

Stochastik (AIN)

Prof. Dr. Barbara Staehle

HTWG Konstanz
Fakultaet für Informatik

SS 2021

Teil I

Beschreibende Statistik

Teil I Beschreibende Statistik

1. Charakterisierung einer Stichprobe

- 1.1 Grundbegriffe
- 1.2 Häufigkeitsverteilung einer Stichprobe
- 1.3 Kennwerte einer Stichprobe

2. Multivariate Statistik (v3 only)

- 2.1 Lineare Korrelation (v3 only)
- 2.2 Lineare Regression (v3 only)

Abschnitt 1

Charakterisierung einer Stichprobe

Statistik - worum geht es?

- **Hauptaufgabe:** Informationen über bestimmte Objekte gewinnen, ohne alle Objekte untersuchen zu müssen.
- **Methode:** Daten über eine Stichprobe erheben, auswerten und Schlussfolgerungen ableiten.
- **Grundbegriffe:**

Statistische Einheiten Objekte, an denen interessierende Größen beobachtet und erfasst werden.

Grundgesamtheit alle statistischen Einheiten, über die man Aussagen gewinnen möchte.

Stichprobe tatsächlich untersuchte Teilmenge der Grundgesamtheit.

Merkmal (Variable) interessierende Größe, die an den statistischen Einheiten in der Stichprobe beobachtet (gemessen, erhoben) wird.

(Merkmals)Ausprägungen die verschiedenen Werte, die jedes Merkmal annehmen kann.

Beispiele zur Terminologie

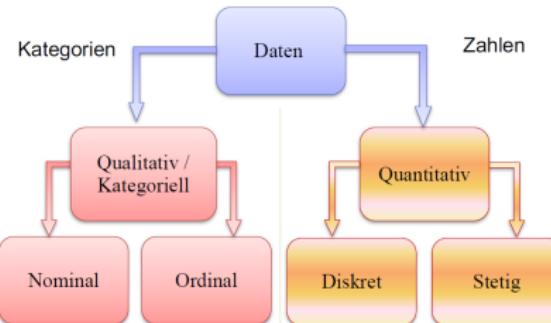
Frage: Was kostet ein WG-Zimmer in Konstanz?

- **Statistische Einheiten:** WG-Zimmer
- **Grundgesamtheit:** alle existierenden WG-Zimmer in Konstanz
- **Stichprobe:** Menge der WG-Zimmer, über die wir Daten erheben können (z.B. Umfragen)
- **Merkmale:** Miete, Größe, Stadtteil, Anzahl Mitbewohner, ...
- **Merkmalsausprägungen:** 100-400 €, 7-15 m², {Paradies, Petershausen, Fürstenberg, . . . , }, 1-5, ...

Frage: Wie „gut“ ist eduroam an der HTWG?

- **Statistische Einheiten:** über eduroam übertragene Pakete
- **Grundgesamtheit:** alle über eduroam übertragene Pakete
- **Stichprobe:** Pakete die während eines Tages von meinem Laptop übertragen werden
- **Merkmale:** Übertragungserfolg, -verzögerung, Paketwiederholungen, Reihenfolge ...
- **Merkmalsausprägungen:** ja/nein, 0.1-10 s, 0-3, ja/nein ...

Eigenschaften von Merkmalen I



Quelle: Oliver Dürr

Ein Merkmal heißt

qualitativ wenn die endlich vielen möglichen Ausprägungen eine **Qualität** oder eine **Kategorie** wiedergeben (und nicht ein Ausmaß, also **keine Zahl**). Beispiele: Geschlecht, Übertragungserfolg, Stadtteil.

quantitativ wenn die Ausprägungen ein Ausmaß bzw. eine Intensität widerspiegeln. Die Ausprägungen sind **Zahlen** (mit oder ohne Maßeinheit). Beispiele: Miete, Zimmergröße, Übertragungsverzögerung, Anzahl Mitbewohner.

Eigenschaften von Merkmalen II

Ein quantitatives Merkmal heißt

diskret wenn es **endliche** viele oder abzählbar unendlich viele Ausprägungen hat, also gezählt werden kann. Beispiele: Anzahl Mitbewohner, Paketwiederholungen

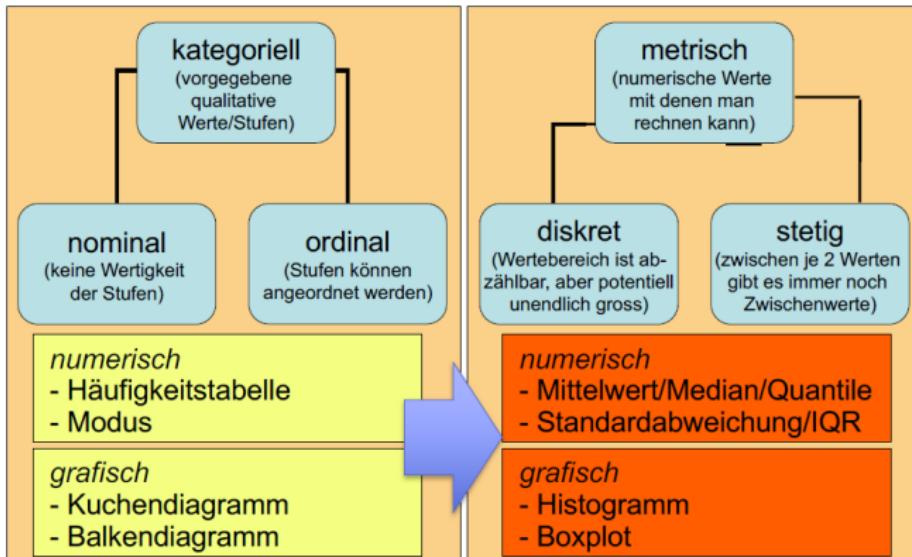
stetig wenn es alle Werte in einem reellen Intervall als Ausprägungen annehmen kann, also **überabzählbar unendlich** viele Ausprägungen gemessen werden können. Beispiel: Miete, Übertragungsverzögerung

Ein qualitatives Merkmal heißt

ordinal wenn sich seine Ausprägungen **in natürlicher Weise** anordnen lassen. Beispiele: wenig, mittel, viel

nominal wenn sich seine Ausprägungen **nicht anordnen** lassen. Beispiele: Stadtteile der WG

Typ-abhängige Darstellung von Merkmalen

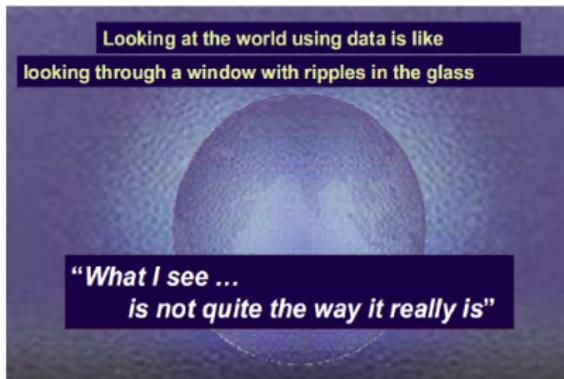


Quelle: Oliver Dürr

Schwerpunkt dieser Vorlesung: Quantitative Merkmale

Immer im Hinterkopf behalten!!

Stichprobe



Population



Quelle: Oliver Duerr, Chris Wild (University of Auckland)

Message to take: Die Untersuchung einer Stichprobe liefert nur ein ungefähres Bild der Wirklichkeit (der Grundgesamtheit). Die Beurteilung der Passgenauigkeit dieses Bildes ist Thema der schließenden Statistik (siehe Kapitel 4).

Im Folgenden verwendete Beispiele & HowTo

- Daten-Download:
 - ▶ Originalquellen (siehe unten): originale (teils große Datensätze)
 - ▶ meine [Homepage](#): verkleinerte Datensätze, CDC User's Guide, verwendete MATLAB-Skripte
- Ist es in Konstanz in den letzten Jahren wärmer geworden?
 - ▶ Stichprobe: tägliche Wetterdaten der Wetterstation Konstanz 1973-2019 (Archivdaten für 2020 noch nicht verfügbar)
 - ▶ Quelle: [Deutscher Wetterdienst](#), Konstanz, Tageswerte, historisch
 - ▶ produkt_klima_Tageswerte_19721101_20181231_02712.txt enthält durch Semikolon getrennte Spalten (csv-Format)
- Kommen erste Babies immer später? (Beispiel nach [Downey, 2014])
 - ▶ Stichprobe: Daten über 3.5 Mio Geburten aus den USA 2014
 - ▶ Quelle: [CDC/NHCS](#), Birth Data Files & User's Guide 2014
 - ▶ Achtung: Datei entpackt ist 5.3 GB groß, daher sollte man zu Testzwecken nur Datei mit dem ersten Zehntel der Daten verwenden
- Daten-Handling: MATLAB, Excel, [Python](#), Java, Octave, ...

Elektronische Helferlein

Bitte keine Statistiken per Hand erstellen (außer zur Übung). Benutzen Sie Taschenrechner oder Softwaretools!!

- TR: z.B. TI-30X Plus Multiview, Casio FX-991DE X ClassWiz
- MATLAB
 - ▶ MATLAB-Beispielcode zu den Bildern der Vorlesung [hier](#) verfügbar
 - ▶ **Vorteile:** Viele Toolboxen, weitverbreitet im technischen Umfeld für Sie dank Campuslizenz gratis erhältlich. Download via [RZ](#)
 - ▶ **Nachteil:** Für nicht-akademische Nutzer teuer; alternativ: Octave
- Python (Anaconda: all-in-one Distribution)
 - ▶ Schweizer Taschenmesser, nützlich auch für anderes
 - ▶ gute Bücher und online-Tutorials: [[Downey, 2014](#), [Haslwanter, 2016](#)]
 - ▶ **Vorteil:** Open Source, frei erhältlich, viele nützliche Pakete

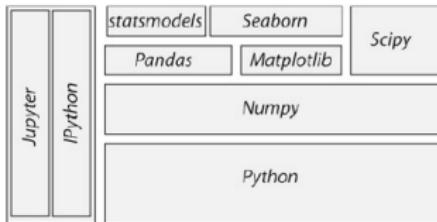
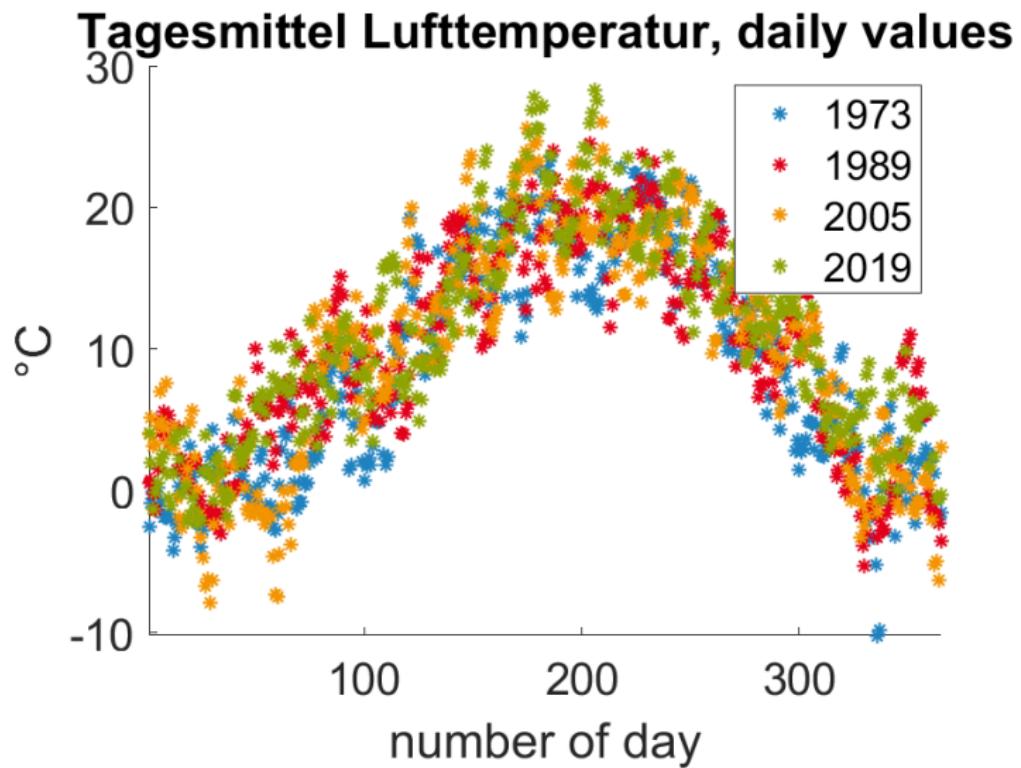


Bild: Wichtige Statistik Python-Pakete [[Haslwanter, 2016](#)]

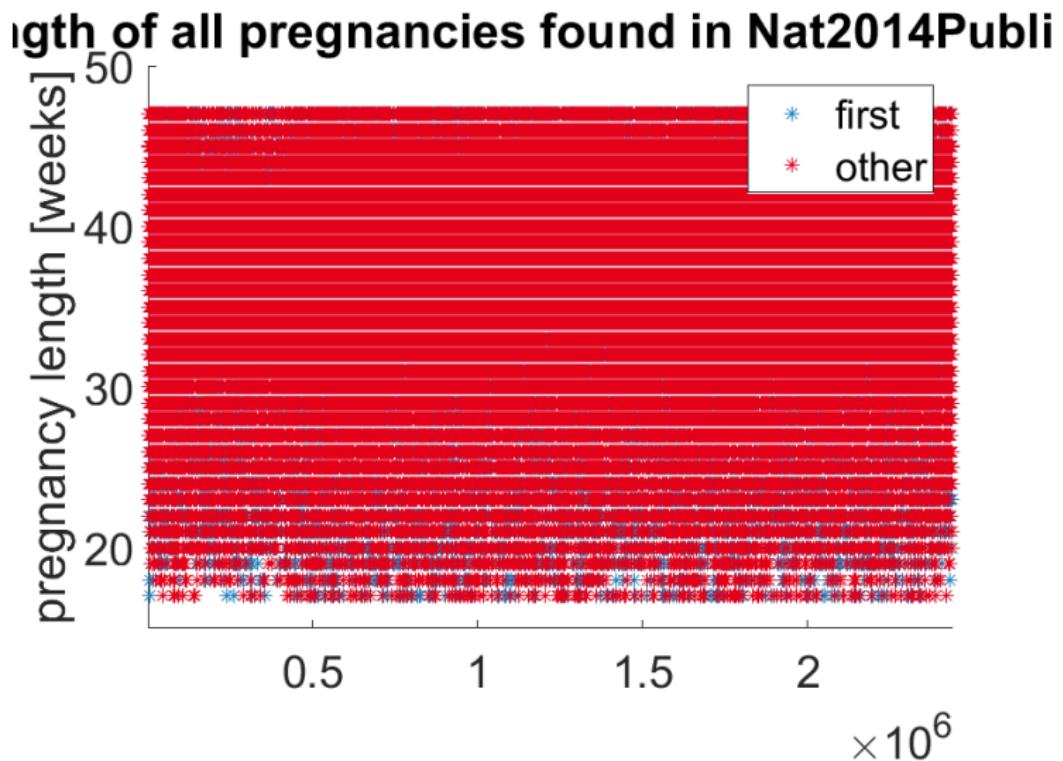
Darstellung und Analyse univariater Daten

- Im Folgenden: statistische Methoden zur Darstellung **univariater** Daten (d.h., Daten, die aus der Beobachtung eines einzigen Merkmals entstehen).
- Erweiterung / Anwendung dieser Methoden auf **multivariate** Daten (d.h., mehrere Merkmale, vor allem deren Zusammenhänge, werden gleichzeitig untersucht): nächster Abschnitt
- **Rohdatenanalyse**: Vergleich der Werte in der **Urliste** (Liste aller erhaltenen Messwerte)
 - ▶ Vorteil: schnell, einfach, erste Trends sind vielleicht schon sichtbar
 - ▶ Nachteil: nicht praktikabel für mehr als 100 Messwerte
- **graphische Rohdatenanalyse**: plotten aller erhaltenen Messwerte in zufälliger Reihenfolge oder z.B. nach Datum
 - ▶ Vorteil: schnell, einfach
 - ▶ Nachteil: unübersichtlich, vor allem für große Datenmengen

Graphische Rohdatenanalyse - Temperatur



Graphische Rohdatenanalyse - Babys



Absolute und relative Häufigkeit

- Analyse der Rohdaten zeigt mehrfach auftretende Werte nicht.
- Besser: verwende **absolute und relative Häufigkeiten**

1. Gegeben: **Urliste** (unsortierte Stichprobe) mit n Messwerten:
 x_1, x_2, \dots, x_n .
2. Ermittle die **verschiedenen auftretenden** Werte als a_1, a_2, \dots, a_k
(eliminiere mehrfach vorkommende).
3. Zähle für jeden Wert a_i dessen **absolute Häufigkeit** h_i in der Urliste.
4. Ermittle die **relative Häufigkeit** als $f_i = \frac{h_i}{n}$.
5. Plausibilitätscheck:

$$\sum_{i=1}^k h_i = n \quad \text{und} \quad \sum_{i=1}^k f_i = 1.$$

Analogie zur Wahrscheinlichkeitsrechnung (WR):

- untersuchtes Merkmal \approx Zufallsvariable.
- relative Häufigkeit von $a_i \approx$ Wahrscheinlichkeit von a_i .
- Zusammenstellung aller möglichen Werte zusammen mit ihrer relativen Häufigkeit \approx Wahrscheinlichkeitsverteilung (der ZV).

Beispiel: Ist der Würfel gezinkt?

Wir möchten herausfinden, ob der in der Spielbank verwendete Würfel gezinkt ist. Daher würfeln wir damit 20 Mal und erhalten folgende „Messwerte“

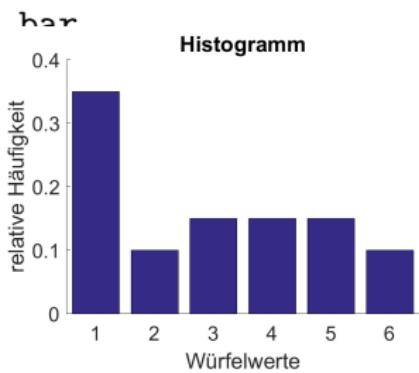
4, 1, 1, 5, 3, 3, 1, 4, 5, 6, 1, 4, 1, 6, 1, 2, 1, 2, 3, 5.

1. Die einzelnen Werte sind $a_1 = 1, a_2 = 2, \dots, a_6 = 6$
2. Die Häufigkeiten ermitteln wir per Hand und Strichliste oder MATLAB und `hist`, `histogram`.

Darstellung von (relativen) Häufigkeiten

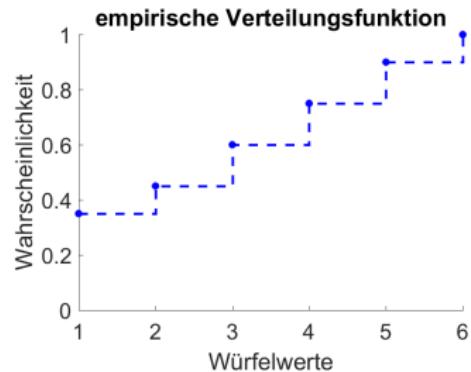
Balken- oder Stabdiagramm

- auch: **Histogramm**
- x-Achse: a_i (Werte der ZV)
- y-Achse: h_i/f_i (absolute oder relative Häufigkeiten) als Säule / Stab / Balken
- MATLAB: `hist`, `histogram`, `bar`

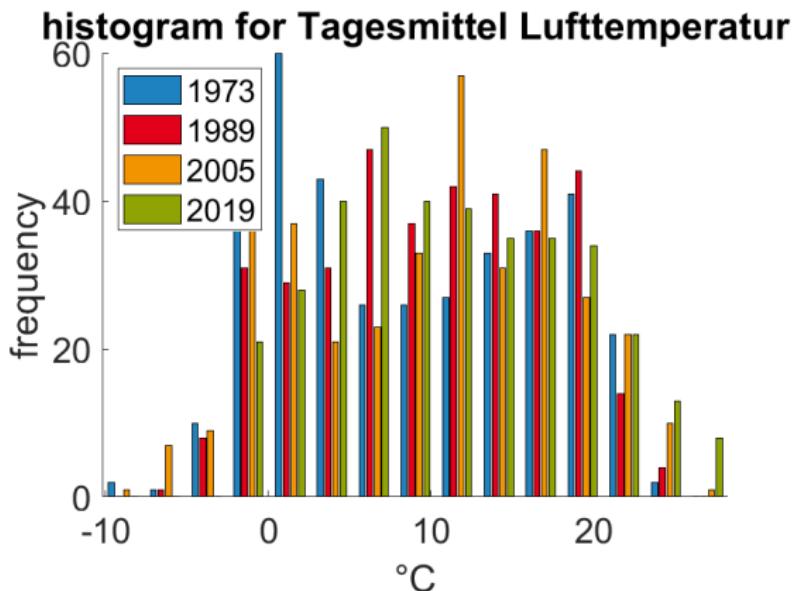


empirische Verteilungsfunktion

- berechne $\bar{F}(x) = \sum_{i: a_i \leq x} f_i$
- \bar{F} ist Stufenfunktion mit $\min > 0$, $\max = 1$, Sprünge bei a_i
- MATLAB: `cumsum`, `stairs`



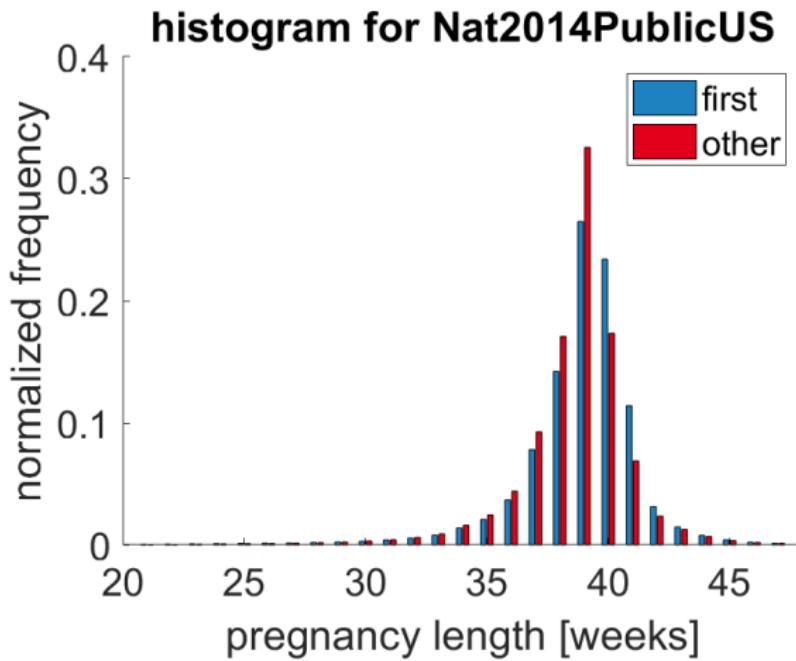
Histogramm - Temperatur



- 365 Messungen für jedes Jahr, daher Vergleich absoluter Häufigkeiten möglich
- Darstellung der Häufigkeit von **Klassen** bzw. **Intervallen** (z.B. $26.5^\circ - 28.5^\circ$) statt einzelner Werte
- 5-20 gleich große Intervalle wählen (Tradeoff Genauigkeit-Lesbarkeit)

Analogie zur WR: Vergleich der Verteilungen mehrerer verschiedener Zufallsvariablen

Histogramm - Babys

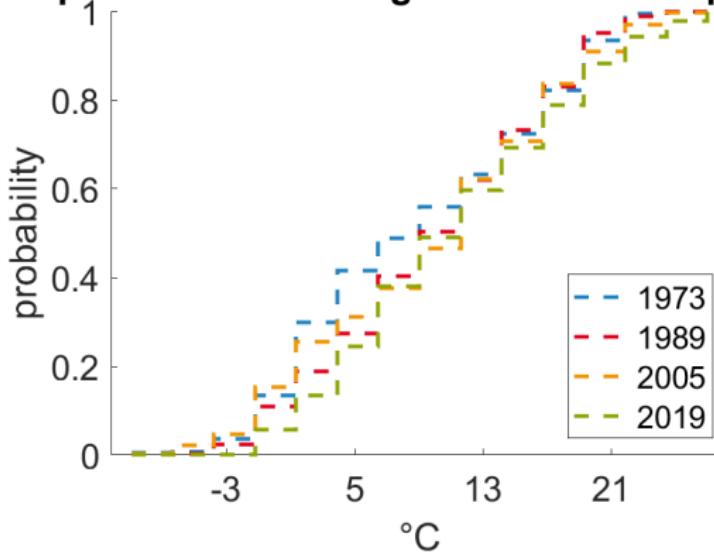


- unterschiedlich große Datensätze, daher Vergleich relativer (statt absoluter) Häufigkeiten
- Wertebereich $\{x \in \mathbb{N} \mid 17 \leq x \leq 47\}$ ohne Klassen darstellbar

Analogie zur WR: Vergleich von Verteilungen

Empirische Verteilungsfunktion - Temperatur

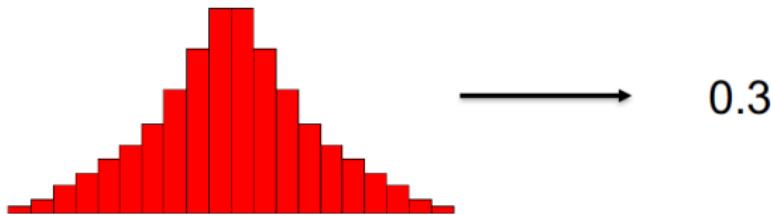
empirical CDF for Tagesmittel Lufttemperatur



- CDF = **cumulative** distribution function
- „empirisch“ weil nur die beobachteten Häufigkeiten aufsummiert werden
- $\bar{F}(x)$ zeigt Wahrscheinlichkeit, dass Temperatur $\leq x$ ist
- eine Kurve liegt unter der anderen: Werte dieser Stichprobe sind größer

Analogie zur WR: Vergleich von Verteilungsfunktionen verschiedener Zufallsvariablen

Statistische Kennzahlen



Quelle: Oliver Dürr

- absolute / relative Häufigkeiten beschreiben Stichprobe **vollständig**
- Kompaktere Beschreibung durch **Kennwerte**:
 - ▶ **Lagekennwerte** geben Information darüber, wo die Werte der Stichprobe typischerweise liegen
⇒ arithmetisches Mittel, Median, Quantile, Modalwert
 - ▶ **Streuungskennwerte** beschreiben, ob die Stichprobenwerte an einer Stelle konzentriert sind oder ob sie stark streuen (stark verteilt sind)
⇒ Varianz, Standardabweichung, Spannweite, Interquartilabstand

Das arithmetische Mittel I



Was will der Autor damit sagen?

Lösung: Im Mittel oder durchschnittlich muss man 4844 Sicker (ca. 970 Tütchen) kaufen, bis man das Album voll hat.

Quelle: Der SPIEGEL, 06/2018

Achtung: Der Mittelwert gibt lediglich das Zentrum der Verteilung an und sagt nichts darüber aus, wie sehr die Werte von ihm abweichen!

Das arithmetische Mittel II

Definition

Sei x_1, \dots, x_n eine Stichprobe mit den verschiedenen Werten a_1, \dots, a_k und den absoluten Häufigkeiten h_1, \dots, h_k bzw. relativen Häufigkeiten f_1, \dots, f_k . Das (arithmetische) Mittel (oder Mittelwert) der Stichprobe ist

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{n} \sum_{i=1}^k h_i a_i = \sum_{i=1}^k f_i a_i.$$

Algorithmus: Alle Werte aufsummieren und durch ihre Anzahl teilen.

Analogie zur WR: Erwartungswert einer Zufallsvariable.

Beispiel: Angenommen Alice hat 5 € im Portemonnaie, Bob gar nichts und Charlie 295 €. Wie viel Geld haben die drei dann im Mittel in ihren Portemonnaies?

Der Median

Definition

Der **Median** (auch **Zentralwert**) einer geordneten Stichprobe x_1, \dots, x_n ist

$$\tilde{x} = \begin{cases} x_{m+1} & \text{falls } n = 2m + 1, \\ \frac{1}{2}(x_m + x_{m+1}) & \text{falls } n = 2m. \end{cases}$$

Bemerkungen:

- \tilde{x} wird als „ x Schlange“ oder „ x Tilde“ gelesen.
- \tilde{x} teilt die Stichprobe in der Mitte: 50% der Stichprobenwerte sind kleiner gleich, 50% sind größer oder gleich \tilde{x} .
- **Algorithmus:**
 1. Ordne die n Stichprobenwerte der Größe nach.
 2. **n ungerade:** \tilde{x} ist der Wert in der Mitte der Liste.
 n gerade: \tilde{x} ist das arithmetische Mittel der zwei Werte in der Mitte der Liste.

Beispiel: Was ist der Median der Portemonnaie-Inhalte 5,0, 295 €?

Median ist robust gegenüber Ausreißern!



»Sollen wir das arithmetische Mittel als durchschnittliche Körpergröße nehmen und den Gegner erschrecken, oder wollen wir ihn einlullen und nehmen den Median?«

Quelle: [Krämer, 2015]

Der Modalwert

Der **Modalwert** gibt den am häufigsten auftretenden Stichprobenwert an.

Vorteil: auch für nicht-numerische Stichproben verwendbar.

Beispiel: Stichprobe „rot, rot, grün, blau, blau, blau, blau, blau“ hat Modalwert „blau“.

Bemerkung: Kommen mehrere Werte am häufigsten vor, heißt die Stichprobe **multimodal** (**bimodal**, falls es zwei Modi gibt)

Beispiel: Aus denen vor einer Turnhalle abgestellten $n = 10$ Schuhen wird jeweils die Größe ermittelt:

32 30 33 31 30 30 32 33 34 45

Bestimmen Sie den Mittelwert, Median und den Modalwert der Stichprobe.

Fazit: statistische **Ausreißer** (Werte die deutlich kleiner bzw. größer sind als alle anderen) beeinflussen den Mittelwert stark, den Median nicht.

Feinere Lagekennwerte: Quantile & Co

Definition

Für die **geordneten** Stichprobenwerte x_1, \dots, x_n und $0 < p < 1$ heißt

$$\tilde{x}_p = \begin{cases} x_{\lceil np \rceil} & \text{falls } np \notin \mathbb{N}, \\ \frac{1}{2}(x_{np} + x_{np+1}) & \text{falls } np \in \mathbb{N} \end{cases}$$

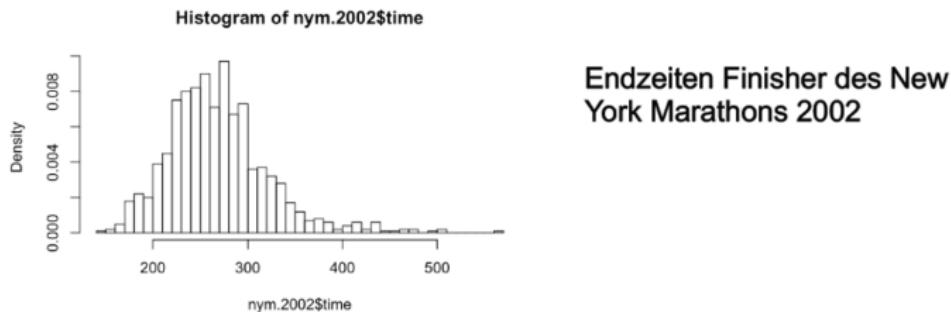
das **p-Quantil**.

- Das 0.5-Quantil ist genau der Median: $\tilde{x}_{0.5} = \tilde{x}$.
- $\tilde{x}_{0.25}, \tilde{x}, \tilde{x}_{0.75}$ werden als **Quartile** bezeichnet.
- $\tilde{x}_{0.1}, \tilde{x}_{0.2}, \dots, \tilde{x}_{0.9}$ heißen **Dezile**.
- $\tilde{x}_{0.01}, \tilde{x}_{0.02}, \dots, \tilde{x}_{0.99}$ heißen **Perzentile**.

Bemerkung: Ein p-Quantil zerlegt die geordnete Stichprobe in zwei Teile: Mindestens ein Anteil p der Stichprobenwerte ist kleiner oder gleich \tilde{x}_p und mindestens ein Anteil $1 - p$ ist größer oder gleich \tilde{x}_p .

Analogie zur WR: Quantil einer Zufallsvariable ist genauso definiert; wird durch Invertierung der Verteilungsfunktion berechnet.

Anwendung von Quantilen



Frage:

- Wie schnell muss ich laufen, so dass ich genau im Mittelfeld liege?

Median = 256.02 Minuten = 4:16 Stunden

- Wie schnell muss ich laufen, so dass ich zu den 10% schnellsten gehöre?

10%-Quantil=201.96 = 3:22 Stunden

Quelle: Oliver Dürr

Funktioniert auch für: Altstadtlauf Konstanz, Halbmarathon Singen, Schwarzwaldmarathon, ...

Beispiele zum Mitdenken (TR verwenden!!)

Sie messen in einem drahtlosen Sensornetz die Größe von $n = 10$ Paketen in Bits und erhalten Messwerte: 14, 24, 22, 19, 18, 36, 15, 29, 41, 17.

Bestimmen Sie den Mittelwert, die Quartile sowie das 90% Quantil.

Varianz und Standardabweichung I

Problem aller Lagewerte: Ungeeignet um **Streuung** zu repräsentieren.

Definition

Für die Stichprobe x_1, \dots, x_n gibt die **(Stichproben-)Varianz** oder **empirische Varianz** an, wie sehr die Stichprobenwerte x_i um ihren Mittelwert \bar{x} streuen:

$$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Als Maß für die Streuung wird auch die Wurzel der Varianz verwendet

$$s = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2},$$

die so genannte **(Stichproben-)Standardabweichung** oder **empirische Standardabweichung**.

Varianz und Standardabweichung II

Bemerkungen:

- Zur Berechnung von s^2 dividiert man durch $n - 1$ und nicht durch n . Grund: die durch $n - 1$ dividierte Summe ist ein besserer Schätzer für die tatsächliche Varianz der Grundgesamtheit als die durch n dividierte Summe.
- Je kleiner s bzw. s^2 , desto stärker sind die Messwerte um die Mittelwert konzentriert. Extremfall: $s^2 = 0$, falls alle Messwerte identisch sind.
- Für die k verschiedenen Werte der Stichprobe a_i und deren Häufigkeiten h_i gilt

$$s^2 = \frac{(\sum_{i=1}^n x_i^2) - n \cdot \bar{x}^2}{n - 1} = \frac{(\sum_{i=1}^k h_i a_i^2) - n \cdot \bar{x}^2}{n - 1}.$$

Analogie zur WR: Varianz und Standardabweichung einer Zufallsvariable sind **ähnlich** definiert: Zur Berechnung der **Varianz einer Zufallsgröße σ** dividiert man durch **n** , zur Berechnung der **Stichprobenvarianz** dividiert man durch **$n - 1$** .

Beispiele zum Mitdenken

Bestimmen Sie die absolute Häufigkeit aller auftretenden Werte.

Berechnen Sie Mittelwert, Median, Varianz und Standardabweichung der Stichprobe

1, 1, 2, 2, 3, 3, 3, 3, 4, 4, 7.

Beispiele zum Mitdenken

Gegeben sind zwei Datensätze. Berechnen Sie Mittelwert, Median, Varianz und Standardabweichung und interpretieren Sie diese

- $dat_1 = -2, -1, 0, 1, 2$
- $dat_2 = -20, -10, 0, 10, 20$

Weitere Streuungsmaße

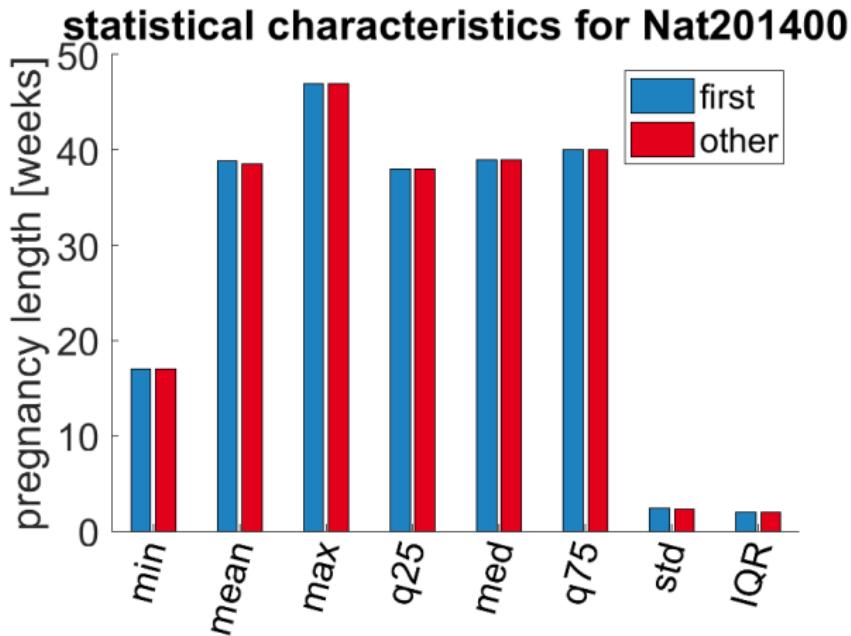
Spannweite $R = x_{max} - x_{min}$ (Differenz von größtem und kleinstem Stichprobenwert); **Vorteil:** einfach zu berechnen, **Nachteil:** starke Beeinflussung durch Ausreißer.

Interquartilabstand $I = \tilde{x}_{0.75} - \tilde{x}_{0.25}$ (Differenz zwischen 75% und 25% Quantil); **Vorteil:** resistent gegen Ausreißer, **Nachteil:** aufwändiger zu berechnen. Abkürzung: IQR.

Beispiel: Berechnen Sie Spannweite und IQR der Stichprobe

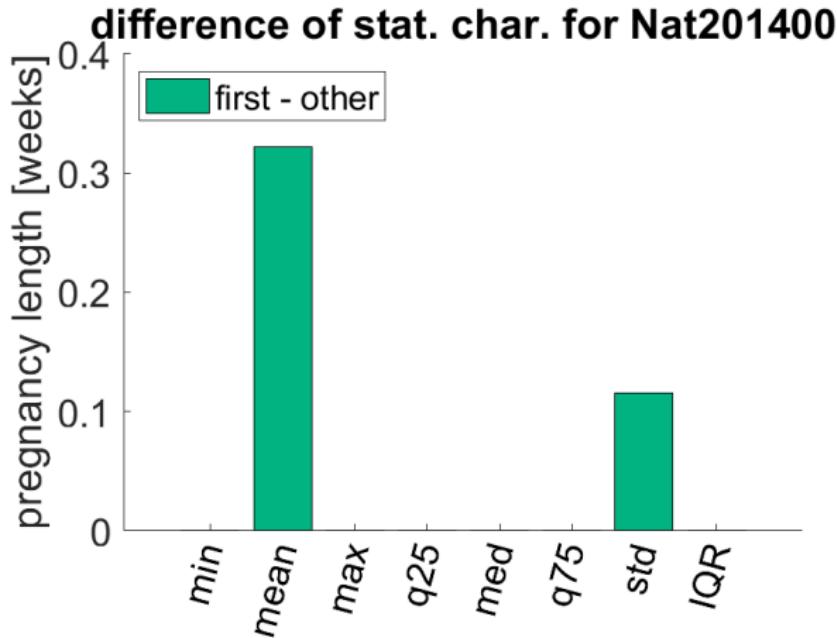
1, 1, 2, 2, 3, 3, 3, 3, 4, 4, 7.

Vergleich statistischer Kennwerte - Babys I



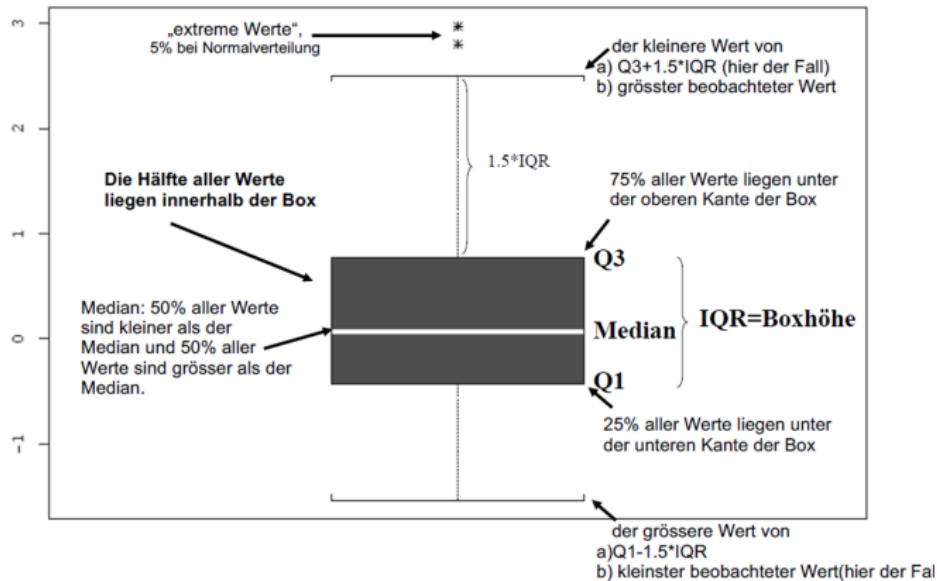
- hinsichtlich der Lage- und Streuungskennwerte scheint kaum ein Unterschied sichtbar

Vergleich statistischer Kennwerte - Babys II



- Analyse der Differenz zeigt, dass erste Babys im Mittel tatsächlich 0.32 Wochen (≈ 2.25 Tage) später kommen.
- Allerdings ist die Standardabweichung höher, daher ist Aussage wenig verlässlich.

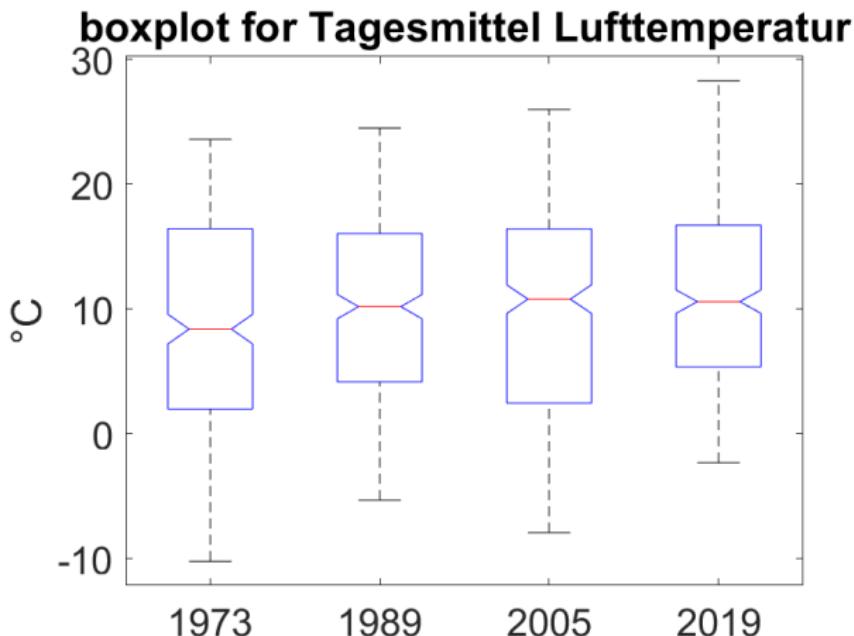
Boxplots



Quelle: Oliver Dürr

Anmerkung: Boxplots stellen viele Informationen dar, müssen aber gut erklärt werden! Python: `seaborn.boxplot` (z.B.), MATLAB: `boxplot`

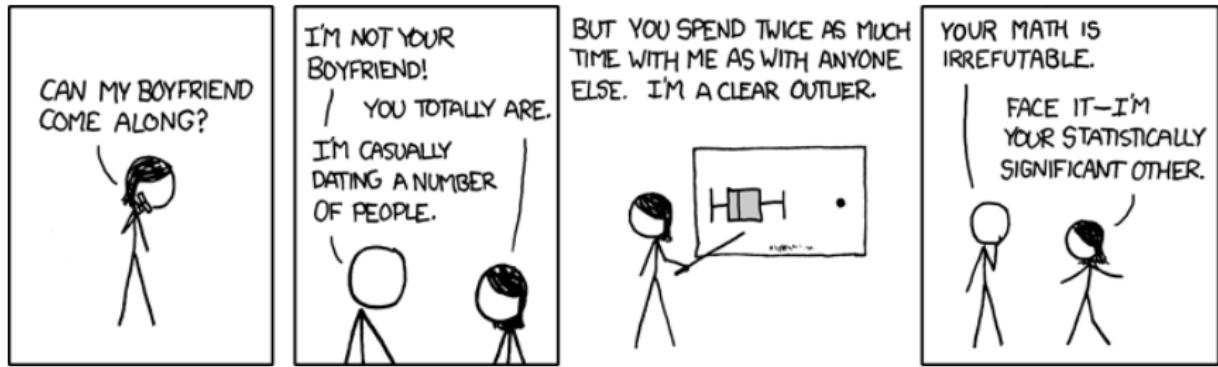
Boxplot - Temperaturen



- MATLAB: Box geht von $\tilde{x}_{0.25}$ bis $\tilde{x}_{0.75}$, Median ist markiert, die „Antennen“ oder „Whiskers“ oben und unten geben das 1.5-fache der IQR an.
- Boxplot für Baby-Problem unübersichtlich aufgrund zu vieler Ausreißer

Fazit: In Konstanz ist es in den letzten Jahren tatsächlich etwas wärmer geworden.

Boxplots in der Anwendung I



Quelle: xkcd.com

Boxplots in der Anwendung II

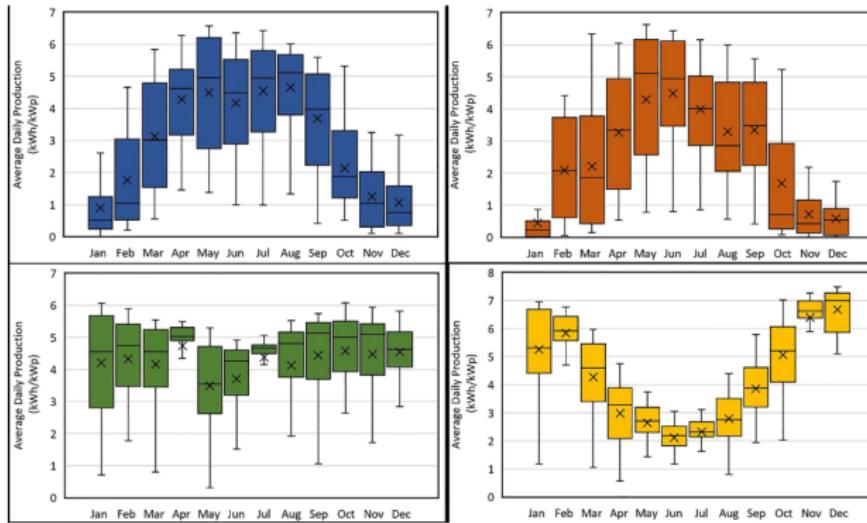


FIGURE 5 Box plot showing the average, minimum, and maximum daily photovoltaic (PV) production in 2016 for each location at an optimized tilt angle (see Table 1). Clockwise from top left: Amsterdam (NL), Oslo (NO), Perth (AU), and São Paulo (BR). The X represents the mean (average), and the line in the middle of the boxes represents the median [Colour figure can be viewed at wileyonlinelibrary.com]

Quelle: Rodriguez, de Santana et al: Feasibility study of solar PV-powered electric cars using an interdisciplinary modeling approach for the electricity balance, CO₂ emissions, and economic aspects: The cases of The Netherlands, Norway, Brazil, and Australia, *Progress in Photovoltaics: Research and Applications*, 2020

Abschnitt 2

Multivariate Statistik (v3 only)

Motivation

Bisher: Betrachtung **eines** Merkmals einer Stichprobe.

Jetzt: Betrachtung **zweier** Merkmale einer Stichprobe.

Ziel: Finden eines Zusammenhangs zwischen den Merkmalen.

Beispiel: Von 15 zufällig ausgewählten erwachsenen Personen werden Größe [cm], Gewicht [kg], monatliches Nettoeinkommen [€] und die Bestzeit über 10km (Laufen) [min] erhoben.

Frage: Gibt es Zusammenhänge zwischen der Körpergröße und den anderen Größen? Wenn ja welche?

Größe	Gewicht	Einkommen	Bestzeit
163	59	3900	55
165	62	2100	52
166	65	3600	45
169	69	2300	40
170	65	4000	40
171	69	5600	56
171	76	2100	51
173	73	5100	45
174	75	3400	35
175	73	1800	52
177	80	2100	50
177	71	2600	40
179	82	4600	42
180	84	3600	36
185	85	2700	38

Streudiagramme

Einfache graphische Möglichkeit zur Darstellung eines Zusammenhangs:
Streudiagramme oder **Punktwolken**.

Methode: Stelle von allen Paaren den x -Wert auf der x -Achse dar, den y -Wert auf der y -Achse.

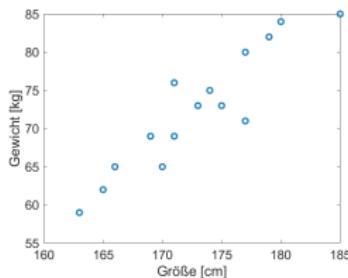


Bild: Größe vs. Gewicht

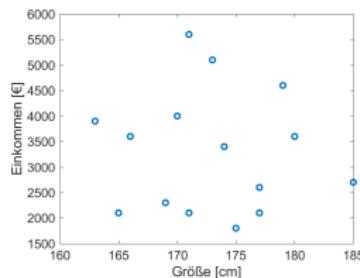


Bild: Größe vs. Einkommen

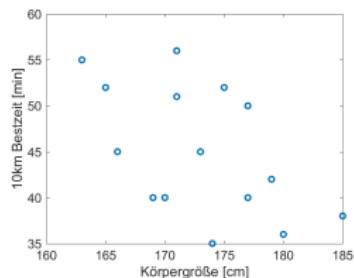


Bild: Größe vs. 10km Bestzeit

Der empirische Korrelationskoeffizient I

Definition

Gegeben seien die Wertepaare $(x_1, y_1), \dots, (x_n, y_n)$ wobei nicht alle x_i gleich sind bzw. nicht alle y_i gleich sind. Die Zahl

$$r_{x,y} = \frac{s_{x,y}}{s_x s_y}$$

heißt **(empirischer) Korrelationskoeffizient**. Dabei ist

$$s_{x,y} = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

die **(empirische) Kovarianz**, \bar{x}, \bar{y} sind die arithmetischen Mittelwerte und

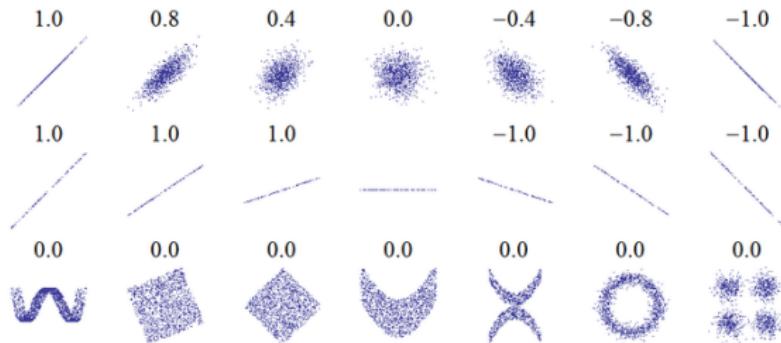
$$s_x = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \quad s_y = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2}$$

sind die **(empirischen) Standardabweichungen** der x_i bzw. der y_i -Werte.

Analogie zur WR: Korrelationskoeffizient und Kovarianz zweier ZV.

Der empirische Korrelationskoeffizient II

Bemerkung: Der empirische Korrelationskoeffizient (Pearson Koeffizient) quantifiziert nur den **linearen** Zusammenhang zwischen zwei Stichproben:



Quelle: Wikipedia

Eigenschaften des Korrelationskoeffizienten

- $-1 \leq r_{x,y} \leq 1$
- Falls $r_{x,y} > 0$, sind die Stichproben **(linear)** positiv korreliert
- Falls $r_{x,y} < 0$, sind die Stichproben **(linear)** negativ korreliert
- Falls $r_{x,y} = 0$, sind die Stichproben **(linear)** unkorreliert
- Falls $|r_{x,y}| = 1$, besteht perfekte lineare Abhangigkeit.

Korrelationskoeffizient am kleinen Beispiel I

Berechnen Sie für die ersten 3 Werte des vorherigen Beispiels den Korrelationskoeffizient zwischen Größe und Einkommen.

Erinnerung: $r_{x,y} = \frac{s_{x,y}}{s_x s_y}$

Größe	Gewicht	Einkommen	Bestzeit
163	59	3900	55
165	62	2100	52
166	65	3600	45

Rechnung: $n = 3, x = g, y = e$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{3}(163 + 165 + 166) = 164\frac{2}{3}$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{3}(3900 + 2100 + 3600) = 3200$$

$$\begin{aligned}s_x &= \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \\&= \sqrt{\frac{1}{2}((163 - 164\frac{2}{3})^2 + (165 - 164\frac{2}{3})^2 + (166 - 164\frac{2}{3})^2)} \\&= \sqrt{\frac{7}{3}} = 1.528\end{aligned}$$

Korrelationskoeffizient am kleinen Beispiel II

Berechnen Sie für die ersten 3 Werte des vorherigen Beispiels den Korrelationskoeffizient zwischen Größe und Einkommen.

Erinnerung: $r_{x,y} = \frac{s_{x,y}}{s_x s_y}$

Größe	Gewicht	Einkommen	Bestzeit
163	59	3900	55
165	62	2100	52
166	65	3600	45

Rechnung:

$$\begin{aligned}s_y &= \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} \\ &= \sqrt{\frac{1}{2}((3900 - 3200)^2 + (2100 - 3200)^2 + (3600 - 3200)^2)} \\ &= \sqrt{930000} = 964.365\end{aligned}$$

$$\begin{aligned}s_{x,y} &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\ &= \frac{1}{2}((163 - 164\frac{2}{3})(3900 - 3200) \\ &\quad + (165 - 164\frac{2}{3})(2100 - 3200) + (166 - 164\frac{2}{3})(3600 - 3200)) \\ &= \frac{1}{2}(-1166\frac{2}{3} - 366\frac{2}{3} + 533\frac{1}{3}) \\ &= -500\end{aligned}$$

$$r_{x,y} = \frac{s_{x,y}}{s_x s_y} = \frac{-500}{1.528 \cdot 964.365} = -0.339$$

Korrelationskoeffizient des ganzen Beispiels

Betrachten Sie das alle 15 Datenpaare des Beispiels und berechnen Sie

- $r_{g,w} = 0.925$
- $r_{g,e} = -0.067$
- $r_{g,b} = -0.530$

Methodik

per Hand Vorteil: funktioniert immer; Nachteil: extrem langwierig

TR x, y in L1, L2 eingeben, dann „2-Var Stats“ (TI)

EXCEL KORREL(A1:A15;B1:B15)

MATLAB corrcoef(dat(:,1),dat(:,2)); liefert Kovarianzmatrix mit Korrelationskoeffizient außerhalb der Hauptdiagonale

Python seaborn.regplot

Größe (g)	Gewicht (w)	Einkommen (e)	Bestzeit (b)
163	59	3900	55
165	62	2100	52
166	65	3600	45
169	69	2300	40
170	65	4000	40
171	69	5600	56
171	76	2100	51
173	73	5100	45
174	75	3400	35
175	73	1800	52
177	80	2100	50
177	71	2600	40
179	82	4600	42
180	84	3600	36
185	85	2700	38

Scheinkorrelationen (Spurious Correlations)

Beispiel: Störche bringen tatsächlich Babys!

(Quelle: R. Matthews, "Storks deliver babies ($p= 0.008$)")

Auswertung von Daten aus 17 EU-Ländern ergibt einen starken linearen Zusammenhang. Bringen Störche also wirklich die Babys?

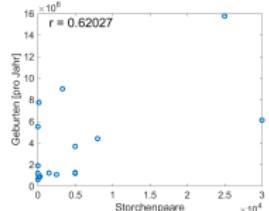
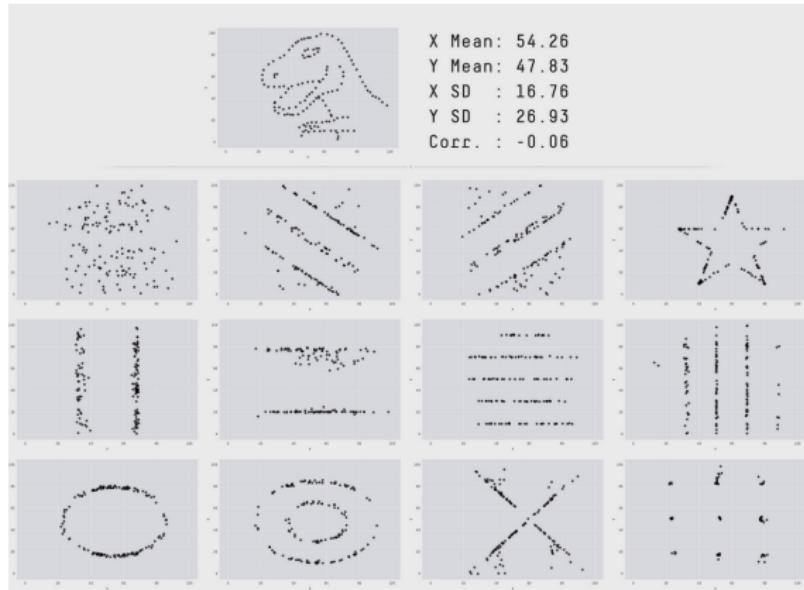


Bild: Störche vs. Geburten

Weitere Beispiele: Schuhgröße vs. Lesefähigkeit von Kindern, aggressiven Verhalten und Dauer der täglichen Beschäftigung mit Computerspielen, Anzahl von Piraten vs. Klimaerwärmung, ...

Traue nie nur dem Korrelationskoeffizienten!



Quelle: Autodesk Research

Erklärung: Die gezeigten Daten haben alle die gleichen Mittelwerte, Standardabweichungen und Korrelationen.

Daher: Daten immer visualisieren, nie **nur** den Kennzahlen trauen!

Korrelation vs. Regression

Korrelationsanalyse Untersuchung des **Ausmaßes des (linearen) Zusammenhangs** zwischen zwei Merkmalen x und y ; beide Merkmale sind **gleichrangig**, d.h., sowohl gemessene x -Werte als auch gemessene y -Werte können streuen.

Regressionsanalyse Untersuchung der **Art des (linearen) Zusammenhangs** zwischen zwei Merkmalen x und y ; beide Merkmale sind **nicht gleichrangig**, man betrachtet y als abhängig von x . Man geht davon aus, dass x festgehalten und exakt messbar ist und nur die y -Werte streuen.

Beispiel: **Korrelationsanalyse** zeigt einen deutlichen Zusammenhang zwischen Größe und Gewicht einer Person.

Regressionsanalyse untersucht, wie sich das Gewicht y in Abhängigkeit von der Größe x verändert und findet idealerweise eine Funktion, um aus einer gegebenen Größe x ein Gewicht y auszurechnen.

Lineare Regression

Lineare Regression ist das einfachstes Regressionsmodell: Ermittlung einer Regressionsgerade, die den mittleren quadratischen Fehler (die mittlere quadratische Abweichung der Messwerte von den Funktionswerten) minimiert (Gauß'sche Methode der kleinsten Quadrate).

Berechnung einer Regressionsgerade

Gegeben Wertepaare $(x_1, y_1), \dots, (x_n, y_n)$

Gesucht Gerade $f(x) = kx + d$

Bedingung minimiere den quadratischen Fehler $\sum_{i=1}^n (y_i - f(x_i))^2$

Lösung Berechne k und d als

$$k = r_{x,y} \frac{s_y}{s_x} = \frac{s_{x,y}}{s_x^2} \quad \text{und} \quad d = \bar{y} - k\bar{x}$$

Legende $r_{x,y}$: empirischer Korrelationskoeffizient, s_x, s_y : Standardabweichungen, \bar{x}, \bar{y} : arithmetische Mittelwerte, $s_{x,y}$: empirische Kovarianz

Lineare Regression am Beispiel I

Erinnerung: Von 15 zufällig ausgewählten erwachsenen Personen werden u.a. Größe [cm] g und Gewicht [kg] w erhoben. Graphische Auswertung und Korrelationsanalyse ($r_{g,w} = 0.925$) zeigten für Größe und Gewicht deutlich positiven Zusammenhang.

Aufgabe: Berechnen Sie (anhand der ersten 3 Werte) die Regressionsgerade, um für eine nicht in der Stichprobe vorhandene Größe ein passendes Gewicht vorherzusagen.

Gegeben $(x_1, y_1) = (163, 59), (x_2, y_2) = (165, 62), (x_3, y_3) = (166, 65)$

Gesucht Den quadratischen Fehler minimierende Gerade $f(x) = kx + d$

mit

$$k = r_{x,y} \frac{s_y}{s_x} = \frac{s_{x,y}}{s_x^2} \quad \text{und} \quad d = \bar{y} - k\bar{x}$$

Nebenrechnung I $n = 3$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i = \frac{1}{3}(163 + 165 + 166) = \mathbf{164\frac{2}{3}}$$

$$\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i = \frac{1}{3}(59 + 62 + 65) = \mathbf{62}$$

Lineare Regression am Beispiel II

Gegeben $(x_1, y_1) = (163, 59), (x_2, y_2) = (165, 62), (x_3, y_3) = (166, 65)$

Gesucht Den quadratischen Fehler minimierende Gerade $f(x) = kx + d$

mit

$$k = r_{x,y} \frac{s_y}{s_x} = \frac{s_{x,y}}{s_x^2} \quad \text{und} \quad d = \bar{y} - k\bar{x}$$

Nebenrechnung II

$$\begin{aligned}s_x &= \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2} \\&= \sqrt{\frac{1}{2} ((163 - 164\frac{2}{3})^2 + (165 - 164\frac{2}{3})^2 + (166 - 164\frac{2}{3})^2)} \\&= \sqrt{\frac{7}{3}} = 1.528\end{aligned}$$

$$\begin{aligned}s_y &= \sqrt{\frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2} \\&= \sqrt{\frac{1}{2} ((59 - 62)^2 + (62 - 62)^2 + (65 - 62)^2)} = \sqrt{9} = 3\end{aligned}$$

$$\begin{aligned}s_{x,y} &= \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \\&= \frac{1}{2} ((163 - 164\frac{2}{3})(59 - 62) + (165 - 164\frac{2}{3})(62 - 62) + (166 - 164\frac{2}{3})(65 - 62)) = \frac{1}{2}(5 + 0 + 4) = 4.5\end{aligned}$$

$$r_{x,y} = \frac{s_{x,y}}{s_x s_y} = \frac{4.5}{1.528 \cdot 3} = 0.982$$

Lineare Regression am Beispiel III

Gegeben $(x_1, y_1) = (163, 59), (x_2, y_2) = (165, 62), (x_3, y_3) = (166, 65)$

Gesucht Den quadratischen Fehler minimierende Gerade $f(x) = kx + d$

mit

$$k = r_{x,y} \frac{s_y}{s_x} = \frac{s_{x,y}}{s_x^2} \quad \text{und} \quad d = \bar{y} - k\bar{x}$$

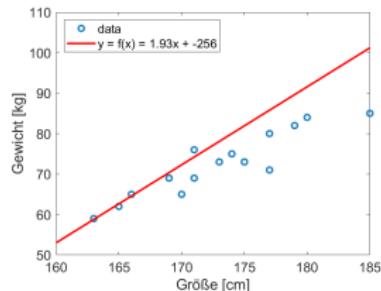
Zusammensetzen

$$\begin{aligned} k &= r_{x,y} \frac{s_y}{s_x} = 0.982 \cdot \frac{3}{1.528} = \mathbf{1.928} \\ d &= \bar{y} - k\bar{x} = 62 - 1.928 \cdot 164 \frac{2}{3} = \mathbf{-255.477} \end{aligned}$$

$$f(x) = kx + d = 1.982x - 255.477$$

Resultat: Gerade passt für die größeren Leute nicht mehr so gut

x_i	y_i	$f(x_i)$	$y_i - f(x_i)$
163	59	58.78	-0.212
165	62	62.644	0.644
166	65	64.572	-0.428
169	69	70.356	1.356
170	69	72.284	7.284



Lineare Regression am Beispiel IV

Methodik

per Hand Vorteil: funktioniert immer; Nachteil: extrem langwierig

EXCEL via Diagramm erstellen, Trendlinie hinzufügen

MATLAB `[r,k,d] = regression(x,y,'one')`

(Deep Learning Toolbox): liefert

Korrelationskoeffizient r , sowie Parameter k und d der Gerade;

`mdl = fitlm(x,y)` (Machine Learning Toolbox): ergibt ein Lineares Modell

TR x, y in L1, L2 eingeben, dann „2-Var Stats“ oder „LinReg ax+b“ (Texas Instruments)

Problem: Wie bewertet man die unterschiedliche gute Passung der Regressionsgeraden?

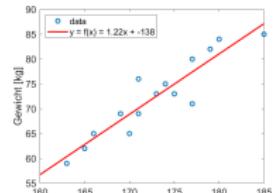


Bild: Größe vs. Gewicht

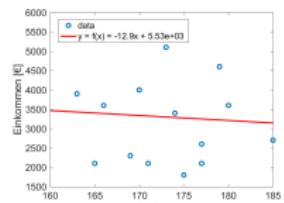


Bild: Größe vs. Einkommen

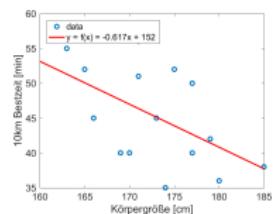


Bild: Größe vs. 10km Bestzeit

Die Qualität der Regressionsgeraden

Bestimmtheitsmaß $R^2 = r_{x,y}^2$ (Quadrat des Korrelationskoeffizienten) sagt aus, welcher Anteil der Variation in der abhängigen Variablen durch die Regressionsgerade erklärt werden kann.

Je größer R^2 , desto besser beschreibt die Gerade den Zusammenhang zwischen x und y .

Winkel zwischen den Regressionsgeraden für x und y
 $f(x) = kx + d$ bzw. $g(y) = k'y + d'$; kleiner Winkel visualisiert hohe Korrelation zwischen x und y , großer Winkel eine kleine Korrelation; Schnittpunkt im Daten-Schwerpunkt.

Probleme: Es gibt immer Datensätze für die R^2 nur eine begrenzte Aussagekraft hat.
 ⇒ lineare Regression ohne **optischen Vergleich von Daten und Gerade** ist grob fahrlässig!!

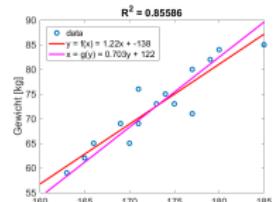


Bild: Größe vs. Gewicht

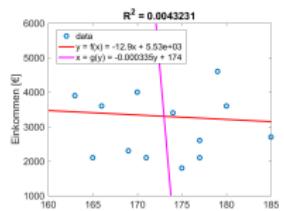


Bild: Größe vs. Einkommen

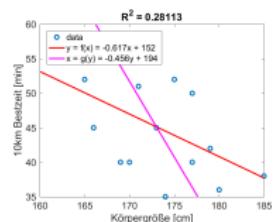


Bild: Größe vs. 10km Bestzeit

Graphische Datenanalyse - noch mehr Werbung

Das von Francis Anscombe konstruierte Anscombe-Quartett zeigt ebenso, dass graphische Datenanalyse wichtig ist (Quelle: [Wikipedia](#)).

Das Anscombe-Quartett									
I		II		III		IV			
x	y	x	y	x	y	x	y		
10,0	8,04	10,0	9,14	10,0	7,46	8,0	6,58		
8,0	6,95	8,0	8,14	8,0	6,77	8,0	5,76		
13,0	7,58	13,0	8,74	13,0	12,74	8,0	7,71		
9,0	8,81	9,0	8,77	9,0	7,11	8,0	8,84		
11,0	8,33	11,0	9,26	11,0	7,81	8,0	8,47		
14,0	9,96	14,0	8,10	14,0	8,84	8,0	7,04		
6,0	7,24	6,0	6,13	6,0	6,08	8,0	5,25		
4,0	4,26	4,0	3,10	4,0	5,39	19,0	12,50		
12,0	10,84	12,0	9,13	12,0	8,15	8,0	5,56		
7,0	4,82	7,0	7,26	7,0	6,42	8,0	7,91		
5,0	5,88	5,0	4,74	5,0	5,73	8,0	6,89		

Quelle: [Wikipedia](#)

Datensätze I-IV
bestehen jeweils
aus 11 (x, y)
Paaren

Eigenschaft	Wert
Mittelwert von x in jedem Fall	9 (exakt)
Varianz von x in jedem Fall	11 (exakt)
Mittelwert von y in jedem Fall	7,50 (auf 2 Stellen)
Varianz von y in jedem Fall	4,122 oder 4,127 (auf 3 Stellen)
Korrelation zwischen x und y in jedem Fall	0,816 (auf 3 Stellen)
Lineare Regression in jedem Fall	$y = 3,00 + 0,500x$ (auf 2 bzw. 3 Stellen)

Quelle: [Wikipedia](#)

Statistische
Kenngrößen (ebenso
 $R^2 = 0.667$) aller vier
Datensätze sind
identisch

Einflussfaktoren auf das Geburtsgewicht von Babys I

Frage: Welche Faktoren haben einen starken Einfluss auf das Geburtsgewicht eines Babys?

Methodik:

- Korrelations- und Regressionsanalyse des [CDC/NHCS-Datensatzes](#). (Verkleinerter Datensatz auch auf meiner [Homepage](#) verfügbar.)
- Alle ungültigen Daten sind herausgefiltert
- Verwendung eines Jitters für den Scatterplot.
- Bildtitel / Achsenbeschriftungen wie in der Original-Datei.

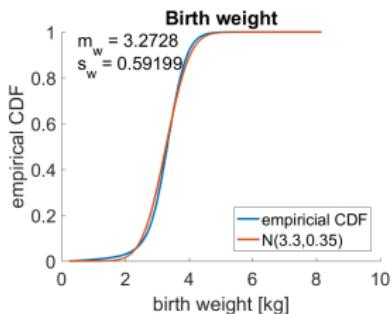
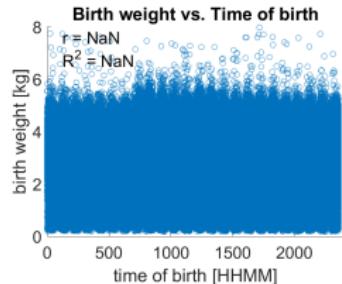
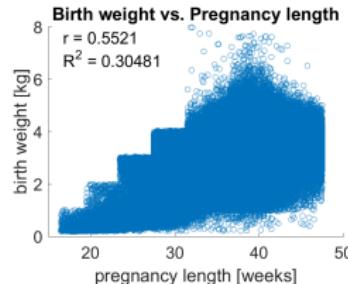
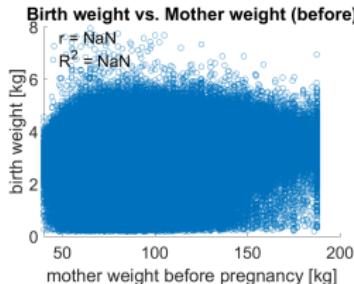
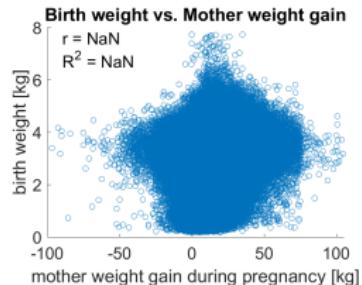
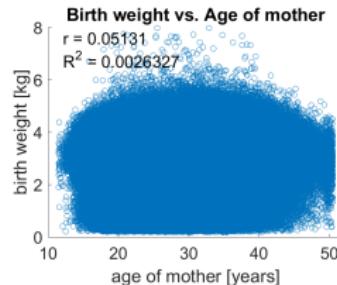
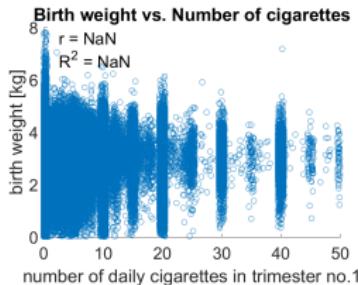


Bild: Geburtsgewicht ist approximativ normalverteilt

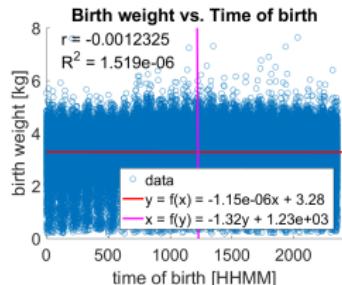
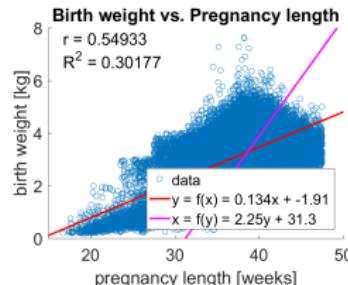
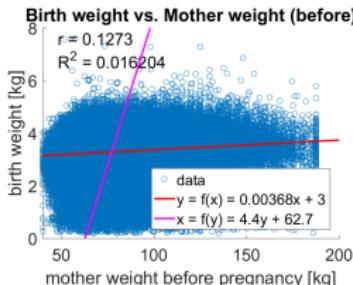
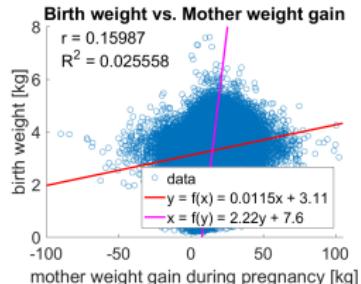
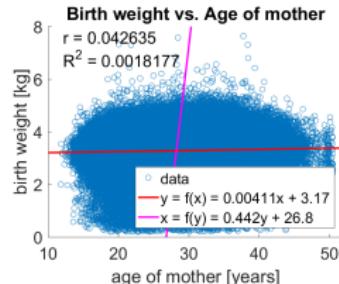
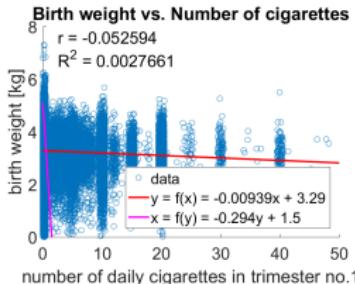
Ihre Aufgabe: Interpretation der Daten!

Einflussfaktoren auf das Geburtsgewicht von Babys II



Auswertung aller Daten (ca. 3.5 Mio Geburten)
daher ergibt Berechnung von r , R^2 manchmal NaN (not a number)

Einflussfaktoren auf das Geburtsgewicht von Babys III



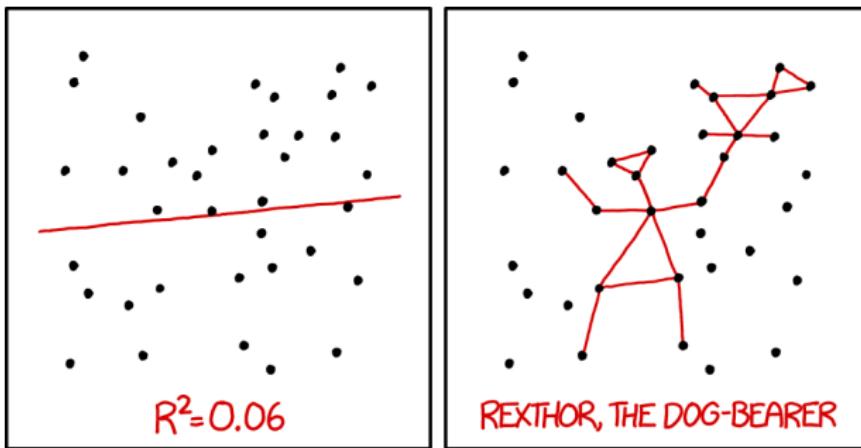
Auswertung eines Teils der Daten (ca. 0.35 Mio Geburten)

Einflussfaktoren auf das Geburtsgewicht von Babys IV

Einflussfaktoren auf das Geburtsgewicht von Babys V

Limitations of linear regression

Achtung: Lineare Regression ist **bei Weitem** nicht das einzige Werkzeug der Regressionsanalyse und liefert in **vielen Fällen** auch kein wirklich passendes Ergebnis!



I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER
TO GUESS THE DIRECTION OF THE CORRELATION FROM THE
SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

Bild: xkcd.com

Beyond linear regression

Weitere Möglichkeiten: (siehe z.B. [Papula, 2016], Wikipedia)

- Verwendung anderer Funktionen für die Regressionskurve, z.B. quadratische Funktionen, Polynome, Logarithmen, Exponentialfunktionen; Bestimmung der Parameter durch die Methode der kleinsten Quadrate
- Multiple lineare Regression: Herleitung eines Zusammenhangs zwischen mehreren (Input-)Parametern und einem Zielparameter
- mathematische komplexere Modelle
- **Fitting:** Finden einer beschreibenden Funktion z.B. durch MATLAB
 - ▶ Vorteil: Mathematik wird erledigt, Güte der Schätzung kann optisch überprüft werden.
 - ▶ Nachteil: A fool with a tool is still a fool.

Verwendete oder empfohlene Literatur I

[Diez et al., 2015] Diez, D. M., Barr, C. D., and Çetinkaya Rundel, M. (2015).

OpenIntroStatistics.

openintro.org.

Online verfügbar unter www.openintro.org.

[Downey, 2014] Downey, A. B. (2014).

Think Stats - Exploratory Data Analysis in Python.

O'Reilly.

Online verfügbar unter www.greenteapress.com.

[Griffith, 2014] Griffith, D. (2014).

Head First Statistics / Statistik von Kopf bis Fuß.

O'Reilly.

Verwendete oder empfohlene Literatur II

[Haslwanter, 2016] Haslwanter, T. (2016).

An Introduction to Statistics with Python.

Springer.

Python Notebooks auf [github](#) verfügbar.

[Krämer, 2015] Krämer, W. (2015).

So lügt man mit Statistik.

Campus Verlag.

[Papula, 2016] Papula, L. (2016).

Mathematik für Ingenieure und Naturwissenschaftler, Band 3.

Springer Vieweg, 7. Auflage.

Als eBook in der HTWG-Bibliothek verfügbar.

[Teschl and Teschl, 2014] Teschl, G. and Teschl, S. (2014).

Mathematik für Informatiker: Band 2: Analysis und Statistik.

Springer Vieweg, 3. Auflage.

Als eBook in der HTWG-Bibliothek verfügbar.