Histogramm
Zweidimensionale Deskriptive Statistik

Peter Büchel

HSLU I

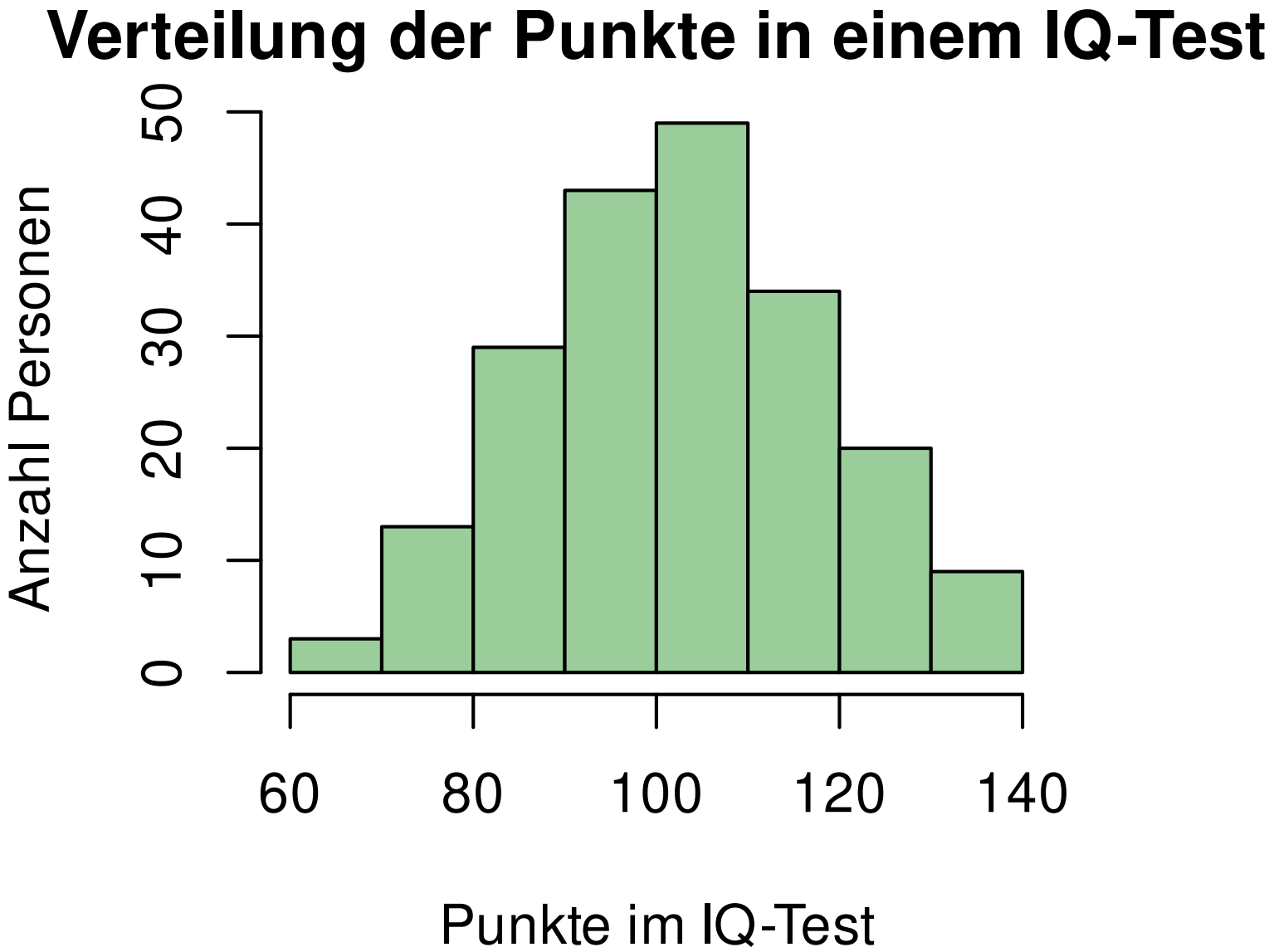
ASTAT: Block 03

Histogramm

- *Histogramm*: Graphischer Überblick über die auftretenden Werte
- Aufteilung des Wertebereichs in k *Klassen* (Intervalle)
- Faustregel:
  - bei weniger als 50 Messungen ist die Klassenzahl 5 bis 7
  - bei mehr als 250 Messungen wählt man 10 bis 20 Klassen
- Zeichne für jede Klasse einen *Balken*, dessen Höhe proportional zur Anzahl Beobachtungen in dieser Klasse ist

Beispiel: IQ-Test

- Abbildung: Histogramm vom Ergebnis eines IQ-Testes von 200 Personen



- Daten wurden hier allerdings simuliert
- Breite der Klassen: 10 IQ-Punkte; für jede Klasse gleich
- Höhe der Balken: Anzahl Personen, die in diese Klasse fallen
- Bsp: ca. 20 Personen fallen in Klasse zwischen 120 und 130 Punkten
- Form dieses Histogrammes typisch für viele Histogramme
  $\rightarrow$ Normalverteilung

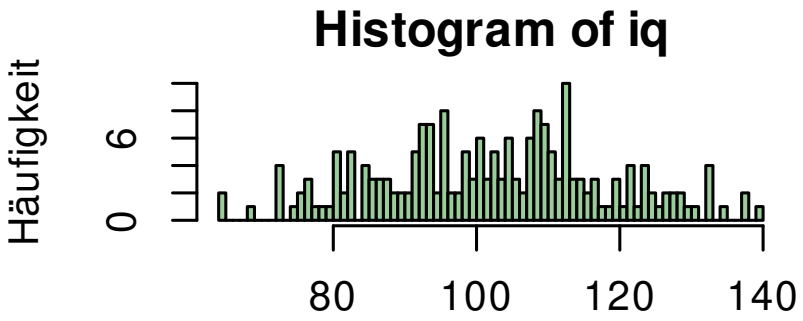
- Code: Der R-Code für das Histogramm oben lautet wie folgt:

```
iq <- rnorm(n = 200, mean = 100, sd = 15)

hist(iq,
     col = "darkseagreen3",
     xlab = "Punkte im IQ-Test",
     ylab = "Anzahl Personen",
     main = "Verteilung der Punkte in einem IQ-Test"
)
```

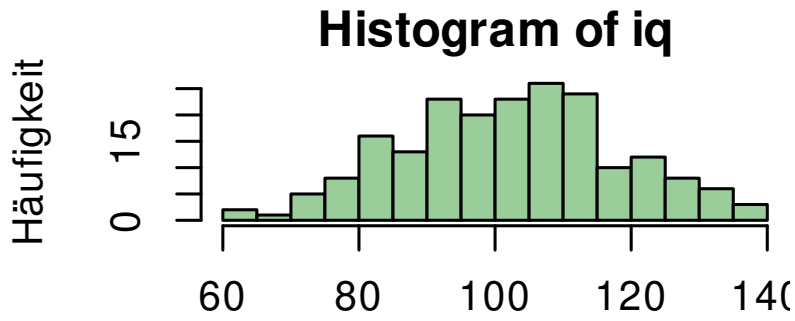
- Befehl `rnorm(n = 200, mean = 100, sd = 15)`: Wählt zufällig 200 normalverteilte Daten (siehe Kapitel Normalverteilung) mit Mittelwert 100 mit Standardabweichung 15 aus
- Befehl `hist(iq, ...)`: Histogramm für die Daten `iq`
- Die weiteren Optionen sollten klar sein:
  - `xlab` steht für x-Label, die Beschriftung der x-Achse
  - `ylab` steht für y-Label, die Beschriftung der y-Achse
  - `col` steht für color
  - `main` steht für Haupttitel

Wahl der Klassen

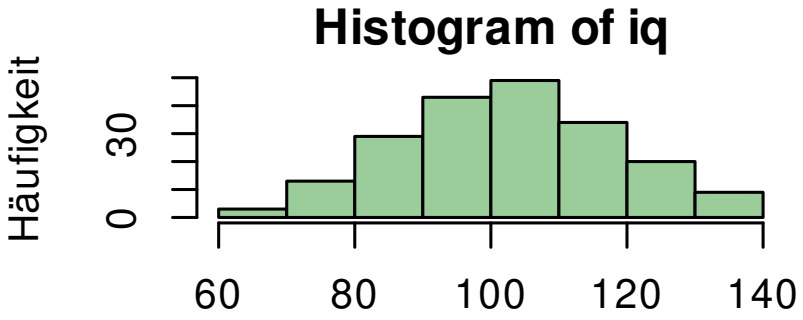
- Wahl der Anzahl Klassen relevant für Aussagekraft eines Histogrammes
- Es gibt keine allgemeine Grundregel, wie man Anzahl Klassen wählt
- Abbildung: IQ Daten von Beispiel mit verschiedener Anzahl Klassen



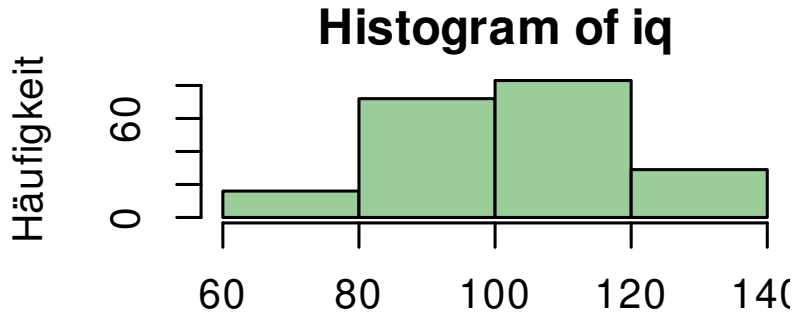
100 Klassen



20 Klassen



Sturges-Regel von R



4 Klassen

- Histogramm links oben: Viel zu detailiert, als dass man ein Muster erkennen könnte
- Histogramm rechts unten zu ungenau
- R: Anzahl Klassen nach der sogenannten *Sturges*-Regel

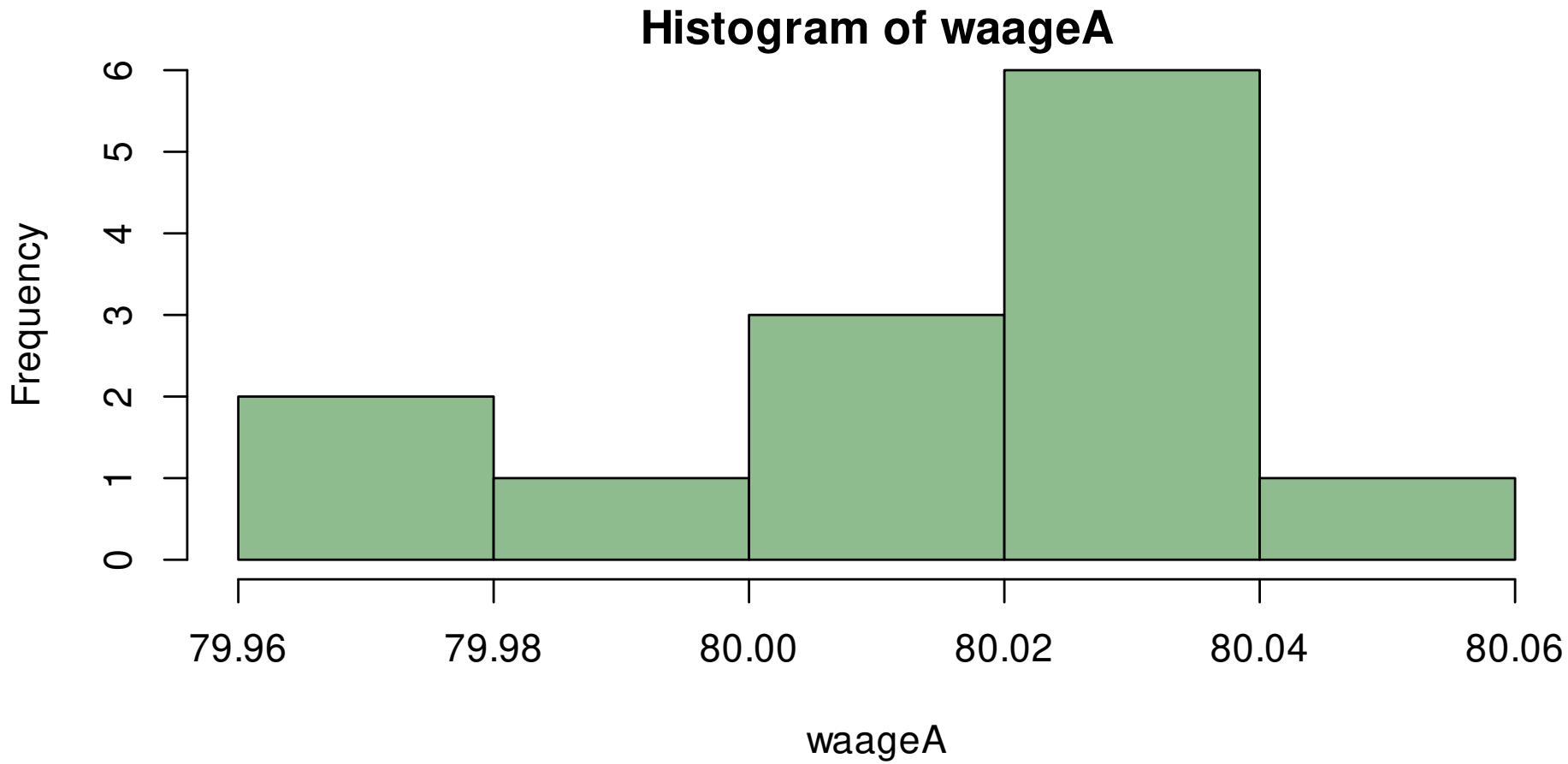
- Code:

```
par(mfrow = c(2, 2))

hist(iq,
     breaks = 100,
     xlab = "100 Klassen",
     ylab = "Häufigkeit",
     col = "darkseagreen3"
)
hist(iq,
     breaks = 20,
     xlab = "20 Klassen",
     ylab = "Häufigkeit",
     col = "darkseagreen3"
)
hist(iq,
     breaks = "sturges", # default R
     xlab = "Sturges-Regel von R",
     ylab = "Häufigkeit",
     col = "darkseagreen3"
)
hist(iq,
     breaks = 3,
     xlab = "4 Klassen",
     ylab = "Häufigkeit",
     col = "darkseagreen3"
)
```

- Befehl `par(mfrow = c(2, 2))`: Die vier Histogramme in 2 Zeilen (erste 2) und 2 Spalten gezeichnet (zweite 2)
- Option `breaks = ...`: Anzahl Klassen festlegen
- Beachte letzte Graphik: `breaks = 3`, aber vier Klassen gezeichnet
- `R` nimmt die Option `breaks = ...` nur als Vorschlag

R: hist()

```
waageA <- c(79.98, 80.04, 80.02, 80.04, 80.03, 80.03, 80.04,
            79.97, 80.05, 80.03, 80.02, 80.00, 80.02)

hist(waageA, col="darkseagreen")
```



Bemerkungen

- Waage *A* 13 Messungen $\rightarrow$ 5 Balken
- R: Interne Regel (Sturges Regel)
- Bedeutung der Anzahlen (Frequency):
  - In 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.
- Linke Grenze wird also *nicht* berücksichtigt, die rechte schon
- Umgekehrt auch möglich $\rightarrow$ Histogramm würde etwas anders aussehen
- Bei grossen Datensätzen spielt das kaum eine Rolle
- Mit Optionen lassen sich auch die Anzahl Klassen festlegen, Überschriften ändern, usw. (siehe Übungen)

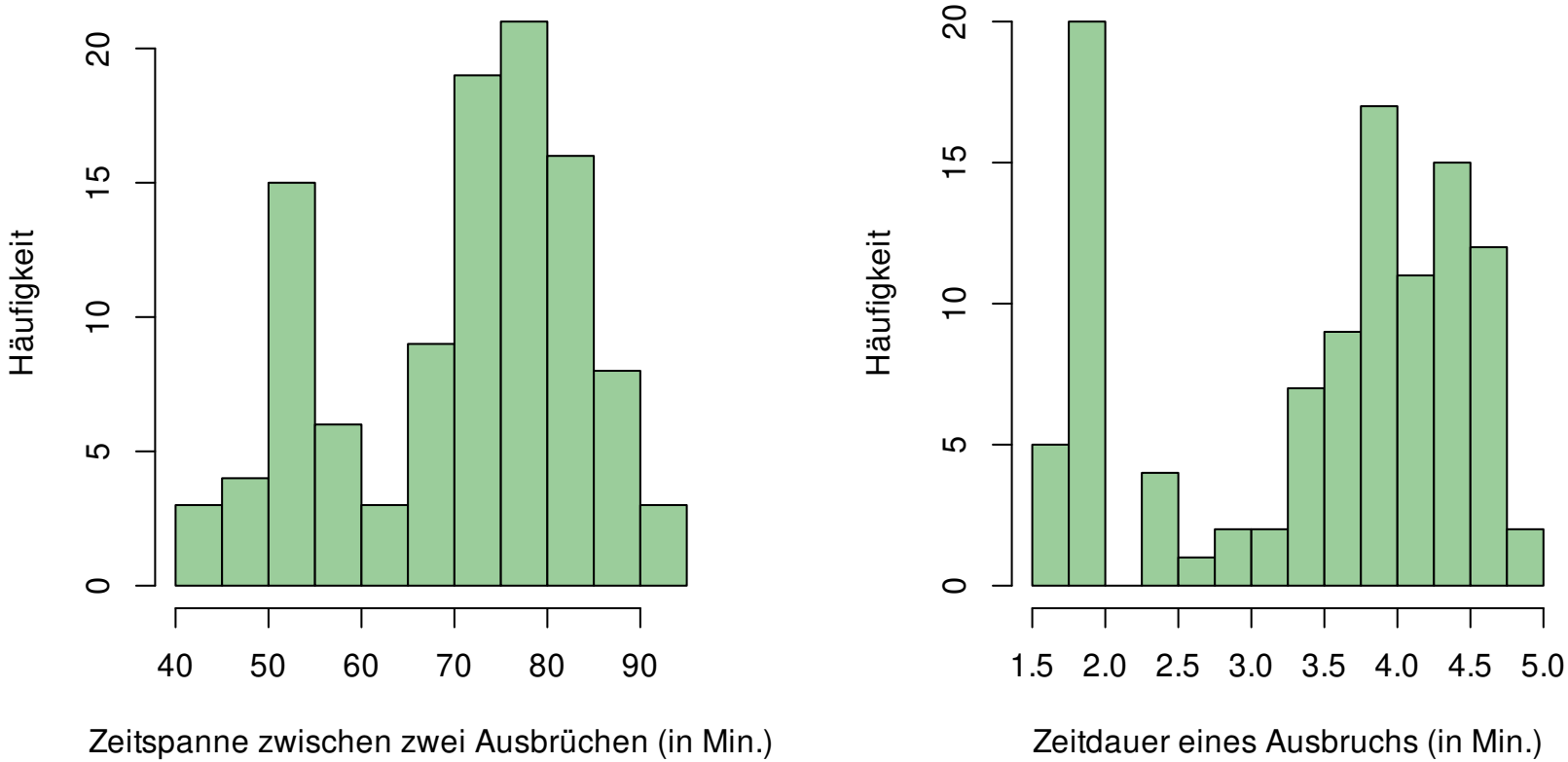
Old Faithful Geysir (Yellowstone NP)

- Geysir Old Faithful (Yellowstone National Park): Bekannte heisse Quelle
- Für Zuschauer und Nationalparkdienst ist die Zeitspanne zwischen zwei Ausbrüchen und die Eruptionsdauer von grossem Interesse
- Von 1.8.1978 - 8.8.1978 insgesamt 107 Messungen von aufeinanderfolgenden Ausbrüchen gemacht
- Daten Datei `geysir.txt`:

```
geysir <- read.table("../Data/geysir.txt")
head(geysir)
```

```
##   X.Tag. Zeitspanne Eruptionsdauer
## 1      1         78            4.4
## 2      1         74            3.9
## 3      1         68            4.0
## 4      1         76            4.0
## 5      1         80            3.5
## 6      1         84            4.1
```

- Abbildung Histogramme:



- Ausbruchsdauer eines Ausbruchs (rechts)
- Zeitspanne zwischen zwei Ausbrüchen (links)

- Bei beiden Histogrammen: *Bimodales* Verhalten sichtbar
- Es gibt zwei „Hügel" im Histogramm:
  - Zeitspanne zwischen zwei Ausbrüchen: Dauer relativ kurz (um die 50 Minuten) oder eher lang (um die 80 Minuten)
  - Zeitdauer zwischen zwei Ausbrüchen nicht „gleichmässig" verteilt
  - Dasselbe Verhalten bei der Zeitdauer eines Ausbruchs: Entweder ist der Ausbruch relativ kurz (um die 1.5-2 Minuten) oder lang (um die 4-4.5 Minuten)

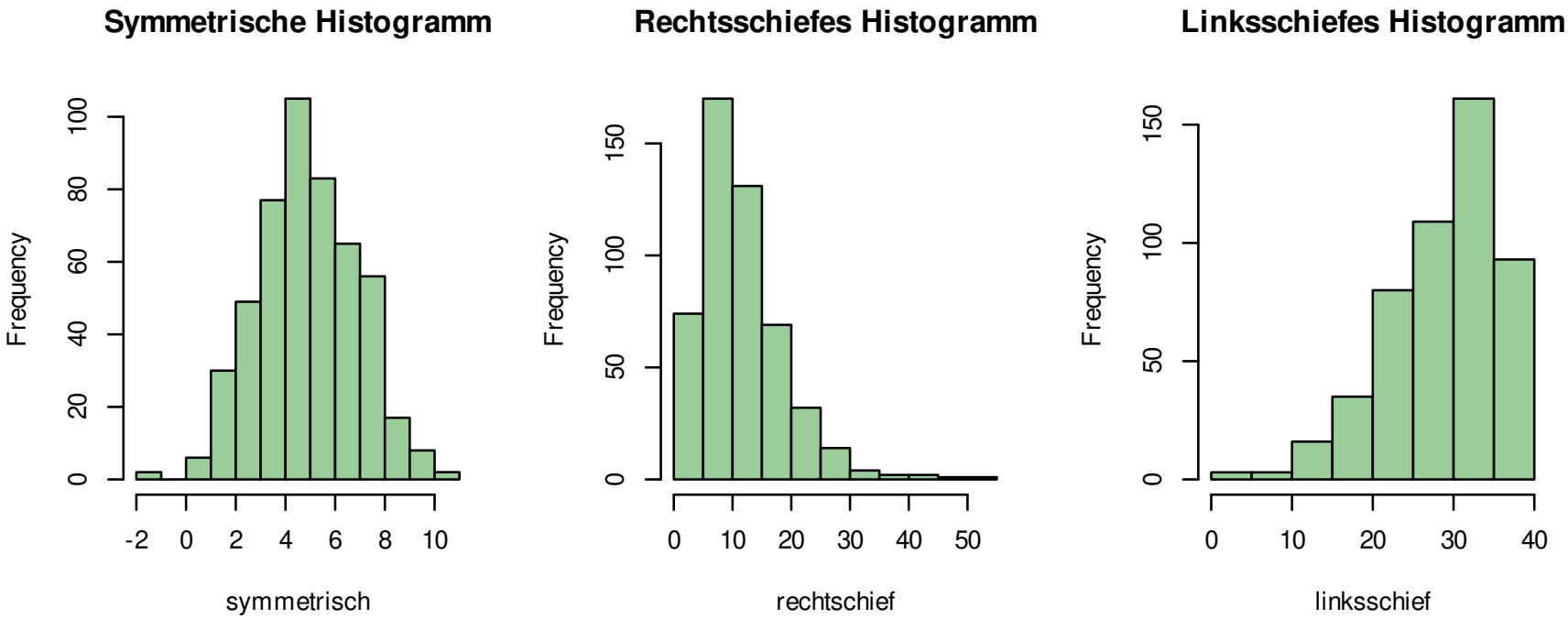
- Frage: Gibt es einen Zusammenhang zwischen Eruptionsdauer und Zeitspanne zwischen zwei Ausbrüchen gibt?

- Oder anders gefragt:
  - Geht es nach einem langen Ausbruch länger bis es wieder einen Ausbruch gibt?
  - Oder kommt ein Ausbruch schon sehr schnell wieder?
  - Oder gibt es gar keinen Zusammenhang?

- Fragen werden im 2. Teil dieses Blockes beantwortet

Schiefe von Histogrammen

- Abbildung:



- Histogramm links ist symmetrisch bezüglich ungefähr 5. Die Daten sind um 5 auf beiden Seiten ähnlich verteilt

- Mittleres Histogramm: Die meisten Daten links im Histogramm
  $\rightarrow$ *rechtsschiefes* Histogramm

- Rechtes Histogramm: Die meisten Daten rechts im Histogramm
  linksschiefen Histogramm

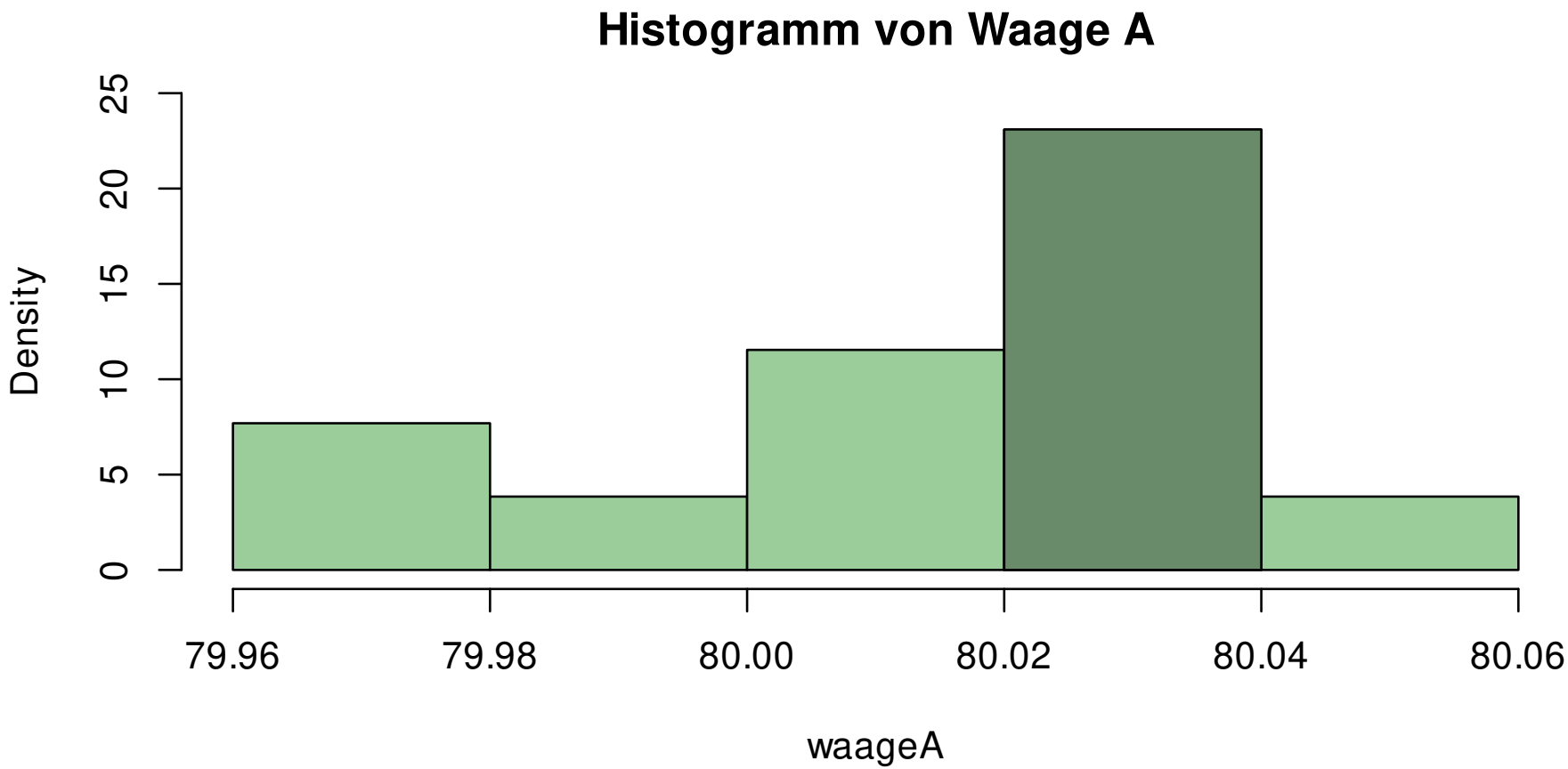
- Bezeichnung „rechts" und „links": Bezieht sich immer auf die Richtung, wo es *weniger* Daten hat

Normiertes Histogramm

- In Histogrammen bisher: Höhe der Balken entspricht Anzahl der Beobachtungen in einer Klasse
- Oft besser und übersichtlicher: Balkenhöhe so wählen, dass die *Balkenfläche* dem prozentualen Anteil der jeweiligen Beobachtungen an der Gesamtanzahl Beobachtungen entspricht
- Gesamtfläche aller Balken muss dann gleich eins sein
- Auf der vertikalen Achse wird die *Dichte* angegeben

Beispiel Waage *A*

- Normiertes Histogramm:



- Dichte der Klasse von $80.02 - 80.04$ ist etwa 23
- Fläche dieses Balkens (dunkelgrüne Fläche in Abbildung):

$$(80.04 - 80.02) \cdot 23 = 0.46$$

- Fläche mit 100 multipliziert: Prozentzahl der Daten, die in diesem Balken liegen
- Also etwa 46 % der Daten befinden sich zwischen 80.02 und 80.04

R-Code

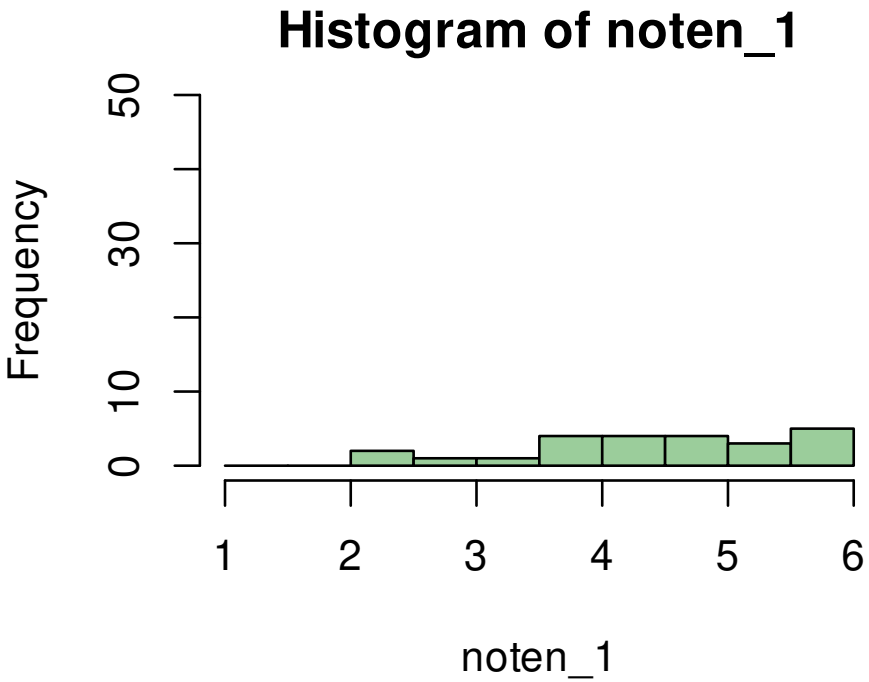
- Code:

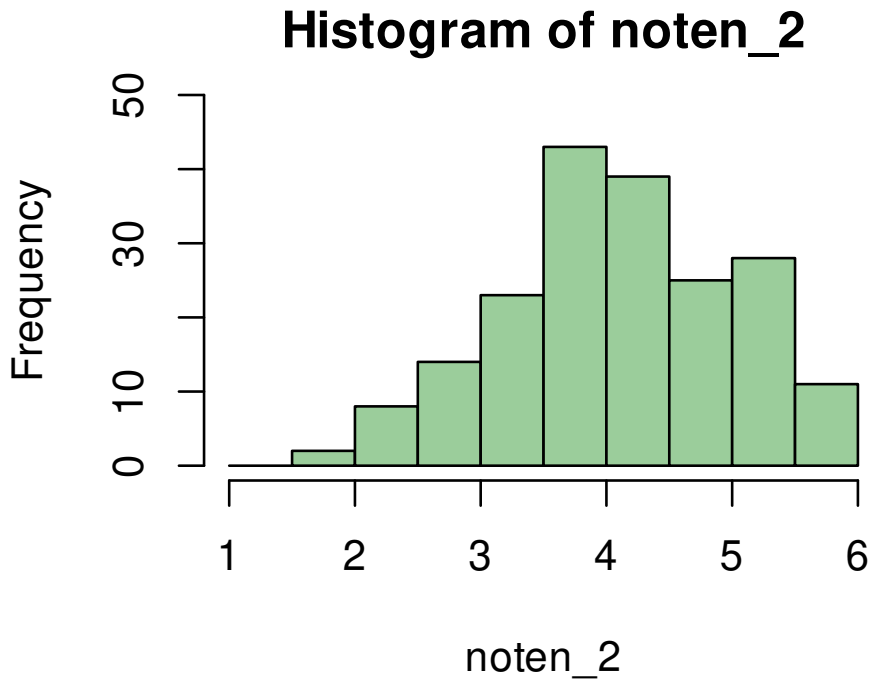
```
hist(waageA,
     freq = F,
     main = "Histogramm von Waage A",
     col = "darkseagreen3",
     ylim = c(0, 25)

)
rect(80.02, 0, 80.04, 23.1, col="darkseagreen4")
```

- Option `freq = F` (*frequency false*): Histogramm wird normiert gezeichnet
- Option `ylim = c(0, 25)`: siehe Skript
- `rect(80.02, 0, 80.04, 23.1, col = "darkseagreen4")`: siehe Skript

Beispiel

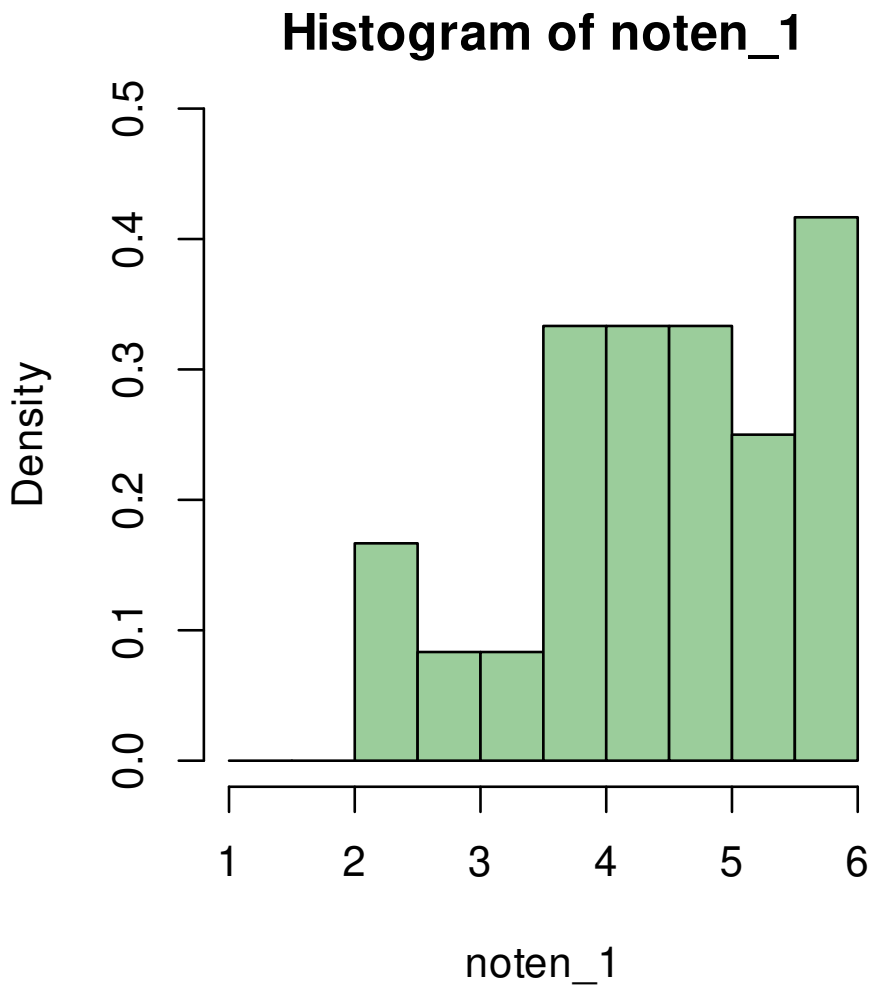
- Betrachten die Schulnoten einer Klasse von 24 Lernenden früher
- Vergleichen die Klasse mit einer (hypothetischen) anderen Klasse mit 194 Lernenden, die dieselbe Prüfung machten
- Histogramme mit der jeweiligen Häufigkeit

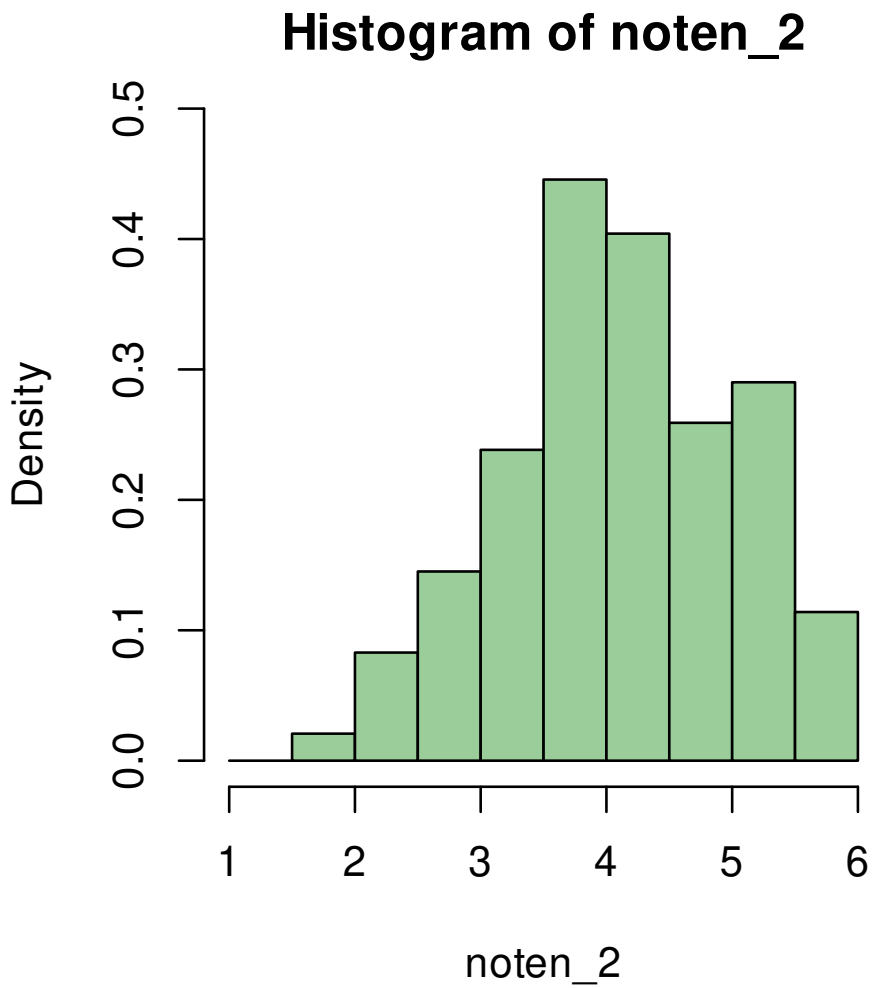




- Nicht vergleichbar

- Normierte Histogramme:

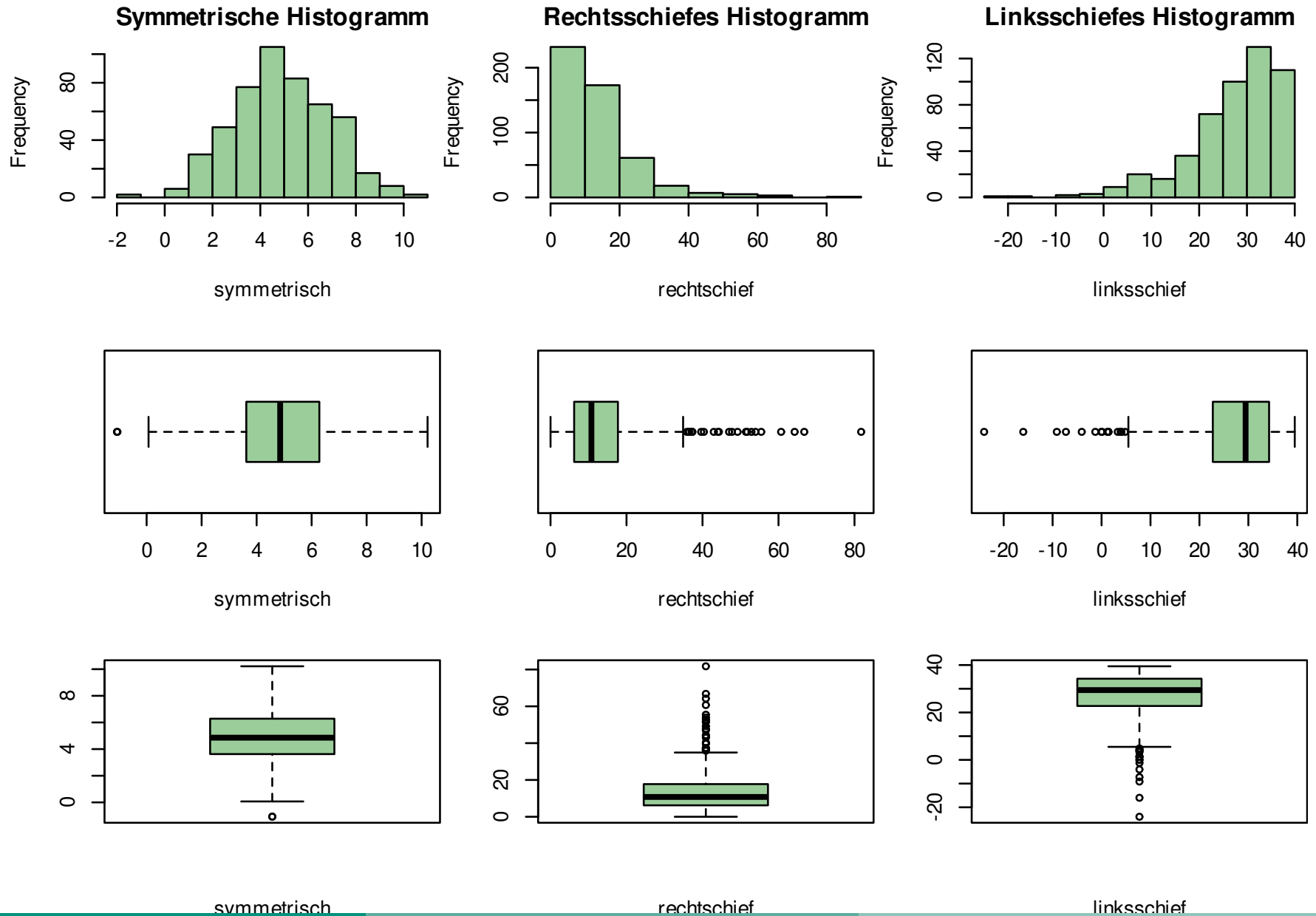




- Klasse 1 hat im oberen Bereich mehr Anteile als Klasse 2
- Vor allem der Balken von 5.5-6 von Klasse 1 ist sehr viel höher als der von Klasse 2
- Klasse 1 prozentual mehr starke Lernende als in Klasse 2
- Klasse 1 hat eher mehr schwächere Lernende als Klasse 2
- Im mittleren Bereich der Noten hat Klasse 2 prozentual mehr Lernende als Klasse 1

Schiefe im Boxplot

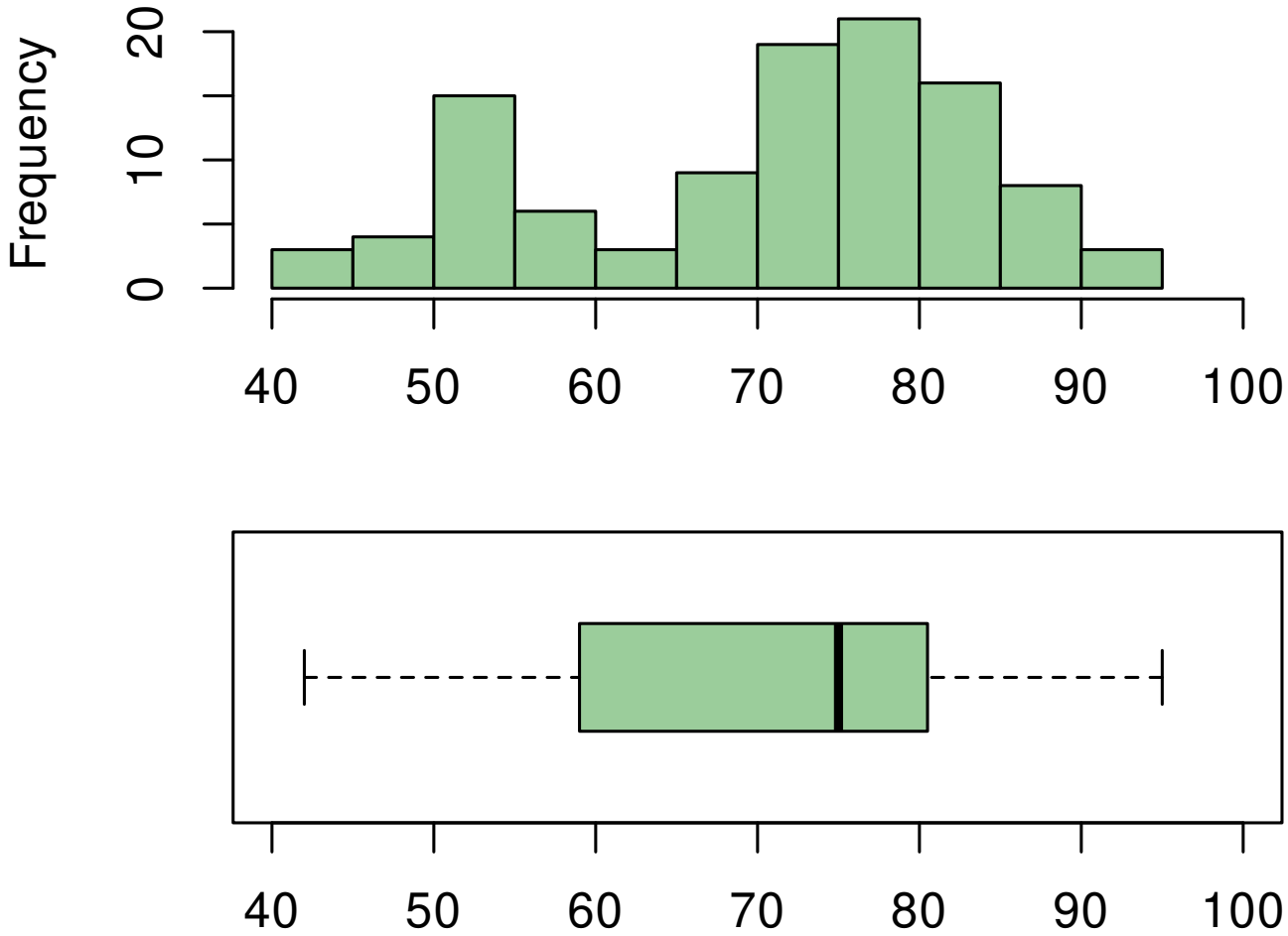
- Abbildung:



- Symmetrisches Diagramm links: Median in der Mitte der Box
- Rechtsschiefes Histogramm (Mitte): Median nicht mehr in der Mitte der Box, sondern nach links verschoben
- Der Abstand vom unteren Quartil zum Median ist kleiner als der Abstand vom Median zum oberen Quartil
- Vom unteren Quartil zum Median viele Daten in kleinem Bereich liegen
- Vom Median zum oberen Quartil braucht es ein viel grösserer Bereich bis 25 % der Daten in diesem Intervall liegen
- Beim linksschiefen Histogramm ist die Sachlage gerade umgekehrt

Beispiel: Old Faithful

- Abbildung: Histogramm und der Boxplot von der Zeitspanne zwischen zwei Ausbrüchen von Old Faithful:



- Daten sind linksschief
- Boxplot: 50 % der mittleren Zeitspannen zwischen 60 und 80 Minuten
- Median liegt bei etwa 75 Minuten
- Daten zwischen dem Median und dem oberen Quartil liegen in einem Bereich von 5 Minuten (von 75-80 Minuten)
- D.h.: In diesem Bereich befinden sich relativ viele Zeitspannen verglichen zu Abstand 15 Minuten vom unteren Quartil zum Median

Boxplot: Bemerkungen

- Im *Boxplot* sind ersichtlich:
  - Lage
  - Streuung
  - Schiefe
- Man sieht aber z.B. *nicht*, ob eine Verteilung mehrere „Peaks" hat

Deskriptive Statistik zweidimensionaler Daten

- Zweidimensionale Daten: An einem Versuchsobjekt werden jeweils *zwei* verschiedene Grössen gemessen
- Beispiel: An einer Gruppe von Menschen wird jeweils die Körpergrösse *und* das Körpergewicht gemessen
- Versuchsobjekt: Menschen, zudem je zwei Messungen gehören:
  - die Körpergrösse
  - das Körpergewicht
- Old Faithful: Versuchsobjekt ist ein Ausbruch, zudem je zwei Messungen gehören:
  - die Eruptionsdauer
  - die Zeit bis zum nächsten Ausbruch

Daten: Weinkonsum - Mortalität

- Datensatz: Untersucht durchschnittlicher Weinkonsum (in Liter pro Person und Jahr) und die Sterblichkeit (Mortalität;Anzahl Todesfälle pro 1000 Perso- nen zwischen 55 und 64 Jahren pro Jahr) aufgrund von Herz- und Kreislauferkrankungen in 18 Ländern
- Tabelle:

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

- Frage: Gibt es einen Zusammenhang zwischen der Sterblichkeitsrate aufgrund von Herzkreislauferkrankung und Weinkonsum?

Graphische Darstellung: Streudiagramm

- Wichtiger Schritt in der Untersuchung zweidimensionaler Daten: Graphische Darstellung
- Meist über ein sogenanntes *Streudiagramm* (engl.: *Scatterplot*)
- Zwei Messungen als Koordinaten von Punkten in einem Koordinatensystem interpretiert und dargestellt

Beispiel: Weinkonsum

- Grösse „Weinkonsum”:

$$x_1,\ x_2,\ \ldots,\ x_{18}$$

- Zugehörige Grösse „Mortalität”:

$$y_1,\ y_2,\ \ldots, y_{18}$$

- Koordinaten der Punkte:

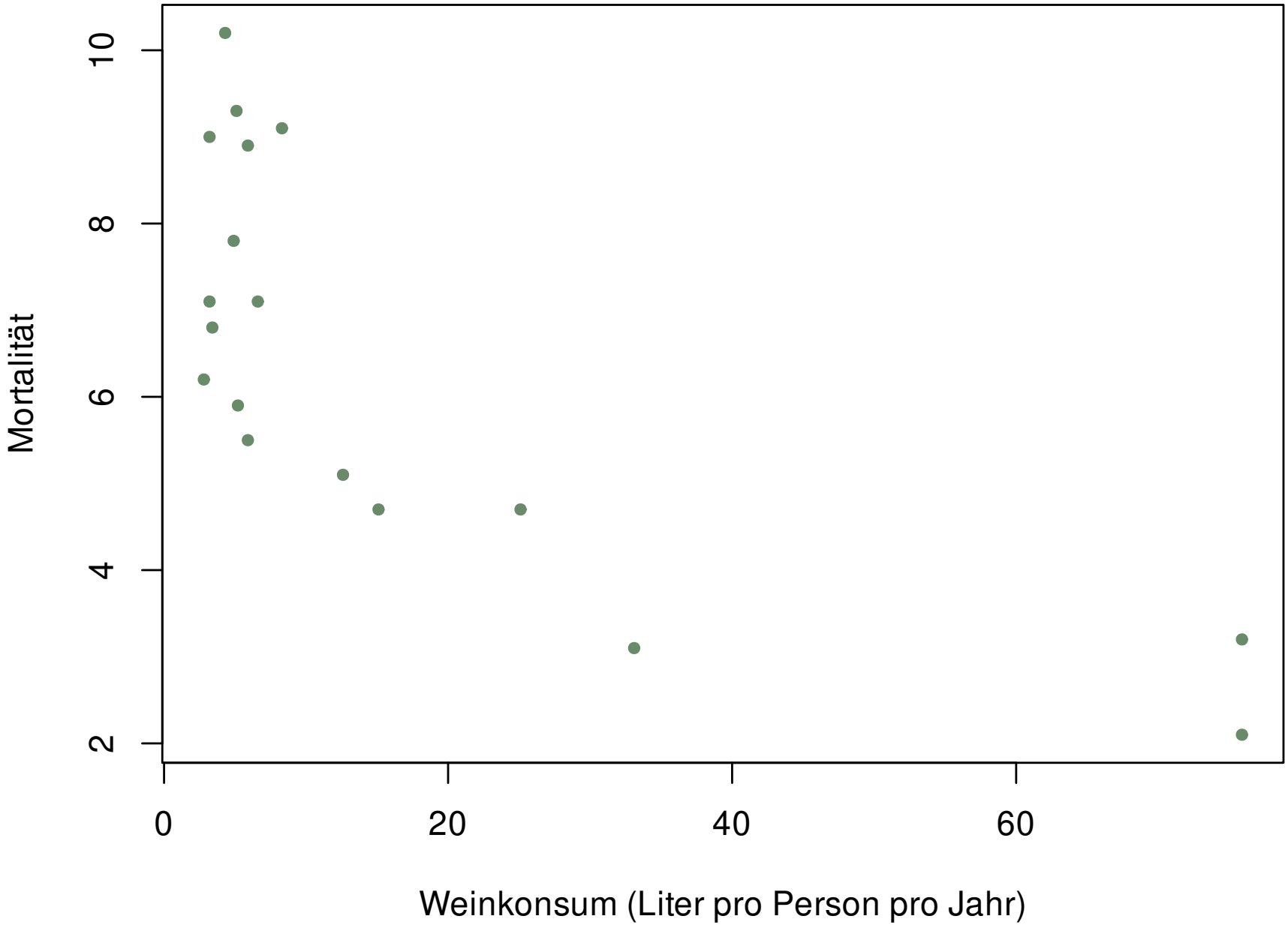
$$(x_1, y_1),\ (x_2, y_2),\ \ldots,\ (x_{18}, y_{18})$$

- Punkt mit den Koordinaten von Norwegen

$$(x_1, y_1) = (2.8,\ 6.2)$$

- Punkte in Koordinatensystem einzeichnen

Zweidimensionales Streudiagramm

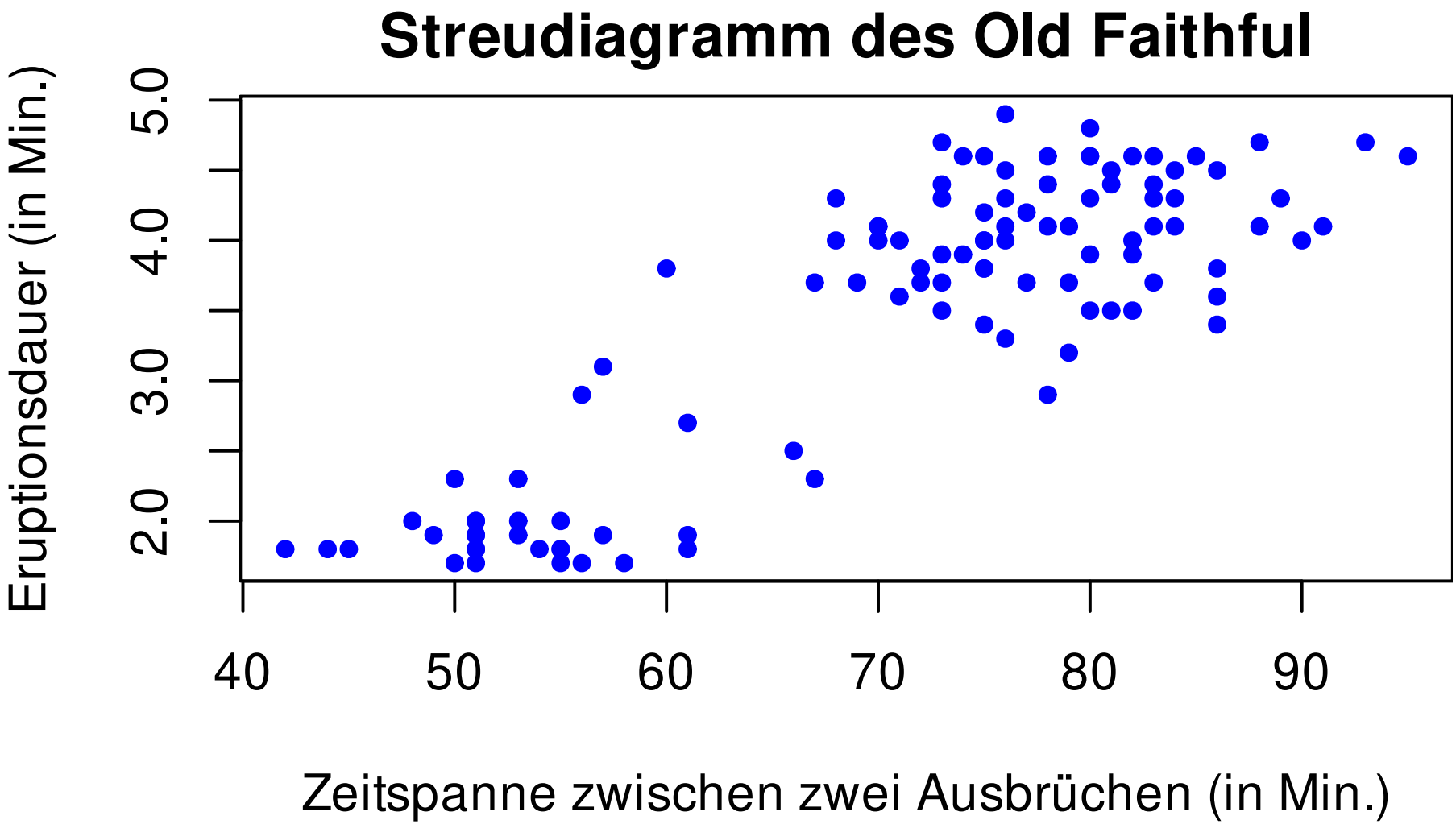


Streudiagramm mit R

- Plot deutet an, dass hoher Weinkonsum weniger Sterblichkeit wegen Herz-Kreislauferkrankungen zur Folge hat
- Kann Zufall sein (keine Kausalität)
- Heisst *nicht*, dass Weinkonsum gesund ist (Leber!)
- R-Befehl

```
wein <- c(2.8, 3.2, 3.2, 3.4, 4.3, 4.9, 5.1, 5.2, 5.9, 5.9,
          6.6, 8.3, 12.6, 15.1, 25.1, 33.1, 75.9, 75.9)

mort <- c(6.2, 9.0, 7.1, 6.8, 10.2, 7.8, 9.3, 5.9, 8.9, 5.5,
          7.1, 9.1, 5.1, 4.7, 4.7, 3.1, 3.2, 2.1)

plot(wein, mort,
     xlab = "Weinkonsum (Liter pro Jahr)",
     ylab = "Mortalität",
     col = "blue",
     pch = 20

)
```

Beispiel Old Faithful



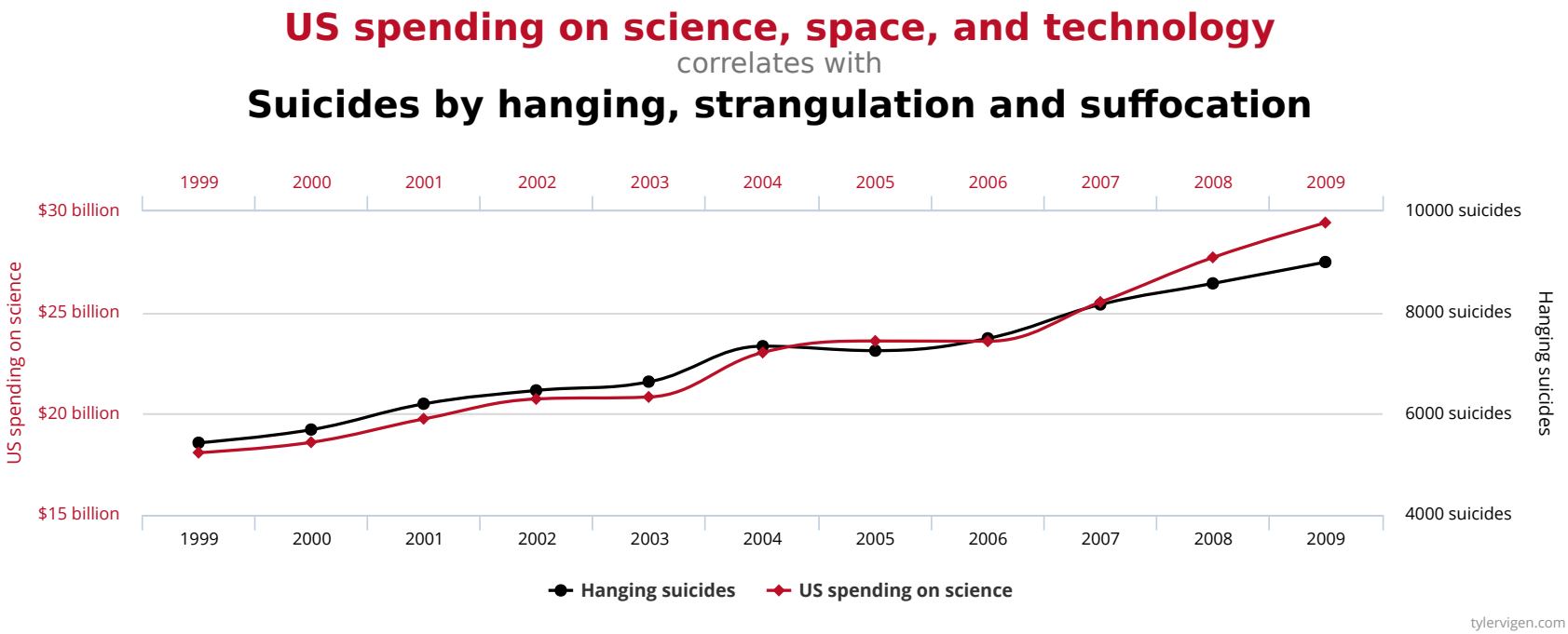
- Zunächst ist die Punktwolke steigend:
  - Je länger die Zeitspanne zwischen den Ausbrüchen, umso länger dauert der Ausbruch
- Im Streudiagramm hat es zwei Gruppen:
  - Eine links unten und eine rechts oben
  - Zeitspanne zwischen zwei Ausbrüchen kurz und die nächste Eruptionsdauer kurz
  - Zeitspanne ist lang und der Eruptionsdauer ist lang
  - Eine mittlere Zeitspanne (um die 70 Minuten) mit einer mittleren Ausbruchsdauer (um die 3 Minuten) gibt es nicht

Abhängigkeit und Kausalität

- Bei Streudiagrammen aufpassen: Abhängigkeit nicht mit Kausalität verwechseln
- Gesetzmässigkeit vorhanden, heisst dies noch lange nicht, dass diese Gesetzmässigkeit auch kausal erklärt werden kann
- Beispiel Old Faithful: „Je länger desto länger"-Sachverhalt festgestellt
- Das dieser aber *erklärt* werden kann, reicht das Streudiagramm nicht
- Dazu müssen andere Methoden verwendet werden

Beispiel

- Abbildung:



- Kurven haben gleiche Form
- Es gibt aber keinen kausalen Zusammenhang zwischen den Kurven
- Mehr Beispiele unter

  `https://www.tylervigen.com/spurious-correlations`

Beispiel

- Abbildung:

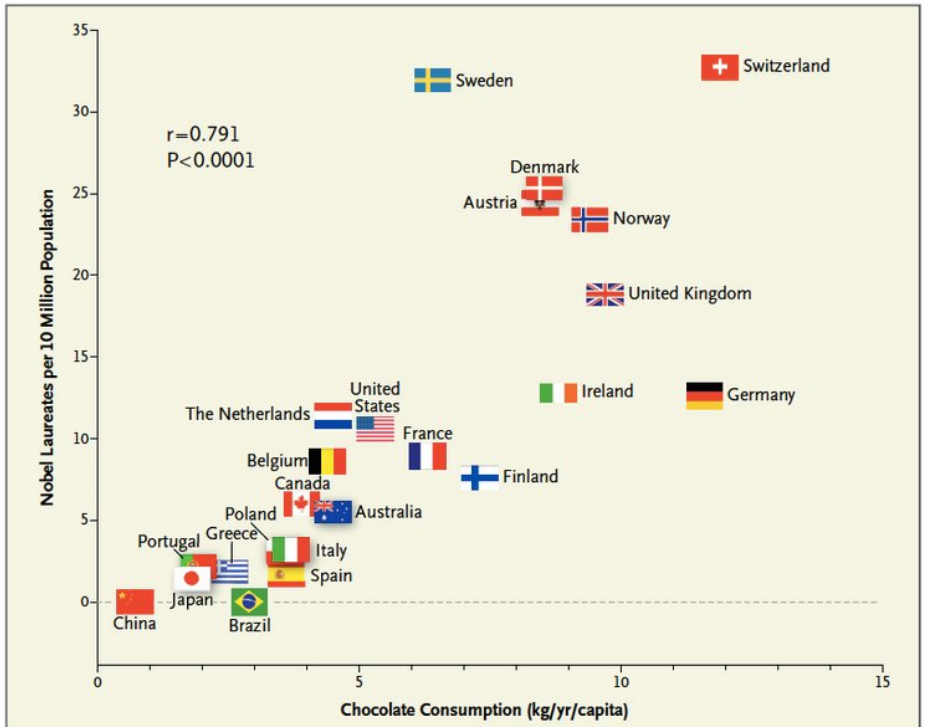


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.

- Auch hier keine kausale Gesetzmässigkeit vorhanden

(Fiktives) Beispiel für Lineare Regression

- Kunde kauft in Buchhandlung 10 Bücher

	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

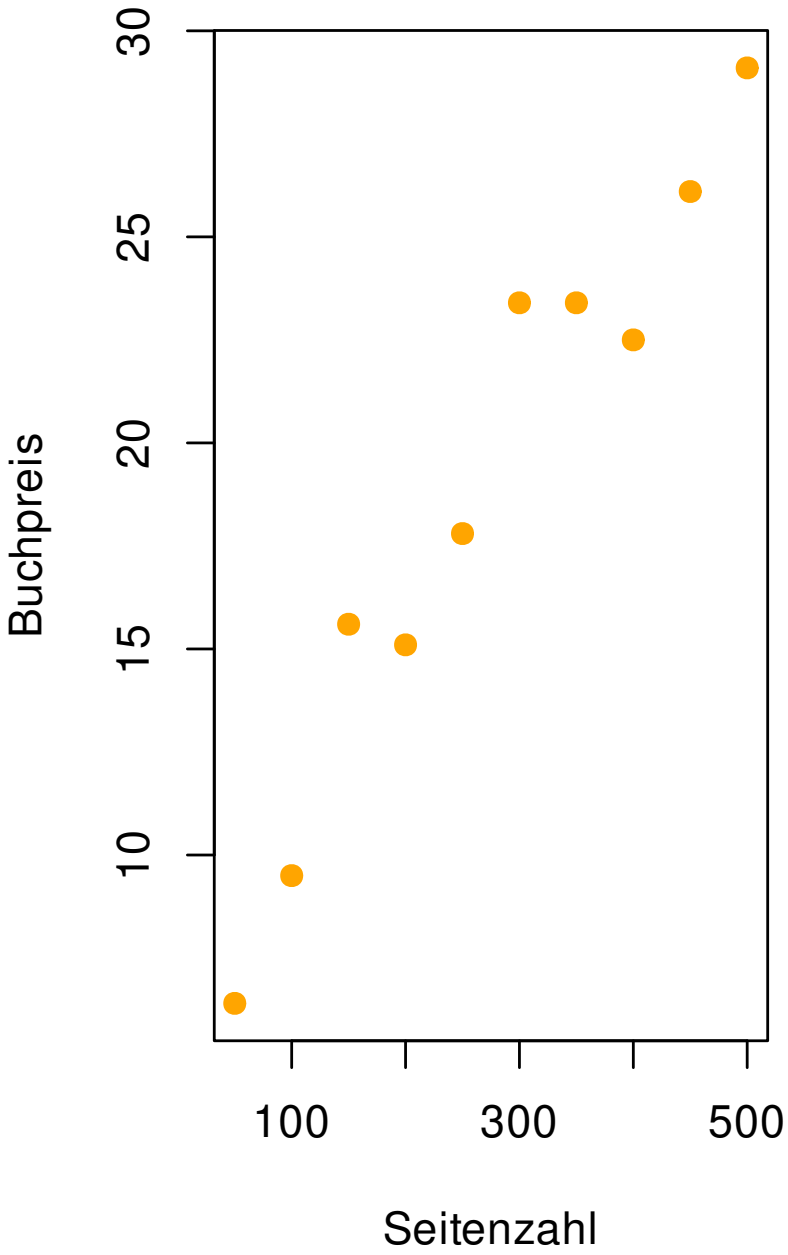
- *Beobachtung*: Je dicker ein Roman ist, desto teurer ist er in der Regel

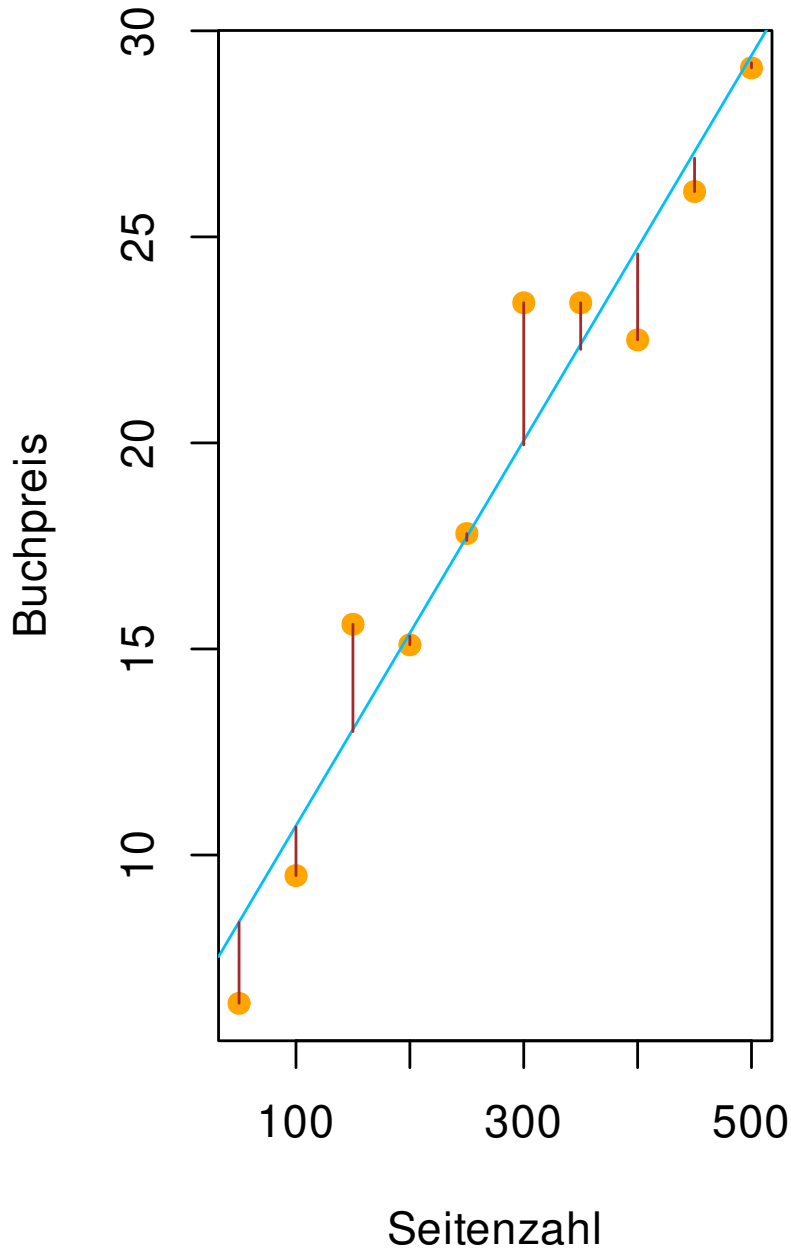
- Fragen:
  - Wieviel kostet eine Seite?
  - Wie teuer ein Buch mit „null" Seiten wäre? $\rightarrow$ Grundkosten für ein Buch
  - Was würde dann voraussichtlich ein Buch mit 375 Seiten kosten? Diese Seitenzahl kommt in der Tabelle nicht vor

- *Ziel:* Formelmässiger Zusammenhang zwischen Buchpreis und Seitenzahl
  - Zusammenhang zwischen Seitenzahl x und Buchpreis y

- Vorhersagen über Buchpreis möglich für Bücher mit Seitenzahlen, die in Liste nicht auftauchen

Streudiagramm und Regressionsgerade





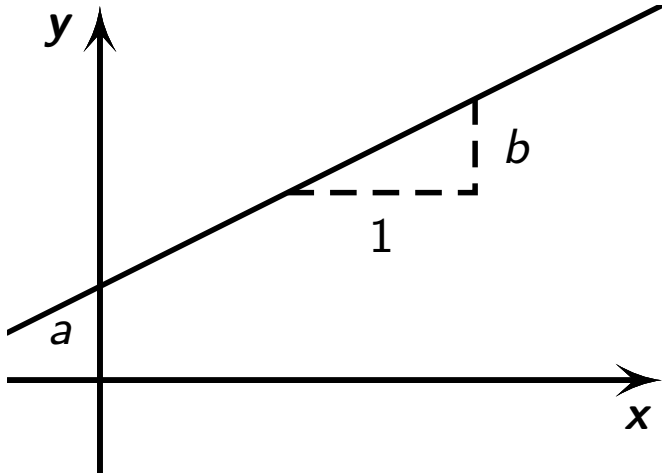
Repetition: Geradengleichung

- Gerade:

$$y = a + bx$$

  - a: y–Achsenabschnitt
  - b: Steigung

- Skizze:



- Interpretation der Steigung: Nimmt x um eine Einheit zu, so ändert sich y um b

Regressionsgerade und Residuum

- Vermutung: Eine Gerade scheint recht gut zu den Daten zu passen
- Diese Gerade hätte die Form:

$$y = a + bx$$

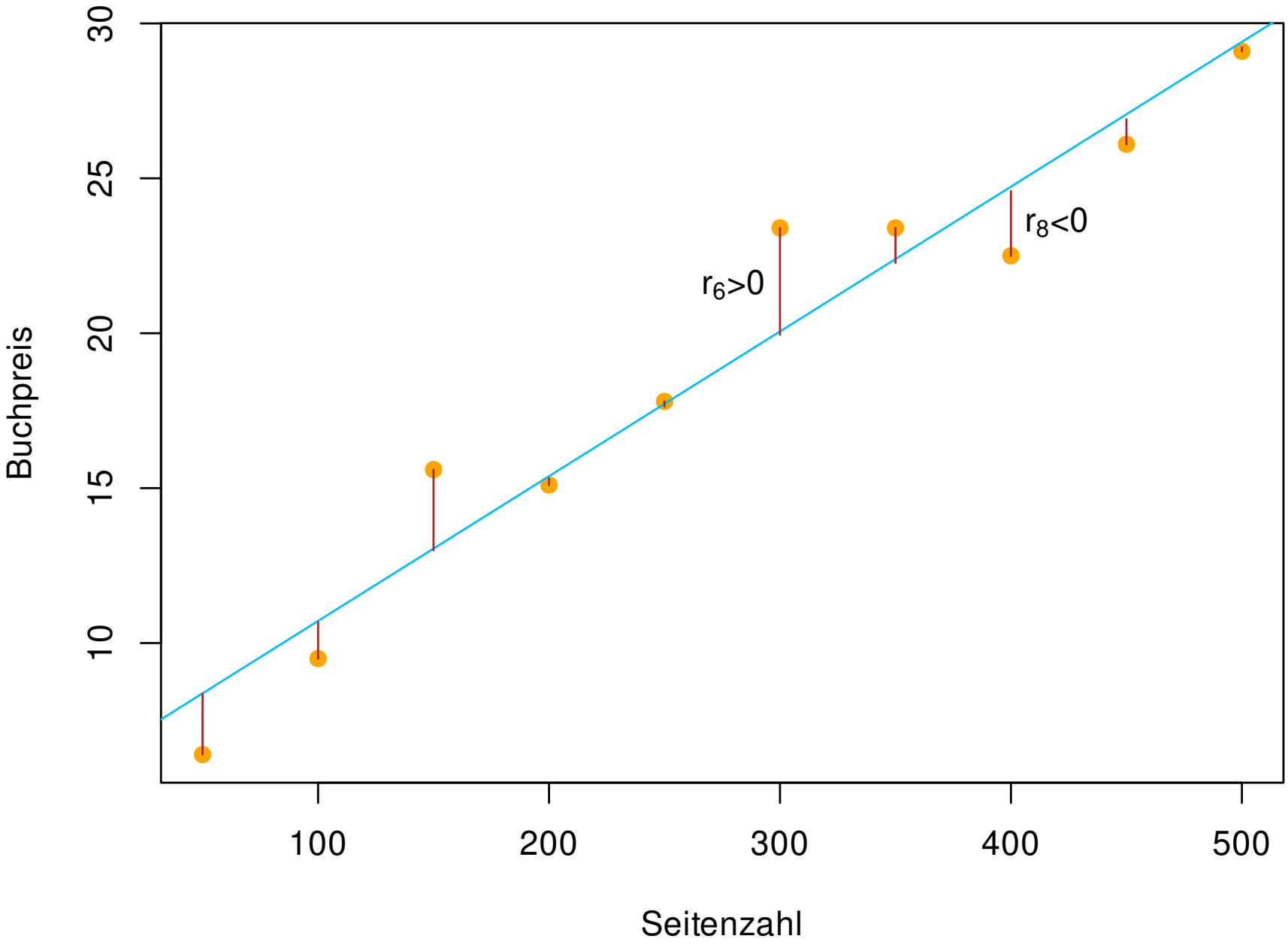
  mit
  - y: Buchpreis
  - x: Seitenzahl
  - Parameter a: Grundkosten des Verlags für ein Buch
  - Parameter b: Kosten pro Seite
- Problem: Gerade finden, die möglichst gut zu allen Punkten passt?
- Was heisst möglichst gut?

- Möglichkeit: Vertikale Abstände zwischen Beobachtung und Gerade zusammenzählen
- Dabei sollte eine kleine Summe der Abstände eine gute Anpassung bedeuten
- Abstände von Messpunkten zu Geraden $\rightarrow$ neuer Begriff:

Residuum

Ein *Residuum* r_i ist die vertikale Differenz zwischen einem Datenpunkt (x_i, y_i) und dem Punkt $(x_i, a + bx_i)$ auf der gesuchten Geraden:

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


Buchpreis
10
15
20
25
30
100
200
300
400
500
Seitenzahl
$r_6>0$
$r_8<0$

- Beispiel: Residuen r_6 und r_8 für *diese* Gerade in Abbildung
- Residuum r_6 positiv, da Punkt überhalb der Gerade
- Entsprechend ist $r_8 < 0$
- Gerade $y = a + bx$ so bestimmen, dass die Summe

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

  minimal wird

- Minimierung von $\sum_i r_i$ hat aber eine *gravierende Schwäche*: Falls Hälfte der Punkte weit über der Geraden, die andere Hälfte weit unter der Geraden liegen: Summe der Abstände etwa null
- Dabei passt die Gerade gar nicht gut zu den Datenpunkten!

Beispiel

- Abbildung:

- Abbildung links: Summe der Residuen ist 0, aber Gerade passt aber überhaupt nicht
- Abbildung rechts: Summe der Residuen ist 0, Gerade passt perfekt
- Aber welches ist die „richtige" Gerade, die am besten zu der Punktwolke passt?
- Verfahren gesucht, das diese Gerade eindeutig festlegt

Methode der kleinsten Quadrate

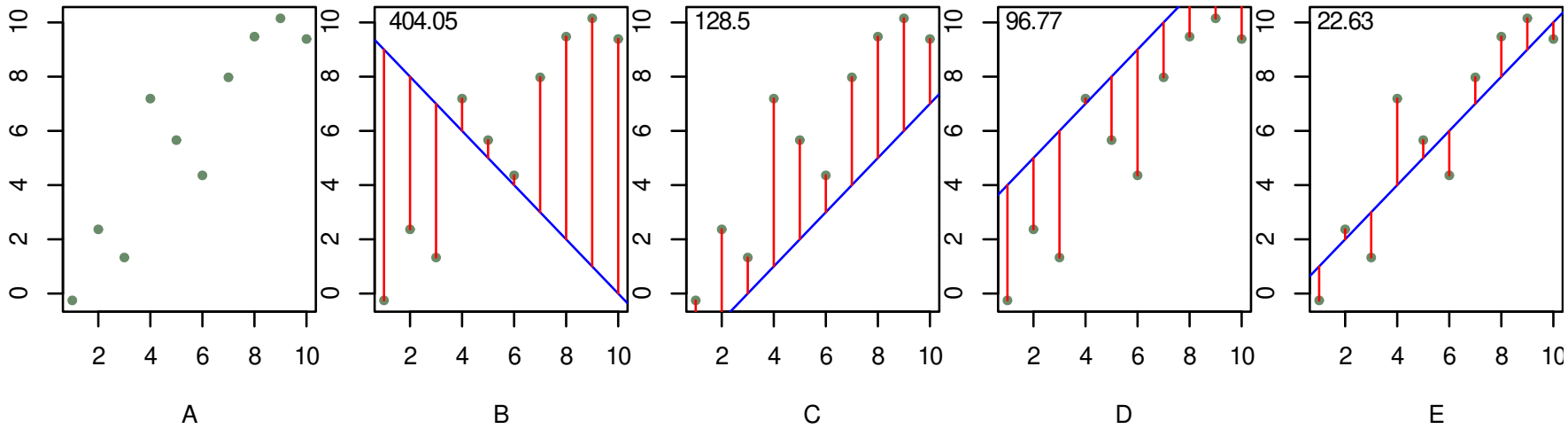
- Eine andere Möglichkeit besteht darin, die Quadrate der Abweichungen aufzusummieren, also

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

- Parameter a und b so wählen, dass diese Summe minimal wird

Graphisch: Methode der kleinsten Quadrate

- Skizze:



- Gesucht: Gerade, die gemäss der kleinsten Methode am besten zur Punktwolke links passt
- Punktwolke steigt: Steigung der Geraden positiv
- Abb. B: fallende Gerade:
  - Gerade passt nicht gut zur Geraden
  - Residuen (rot) sind sehr lang
  - Oben links: Summe der Quadrate der (roten) Residuen: 404.05

- Abb. C: steigende Gerade:
  - Gerade zwar steigend aber zu tief
  - Residuen (rot) immer noch lang, aber besser als bei Abb. A
  - Oben links: Summe der Quadrate der Residuen: 128.25
- Abb. D: steigende Gerade:
  - Gerade zwar steigend aber zu hoch
  - Residuen (rot) immer noch lang, aber besser als bei Abb. A
  - Oben links: Summe der Quadrate der Länge der Residuen: 96.77
- Abb. E: steigende Gerade:
  - Gerade passt gut zur Punktwolke
  - Residuen (rot) klein, verglichen zu den anderen Abb.
  - Oben links: Summe der Quadrate der Länge der Residuen: 22.63

Buchbeispiel

- R berechnet für Beispiel die Werte $a = 6.04$ und $b = 0.047$
  - Grundkosten des Verlags sind also rund 6 SFr. (Preis des Buches für 0 Seiten)
  - Pro Seite verlangt der Verlag rund 5 Rappen

Bestimmung der Parameter a und b

- *Frage:* Wie berechnet der Computer die Parameter a und b?

- Parameter a, b minimieren (Methode der Kleinsten-Quadrate)

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

Die Lösung dieses Optimierungsproblem ergibt:

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

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

wobei $\overline{x}$ und $\overline{y}$ die Mittelwerte der jeweiligen Daten

- Diese Gerade $y = a + bx$ wird auch *Regressionsgerade* genannt

Lineare Regression mit R

```
seiten <- 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)

lm(preis ~ seiten)

##
## Call:
## lm(formula = preis ~ seiten)
##
## Coefficients:
## (Intercept)       seiten
##     6.04000      0.04673
```

- Der Befehl `lm()` steht für „linear model"
- Mit Befehl `lm(y~x)` passt R ein Modell von der Form $y = a + bx$ an die Daten an
- R findet also $a = 6.04$ und $b = 0.0467$

Plotten der Regressionsgerade

- Diese Gerade wird in R wie folgt gezeichnet:

```
seiten <- 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(seiten, preis,
     col = "orange",
     pch = 19,
     xlab = "Seitenzahl",
     ylab = "Buchpreis"
)

abline(lm(preis ~ seiten), col = "deepskyblue")
```

Beispiel: Buchpreis

- Mit diesem Modell: Preis für Bücher mit Seitenzahlen berechnen, die in der Tabelle nicht vorkommen
- Wieviel würde nach diesem Modell ein Buch von 375 Seiten kosten?
- $x = 375$ in die Geradengleichung oben einsetzen:

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

- Das Buch dürfte also etwa CHF 23.60 kosten
- Dieses Modell ist allerdings nur begrenzt gültig
- Vor allem bei *Extrapolationen* muss man vorsichtig sein
- Möglich: Was kostet ein Buch mit einer Million Seiten?
- Oder ein Buch mit -100 Seiten? $\rightarrow$ Nicht realistisch!

Beispiel: Körpergrösse Vater-Sohn

- Vermutung: Zusammenhang zwischen der Körpergrösse der Väter und der Grösse der Söhne
- 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

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

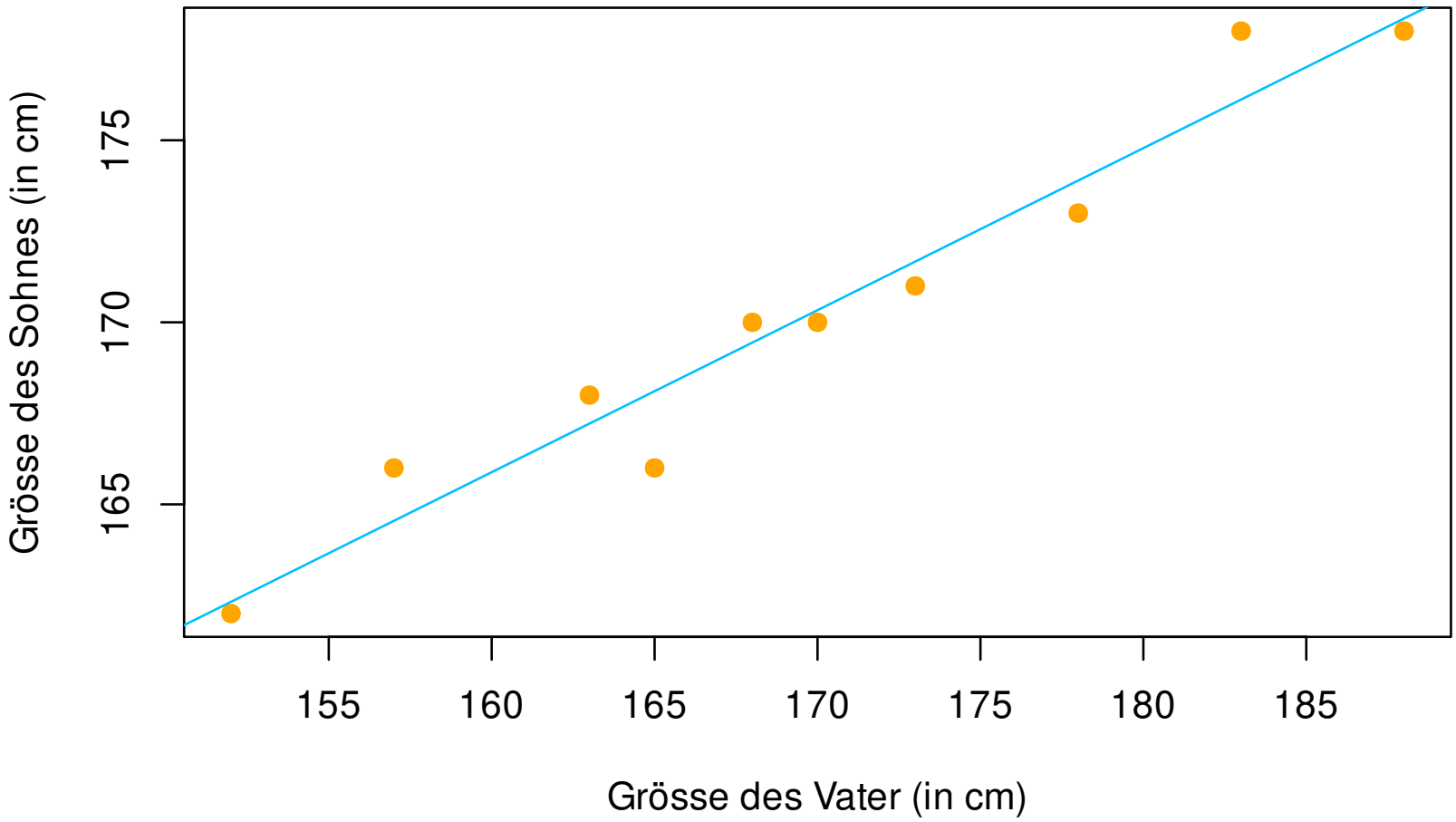
- Es *scheint* einen Zusammenhang zu geben: Je grösser der Vater, desto grösser der Sohn
- Streudiagramm: Möglicher linearer Zusammenhang besteht

- Die Punktwolke „folgt” der Geraden

$$y = 0.445x + 94.7$$

(mit der Methode der Kleinsten Quadrate aus den Daten)

- Streudiagramm:



- Möglich: In Tabelle nicht vorkommende Grösse von 180 cm des Vater, den zu erwartenden Wert für die Grösse seines Sohnes berechnen:

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

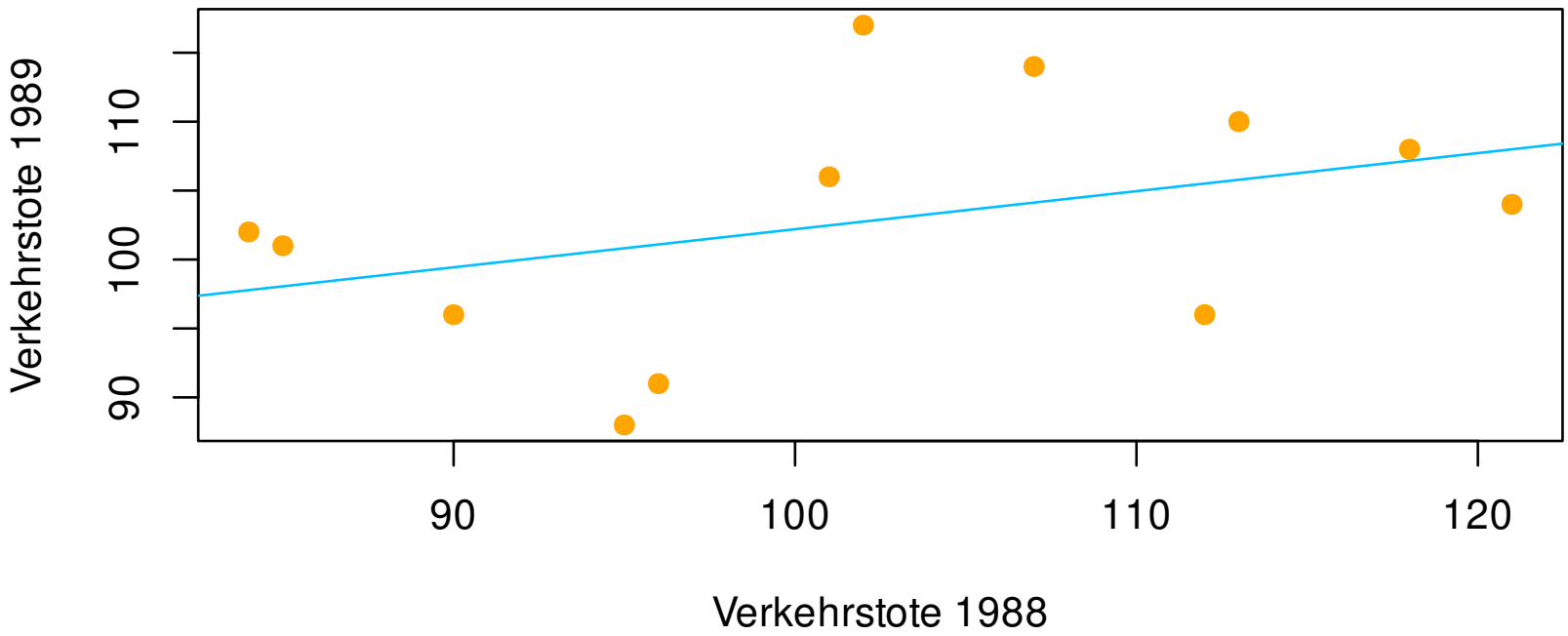
- Achtung: Formel nicht dort anwenden, wo man es *nicht* darf
- Für $x = 0$ erhält man einen Wert von 94.7
- Was heisst dies aber? Wenn der Vater 0 cm gross ist, so ist der Sohn ungefähr 95 cm gross $\rightarrow$ Macht keine Sinn!

Beispiel: Autounfälle

- Tabelle: Zusammenhang zwischen den Zahlen der Verkehrstoten her, die es 1988 und 1989 in zwölf Bezirken in den USA geben hat

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

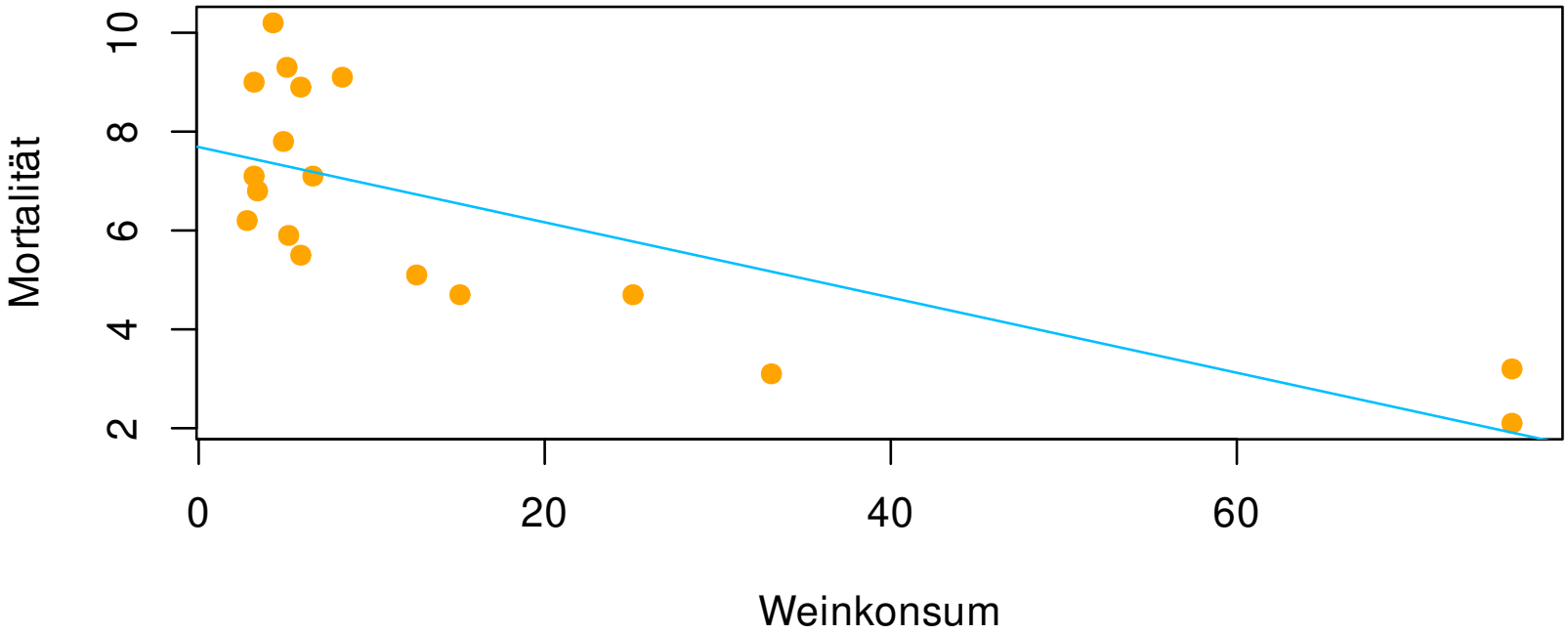
- Es besteht kein offensichtlicher Zusammenhang
- Streudiagramm: kein offensichtlicher Zusammenhang



- Zu erwarten, da es zwischen den Verkehrstoten der einzelnen Bezirke keinen Zusammenhang gibt
- In Abbildung ist noch die Regressionsgerade eingezeichnet
- Können sie zwar berechnen/einzeichnen, *aber diese macht hier gar keinen Sinn*
- *Immer* Berechnung und Plot vergleichen

Beispiel: Weinkonsum

- Schon gesehen: Sterblichkeit vs. Weinkonsum



- Regressionsgerade

$$y = 7.68655 - 0.07608x$$

- Zusammenhang der Daten nicht linear ist (folgt eher einer Hyperbel)
- Regressionsgerade sagt wenig über den wahren Zusammenhang aus