Stochastik

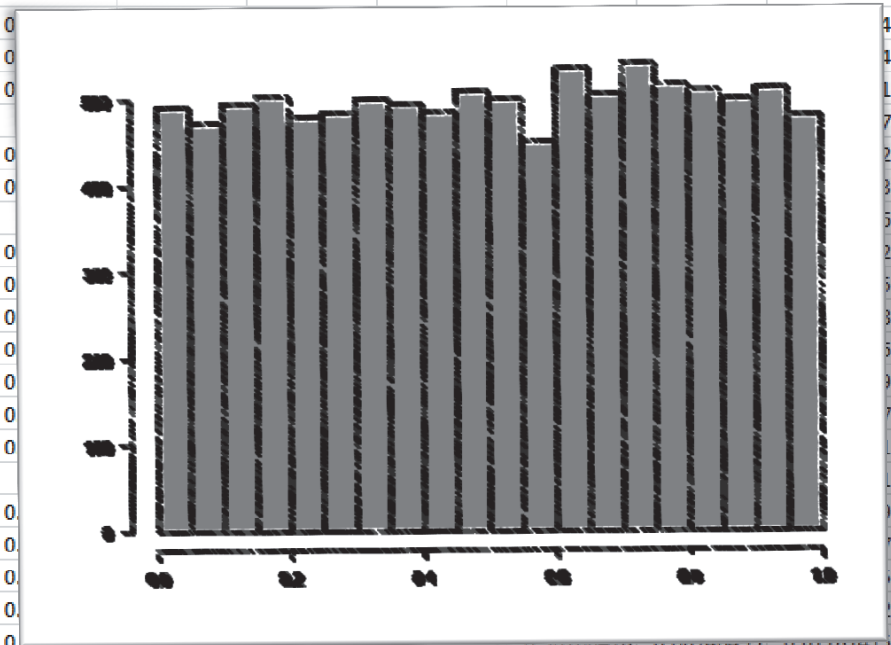
Deskriptive Statistik

Mirko Birbaumer

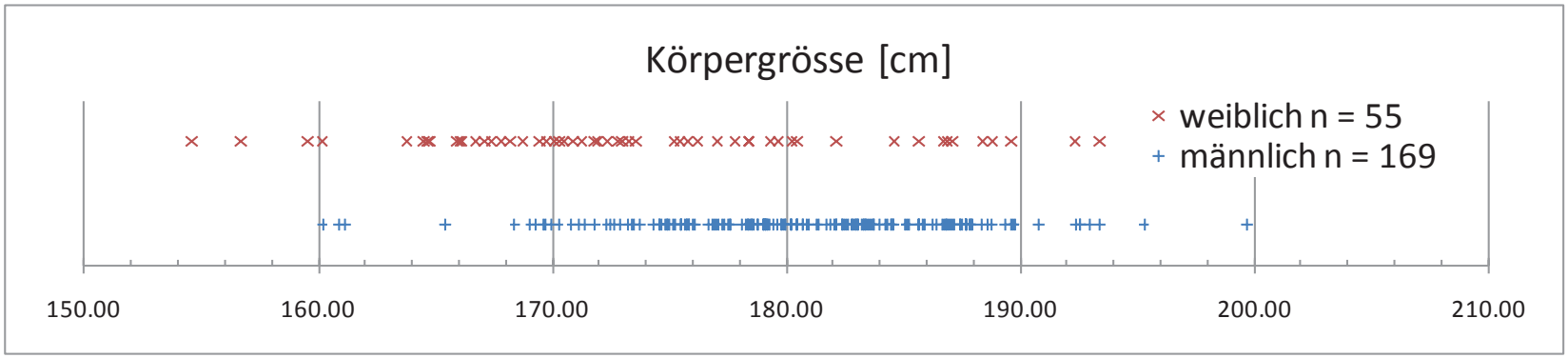
Hochschule Luzern Technik & Architektur

1. Graphische Darstellungen: 1 Dimension
   - Eindimensionales Streudiagramm
   - Histogramm
   - Boxplot
   - Empirische kumulative Verteilungsfunktion
2. Graphische Darstellungen: 2 Dimensionen
   - Streudiagramm
3. Lineare Regression
   - Beispiel: Hubble's Datensatz
     - Absorptionslinie
     - Rotverschiebung
     - Beispiel einer Rotverschiebung
     - Distanzmessung
     - Streudiagramm
     - Big Bang
4. Empirische Korrelation

Graphische Darstellungen

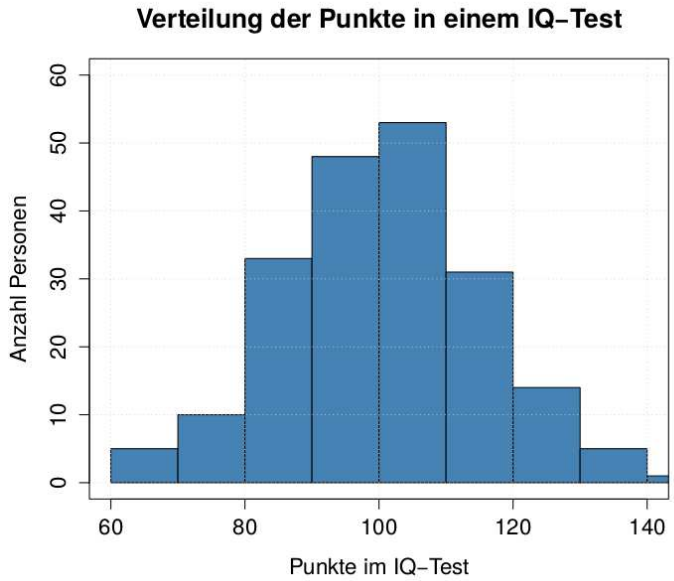


Eindimensionales Streudiagramm



- Guter Überblick, falls nicht zu viele Daten vorhanden sind
- Achtung bei diskret verteilten Daten (Punkte liegen aufeinander!)

Beispiel Histogramm: IQ-Test



- Histogramm von IQ-Test Ergebnis von 200 Personen
- Breite der Klassen: 10 IQ-Punkte ; für jede Klasse gleich
- Höhe der Balken gibt die Anzahl Personen an, die in diese Klasse fallen
- Beispiel: ca. 14 Personen fallen in die Klasse zwischen 120 - 130 IQ-Punkten

Histogramm

- Mit einem **Histogramm** erhalten wir einen graphischen Überblick über die auftretenden Werte
- Aufteilung des Wertebereichs in k **Klassen** (Intervalle)
- **Faustregel** für die Anzahl Klassen k bei n Datenpunkten : z.B. „Sturges Rule“

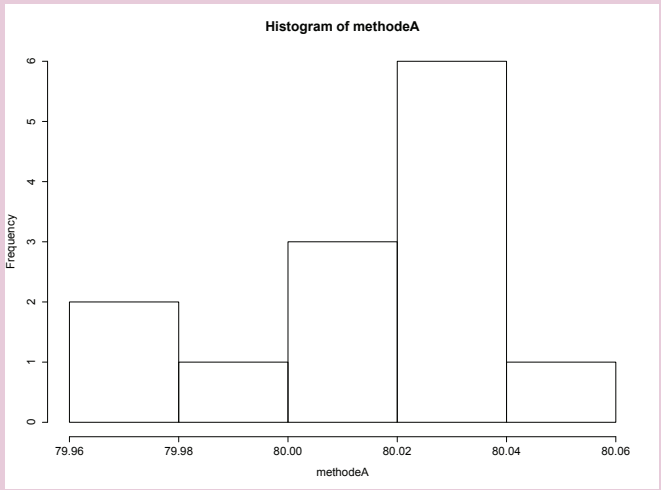
$$k = 1 + 3.3 \log_{10}(n)$$

- 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

Histogramm mit R

R-Befehl: hist()

```
> hist(methodeA)
```



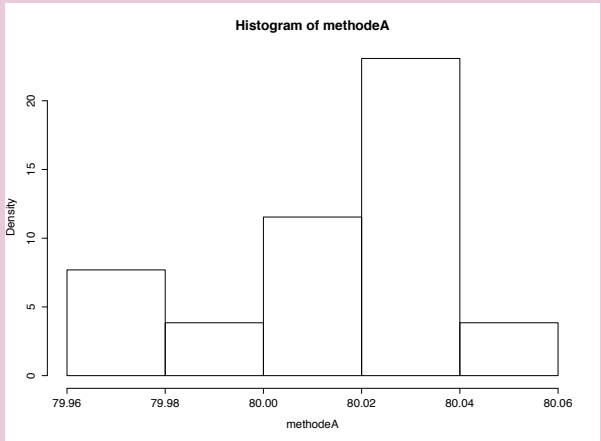
- Methode A enthält 13 Messungen: man wählt 5 Balken (Sturges-Regel: $k = 1 + 3.3 \cdot \log_{10} 13 \approx 5$)

- 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.

Histogramm: Dichte

R-Befehl: hist(...,freq=F)

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



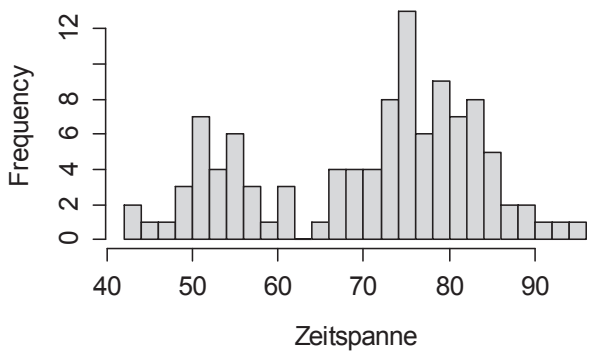
- Gesamtfläche der Balken muss eins sein und die Fläche eines Balken ist proportional zur relativen Häufigkeit
- Der Balken zwischen 80.02 und 80.04 beinhaltet also etwa $0.02 \cdot 20 \approx 0.4$ oder rund 40% der Daten

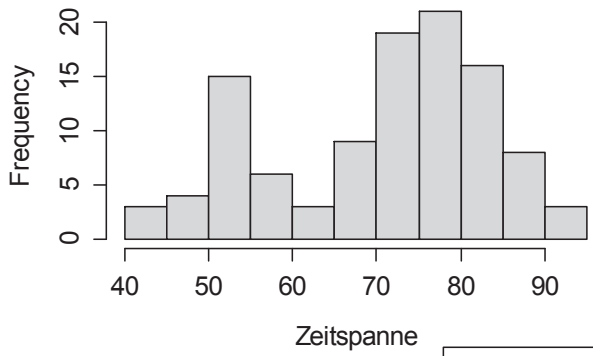
Old Faithful Geysir (Yellowstone NP): Daten

- **Zeitspanne** [min] zwischen Ausbrüchen
- **Eruptionsdauer** [min]
- Daten finden Sie auf ILIAS

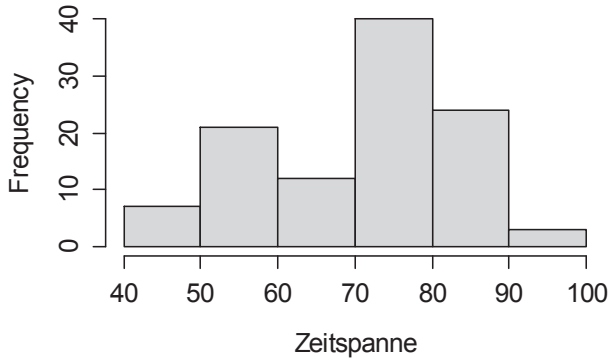
	A	B	C	D
1	Tag	Zeitspanne	Eruptionsdauer	
2	1	78	4.4	
3	1	74	3.9	
4	1	68	4	
5	1	76	4	
6	1	80	3.5	
7	1	84	4.1	
8	1	50	2.3	
9	1	93	4.7	
10	1	55	1.7	
11	1	76	4.9	
12	1	58	1.7	
13	1	74	4.6	
14	1	75	3.4	
15	2	80	4.3	
16	2	56	1.7	
17	2	80	3.9	
18	2	69	3.7	
19	2	57	3.1	
20	2	90	4	
21	2	42	1.8	
22	2	91	4.1	
23	2	51	1.8	

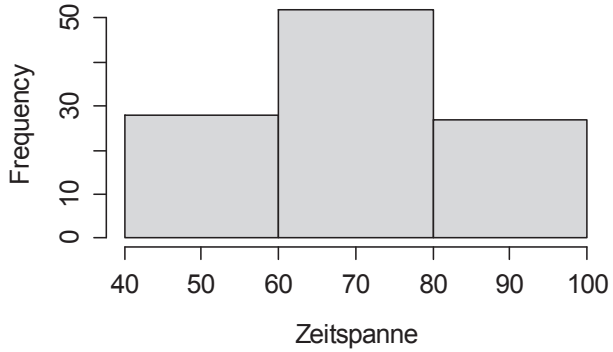
Histogramme für die Zeitspanne (verschiedene Anzahl Klassen)





Resultat hängt von Anzahl Klassen ab!





Interaktiv Klassen verändern (bei anderen Daten): `http://www.amstat.org/publications/jse/v6n3/applets/Histogram.html`

Histogramme mit **R**

- Histogramm mit 20 Klassen und **absoluter Häufigkeit**:

R-Befehl: hist()

```
> hist(geysir[,"Zeitspanne"],breaks=20)
```

- Histogramm mit 20 Klassen und **relativer Häufigkeit**:

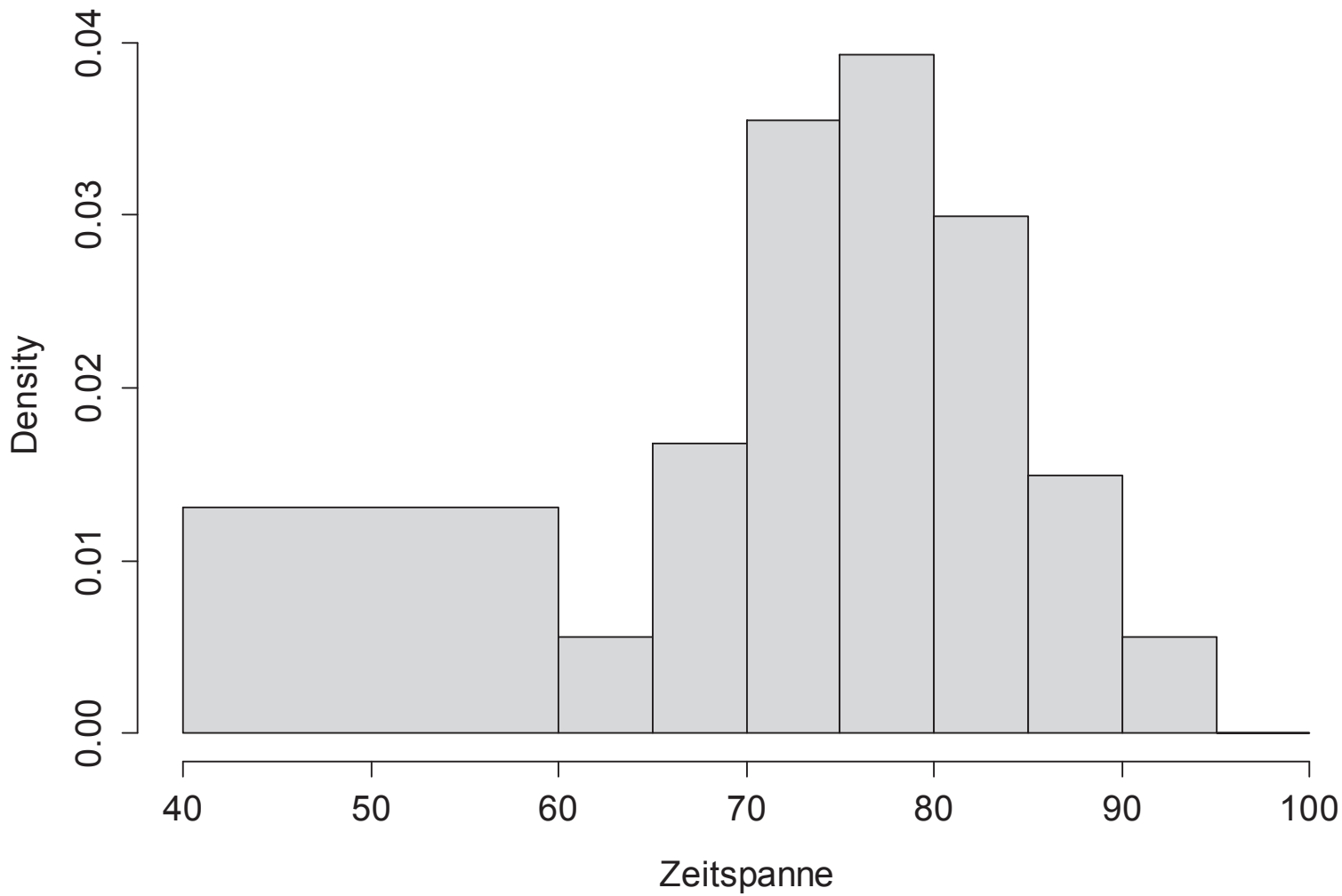
R-Befehl: hist()

```
> hist(geysir[,"Zeitspanne"],breaks=20,freq=FALSE)
```

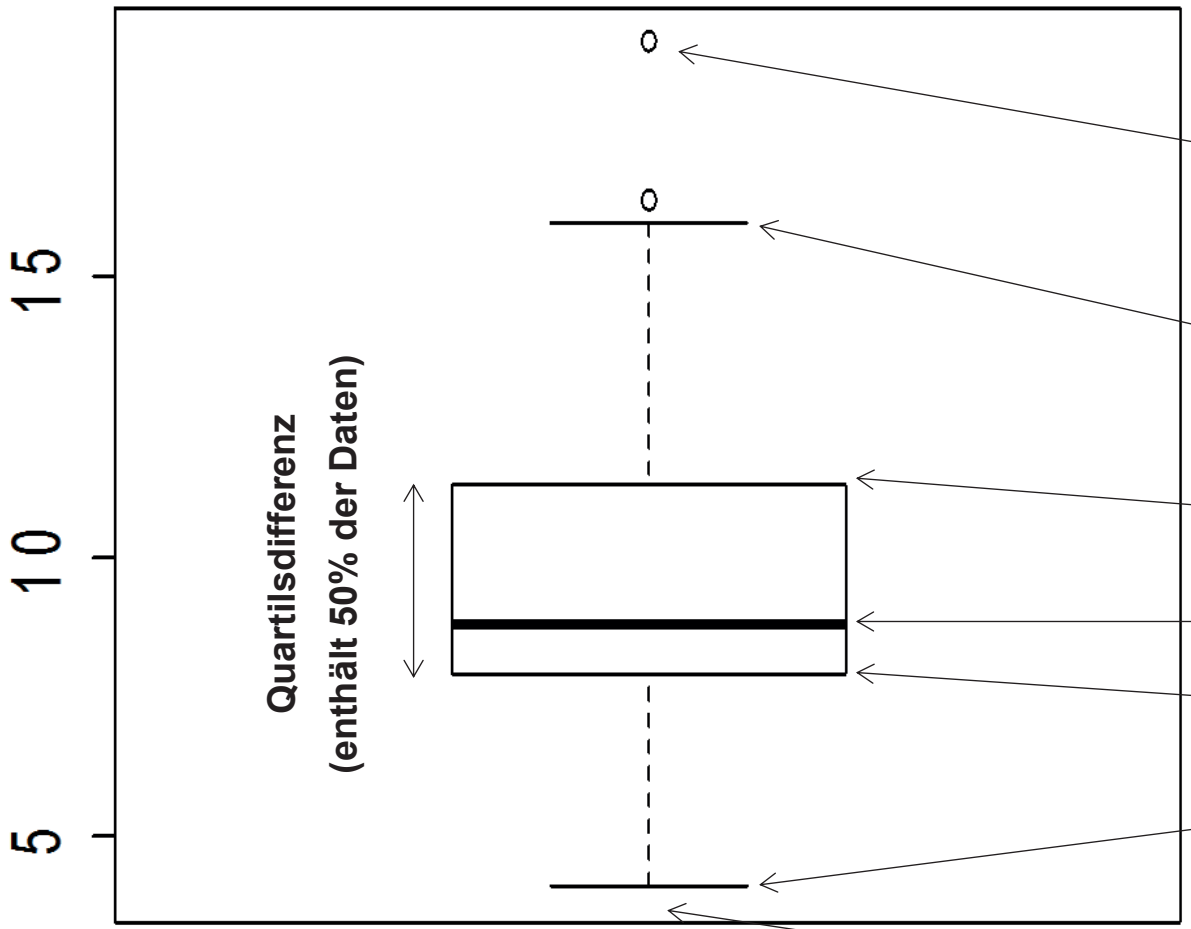
  - Dividieren Anzahl Beobachtungen in einer Klasse durch die Gesamtzahl der Beobachtungen $\rightarrow$ prozentualer Anteil einer Klasse zur Gesamtbeobachtung (relative Häufigkeit)

  - Man wählt Höhe der Balken so, dass **Fläche eines Balken** proportional zur (relativen) Häufigkeit in einer Klasse und Gesamtfläche aller Balken gleich 1 ist

Histogramm der Zeitspanne mit unterschiedlicher Intervallbreite



Boxplot: Schematischer Aufbau



Ausreisser
(falls vorhanden)

Grösste «normale» Beobachtung

Oberes Quartil

Median

Unteres Quartil

Kleinste «normale» Beobachtung

Ausreisser
(falls vorhanden)

Boxplot: Schematischer Aufbau

- Die **grösste normale Beobachtung** ist definiert als die grösste Beobachtung, die höchstens

$$1.5 \cdot \text{Quartilsdifferenz}$$

  vom oberen Quartil entfernt ist

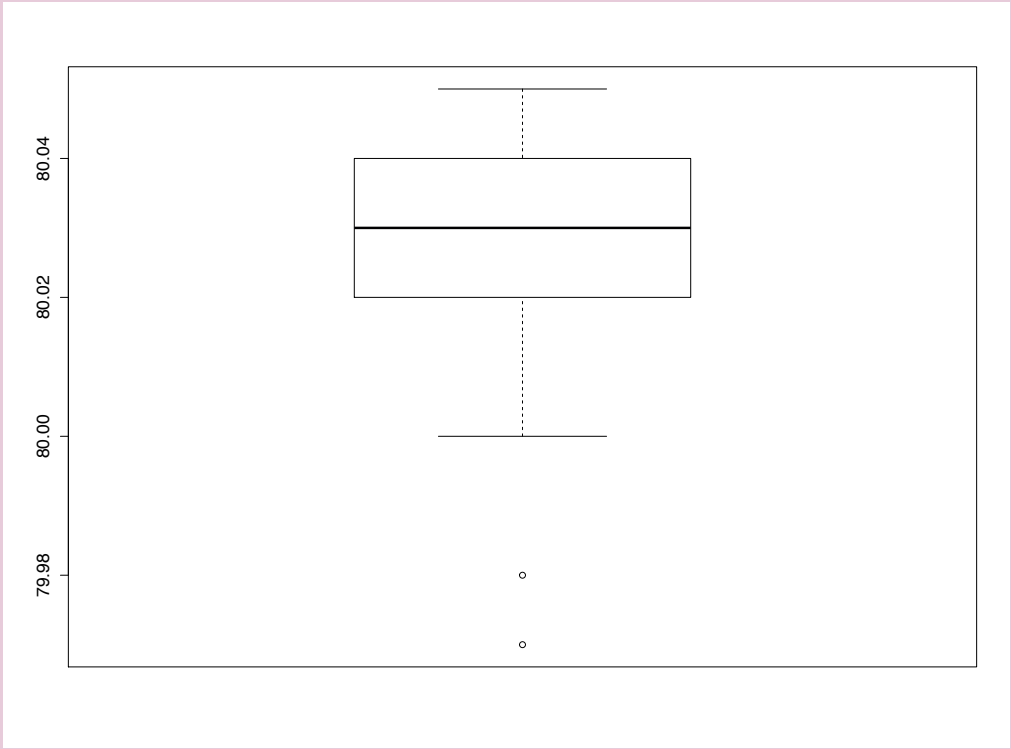
- Die **kleinste normale Beobachtung** ist entsprechend analog definiert mit dem unteren Quartil

- **Ausreisser** sind Punkte, die ausserhalb dieser Bereiche liegen

Boxplot mit R

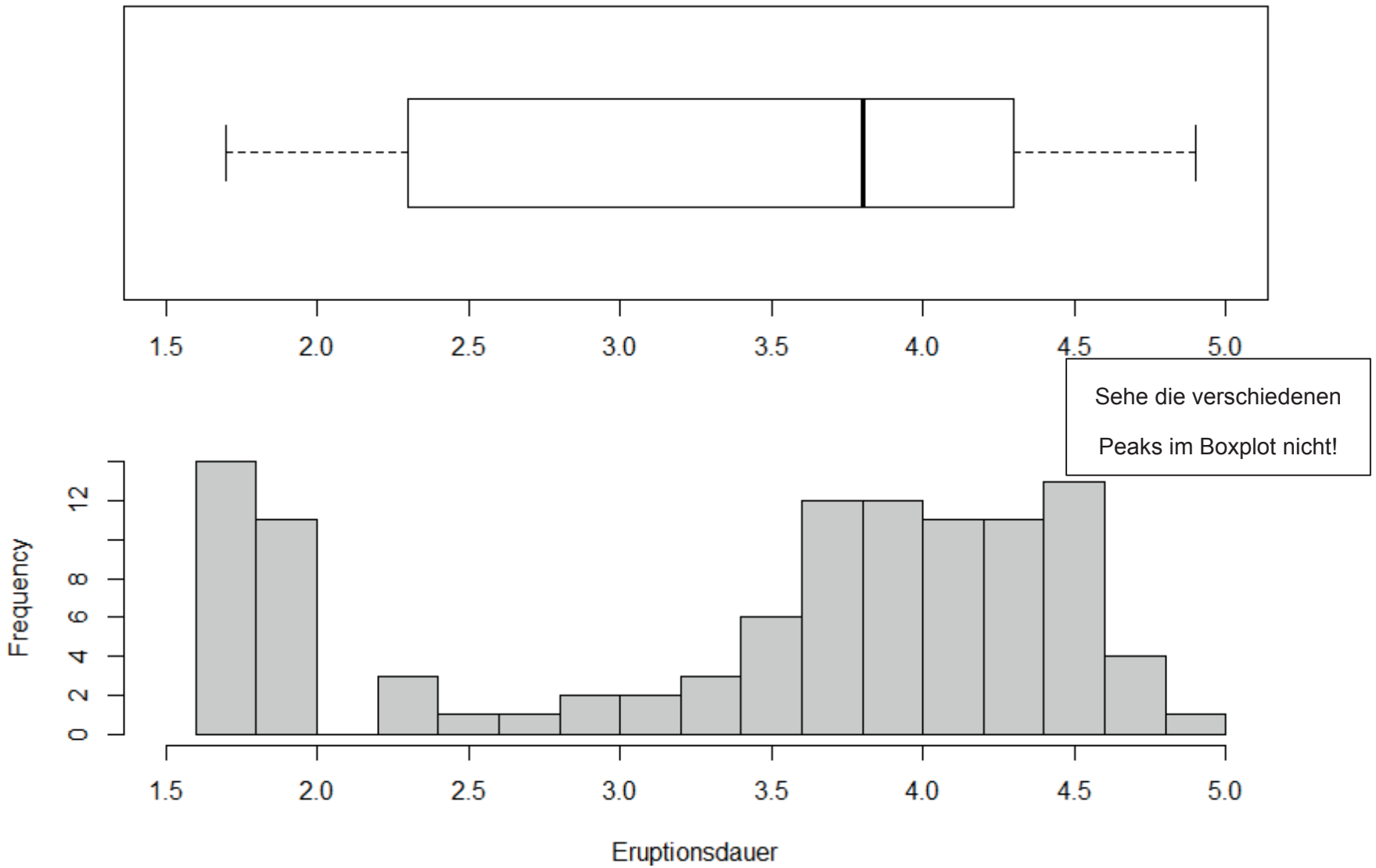
R-Befehl: boxplot()

```
> boxplot(methodeA)
```



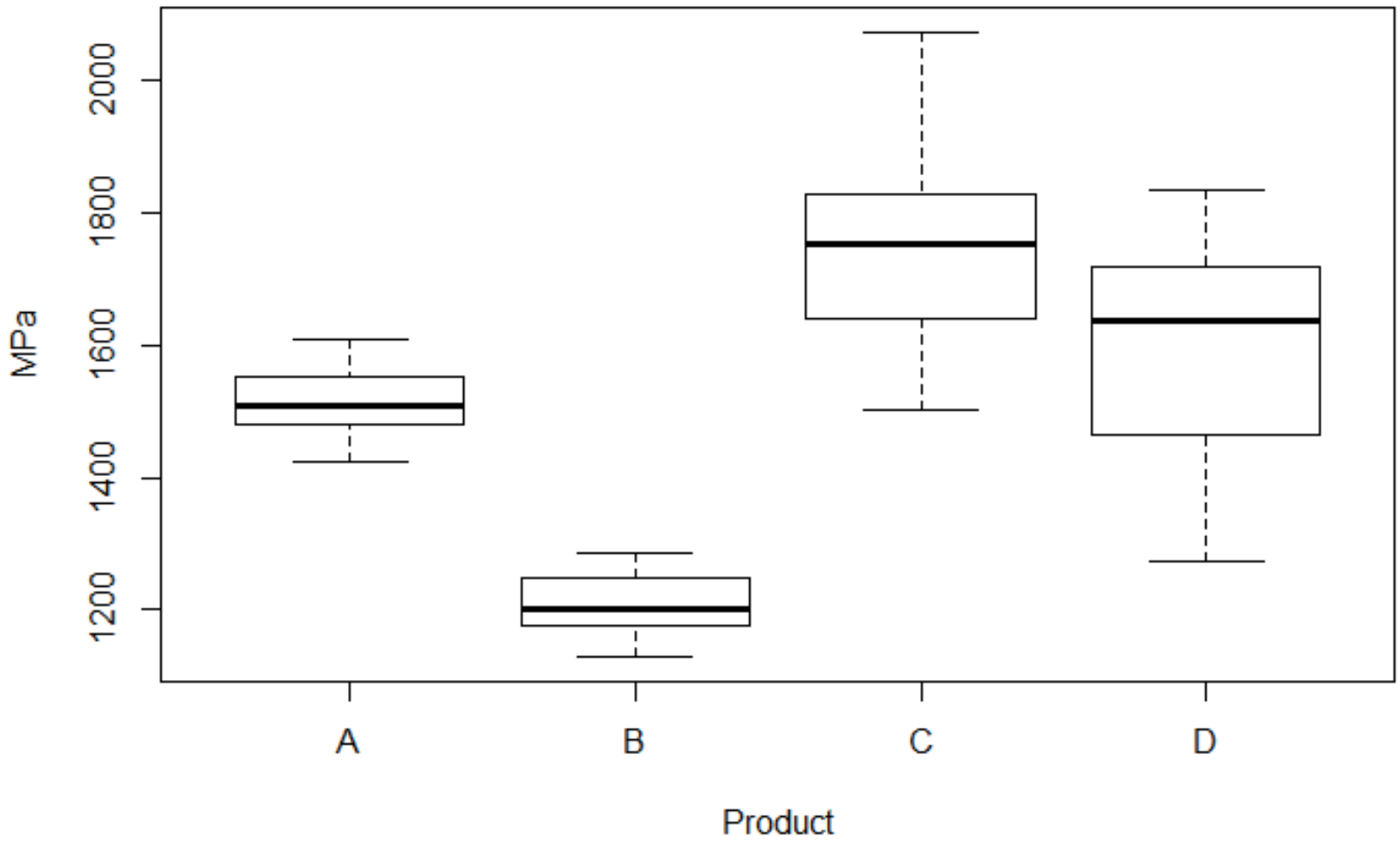
(1)

Boxplot und Histogramm der Eruptionsdauer

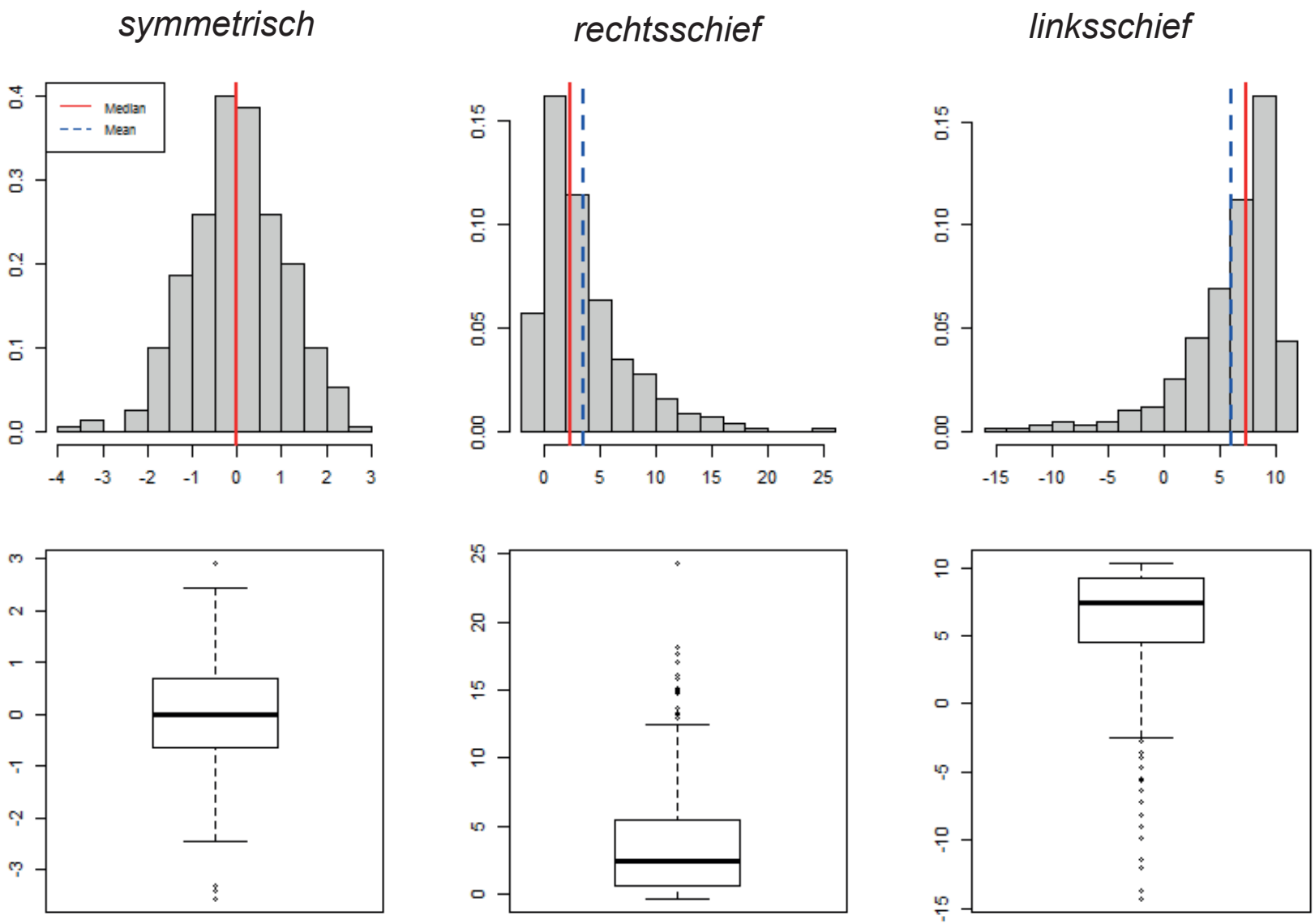


Mehrere Boxplots

Mit mehreren Boxplots kann man einfach und schnell die Verteilung von verschiedenen Gruppen (Methoden, Produkte, ...) vergleichen



Schiefe



Boxplot: Bemerkungen

Im **Boxplot** sind ersichtlich:

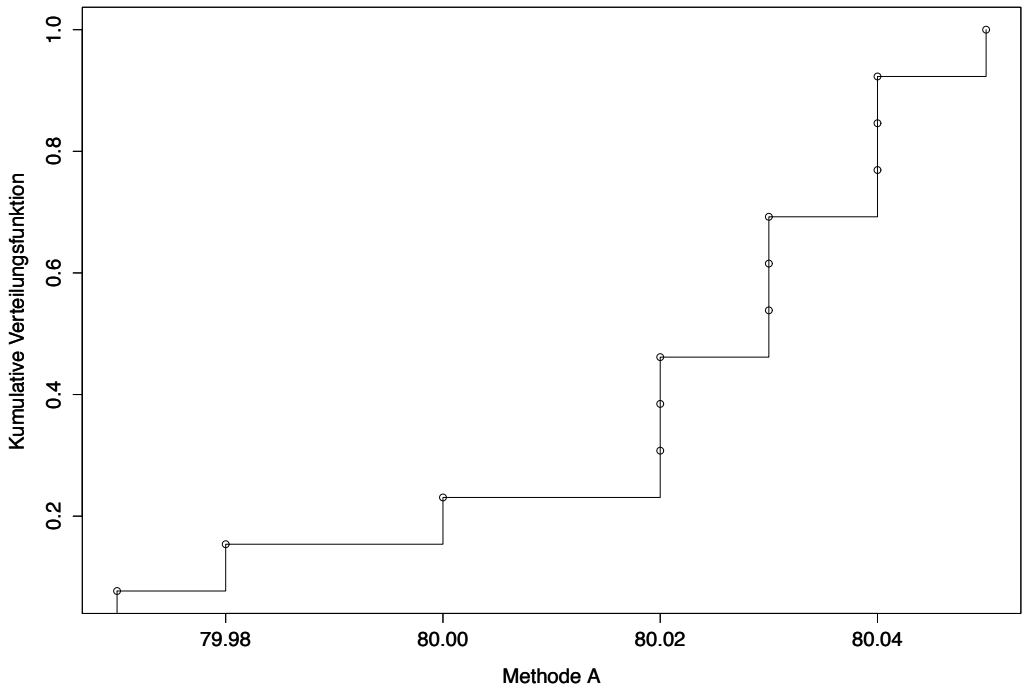
- Lage
- Streuung
- Schiefe

Man sieht aber z.B. **nicht**, ob eine Verteilung mehrere „Peaks“ hat.

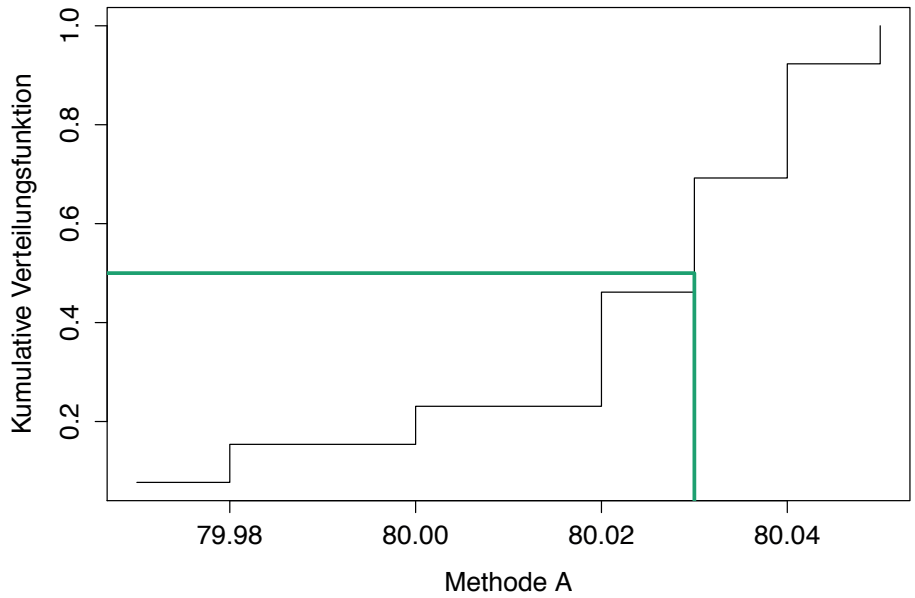
Empirische kumulative Verteilungsfunktion

- Die **empirische kumulative Verteilungsfunktion** $F_n(\cdot)$ ist eine Treppenfunktion, die wie folgt erzeugt wird:
  - links von $x_{(1)}$ ist die Funktion gleich null
  - bei jedem $x_{(i)}$ wird ein Sprung der Höhe $\frac{1}{n}$ gemacht
  - falls ein Wert mehrmals vorkommt, ist Sprung entsprechendes Vielfache von $\frac{1}{n}$

- Beispiel: kumulative Verteilungsfunktion der Methode A



Beispiel: Schmelzwärme mit Methode A



- Links von 79.97 ist Funktion 0 : keine kleineren Beobachtungswerte
- Bei 79.97 : Funktion macht einen Sprung auf $\frac{1}{13} \approx 0.077$
- Zeichnen von 0.5 horizontale Linie $\rightarrow$ grüne Linie in Abbildung schneidet kumulative Verteilungsfunktion bei 80.03 : entspricht gerade dem **Median**
- Dort, wo die kumulative Funktion steil, liegen viele Beobachtungswerte

Empirische kumulative Verteilungsfunktion mit R

R-Befehl: plot()

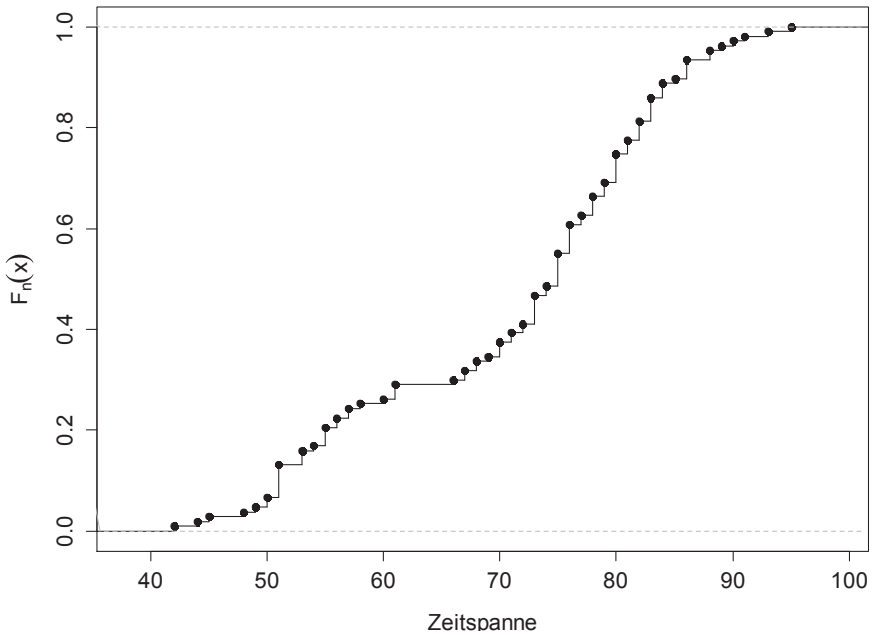
```
> n <- length(methodeA)
> plot(sort(methodeA), (1:n)/n , type="s", ylim=c(0,1),
ylab="Kumulative Verteilungsfunktion", xlab="Methode A")
```

Empirische kumulative Verteilungsfunktion: allgemein

- Empirische kumulative Verteilungsfunktion ist definiert als der **Anteil der Punkte kleiner als ein bestimmter Wert**

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

- Graphik der Kumulativen Verteilungsfunktion für die Zeitspanne im Geysir-Datensatz



Sprunghöhe $1/n$ bei Beobachtungen x_i (bzw. ein Vielfaches davon, wenn es mehrere Beobachtungen mit dem gleichen Wert x_i gibt).

Deskriptive Statistik: 2 Dimensionen

- Wir betrachten nun paarweise beobachtete Daten: zwei Messgrössen pro Messeinheit

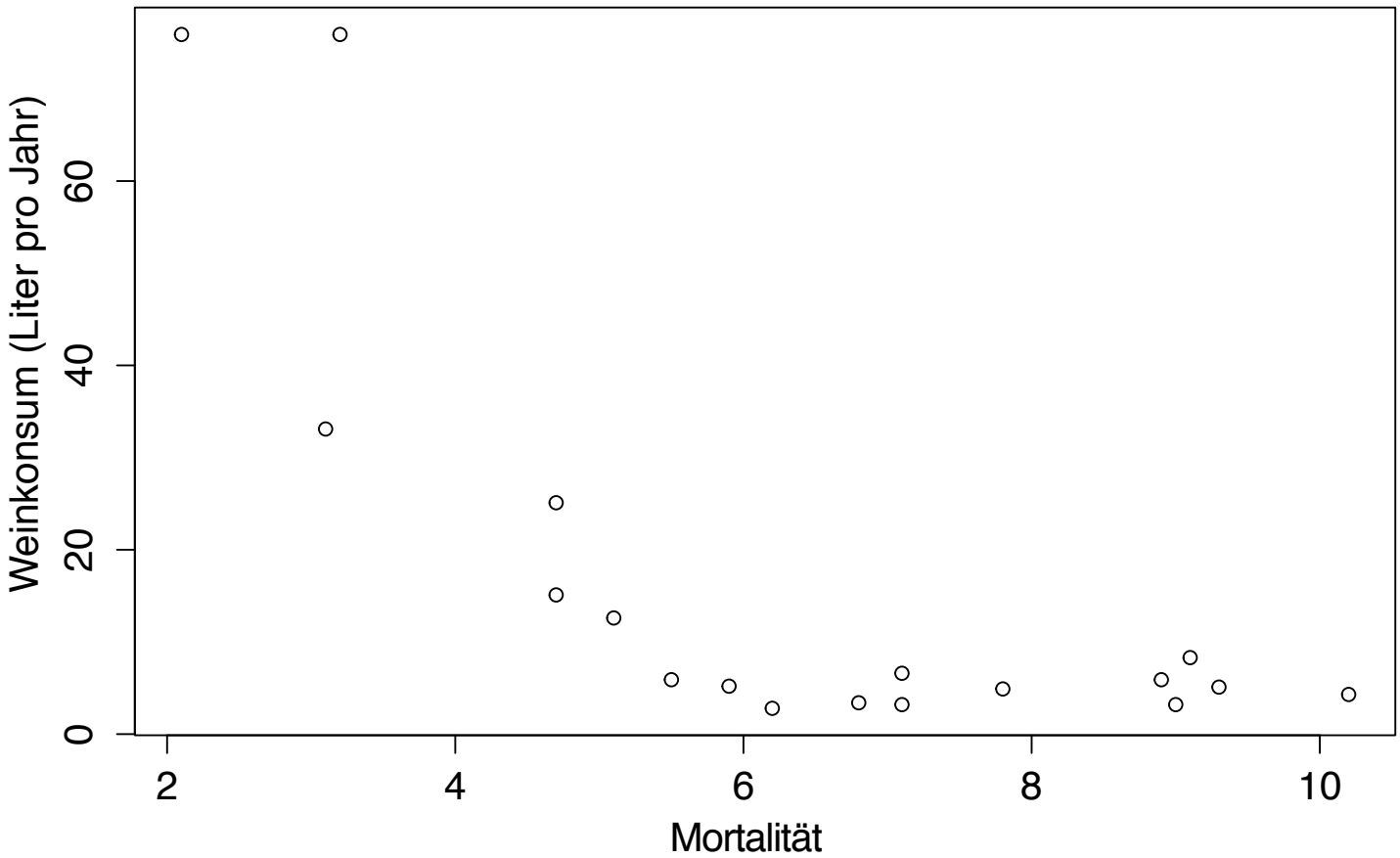
$$x_1, \ldots, x_n$$
$$\updownarrow \qquad \updownarrow$$
$$y_1, \ldots, y_n$$

- Weinkonsumation (Liter pro Person pro Jahr) und Mortalität aufgrund von Herzkreislauferkrankung (Todesfälle pro 1000) in 18 Ländern
- Zum Beispiel die Eruptionsdauer (y_i) und die Zeitspanne (x_i) zum vorangehenden Ausbruch des Old Faithful Geysir

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

Zweidimensionales Streudiagramm

Am Beispiel des Weinkonsums und der Mortalität



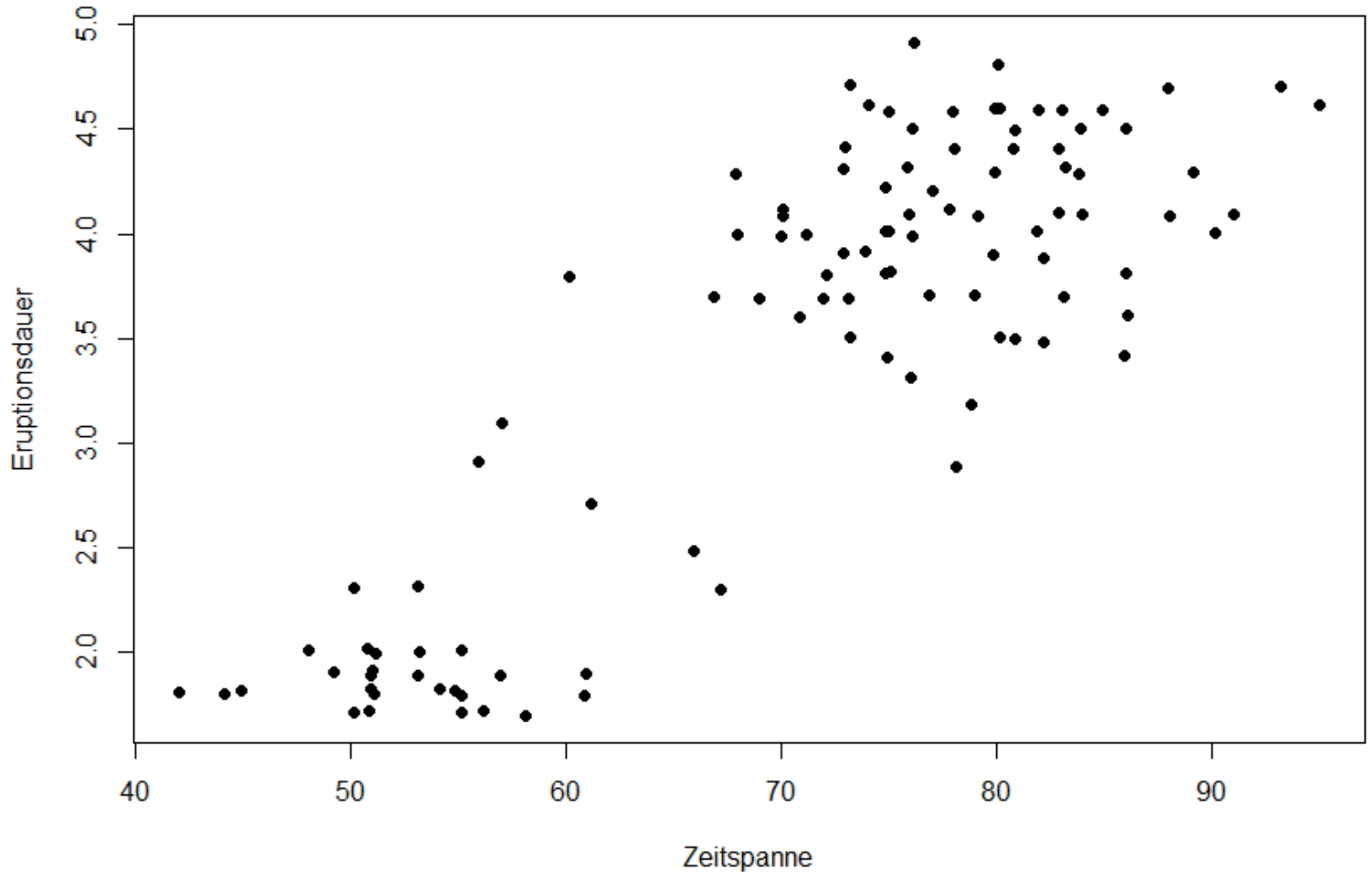
Streudiagramm mit **R**

- 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

R-Befehl: plot()

```
> wein <- c(2.8,3.2,...,75.9)
> mort <- c(6.2,9.0,...,2.1)
> plot(wein,mort,xlab="Weinkonsum (Liter pro Jahr und Person)"
ylab="Mortalität")
```

Am Beispiel Old Faithful



Frage: Lohnt sich das lange Warten auf den nächsten Ausbruch?

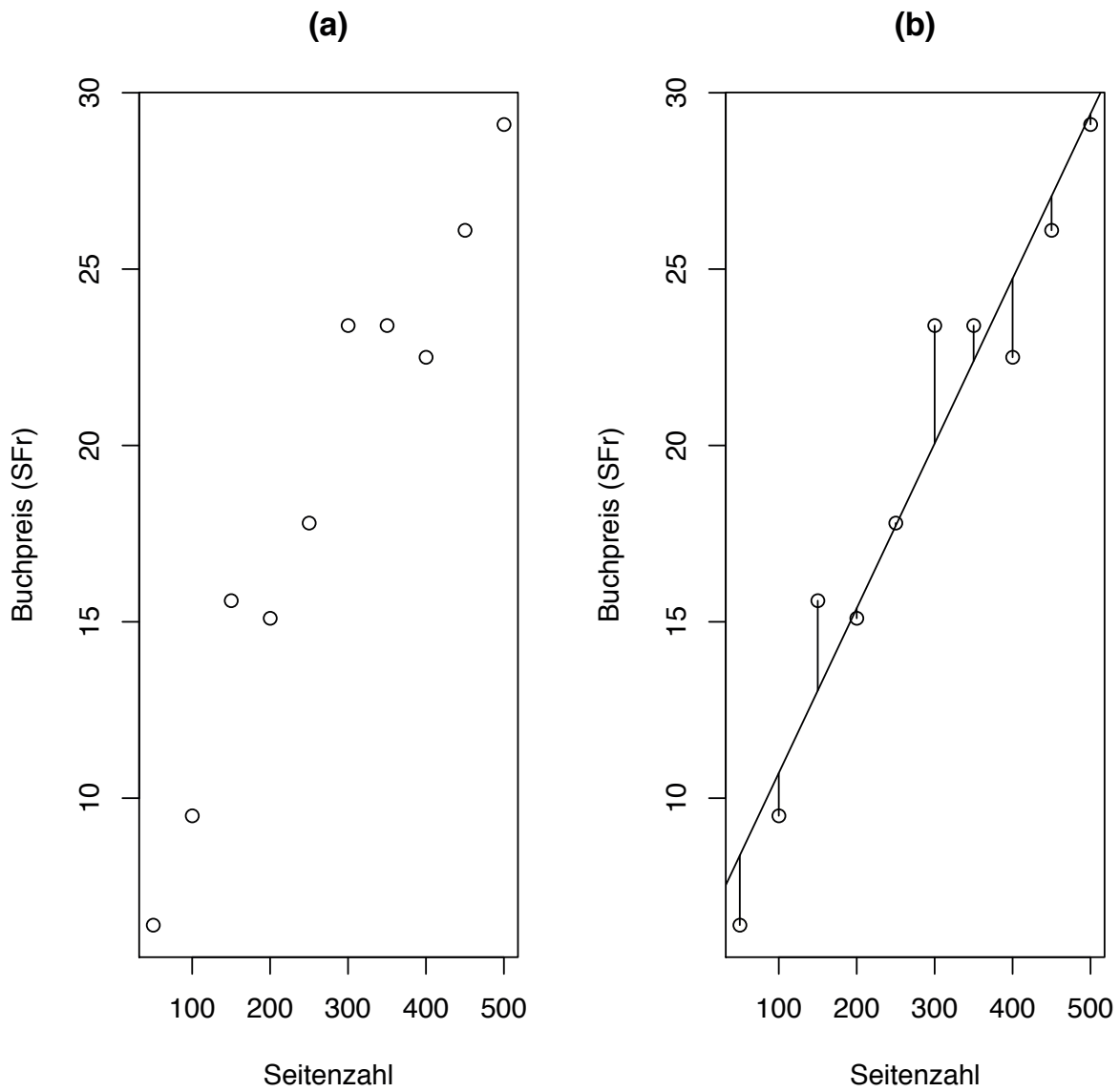
(Fiktives) Beispiel für Lineare Regression

- Wir gehen in eine Buchhandlung und kaufen 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 (Hardcover) ist, desto teurer ist er in der Regel; es gibt Zusammenhang zwischen Seitenzahl x und Buchpreis y
- **Ziel:** Formelmässiger Zusammenhang zwischen Buchpreis und Seitenzahl. Vorhersagen für Bücher mit Seitenzahlen, die wir nicht beobachtet haben.

Streudiagramm und Regressionsgerade



Regressionsgerade und Residuum

- Vermutung: eine Gerade scheint 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: Grundkosten des Verlags, b : Kosten pro Seite)

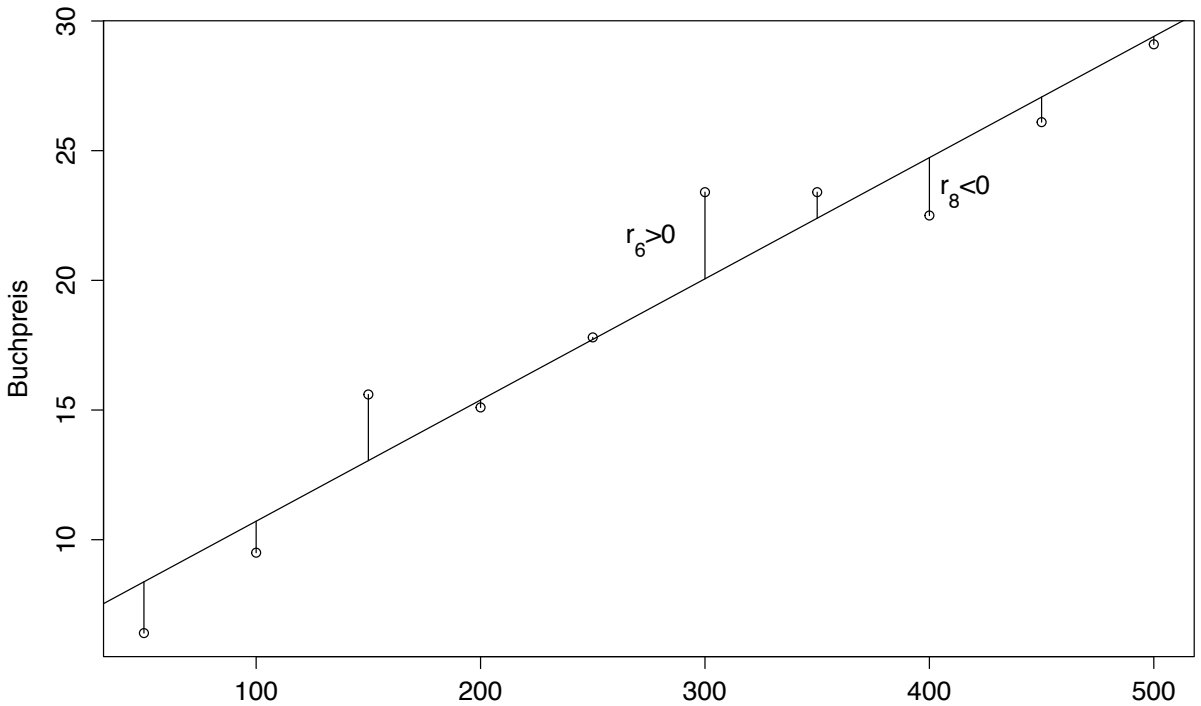
- Wie könnten wir Gerade finden, die möglichst gut zu allen Punkten passt?
- Wir könnten vertikale Abstände zwischen Beobachtung und Gerade zusammenzählen
- Dabei sollte eine kleine Summe der Abstände eine gute Anpassung bedeuten

Residuum

Abstand von Messpunkt zu Geraden: Residuum

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

$$r_i = y_i - a - bx_i$$



Residuum und Regressionsgerade

- Beispiel: Residuen r_6 und r_8 für *diese* Gerade in Abbildung
- Residuum r_6 positiv, da Punkt überhalb der Geraden. 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!

Methode der kleinsten Quadrate

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

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

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

- R berechnet für Beispiel die Werte $a = 6.04$ und $b = 0.047$
  - Die 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?
- Die 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**

R-Befehl: lm()

```
> 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
```

- Der Befehl `lm()` steht für „linear model“
- Mit dem Befehl `lm(y~x)` passt **R** ein Modell von der From $y = a + bx$ 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:

R-Befehl: Regressionsgerade

```
> 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))
```

Beispiel: Buchpreis

- Mit diesem Modell können wir auch Bücher mit Seitenzahlen berechnen, die in der Tabelle nicht vorkommen
- Wieviel 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, wieviel ein Buch mit einer Million Seiten kostet, aber dieser Betrag entspricht dann sicher nicht mehr der Realität
- Oder ein Buch mit -100 Seiten?

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

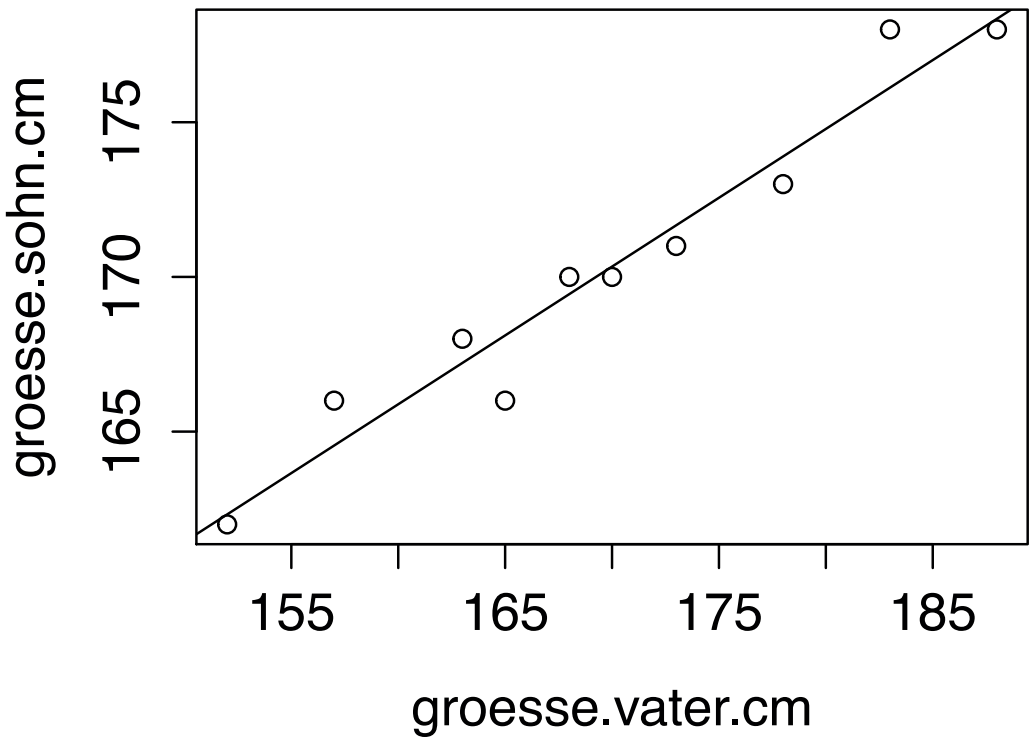
- Es *scheint* hier tatsächlich einen Zusammenhang zu geben: je grösser der Vater, desto grösser der Sohn
- Streudiagramm: möglicher linearer Zusammenhang besteht

Beispiel: Körpergrösse Vater-Sohn

- Die Punktwolke „folgt“ der Geraden

$$y = 0.445x + 94.7$$

Parameter mit der Methode der Kleinsten Quadrate aus den Daten berechnet



Beispiel: Körpergrösse Vater-Sohn

- Wir können also für die in der Tabelle nicht vorkommende Grösse von 180 cm des Vater, den zu erwartenden Wert für die Grösse seines Sohnes berechnen:

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

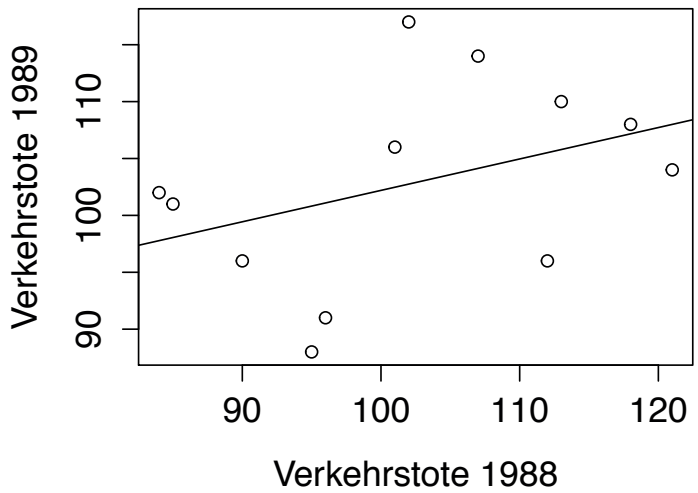
- Achtung: Formel nicht dort anwenden, wo keine Daten vorhanden (Extrapolation)
- Für $x = 0$ erhalten wir einen Wert von 94.7
- Was heisst dies aber? Wenn der Vater 0 cm gross ist, so ist der Sohn ungefähr 95 cm gross und das macht keinen Sinn.

Beispiel: Autounfälle

- Tabelle stellt einen 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

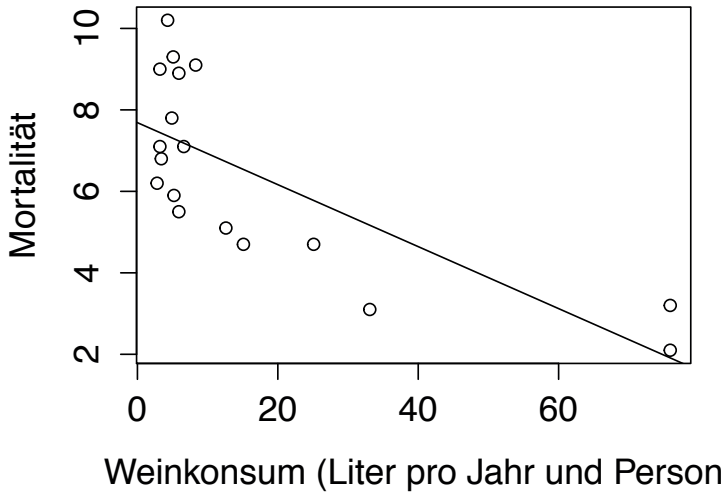


Beispiel: Autounfälle

- Dies war aber auch zu erwarten, wenn wir vernünftigerweise annehmen, dass 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)
- Die Regressionsgerade sagt hier wenig über den wahren Zusammenhang aus

Edwin Hubble

1889 Edwin Hubble kommt am 20. November in Missouri (USA) zur Welt.

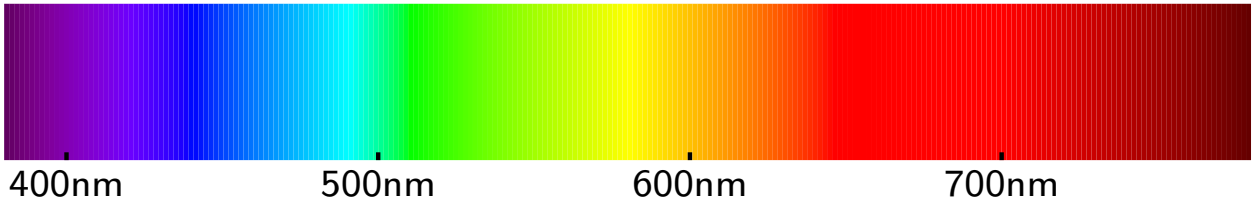
1907-1913 Das Studium der Mathematik, Astronomie und Philosophie an der Universität von Chicago schließt Hubble 1910 mit dem Bachelor of Science ab.

1914-1917 Doktorarbeit, die ihm 1917 die Auszeichnung zum Doktor der Philosophie beschert.

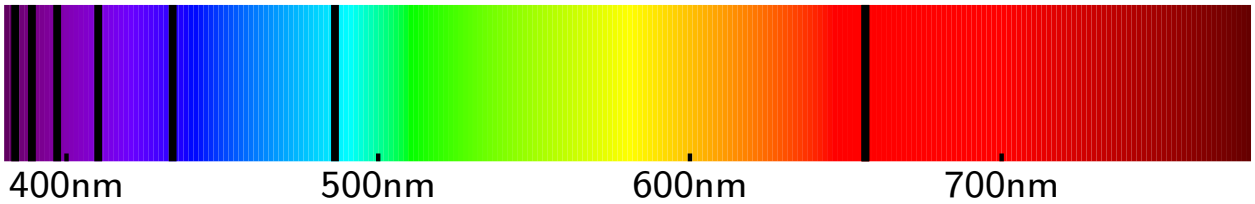
1919-1923 Beobachtung anderer Galaxien und Expansion des Universums aufgrund des Phänomens der Rotverschiebung am Mount Wilson Observatorium.

1953 Am 28. September stirbt Hubble in San Marino (Kalifornien).

- **Spektrum**: Als elektromagnetisches Spektrum bezeichnet man die Gesamtheit aller elektromagnetischen Wellen verschiedener Energien.
  Beispiel: Lichtspektrum, welches ohne technische Hilfsmittel über das menschliche Auge wahrgenommen werden kann. Der Wellenlängenbereich des Lichtspektrums reicht von ungefähr 380nm bis 780nm



- **Absorptionslinie**: dunkle Linien im kontinuierlichen Spektrum einer Lichtquelle, die infolge Absorption des Lichts durch Materie entstehen. Das Vorkommen von Absorptionslinien im Spektrum erlaubt uns Rückschlüsse auf die Chemie, die Temperatur, den Druck und die Bewegung der absorbierenden Gasschicht zu ziehen.
  Beispiel: Lichtspektrum mit Absorptionslinien von Wasserstoff.



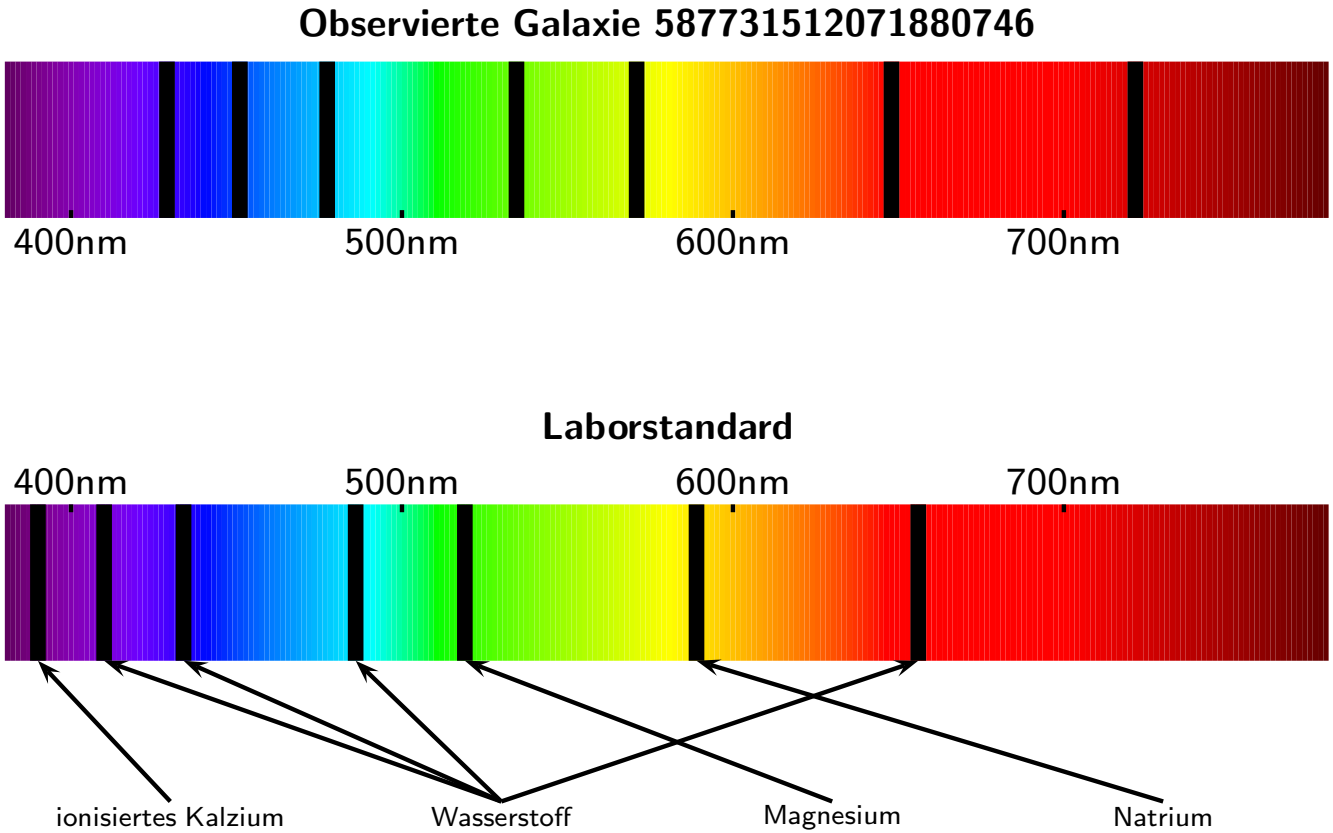
- **Dopplereffekt**: Verschiebung der Spektrallinien nach Rot bei Entfernen einer Lichtquelle von einem Beobachtungsort, Blauverschiebung bei Annähern
- **Rotverschiebung z**: Als Rotverschiebung z elektromagnetischen Wellen wird die Verlängerung der gemessenen Wellenlänge λ gegenüber der ursprünglich emittierten Strahlung λ_e bezeichnet:

$$\boxed{z = \frac{\lambda - \lambda_e}{\lambda_e} = \frac{\lambda}{\lambda_e} - 1}$$

  Gemessen wird die Rotverschiebung meist anhand der Verschiebung von Spektrallinien, d.h. Emissionen oder Absorptionen atomar oder molekular festliegender Frequenzen.
- Die **Geschwindigkeit** u der sich entfernenden Lichtquelle (z.B. Galaxie) kann nun aufgrund der relativistischen Dopplerverschiebung ermittelt werden. Für kleine Geschwindigkeiten u gilt näherungsweise folgende Beziehung zwischen Rotverschiebung z und Fluchtgeschwindigkeit

$$\boxed{u \approx z \cdot c}$$

Rotverschiebung der Galaxie 587731512071880746:



Ionisiertes Kalzium: $\lambda = 429\text{nm}$, $\lambda_e = 390\text{nm} \Rightarrow z = \frac{429}{390} - 1 = 0.1$
Fluchtgeschwindigkeit: $\approx 10\%$ der Lichtgeschwindigkeit

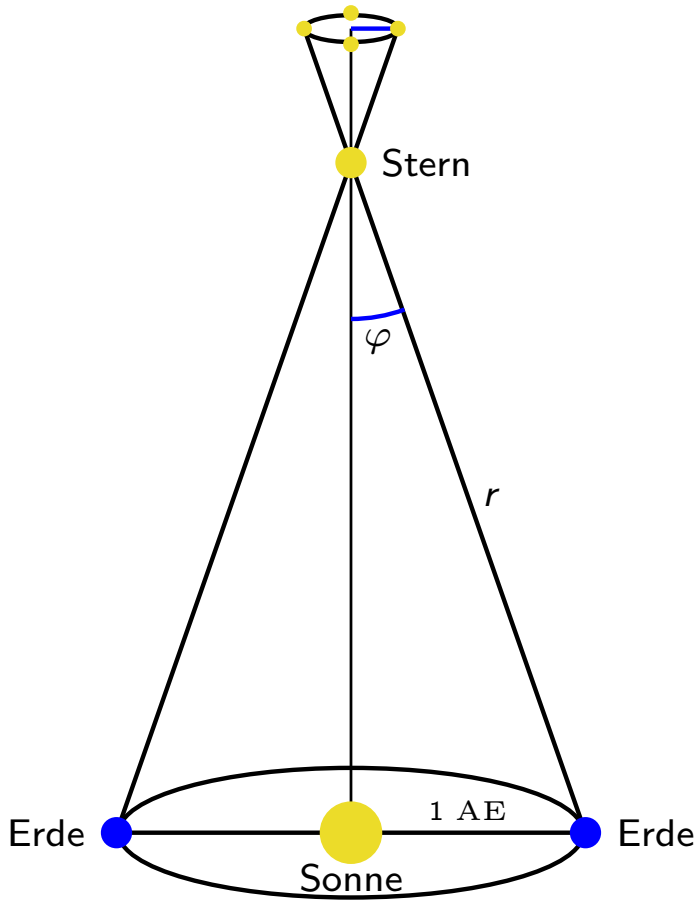
Distanzmessung: Trigonometrische Parallaxe

- Die Entfernung Erde-Stern r ist gegeben durch

$$\varphi = \frac{1AE}{r}$$

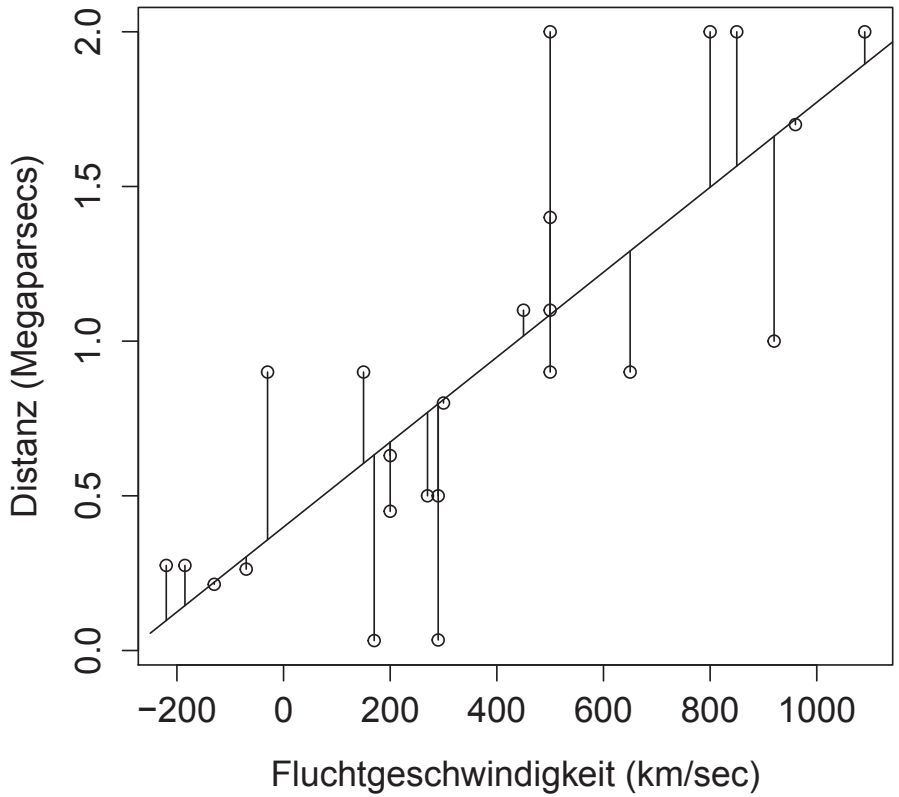
wobei AE die astronomische Einheit (mittlere Distanz Erde-Sonne) bezeichnet, die Entfernung r in parsec (pc) und die Parallaxe φ in Bogensekunden gegeben ist.

- Ein Parsec ist die Entfernung, aus der der mittlere Abstand der Erde zur Sonne unter einem Winkel von einer Bogensekunde erscheint: $3.086 \cdot 10^{16}$ m.



Zusammenhang zwischen Distanz und Fluchtgeschwindigkeit von Galaxie

Nebel	Geschwindigkeit (km/sec)	Distanz (Mparsec)
S. Mag.	170	0.032
L. Mag. 2	290	0.034
NGC 6822	-130	0.214
NGC 598	-70	0.263
NGC 221	-185	0.275
NGC 224	-220	0.275
NGC 5457	200	0.450
NGC 4736	290	0.500
NGC 5194	270	0.500
NGC 4449	200	0.630
NGC 4214	300	0.800
NGC 3031	-30	0.900
NGC 3627	650	0.900
NGC 4626	150	0.900
NGC 5236	500	0.900
NGC 1068	920	1.000
NGC 5055	450	1.100
NGC 7331	500	1.100
NGC 4258	500	1.400
NGC 4151	960	1.700
NGC 4382	500	2.000
NGC 4472	850	2.000
NGC 4486	800	2.000
NGC 4649	1090	2.000



Lineare Regression: Parameterschätzung

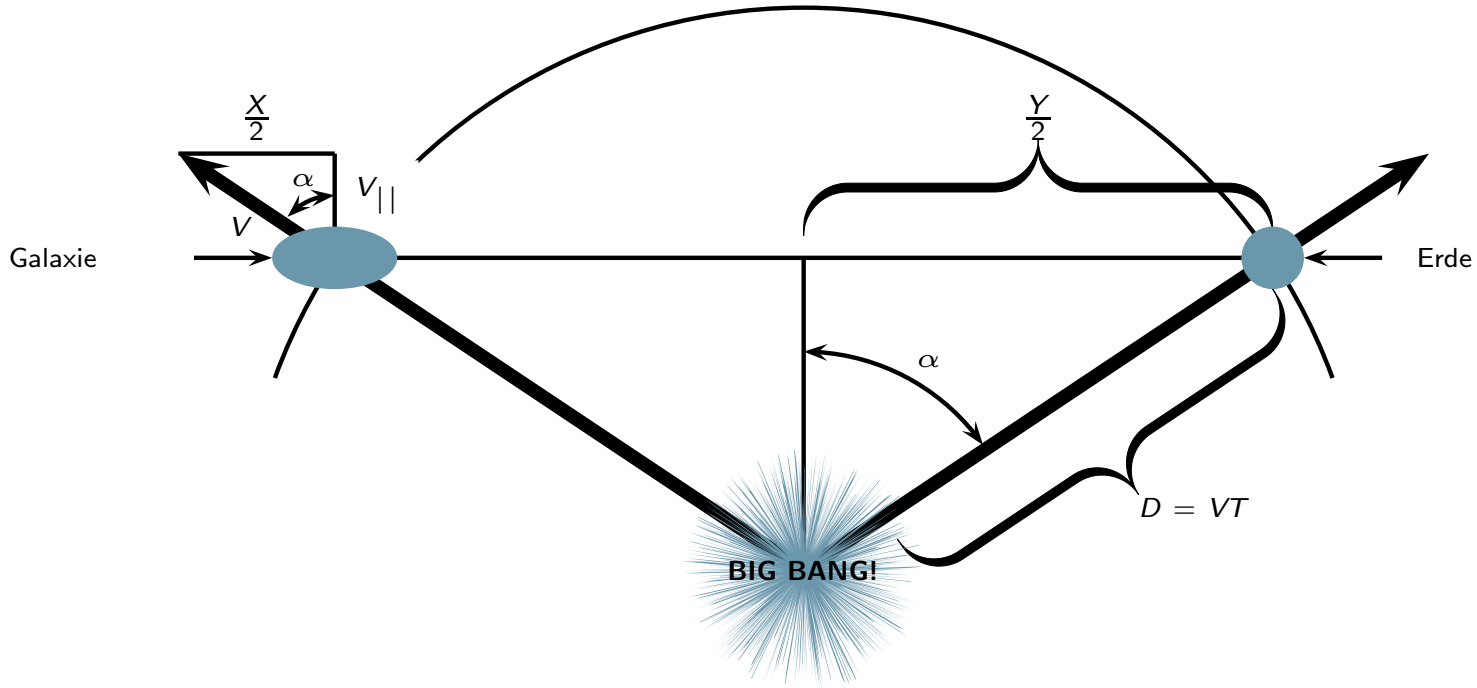
- Wir nehmen an, es besteht ein *linearer Zusammenhang* zwischen Distanz y und Fluchtgeschwindigkeit x. Lineares Modell:

$$y = \beta_0 + \beta_1 x$$

- Parameterschätzung mit der Methode der kleinsten Quadrate:

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}\,. \end{aligned}$$

- Hubble-Datensatz: $\hat{\beta}_1 = 0.00137$ und $\hat{\beta}_0 = 0.39910$.



$$\frac{Y/2}{VT} = \frac{X/2}{V} = \sin(\alpha) \Rightarrow Y = TX$$

- Alter des Universums T entspricht dem Parameter β_1. Einheit von β_1 : megaparsec-Sekunde pro Kilometer $\longrightarrow$ 979.8 Milliarden Jahre
- Alter des Universums: $0.00137 \cdot 979.8 = 1.34$ Milliarden Jahre

Wie gut passt die Regressionsgerade?

- Die Regressionsgerade können wir (fast) immer bestimmen
- Regressionsgerade sagt manchmal wenig über die wirkliche Verteilung der Punkte im Streudiagramm aus: z.B. wenn
  - die Punkte scheinbar gar keiner Gesetzmässigkeit folgen
  - die Punkte folgen einer nichtlinearen Gesetzmässigkeit folgen
- Wie können wir nun aber feststellen, ob ein linearer Zusammenhang der Daten besteht oder nicht?
- Möglichkeit: Datensatz graphisch darstellen
- Wert angeben, der den Zusammenhang numerisch beschreibt (empirische Korrelation)

Empirische Korrelation

Numerische Zusammenfassung der linearen Abhängigkeit von zwei Grössen:

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

Berechnung von Korrelation mit **R**

- Für unser Seitenzahl-Preis-Beispiel erhalten wir mit **R**

R-Befehl: cor()

```
> cor(seitenzahl,buchpreis)
[1] 0.9681122
```

- Der Wert ist also sehr nahe bei 1 und somit besteht ein starker linearer Zusammenhang
- Dazu ist der Wert positiv, was einem „je mehr, desto mehr“ Zusammenhang entspricht

Empirische Korrelation: Beispiele

- Beispiel der Körpergrösse von Vater und Sohn: erwarten hohen Korrelationskoeffizienten, da Daten nahe der Regressionsgerade

  $\rightarrow \quad 0.973$

- Verkehrsunfällen: keinen Zusammenhang und erwarten tiefen Korrelationskoeffizienten

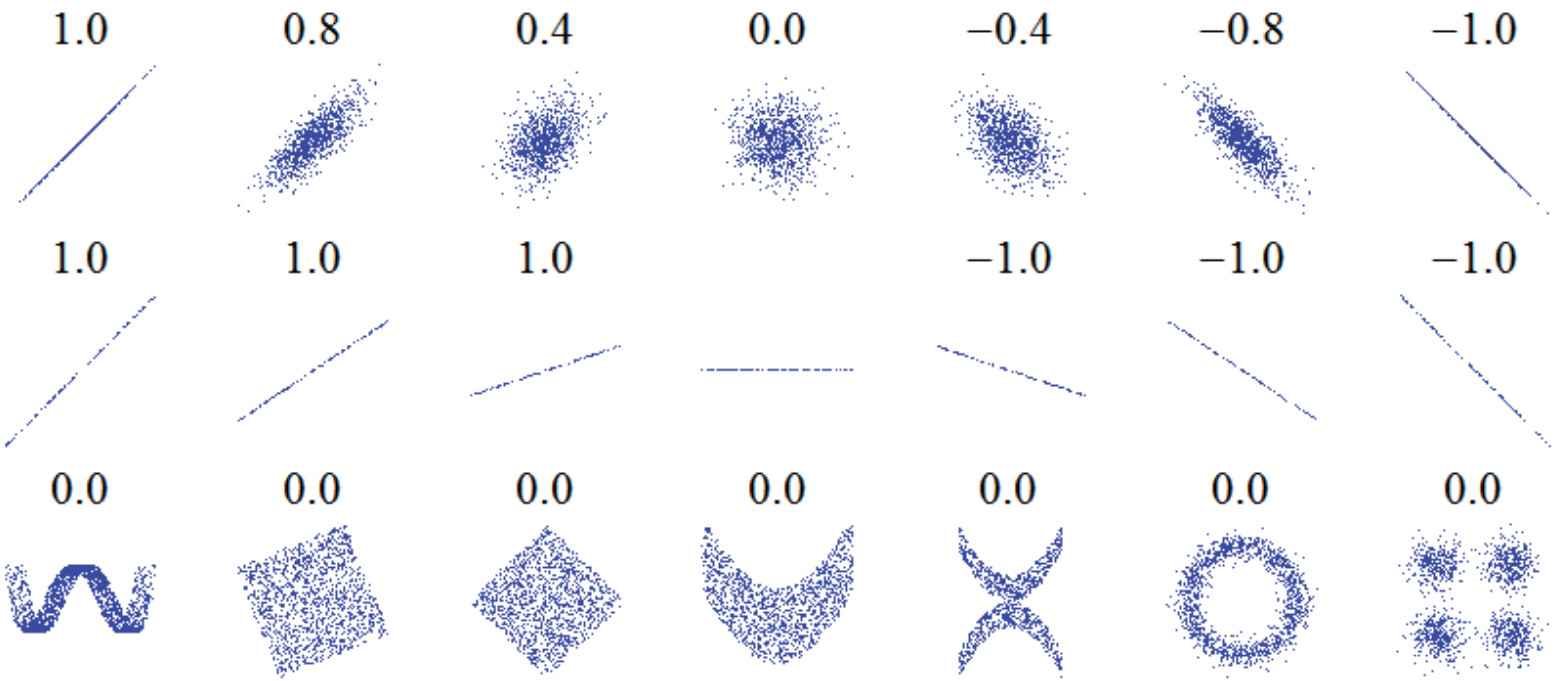
  $\rightarrow \quad 0.386$

- Weinkonsum: wir erwarten negativen Korrelationskoeffizienten, da mit steigendem Weinkonsum die Mortalität sinkt:

  $\rightarrow \quad -0.746.$

Empirische Korrelation: Bemerkungen

- Korrelation misst „nur“ den **linearen Zusammenhang**
- Man sollte daher die Daten immer auch anschauen, statt sich „blind“ auf Kennzahlen zu verlassen



Empirische Korrelation: Bemerkungen