wir haben zwei Ausreisser 79.97 und 79.98.

v. Die Punkte a) und b) hatten wir schon bei den Quantilen berechnet. Der Boxplot stellt somit unsere Berechnungen graphisch dar.

Der Boxplot ist vor allem dann geeignet, wenn man die Verteilungen der Daten in verschiedenen Gruppen (die im Allgemeinen verschiedenen Versuchsbedingungen entsprechen) vergleichen will.

Beispiel 2.1.9

Bei unserem Einführungsbeispiel der Schmelzwärme haben wir zwei Methoden zu deren Bestimmung verwendet. Somit können wir die Boxplots auch gegenüberstellen und die Methoden miteinander vergleichen (siehe Abbildung 2.3).

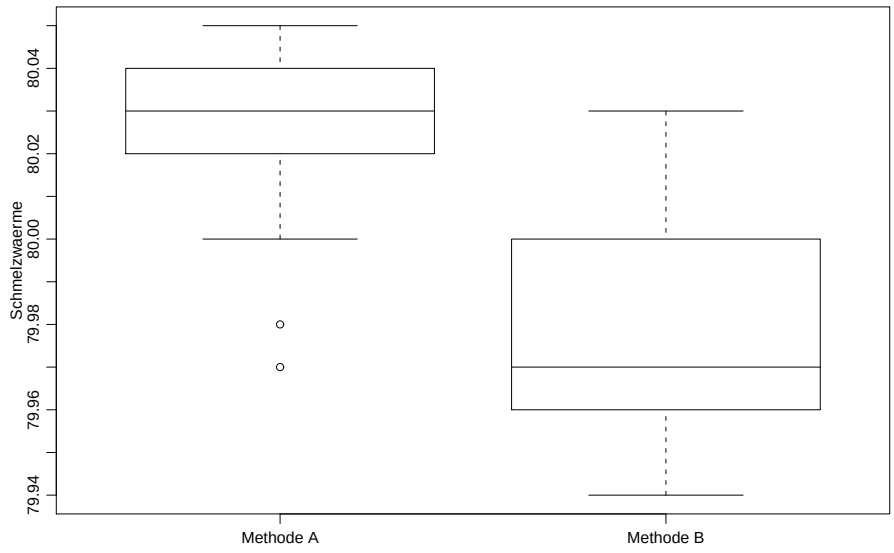


Abbildung 2.3.: Boxplots für die zwei Methoden zur Bestimmung der Schmelzwärme von Eis.

Bemerkungen:

i. Methode A liefert die grösseren Werte als Methode B, da der Median von A grösser ist.

ii. Die Daten von Methode A haben weniger Streuung als die Daten von Methode B, da das Rechteck weniger hoch ist (Quartilsdifferenz!).

□

Empirische kumulative Verteilungsfunktion

Eine weitere graphische Darstellung der Daten ist die *empirische kumulative Verteilungsfunktion*. Diese Darstellung hat den Vorteil gegenüber einem Histogramm, dass man den Median sehr leicht ablesen kann.

Die *empirische kumulative Verteilungsfunktion* $F_n(\cdot)$ ist eine Treppenfunktion, die wie folgt erzeugt wird: links von $x_{(1)}$ ist die Funktion gleich null und bei jedem $x_{(i)}$ wird ein Sprung der Höhe $1/n$ gemacht (falls ein Wert mehrmals vorkommt, ist der Sprung das entsprechende Vielfache von $1/n$). Im folgenden Beispiel wird dieses Vorgehen konkret durchgeführt.

Beispiel 2.1.10 Methode A der Schmelzwärme

In Abbildung 2.4 ist die kumulative Verteilungsfunktion der Methode A aufgezeichnet. Sie entsteht wie folgt:

- Links von 79.97 ist die Funktion 0, da es keinen kleineren Beobachtungswert hat.
- Bei 79.97 macht die Funktion einen Sprung auf $n = 1/13 \approx 0.077$.
- Die Funktion bleibt dann gleich bis 80.00, da es vorher keinen zusätzlichen Beobachtungswert gibt. Bei 80.00 macht die Funktion wieder einen Sprung um 0.077 nach oben, weil es dort einen Messwert hat.
- Bei 80.02 macht die Funktion einen Sprung um $3 \cdot 0.077$ nach oben, da es dort 3 Beobachtungswerte gibt.
- usw.
- Bei 80.05 machen wir unseren letzten Sprung und der Funktionswert wird 1.

Was können wir aus der kumulativen Verteilungsfunktion herauslesen?

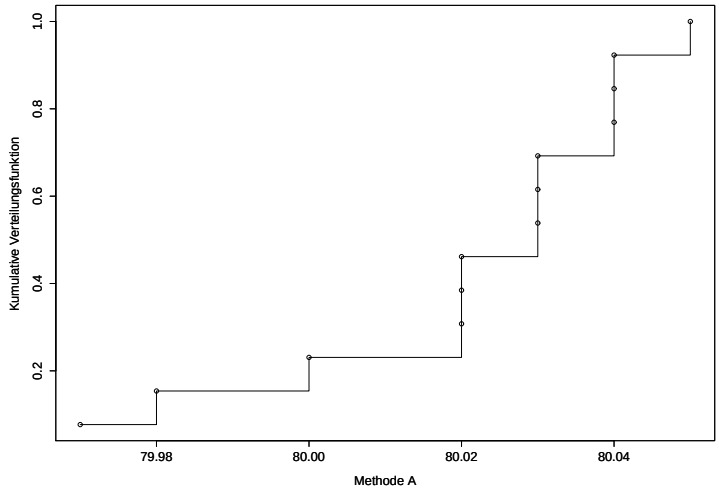


Abbildung 2.4.: Empirische kumulative Verteilungsfunktion der Messungen der Schmelzwärme von Eis mit Methode A.

- Bei 0.5 auf der vertikalen Achse haben wir gerade die Hälfte aller Werte aufsummiert. Zeichnen wir von 0.5 eine horizontale Linie (siehe grüne Linie in Abbildung 2.5, wird die kumulative Verteilungsfunktion bei 80.03 geschnitten. Das entspricht gerade dem Median.
- Dort, wo die kumulative Funktion steil ist, hat es auch viele Beobachtungswerte. Das heisst, die meisten Beobachtungswerte liegen hier zwischen 80.02 und 80.04. Die Werte entsprechen aber gerade dem unteren und oberen Quartil. (Man vergleiche die Funktion mit dem zugehörigen Boxplot.)

□

Allgemein

Die **kumulative Verteilungsfunktion** ist definiert als

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

Mit R lässt sich die kumulative Verteilungsfunktion in der Abbildung 2.5 folgendermassen aufzeichnen:

```
n <- length(methodeA)
```

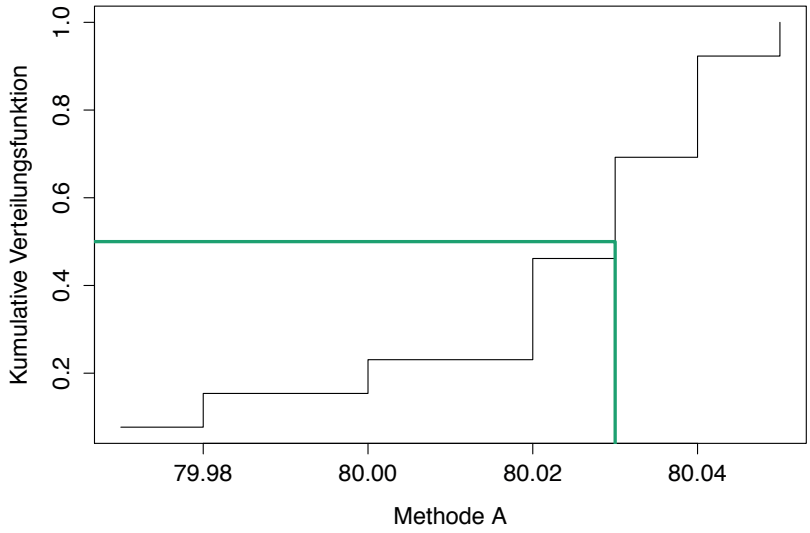
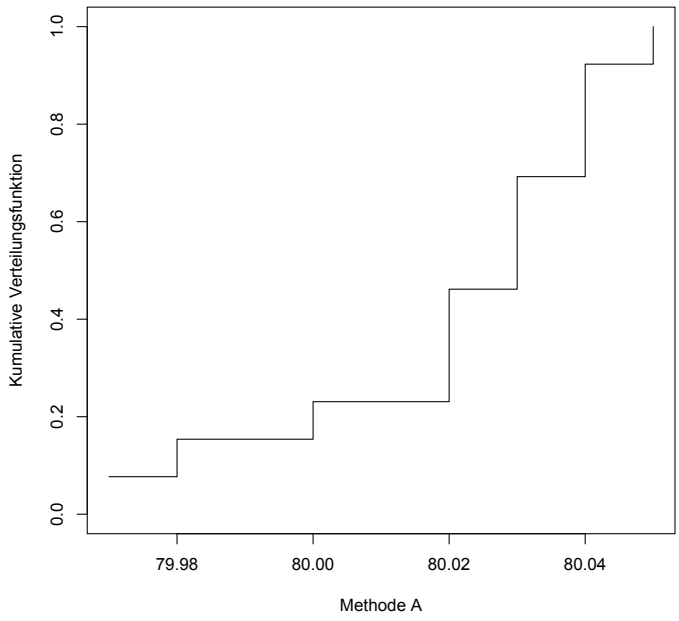


Abbildung 2.5.: Empirische kumulative Verteilungsfunktion der Messungen der Schmelzwärme von Eis mit Methode A.

```
plot(sort(methodeA), (1:n)/n, type = "s", ylim = c(0, 1),
    ylab = "Kumulative Verteilungsfunktion", xlab = "Methode A")
```



2.2. Deskriptive Statistik zweidimensionaler Daten

Bei zweidimensionalen Daten werden an einem Versuchsobjekt jeweils *zwei* verschiedene Grössen gemessen. So wird beispielsweise an einer Gruppe von Menschen jeweils die Körpergrösse *und* das Körpergewicht gemessen.

Beispiel 2.2.1 Weinkonsum und Mortalität

Wir betrachten als Beispiel einen Datensatz (siehe Tabelle 2.2), der den durchschnittlichen Weinkonsum (in Liter pro Person und Jahr) und die Sterblichkeit (Mortalität) aufgrund von Herz- und Kreislauferkrankungen (Anzahl Todesfälle pro 1000 Personen zwischen 55 und 64 Jahren pro Jahr) in 18 industrialisierten Ländern umfasst (A.S.St.Leger, A.L.Chochrane, and F.Moore, "Factors Associated with Cardiac Mortality in Developed Countries with Particular Reference to the Consumption of Wine."*Lancet, 1979*). Es stellt sich nun die Frage, ob diese Daten suggerieren, dass es einen Zusammenhang zwischen der Sterblichkeitsrate aufgrund von Herzkreislauferkrankung und Weinkonsum gibt.

Land	Weinkonsum	Mortalität Herzerkrankung
Norwegen	2.8	6.2
Schottland	3.2	9.0
Grossbritannien	3.2	7.1
Irland	3.4	6.8
Finnland	4.3	10.2
Kanada	4.9	7.8
Vereinigte Staaten	5.1	9.3
Niederlande	5.2	5.9
New Zealand	5.9	8.9
Dänemark	5.9	5.5
Schweden	6.6	7.1
Australien	8.3	9.1
Belgien	12.6	5.1
Deutschland	15.1	4.7
Österreich	25.1	4.7
Schweiz	33.1	3.1
Italien	75.9	3.2
Frankreich	75.9	2.1

Tabelle 2.2.: Weinkonsumation (Liter pro Person pro Jahr) und Mortalität aufgrund von Herzkreislauferkrankung (Todesfälle pro 1000) in 18 Ländern.

Ein kurzer Blick auf die Tabelle zeigt, dass ein höherer Weinkonsum eher weniger Todesfälle wegen Herz- und Kreislaufkrankheiten zur Folge hat.

□

2.2.1. Graphische Darstellung: Streudiagramm

Ein wichtiger Schritt in der Untersuchung zweidimensionaler Daten ist die graphische Darstellung. Dies geschieht meist über ein sogenanntes *Streudiagramm* (engl.: *Scatterplot*). Dabei werden jeweils zwei Messungen als Koordinaten von Punkten in einem Koordinatensystem interpretiert und dargestellt.

In unserem Beispiel stellt ein Land eine Versuchseinheit dar, und es wird die Grösse „Weinkonsum“ $x_1, \ldots, x_{18}$ und die Grösse „Mortalität“ $y_1, \ldots, y_{18}$ gemessen. Wenn wir die Daten in der Form $(x_1, y_1), \ldots, (x_{18}, y_{18})$ schreiben, interessiert man sich in erster Linie für die Zusammenhänge und Abhängigkeiten zwischen den Variablen x und y. Die Abhängigkeit zwischen den beiden Messgrössen kann man aus dem *Streudiagramm* ersehen, welches die Daten als Punkte in der Ebene darstellt: Die i-te Beobachtung (i-tes Land) entspricht dem Punkt mit Koordinaten (x_i, y_i). Die Abbildung 2.6 zeigt das Streudiagramm für die Messgrössen „Weinkonsum“ $(x_1, x_2, \ldots, x_{18})$ und „Mortalität“ $(y_1, y_2, \ldots, y_{18})$. Man sieht einen klaren monoton fallenden Zusammenhang: Länder mit hohem Weinkonsum haben also eine Tendenz zu einer tieferen Mortalitätsrate wegen Herz- und Kreislaufkrankheiten.

Das Streudiagramm in Abbildung 2.6 wurde mit R erstellt.

```
wein <- c(2.8, 3.2, ..., 75.9)

mort <- c(6.2, 9, ..., 2.1)

plot(wein, mort, xlab = "Weinkonsum (Liter pro Jahr und Person)",
    ylab = "Mortalitaet")
```

Bemerkungen:

i. Der Schluss, dass hoher Weinkonsum gesund ist, ist *falsch*. Es *scheint*, dass höherer Weinkonsum zu weniger Toten wegen Herz- und Kreislaufkrankheiten führt. Der Einfluss des höheren Weinkonsums auf andere Körperorgane (z.B. Leber) oder auf die Anzahl Verkehrsunfälle, wird hier *nicht* untersucht.

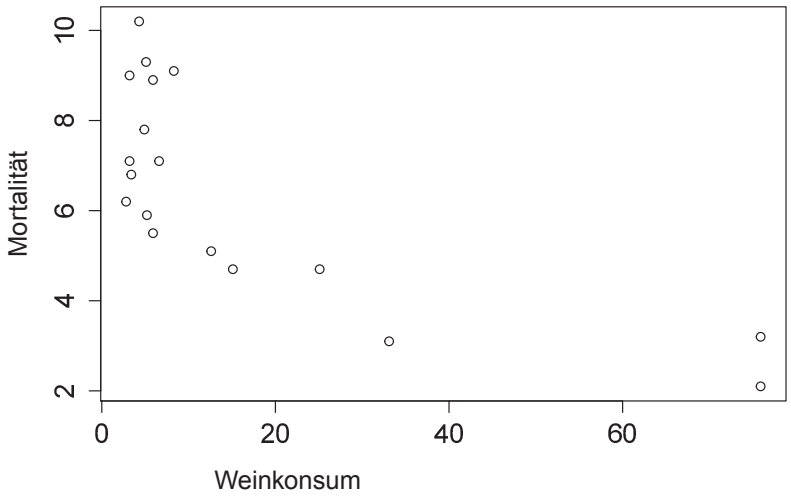


Abbildung 2.6.: Streudiagramm für die Mortalität und den Weinkonsum in 18 industrialisierten Ländern.

ii. Aus den Zahlwerten und Scatterplot lässt sich ein Zusammenhang *erahnen*, aber es muss keinen *kausalen* Zusammenhang zwischen den Messreihen haben. Das Streudiagramm kann zufällig so aussehen, also muss der Weinkonsum keinen Einfluss auf die Sterblichkeit haben.

2.2.2. Einfache lineare Regression

Wir haben im vorangehenden Beispiel eine negative (je mehr desto weniger) Abhängigkeit zwischen Mortalität und Weinkonsum festgestellt. Oft ist diese Abhängigkeit sehr einfach, nämlich *linear*.

Beispiel 2.2.2 Zusammenhang Seitenzahl-Preis eines Buches

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 Dicken 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. Mit diesen Daten erstellen wir Tabelle 2.3. Aus der Tabelle ist tatsächlich ersichtlich, dass dickere Bücher tendenziell mehr kosten. Wenn wir einen formelmässigen Zusammenhang zwischen Buchpreis und

	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 2.3.: Zusammenhang zwischen Buchpreis und Seitenzahl (fiktiv).

Seitenzahl hätten, könnten wir Vorhersagen über den Preis für Bücher mit Seitenzahlen machen, die wir nicht beobachtet haben. Was würde dann voraussichtlich ein Buch mit 375 Seiten kosten? Oder wir könnten herausfinden, wie teuer ein Buch mit „null“ Seiten wäre. Das wären die Grundkosten des Verlags, die unabhängig von der Seitenzahl anfallen: Einband, administrativer Aufwand für jedes Buch, etc. Wie könnten wir diesen Zusammenhang mit einer Formel beschreiben? Das Streudiagramm in Abbildung 2.7 zeigt diesen Zusammenhang graphisch deutlicher auf.

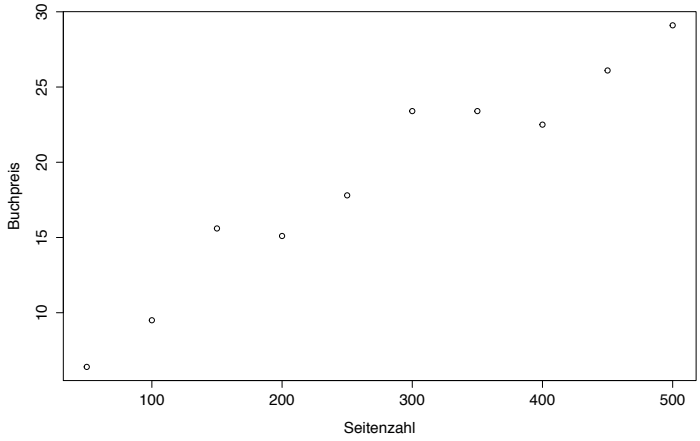


Abbildung 2.7.: Streudiagramm Seitenzahl - Buchpreis

Auf den ersten Blick scheint eine Gerade recht gut zu den Daten zu passen. Diese Gerade hätte die Form

$$y = a + bx$$

mit y dem Buchpreis und x der Seitenzahl sind. Der Parameter a beschreibt dann die Grundkosten des Verlags und der Parameter b entspricht den Kosten pro Seite.

□

Methode der kleinsten Quadrate

Versuchen wir mit einem Lineal eine Gerade durch *alle* Punkte in Abbildung 2.7 zu legen, so werden wir feststellen, dass das nicht möglich ist (siehe Abbildung 2.8). Die Punkte folgen also nur *ungefähr* einer Geraden.

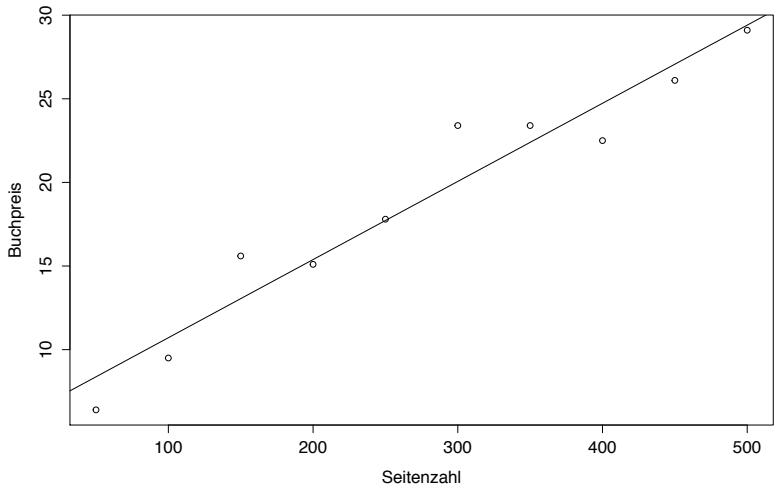


Abbildung 2.8.: Gerade durch das Streudiagramm

Da stellt sich uns die Frage: „Wie können wir eine Gerade finden, die *möglichst gut* zu allen Punkten passt?“ Damit stellt sich die nächste Frage: Was heisst „möglichst gut“? Hier gibt es verschiedene Möglichkeiten. Es treten ähnliche Schwierigkeiten auf, wie bei der Bestimmung der Varianz. Die naheliegendste Möglichkeit ist nicht die optimalste.

Wir könnten die Gerade so wählen, dass wir die vertikalen Differenzen zwischen Beobachtung und Gerade (siehe Abbildung 2.9) zusammenzählen und davon ausgehen, dass eine kleine Summe der Abstände eine gute Anpassung bedeutet.

Wir bezeichnen die vertikale Differenz zwischen einem Beobachtungspunkt (x_i, y_i) und der Geraden (der Punkt auf der Geraden hat die Koordinaten $(x_i, a + bx_i)$) als **Residuum**:

$$r_i = y_i - (a + bx_i) = y_i - a - bx_i$$

Für unser Beispiel sind die Residuen r_6 und r_8 für *diese* Gerade in Abbildung 2.9 dargestellt. Das Residuum r_6 ist positiv, da der Punkt oberhalb der Geraden liegt. Entsprechend ist $r_8 < 0$.

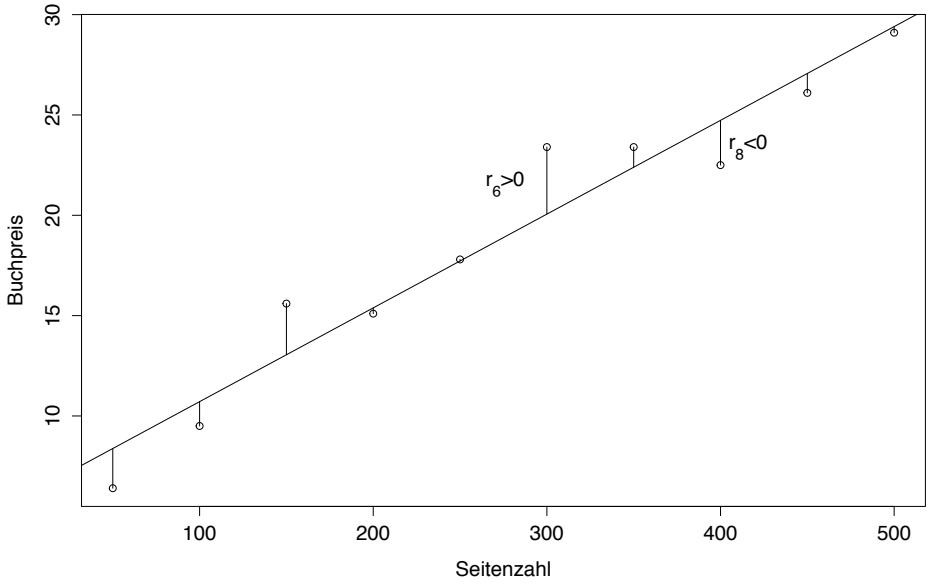


Abbildung 2.9.: Residuen

Wir möchten also die Gerade $y = a + bx$ so bestimmen, dass die Summe

$$r_1 + r_2 + \ldots + r_n = \sum_i r_i$$

minimal wird. Diese Methode hat aber eine gravierende Schwäche: Wenn die Hälfte der Punkte weit über der Geraden, die andere Hälfte weit unter der Geraden liegen, so ist die Summe der Abweichungen (Residuen) etwa null. Dabei passt die Gerade gar nicht gut zu den Datenpunkten. Die positiven Abweichungen heben sich nur mit den negativen Abweichungen auf. Wir müssen also das Vorzeichen der Abweichungen eliminieren, bevor wir zusammenzählen. Eine Möglichkeit besteht darin, den Absolutbetrag der Abweichungen aufzusummieren, also $\sum_i |r_i|$, und diese Summe zu minimieren. Da es sich aber mit Absolutbeträgen nicht besonders bequem rechnen lässt (zum Beispiel, wenn man Ausdrücke mit Absolutbeträgen ableiten möchte),

halten wir nach einer anderen Möglichkeit Ausschau. Die besteht darin, die Quadrate der Abweichungen aufzusummieren, also

$$r_1^2 + r_2^2 + \cdots + r_n^2 = \sum_i r_i^2$$

Die Parameter a und b sind so zu wählen, dass diese Summe minimal wird. 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 Summe der Quadrate der vertikalen Abweichungen minimal ist. Dieses Vorgehen ist unter dem Namen **Methode der kleinsten Quadrate** bekannt. In unserem Fall erhalten wir mit R die Werte $a = 6.04$ und $b = 0.047$.

```
seitenzahl <- c(seq(50, 500, 50))

buchpreis <- c(6.4, 9.5, 15.6, 15.1, 17.8, 23.4, 23.4, 22.5,
    26.1, 29.1)

lm(buchpreis ~ seitenzahl)


##
## Call:
## lm(formula = buchpreis ~ seitenzahl)
##
## Coefficients:
## (Intercept)   seitenzahl
##     6.04000      0.04673
```

Die Geradengleichung lautet

$$y = 6.04 + 0.04673x$$

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

Bemerkungen:

i. Der Befehl `lm()` steht für „linear model“.

ii. Mit dem Befehl `lm(y~x)` passt R ein Modell von der Form $y = a + bx$ an die Daten an.

iii. Diese Gerade wird auch *Regressionsgerade* genannt.

Beispiel 2.2.3

Wie viel würde nach diesem Modell ein Buch von 375 Seiten kosten? Dazu setzen wir $x = 375$ in die Geradengleichung oben ein und erhalten

$$y = 6.04 + 0.04673 \cdot 375 \approx 23.60$$

Das Buch dürfte also etwa CHF. 23.60 kosten. Dieses Modell ist allerdings nur begrenzt gültig. Vor allem bei *Extrapolationen* muss man vorsichtig sein. Wir könnten schon ausrechnen, wie viel ein Buch mit einer Million Seiten kostet, aber dieser Betrag entspricht dann sicher nicht mehr der Realität.

□

Diese Gerade in Abbildung 2.8 auf Seite 33 wird in R wie folgt gezeichnet:

```
seite <- c(seq(50, 500, 50))

preis <- c(6.4, 9.5, 15.6, 15.1, 17.8, 23.4, 23.4, 22.5,
    26.1, 29.1)

plot(seite, preis, xlab = "Seitenzahl", ylab = "Buchpreis")

abline(lm(preis ~ seite))
```

Wie berechnet R die Parameter a und b?

Die Parameter a und b werden wie folgt bestimmt:

Die Parameter a und b sollen den folgenden Ausdruck minimieren (Methode der Kleinsten-Quadrate)

$$\sum_{i=1}^{n} (y_i - (a + bx_i))^2$$

Die Lösung dieses Optimierungsproblems ergibt

$$\hat{b} = \frac{\sum_{i=1}^{n} (y_i - \overline{y})(x_i - \overline{x})}{\sum_{i=1}^{n} (x_i - \overline{x})^2}$$

$$\hat{a} = \overline{y} - \hat{b}\,\overline{x}$$

wobei $\overline{x}$ und $\overline{y}$ die entsprechenden Durchschnitte sind. $\hat{a}$ und $\hat{b}$ sind die Schätzer

von den Parametern a und b, also die Werte, für welche $\sum_{i=1}^{n}(y_i - (a + bx_i))^2$ am kleinsten wird.

Bemerkungen:

i. Wie man auf die Berechnung von a und b kommt, leiten wir hier nicht her. Nur soviel zur Idee: da

$$\sum_i r_i^2 = \sum_{i=1}^{n}(y_i - (a + bx_i))^2$$

minimal werden muss, muss auch die Ableitung von $\sum_i r_i^2$ nach a *und* nach b gleich 0 sein. Wir erhalten also ein Gleichungssystem bestehend aus zwei Gleichungen und zwei Unbekannten:

$$\frac{\partial}{\partial a}\sum_{i=1}^{n}(y_i - (a + bx_i))^2 = \sum_{i=1}^{n} 2\,(y_i - a - bx_i) \stackrel{!}{=} 0$$
$$\frac{\partial}{\partial b}\sum_{i=1}^{n}(y_i - (a + bx_i))^2 = \sum_{i=1}^{n} 2\,(y_i - a - bx_i)\cdot x_i \stackrel{!}{=} 0$$

Die algebraischen Umformungen, die zu den Schätzern von a und b führen, sind dann etwas mühsam.

ii. Die Berechnungen von $\hat{a}$ und $\hat{b}$ werden wir immer mit R berechnen.

Beispiel 2.2.4

Es ist zu vermuten, dass es einen Zusammenhang zwischen der Körpergrösse der Väter und der Körpergrösse der Söhne gibt. Der britische Statistiker Karl Pearson trug dazu um 1900 die Körpergrösse von 10 (in Wahrheit waren 1078) zufällig ausgewählten Männern gegen die Grösse ihrer Väter auf. Dabei erhielt er die Daten von Tabelle 2.4.

Grösse des Vaters	152	157	163	165	168	170	173	178	183	188
Grösse des Sohnes	162	166	168	166	170	170	171	173	178	178

Tabelle 2.4.: Grössenvergleich von Vätern und Söhnen

Es *scheint* hier tatsächlich einen Zusammenhang zu geben: je grösser der Vater, desto grösser der Sohn. Wenn wir noch das Streudiagramm aufzeichnen (siehe Abbildung 2.10), sehen wir, dass ein (möglicher) linearer Zusammenhang besteht: Die Punktwolke „folgt“ der Geraden $y = 0.445x + 94.7$, wobei wir die Parameter mit der Methode der Kleinsten Quadrate aus den Daten berechnet haben.

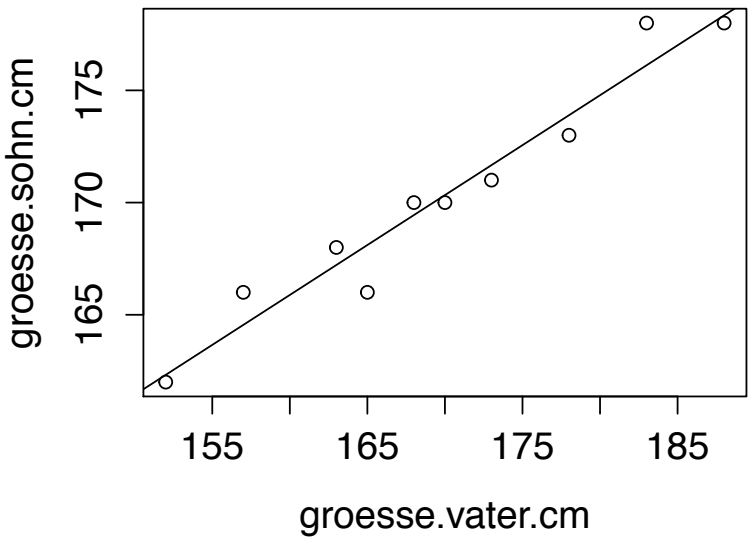


Abbildung 2.10.: Streudiagramm Körpergrössen Väter-Söhne

Wir können also für die in der Tabelle 2.4 nicht vorkommende Grösse von 180 cm des Vaters den zu erwartenden Wert für die Grösse seines Sohnes berechnen.

$$y = 0.445 \cdot 180 + 94.7 \approx 175\,\text{cm}$$

Wir müssen bei dieser Formel allerdings aufpassen, dass wir sie nicht dort anwenden, wo wir sie gar nicht dürfen. So erhalten wir für $x = 0$ einen Wert von 94.7. Was bedeutet dies aber? Wenn der Vater 0 cm gross ist, so wäre der Sohn gemäss dieses Modells ungefähr 95 cm gross, und das macht keinen Sinn.

□

Beispiel 2.2.5

Die folgende Tabelle stellt einen Zusammenhang zwischen den Zahlen der Verkehrstoten her, die es 1988 und 1989 in zwölf Bezirken in den USA geben hat.

Aus der Tabelle ist kein offensichtlicher Zusammenhang ersichtlich. Betrachten wir das Streudiagramm in Abbildung 2.11, so sehen wir, dass kein Zusammenhang besteht. Dies war aber auch zu erwarten, wenn wir vernünftigerweise davon ausgehen

Bezirk	1	2	3	4	5	6	7	8	9	10	11	12
Verkehrstote 1988	121	96	85	113	102	118	90	84	107	112	95	101
Verkehrstote 1989	104	91	101	110	117	108	96	102	114	96	88	106

Tabelle 2.5.: Verkehrstoten in aufeinanderfolgenden Jahren

können, dass es zwischen den Verkehrstoten der einzelnen Bezirke keinen Zusammenhang gibt. In Abbildung 2.11 ist die Regressionsgerade eingezeichnet. Diese können wir zwar berechnen und einzeichnen. Allerdings macht diese hier keinen Sinn, da es keinen linearen Zusammenhang zwischen den Messgrössen gibt.

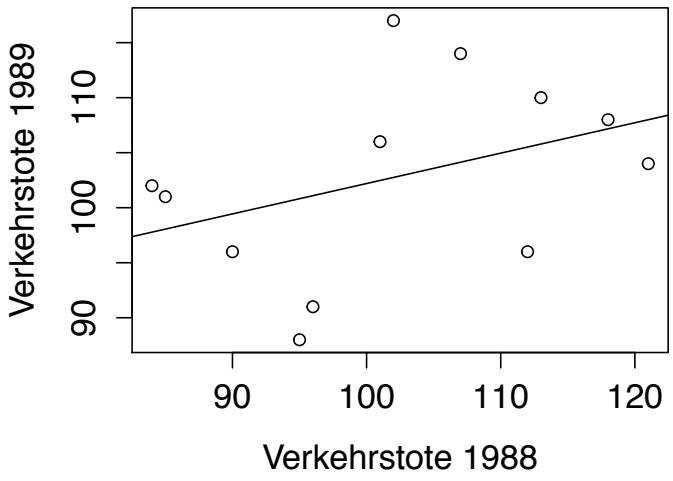


Abbildung 2.11.: Verkehrstote

Wir werden im nächsten Abschnitt eine Grösse kennenlernen, mit der wir eine Aussage treffen können, wie stark der lineare Zusammenhang zwischen Messgrössen ist.

□

Beispiel 2.2.6

Als weiteres Beispiel betrachten wir wieder die Erhebung, die den Zusammenhang zwischen Weinkonsum und der Sterblichkeit untersucht. Legen wir den Daten ein

lineares Modell zu Grunde

$$y = a + bx$$

wobei x den jährlichen Weinkonsum pro Person und y die Mortalität pro 1000 Personen bezeichnet. Dann können wir aufgrund der Datenpunkte die Parameter a und b mit Hilfe der Methode der Kleinsten Quadrate schätzen und erhalten die Regressionsgerade

$$y = 7.68655 - 0.07608x$$

Betrachten wir allerdings das Streudiagramm mit der Regressionsgerade (siehe Abbildung 2.12), so stellen wir fest, dass der Zusammenhang zwischen den Messgrössen nicht linear ist. Das Streudiagramm deutet eher auf eine Hyperbelstruktur hin.

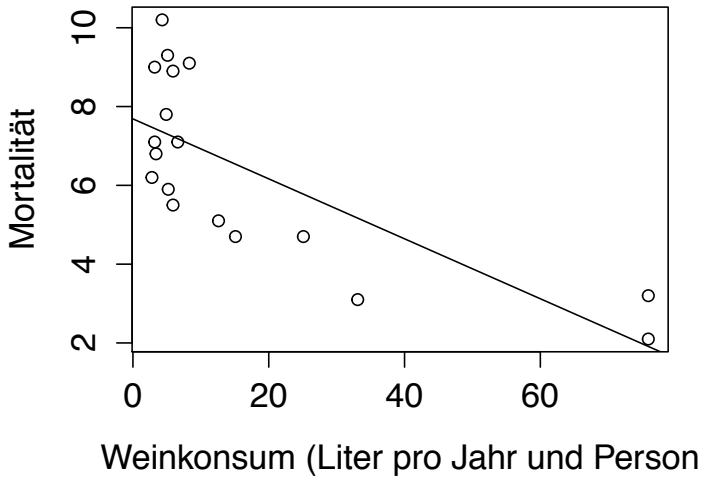


Abbildung 2.12.: Regressionsgerade Weinkonsum-Sterblichkeit

Die Regressionsgerade sagt hier also wenig über den wahren Zusammenhang aus.

□

Die Regressionsgerade können wir (fast) immer bestimmen. In den letzten beiden Beispielen haben wir aber gesehen, dass die Regressionsgerade sehr wenig über die wirkliche Verteilung der Punkte im Streudiagramm aussagt. Dafür gibt es zwei Gründe

- Die Punkte folgen scheinbar gar keiner Gesetzmässigkeit
- Die Punkte folgen einer nichtlinearen Gesetzmässigkeit

Wie können wir nun aber feststellen, ob ein linearer Zusammenhang der Daten besteht oder nicht? Eine Möglichkeit ist sicher, die Situation graphisch zu betrachten, wie wir das eben gemacht haben. Wir können aber auch einen Wert angeben, der den Zusammenhang numerisch beschreibt.

2.2.3. Empirische Korrelation

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

Empirische Korrelation

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

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. Die empirische Korrelation hat folgende Eigenschaften

1. Ist $r = +1$, dann liegen die Punkte auf einer steigenden Geraden ($y = a + bx$ mit $a \in \mathbb{R}$ und ein $b > 0$) und umgekehrt.
2. Ist $r = -1$, dann liegen die Punkte auf einer fallenden Geraden ($y = a + bx$ mit $a \in \mathbb{R}$ und ein $b < 0$) und umgekehrt.
3. Sind x und y unabhängig (d.h. es besteht kein Zusammenhang), so ist $r = 0$.

   Die Umkehrung gilt im allgemeinen nicht: $r = 0$ heisst *nicht*, dass x und y unabhängig voneinander sind (siehe Abbildung 2.13 auf Seite 43)

Die ersten beiden Eigenschaften lassen sich einfach nachvollziehen: man setzt im Ausdruck für den Korrelationskoeffizienten $y_i = a + bx_i$ und $\overline{y} = a + b\overline{x}$ ein. Dann ergibt sich

$$\begin{aligned} r &= \frac{\sum_{i=1}^{n}(x_i - \overline{x})(bx_i + a - (b\overline{x} + a))}{\sqrt{(\sum_{i=1}^{n}(x_i - \overline{x})^2) \cdot (\sum_{i=1}^{n}(bx_i + a - (b\overline{x} + a))^2)}} \\ &= \frac{b \cdot \sum_{i=1}^{n}(x_i - \overline{x})(x_i - \overline{x})}{|b| \cdot \sqrt{(\sum_{i=1}^{n}(x_i - \overline{x})^2) \cdot (\sum_{i=1}^{n}(x_i - \overline{x})^2)}} \\ &= \text{sign}(b)\,, \end{aligned}$$

wobei $\text{sign}(b)$ das Vorzeichen von b bezeichnet, also $\text{sign}(b) = 1$ falls $b > 0$ und $\text{sign}(b) = -1$ falls $b < 0$.

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 Abbildung 2.13 auf der nächsten Seite.

Beispiel 2.2.7

Für unser Seitenzahl-Preis-Beispiel erhalten wir mit R

```
cor(seitenzahl, buchpreis)

## [1] 0.9681122
```

Der Wert liegt also sehr nahe bei 1 und somit besteht ein enger linearer Zusammenhang. Dazu ist der Wert positiv, was einem „je mehr, desto mehr“, also einem positiven linearen Zusammenhang entspricht.

□

Beispiel 2.2.8

Auch im Beispiel der Körpergrösse von Vater und Sohn erwarten wir einen hohen Korrelationskoeffizienten. Wir erhalten 0.973.

□

Beispiel 2.2.9

Bei den Verkehrsunfällen haben wir keinen Zusammenhang und erwarten einen Korrelationskoeffizienten nahe null. Er beträgt 0.386.

□

Beispiel 2.2.10

Auch beim Weinkonsum erwarten wir einen negativen Korrelationskoeffizienten, da mit steigendem Weinkonsum die Mortalität sinkt und der nahe bei null liegt. Er beträgt -0.746. Ohne die Daten in einem Streudiagramm darzustellen, würde man aufgrund dieses Wertes fälschlicherweise auf einen starken negativen linearen Zusammenhang schliessen.

□

Beispiel 2.2.11

In Abbildung 2.13 sind 21 verschiedene Datensätze dargestellt, die je aus gleich vielen Beobachtungspaaren (x_i, y_i) mit den entsprechenden Punkten im Streudiagramm bestehen. Über jedem Datensatz steht jeweils die zugehörige empirische Korrelation.

Bei perfektem linearen Zusammenhang ist die empirische Korrelation $+1$ oder -1 (je nachdem ob die Steigung positiv oder negativ; siehe zweite Zeile in Abbildung 2.13). Je mehr die Punkte um den linearen Zusammenhang streuen, desto kleiner wird der Betrag der empirischen Korrelation (siehe erste Zeile).

Da die empirische Korrelation nur den *linearen* Zusammenhang misst, kann es einen (nichtlinearen) Zusammenhang zwischen den beiden Variablen x und y geben, auch wenn die empirische Korrelation null ist (siehe unterste Zeile in Abbildung 2.13).

□

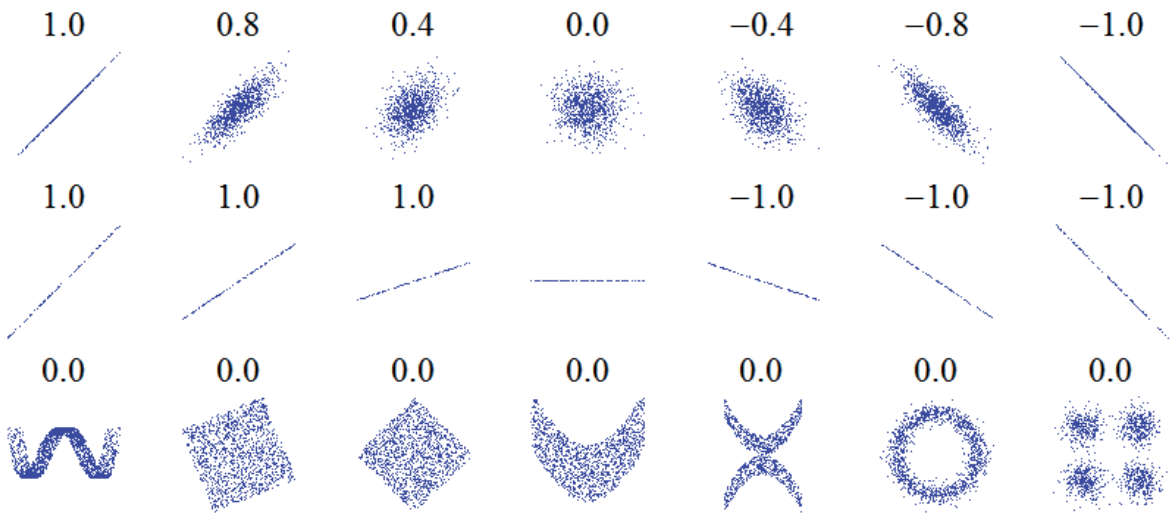


Abbildung 2.13.: 21 verschiedene Datensätze und deren empirische Korrelationskoeffizienten.

Lernziele

- ☐ Sie kennen Methoden der deskriptiven Statistik, können diese interpretieren und folgende Grössen ausrechnen: arithmetisches Mittel, Standardabweichung, Varianz, Quantil, Median 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.