

STOCHASTIK (AIN)

Prof. Dr. Barbara Staehle
WS 2025/26

HTWG Konstanz
Fakultät für Informatik

Teil I

BESCHREIBENDE STATISTIK

TEIL I BESCHREIBENDE STATISTIK

1. Charakterisierung einer Stichprobe

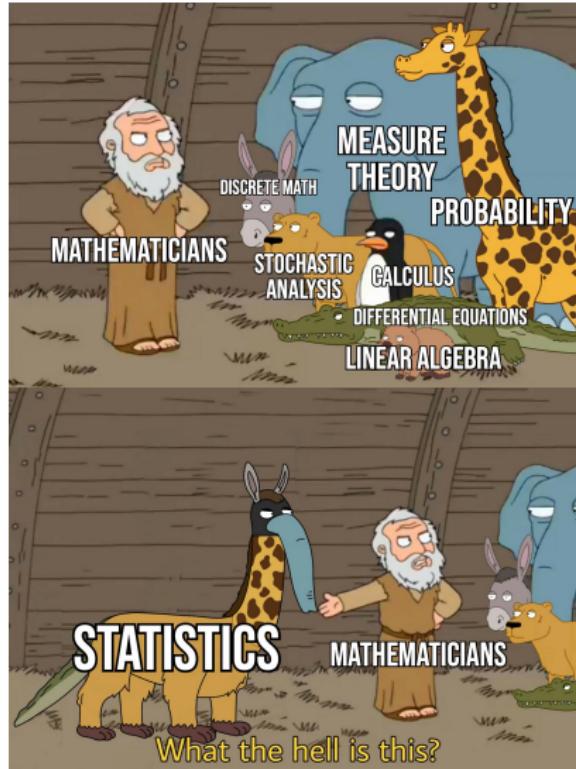
1.1 Grundbegriffe

2. Univariate Statistik

2.1 Häufigkeitsverteilung einer Stichprobe

2.2 Kennwerte einer Stichprobe

VORURTEILE GEGENÜBER STATISTIK I



Quelle: https://bookdown.org/mike/data_analysis/

VORURTEILE GEGENÜBER STATISTIK (QUELLE: MARTIN BECKER, UdS) II

Volksweisheiten und populäre Zitate:

- "Statistik ist Mathematik und in Mathe war ich immer schlecht..."
- "Mit Statistik kann man alles beweisen!"
- "Ich glaube nur der Statistik, die ich selbst gefälscht habe."
(meist Churchill zugeschrieben, aber eher Fake-News von Goebbels)
- "There are three kinds of lies: lies, damned lies, and statistics." (häufig Benjamin Disraeli oder Mark Twain zugeschrieben)

Tatsächlich aber

- verwenden die meisten statistischen Methoden nur Grundrechenarten.
- ist **gesunder Menschenverstand** viel wichtiger als mathematisches Know-How.
- sind nicht die statistischen Methoden an sich schlecht oder falsch, sondern deren falsche oder in korrekte Verwendung.
- werden viele (korrekte) Ergebnisse statistischer Untersuchungen einfach falsch, ungenau, übergenau oder problematisch dargestellt und interpretiert.
- hilft **Informatik Know-How** bei der Datenanalyse!

STATISTIK - WORUM GEHT ES?

- **Hauptaufgabe:** Informationen über bestimmte Objekte gewinnen, ohne alle Objekte untersuchen zu müssen.
⇒ besonders interessant und wichtig für Machine Learning & Co!
- **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 I

Studentische 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:** 400-800 €, 8-15 m², {Paradies, Petershausen, Fürstenberg, ...}, 1-5, ...



Quelle: www.cheatsheet.com

BEISPIELE ZUR TERMINOLOGIE II

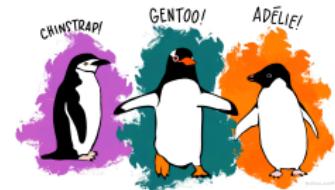
Wie unterscheiden sich unterschiedliche Arten von Pinguinen?

- **Statistische Einheiten:** Pinguine
 - **Grundgesamtheit:** alle Pinguine, die auf der Welt leben
 - **Stichprobe:** 342 Pinguine, die zwischen 2007 und 2010 von Forschenden der Palmer Station in der Antarktis untersucht wurden:
allisonhorst.github.io/palmerpenguins/
[.../penguins_raw.csv](http://allisonhorst.github.io/palmerpenguins/_data/penguins_raw.csv)
 - **Merkmale:** Art, Insel, Schnabelänge, Schnabelbreite, Flossenlänge, Körpergewicht, Geschlecht, ...
 - **Merkmalsausprägungen:** {Adelie, Gentoo, Chinstrap}, {Biscoe, Dream, Torgersen}, {37, 37.1, ..., 52}, {17, 17.1, ..., 22}, {3000, 3001, ..., 6000} {female, male, NA}, ...



Quelle:

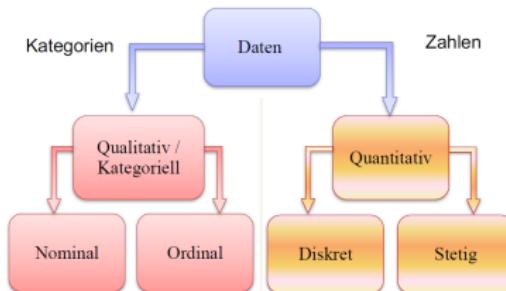
allisonhorst.github.io/palmerpenguins/



Quelle:

allisonhorst.github.io/palmerpenguins/

EIGENSCHAFTEN VON MERKMALEN I



Quelle: Oliver Dürr

Ein Merkmal heißt

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

quantitativ (manchmal auch metrisch) wenn die Ausprägungen ein Ausmaß bzw. eine Intensität widerspiegeln. Die Ausprägungen sind **Zahlen** (mit oder ohne Maßeinheit). **Beispiele:** Miete, Zimmergröße, Anzahl Mitbewohner, Schnabellänge, Schnabelbreite, Flossenlänge, Körpergewicht

EIGENSCHAFTEN VON MERKMALEN II

Ein qualitatives Merkmal heißt

ordinal wenn sich seine Ausprägungen **in natürlicher Weise anordnen** lassen. **Beispiele:** Klausurnoten, militärische Dienstgrade, Unwetter-, Lawinenwarnstufen,

nominal wenn sich seine Ausprägungen **nicht anordnen** lassen. **Beispiele:** Stadtteile, Augenfarbe, Blutgruppe, Studienfach

Ein quantitatives Merkmal heißt

diskret wenn es **endliche viele oder abzählbar unendlich viele** Ausprägungen hat, also gezählt werden kann. **Beispiele:** Anzahl Mitbewohner, Anzahl Hörer:innen einer Vorlesung, Ergebnis eines Würfelwurfs

stetig wenn es alle Werte in einem reellen Intervall als Ausprägungen annehmen kann, also (rein theoretisch) **überabzählbar unendlich viele** Ausprägungen gemessen werden können. **Beispiele:** Miete, Körpergröße, Windgeschwindigkeit

NOMINALE, ORDINALE, DISKRETE UND STETIGE MERKMALE (QUELLE: LECHAT) I

Merkmalstyp	Definition	Eigenschaften	Beispiele	Statistische Auswertung
Nominal	Kategorien ohne Reihenfolge oder numerische Bedeutung.	Keine Ordnung, keine Abstände oder Nullpunkt, nur Gleichheit/Ungleichheit prüfbar.	Haarfarbe (blond, brünett), Geschlecht (m/w/divers), Postleitzahl, Automarke.	Häufigkeiten, Chi-Quadrat-Test.
Ordinal	Kategorien mit natürlicher Reihenfolge, aber keine gleichmäßigen Abstände.	Reihenfolge existiert ($A > B$), keine quantifizierbaren Abstände, kein Nullpunkt.	Schulnoten (1–6), Unwetterwarnstufen (Gelb/Orange/Rot), Zufriedenheit (sehr zufrieden → unzufrieden).	Median, Rangkorrelation (Spearman), Häufigkeiten.
Diskret	Abzählbare, ganze Werte (meist durch Zählen).	Nur ganze Zahlen oder abzählbare Schritte, keine Zwischenwerte, oft Ratioskala.	Anzahl der Kinder pro Familie, Würfelergebnis (1–6), Anzahl der Fehler in einer Prüfung.	Mittelwert (wenn Ratio), Poisson-Verteilung, Häufigkeiten.
Stetig	Unendlich viele Werte in einem Intervall (durch Messen).	Jeder Wert dazwischen ist möglich (z. B. 1,537 m), oft Ratioskala.	Körpergröße (175,3 cm), Gewicht (68,47 kg), Zeit (42,387 Sekunden), Temperatur in °C.	Mittelwert, Standardabweichung, Normalverteilung, Regression.

NOMINALE, ORDINALE, DISKRETE UND STETIGE MERKMALE (QUELLE: LECHAT) II

Wie unterscheidet man die Merkmalstypen?

1. **Qualitativ (Nominal/Ordinal) vs. Quantitativ (Diskret/Stetig):**
 - Qualitativ: Kategorien (z. B. "rot", "zufrieden").
 - Quantitativ: Zahlen (z. B. "3 Äpfel", "175,3 cm").
2. **Nominal vs. Ordinal:** Gibt es eine natürliche Reihenfolge?
 - Nein → Nominal (z. B. Haarfarbe).
 - Ja → Ordinal (z. B. Schulnoten).
3. **Diskret vs. Stetig:** Kann das Merkmal jeden beliebigen Wert in einem Intervall annehmen?
 - Nein (nur ganze Zahlen) → Diskret (z. B. Anzahl der Kinder).
 - Ja (auch Nachkommastellen) → Stetig (z. B. Körpergröße).

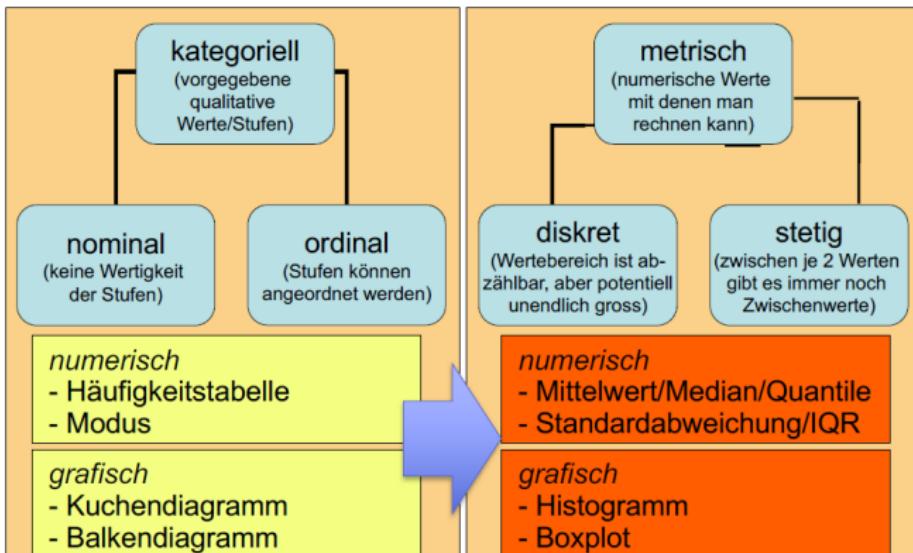
Merksätze

- Nominal/Ordinal = Qualitative Merkmale (Kategorien).
- Diskret/Stetig = Quantitative Merkmale (Zahlen).
- Diskret = Zählen (z. B. „Wie **viele**?“).
- Stetig = Messen (z. B. „Wie **viel**?“).

BEISPIELE ZUM MITDENKEN (QUELLE: LECHAT)

Merkmal	Typ	Begründung
Blutgruppe (A, B, AB, o)	Qualitativ, Nominal	Keine Reihenfolge, nur Kategorien
Schulnote (1–6)	Qualitativ, Ordinal	Reihenfolge
Anzahl der Geschwister	Quantitativ, Diskret	Ganze Zahlen, abzählbar
Körpergröße (in cm)	Quantitativ, Stetig	Reelle Zahlen, überabzählbar
Windstärke (Beaufort 0–12)	Qualitativ, Ordinal	Reihenfolge, aber Abstände möglich
Windgeschwindigkeit (km/h)	Quantitativ, Stetig	Jeder Wert möglich
Basketball-Punkte	Quantitativ, Diskret	Ganze Zahlen, abzählbar
Platz bei der Basketball-EM	Qualitativ Ordinal	Keine Zahl, sondern Reihenfolge
Marathon-Bestzeit	Quantitativ, Stetig	kann jeden beliebigen Wert annehmen
Postleitzahl	Qualitativ, Nominal	PLZ sind bedeutungslos

TYP-ABHÄNGIGE DARSTELLUNG VON MERKMALEN



Quelle: Oliver Dürr

Schwerpunkt dieser Vorlesung / der Informatik / analytischer Methoden / des Machine Learnings: Quantitative Merkmale

IMMER IM HINTERKOPF BEHALTEN!!



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

- Ist es in Konstanz in den letzten Jahren wärmer geworden?
 - Stichprobe: tägliche Wetterdaten der Station Konstanz seit 1973
 - Quelle: Deutscher Wetterdienst; Konstanz, Tageswerte, historisch
 - csv = comma separated values, durch Semikolon getrennte Spalten
- Kommen erste Babys immer später? (Beispiel nach [Downey, 2014] und [Downey, 2025])
 - Stichprobe: Daten von 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
- Eigenschaften von Pinguinen (Idee von [Downey, 2025])
 - Stichprobe: data for 344 penguins. There are 3 different species of penguins in this dataset, collected from 3 islands in the Palmer Archipelago, Antarctica.
 - Quelle: allisonhorst.github.io/palmerpenguins (API für R)
 - Rohdaten: [penguins_raw.csv](https://allisonhorst.github.io/palmerpenguins/penguins_raw.csv)
- Daten-Handling: MATLAB, Python, Excel, Java, Octave, R ...
Code zu den Bildern der Vorlesung überwiegend hier
<https://github.com/barbara2342/htwgstochastik-goes-digital>
verfügbar.

ELEKTRONISCHE HELFERLEIN

Nur für wenige Werte Statistiken per Hand erstellen (das muss man aber auch können)! Für viele Daten:

- TR: z.B. TI-30X Plus Multiview, Casio FX-991DE X ClassWiz
- **MATLAB**
 - viel Beispielcode verfügbar, weil mein Lieblingstool
 - **Vorteile:** Viele Toolboxen, weitverbreitet im technischen Umfeld für Sie dank Campuslizenz (bis 03/2026) gratis. Download via [RZ](#)
 - **Nachteil:** normalerweise (für uns ab 04/26) teuer; freie Alternative: [Octave](#)
- **Python** (Anaconda: all-in-one Distribution, Download [hier](#))
 - Schweizer Taschenmesser, nützlich auch für anderes, in vielen Hauptstudiumsvorlesungen verwendet, Beispielcode verfügbar
 - gute Bücher und online-Tutorials: [[Downey, 2014](#), [Haslwanter, 2016](#)]
 - **Vorteil:** Open Source, frei erhältlich, viele nützliche Pakete

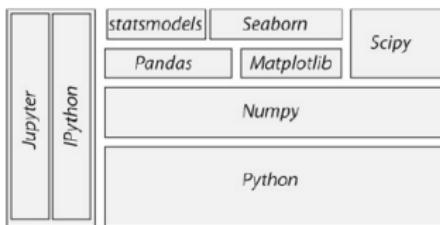


Bild 1: Wichtige Statistik Python-Pakete [Haslwanter, 2016]

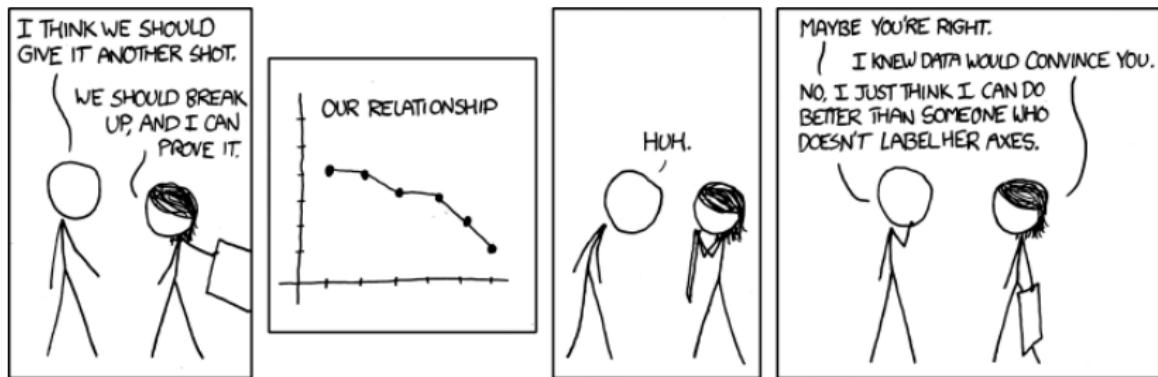
DARSTELLUNG UND ANALYSE UNIVARIATER DATEN

- 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

In beiden Fällen wichtige erste Schritte

- **Rohdatenanalyse:** Vergleich der Werte in der **Urliste** (Liste aller erhaltenen Messwerte ansehen, Datei mit einem Editor öffnen)
 - Vorteil: schnell, einfach, erste Trends sind vielleicht schon sichtbar, offensichtliche Fehler oder Seltsamkeiten sind erkennbar
 - Nachteil: nicht praktikabel für großen Datensätze
- **graphische Rohdatenanalyse:** plotten aller erhaltenen Messwerte in zufälliger Reihenfolge oder z.B. nach Datum
 - Vorteil: schnell, einfach
 - Nachteil: unübersichtlich, vor allem für große Datenmengen

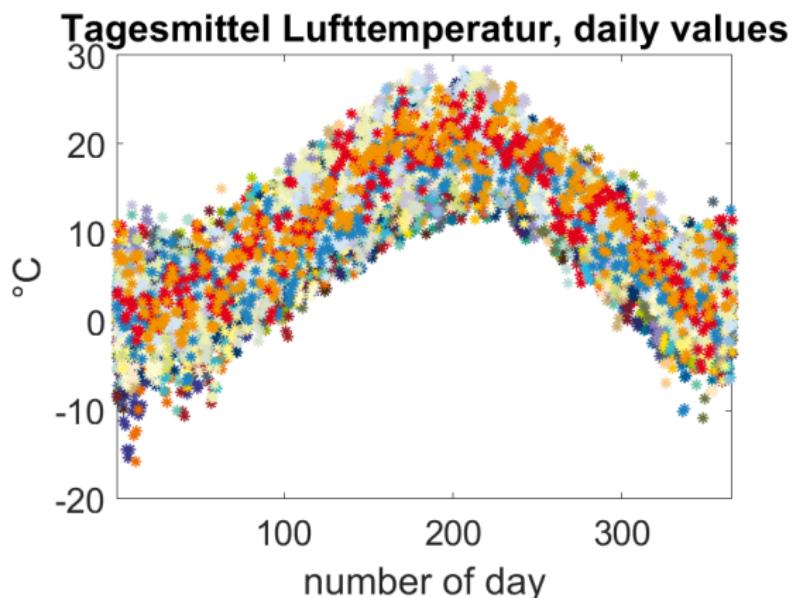
EIN WICHTIGER HINWEIS, BEVOR SIE IHRE ERSTE GRAPHIK ERSTELLEN



Quelle: xkcd.com

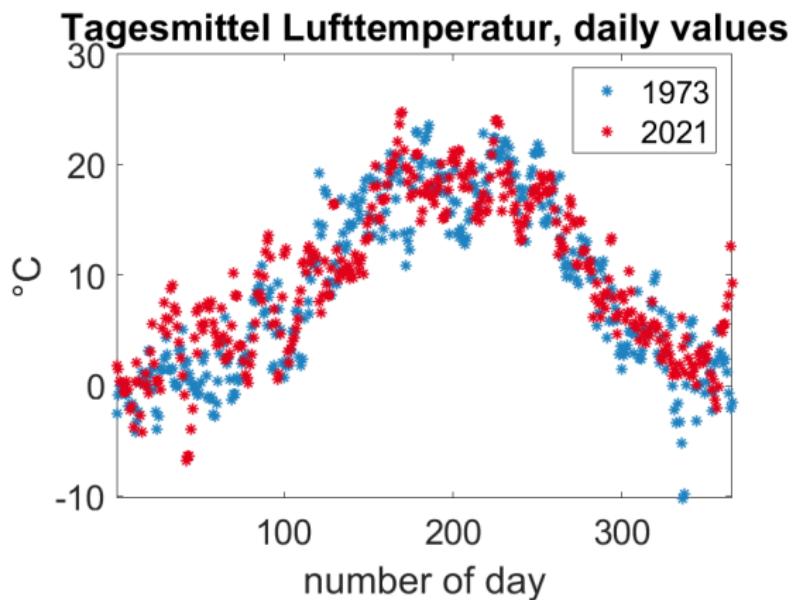
Label your axis!!
Beschriften Sie IMMER die Achsen jedes Plots und jeder Graphik!!

GRAPHISCHE ROHDATENANALYSE - TEMPERATUR (1973-2023)



Bewertung: Offensichtlich zu viele Datenpunkte, keine gute Darstellung.
Aber: die Zuordnung der Temperaturen zu den Tagen des Jahres passt offensichtlich.

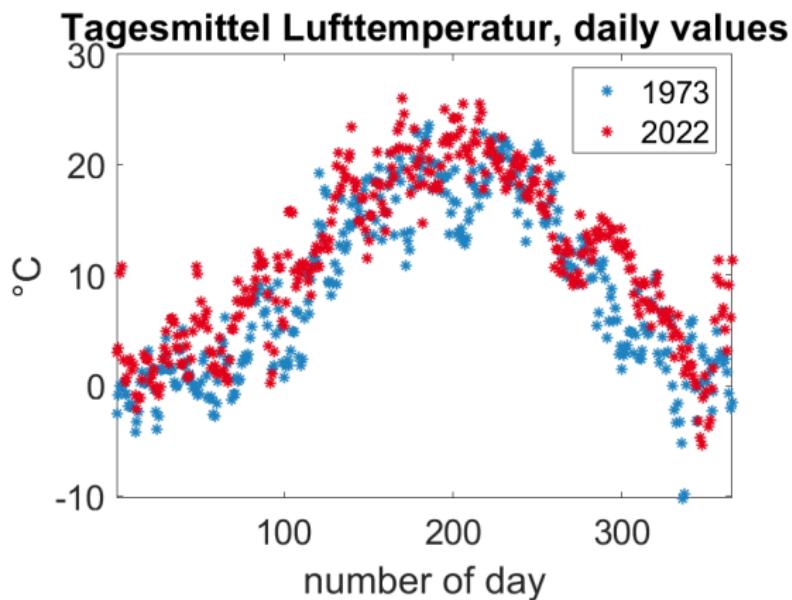
GRAPHISCHE ROHDATENANALYSE - TEMPERATUR (1973 VS. 2021)



Aussage: Es gibt keinen Klimawandel, Temperaturen ähnlich.

Bewertung: Quatsch! Vergleich von nur zwei Jahren ist grob fahrlässig
(zufällige Schwankungen überdecken langfristige Effekte).

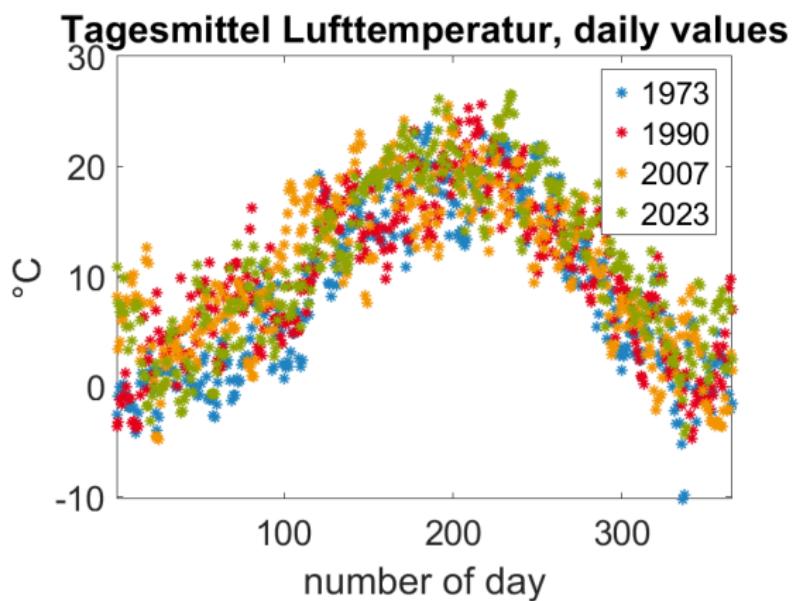
GRAPHISCHE ROHDATENANALYSE - TEMPERATUR (1973 VS. 2022)



Aussage: Klimawandel ist auch in Konstanz offensichtlich!

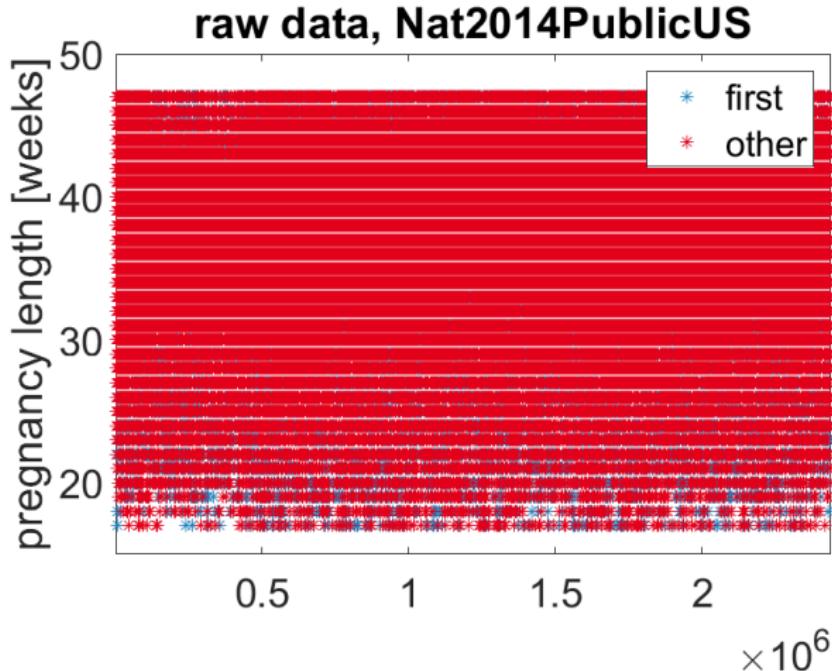
Bewertung: Tendenz erkennbar, aber Vergleich von nur zwei Jahren ist grob fahrlässig (zufällige Schwankungen überdecken langfristige Effekte).

GRAPHISCHE ROHDATENANALYSE - TEMPERATUR (VIER JAHRE)



Bewertung: Vergleich von vier Jahren macht die Grafik unübersichtlicher, aber ehrlicher. Darstellungsform erlaubt aber wenig Aussagen.

GRAPHISCHE ROHDATENANALYSE - BABYS



Bewertung: Mehr als 5 Millionen Datenpunkte in einem Bild erlauben keinerlei Aussagen, außer einer generellen Bewertung der SS-Wochen.

ABSOLUTE UND RELATIVE HÄUFIGKEIT

- Rohdatenanalyse ist ungeeignet für viele Datenpunkte und mehrfach auftretende Werte.
 - **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 einem Casino verwendete Würfel gezinkt ist. Daher protokollieren wir 20 Würfelwerte und erhalten folgende „Messwerte“ x_1, x_2, \dots, x_{20} :

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

1. Auftretende Werte: $a_1 = 1, a_2 = 2, \dots, a_6 = 6$
2. Ermittle Häufigkeiten per Hand oder **MATLAB**: `hist`, `histogram`,
Python: `scipy.stats.relfreq`, `numpy.histogram`,
`numpy.unique`, `Counter`, ...

IST DER WÜRFEL GEZINKT? ANTWORT VON CLAUDE

Das ist eine interessante statistische Frage! Lass mich das analysieren:

Code: [kap1_claude_wuerfelFairnessTest.py](#)

Kurze Antwort: Mit den vorliegenden Daten kann man nicht definitiv sagen, ob der Würfel gezinkt ist. Hier die wichtigsten Punkte:

Auffälligkeiten: Die 1 kommt 7x vor (35% der Würfe) - bei einem fairen Würfel würde ungefähr 3 mal die 3 erwartet. Das ist eine deutliche Abweichung, aber...

Statistische Realität:

- **20 Würfe sind zu wenig** für eine zuverlässige Aussage
- Bei so wenigen Würfen sind auch extreme Abweichungen durch puren Zufall möglich
- Der χ^2 -Test wird zeigen, ob die Abweichung statistisch signifikant ist

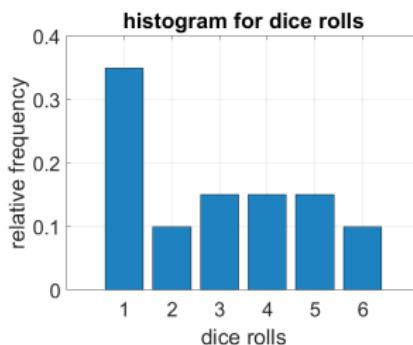
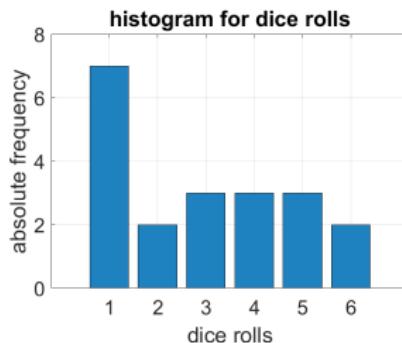
Fazit:

- **Verdacht:** Ja, die Häufung der 1 ist verdächtig
- **Beweis:** Nein, könnte auch Pech/Glück sein
- **Empfehlung:** Mindestens 60-100 weitere Würfe für eine verlässliche Aussage

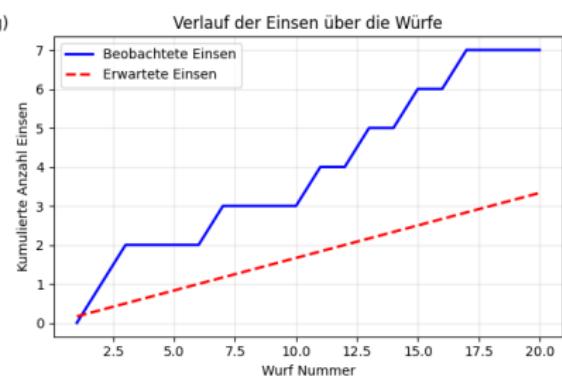
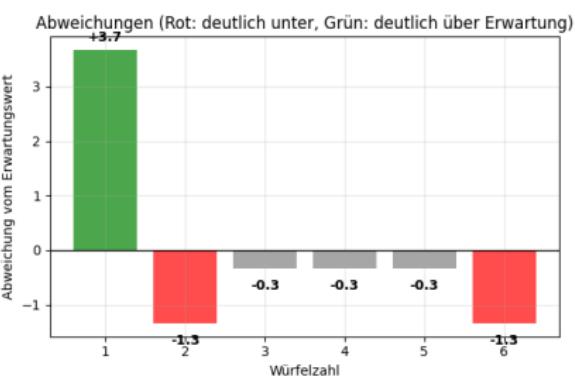
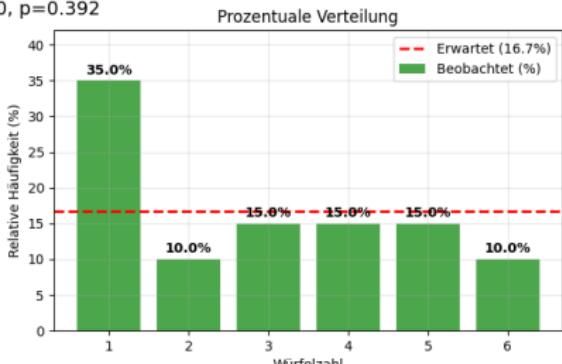
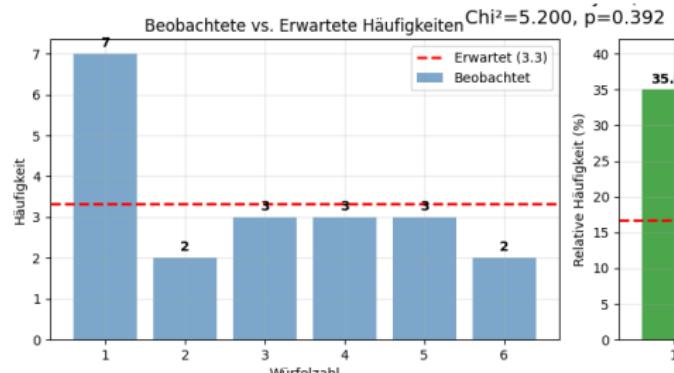
DARSTELLUNG VON (RELATIVEN) HÄUFIGKEITEN I

Balken- oder Stabdiagramm

- auch: Histogramm
- x-Achse: a_i (vorkommende Werte) sortiert nach Größe
- y-Achse: Säule, Stab, Balken für
 - h_i (absolute Häufigkeiten)
 - f_i (relative Häufigkeiten)
- **MATLAB:** hist, histogram, bar, stem, **Python:** matplotlib.pyplot.hist, matplotlib.pyplot.bar
- Geschmackssache: Darstellung von relativen Häufigkeiten als Dezimalbruch (z.B 0.5) oder Prozentzahl (z.B. 50%)



DARSTELLUNG VON (RELATIVEN) HÄUFIGKEITEN II, AUTOR: CLAUDE



Code siehe [kap1_claude_wuerfelFairnessTest.py](#); Antwort auf Prompt „bitte noch als barplot visualisieren“; IMHO etwas überfordernd ...

DARSTELLUNG VON (RELATIVEN) HÄUFIGKEITEN III

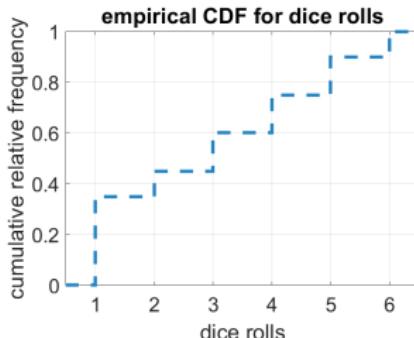
empirische Verteilungsfunktion

- x-Achse: a_i (vorkommende Werte) sortiert nach Größe
- berechne

$$\bar{F}(x) = \sum_{i:a_i \leq x} f_i$$

auf Deutsch: Summiere die relativen Häufigkeiten kumuliert (= in jedem Schritt auf)

- \bar{F} ist Stufenfunktion mit min>0, max=1, Sprünge bei a_i

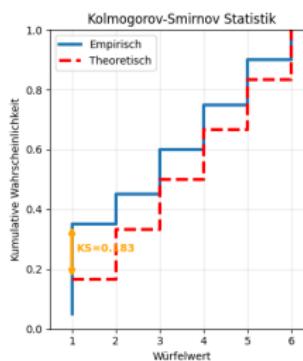
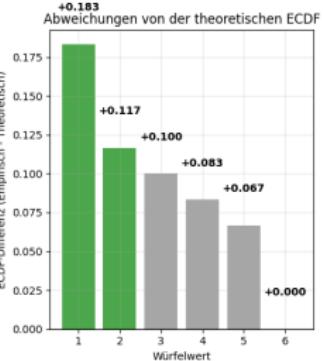
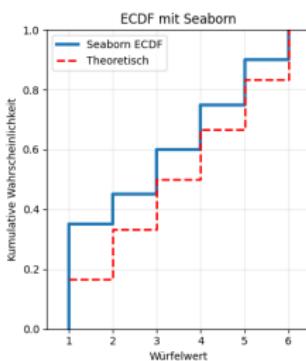
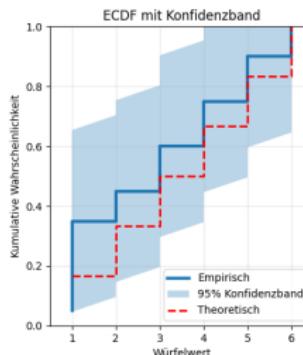
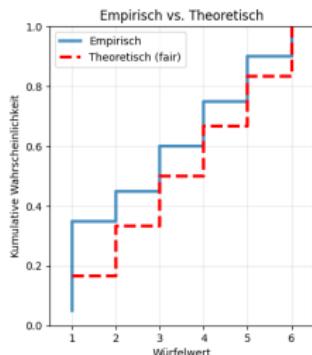
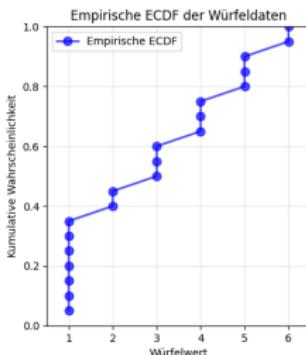


MATLAB: `cumsum`, `stairs`, `cdfplot`

Python:
`matplotlib.pyplot.stairs`,
`matplotlib.pyplot.step`,
`seaborn.ecdfplot`, ...

Würfelwert	1	2	3	4	5	6
relative Häufigkeit f_i	0.35	0.1	0.15	0.15	0.15	0.1
kumulierte relative Häufigkeit $F(i)$	0.35	0.45	0.6	0.75	0.9	1

DARSTELLUNG VON (RELATIVEN) HÄUFIGKEITEN IV, AUTOR: CLAUDE



Code siehe [kap1_claude_wuerfelECDF.py](#); Antwort auf Prompt „Wie erzeuge ich von den werten eine ecdf“; IMHO: nicht sehr hübsch:
Grafik links oben ist falsch, weil keine Treppe, ECDFs zeigen keine Werte < 1 und größer an (Treppe beginnt und endet im Nichts),
außerdem ist Menge der Plots und verwendeten Methoden etwas überfordernd

HISTOGRAMM - TEMPERATUR

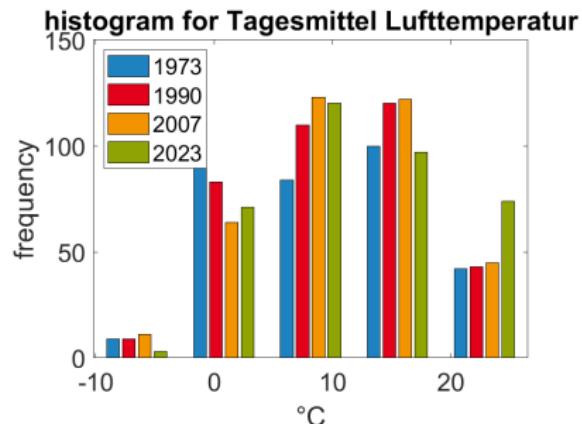


Bild 2: Histogramm mit 5 Intervallen

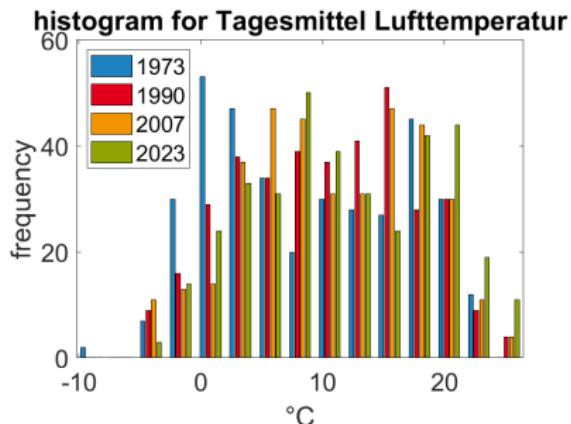
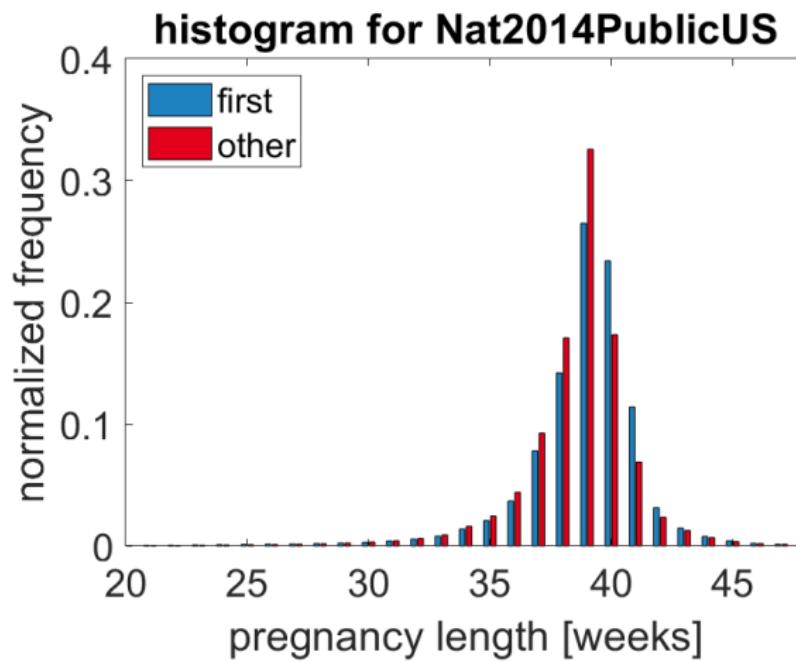


Bild 3: Histogramm mit 15 Intervallen

- 365 Messungen für jedes Jahr, daher Vergleich absoluter Häufigkeiten möglich
- **Aber:** Zu viele verschiedene Werte, daher Zusammenfassung zu **Klassen** bzw. **Intervallen** (z.B. $26.5^\circ - 28.5^\circ$)
- Geschmacksfrage: z.B. 5 oder z.B. 20 gleich große Intervalle wählen (Tradeoff Genauigkeit-Lesbarkeit)

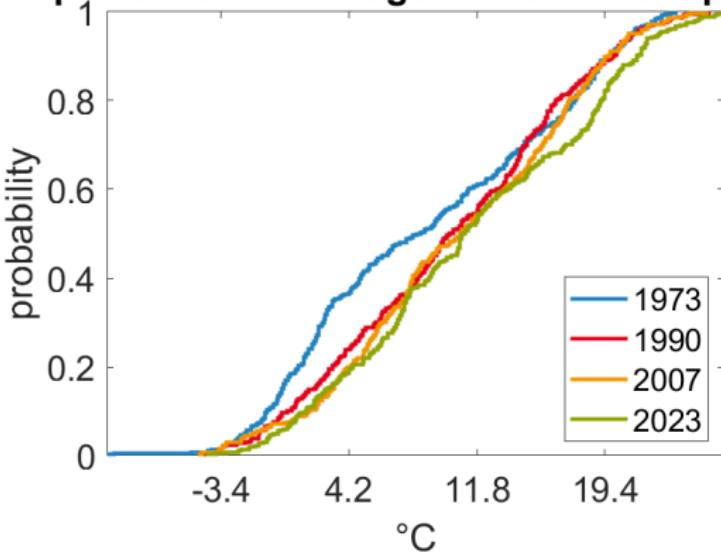
HISTOGRAMM - BABYS



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

EMPIRISCHE VERTEILUNGSFUNKTION - TEMPERATUR

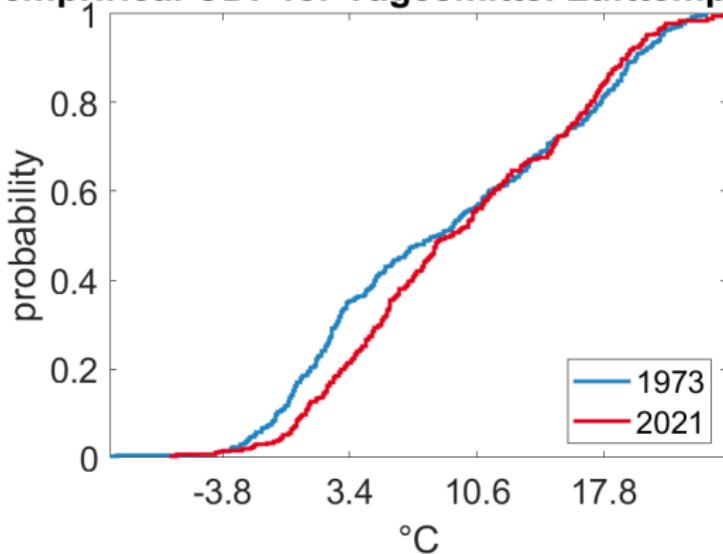
empirical CDF for Tagesmittel Lufttemperatur



- 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

ECDF TEMPERATUR: VARIANTE ES GIBT KEINEN KLIMAWANDEL

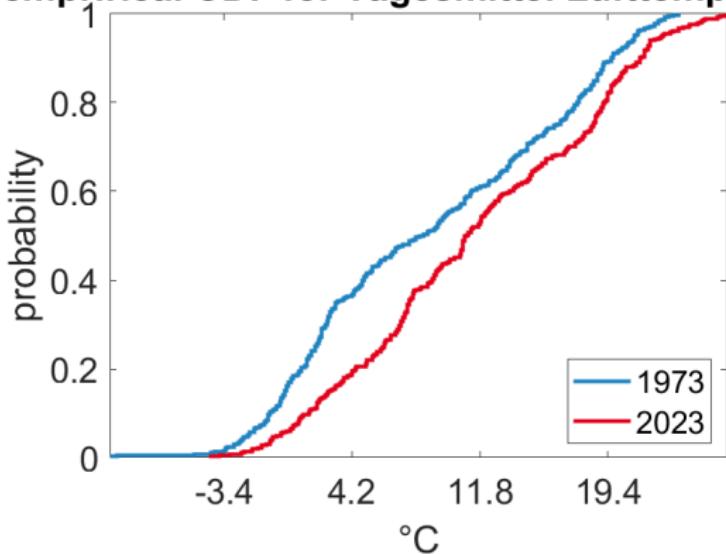
empirical CDF for Tagesmittel Lufttemperatur



- Kurven von 1973 und 2021 sind kaum unterscheidbar
- gibt es den Klimawandel echt nicht??
- **Nein!** Klima \neq Wetter, 2021 war ein besonders kaltes Jahr (später mehr).
- Wie immer: Vergleich von nur zwei Jahren ist grob fahrlässig.

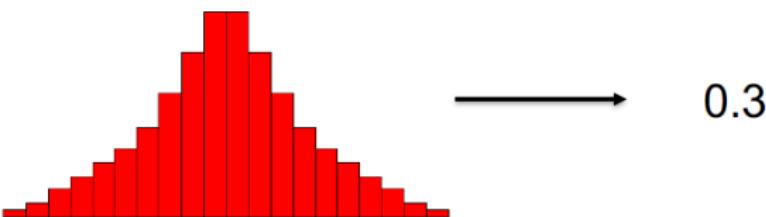
ECDF TEMPERATUR: VARIANTE KLIMAWANDEL DEUTLICH SICHTBAR

empirical CDF for Tagesmittel Lufttemperatur



- Kurve für 2023 liegt deutlich unter der von 1973: Temperaturen 2023 tendenziell höher.
- Aber wie immer: Vergleich von nur zwei Jahren ist grob fahrlässig!

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



Quelle: Der SPIEGEL, 06/2018

Lösung: Im Mittel oder durchschnittlich muss man 4844 Sticker (ca. 970 Tütchen) kaufen, bis man das Album voll hat. Grund: Man erwirbt viele doppelte Aufkleber

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?

Python: `numpy.mean`, **MATLAB:** `mean`

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.

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

Python: `numpy.median`, **MATLAB:** `median`

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 qualitative Merkmale 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: Bestimmen Sie den Mittelwert, Median und den Modalwert der Größen der vor einer Turnhalle abgestellten $n = 10$ Schuhe:

32 30 33 31 30 30 32 33 34 45

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

Python: `statistics.mode`, `statistics.multimode`, **MATLAB:** `mode`

AUGEN AUF BEI DER MODALWERTBERECHNUNG!!

Achtung: Manche Tools gehen schlampig mit multimodalen Stichproben um!

Daher: trauen Sie nie blind einem Wert, der aus einem Tool fällt!



Quelle: clipartmag.com

Beispiele:

- Stichprobe 1: $s_1 = [8, 5, 6, 3, 3, 4, 9, 9, 2, 0, 0]$
- Stichprobe 2: $s_2 = [10, 9, 8, 7, 6, 5, 4, 3, 2, 1, 0]$

FEINERE LAGEKENNWERTE: QUANTILE & CO

DEFINITION

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

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

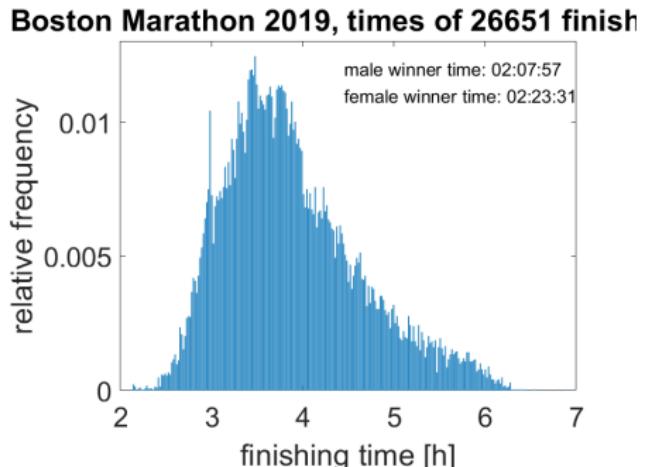
das ***p*-Quantil**.

- 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: SPORT



Datenquelle: www.kaggle.com

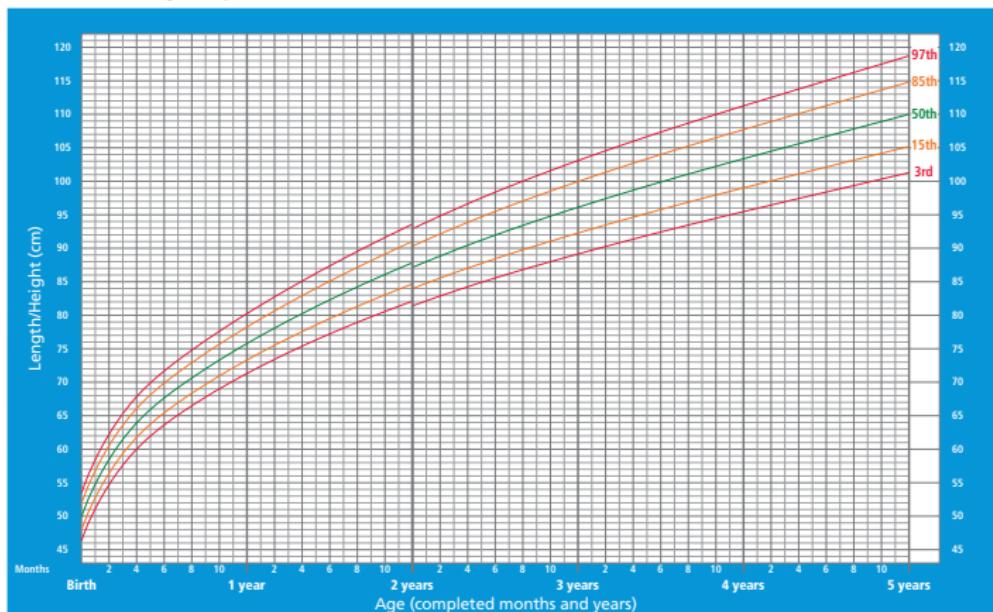
Fragen:

- Wie schnell muss ich laufen, so dass ich genau im Mittelfeld liege?
Median = 03:45:34 h
- Wie schnell muss ich laufen, dass ich zu den schnellsten 10% gehöre?
10%-Quantil = 03:00:14 h

ANWENDUNG VON QUANTILEN: WACHSTUMSKURVEN FÜR KINDER

Length/height-for-age BOYS

Birth to 5 years (percentiles)



WHO Child Growth Standards

Quelle: [WHO](#)

AUGEN AUF BEI DER QUANTILBERECHNUNG!! I

Achtung: Für Quantilberechnung existiert kein eindeutig richtiger Algorithmus [Hyndman and Fan, 1996].

Daher: trauen Sie nie blind einem Wert, der aus einem Tool fällt! Implementieren Sie die Methode selbst, lesen Sie die Doku und **lassen Sie sich auf mehrere richtige Möglichkeiten ein, verstehen Sie aber auch deren Berechnung!**



Quelle: clipartmag.com

Links zur Doku:

- **Python:** `numpy.quantile` und `numpy.percentile` (nimmt als Input Werte aus $[0, 100]$, sonst identisch zu `quantile`)
`percentile`-Seite erklärt die 9 unterschiedlichen Algorithmen
- **MATLAB:** `quantile` (2 verschiedene Algorithmen)
- **R:** `quantile` (9 verschiedene Algorithmen)
- **Excel:** `QUANTIL.EXKL`, `QUANTIL.INKL`

Beispiel: Auch Experten haben Probleme: Anfrage von [Oliver Thomaschewski](#), Data Science-Student in Berlin (siehe nächste Folie)

AUGEN AUF BEI DER QUANTILBERECHNUNG!! II

Anfrage von Oliver Thomaschewski: Als Data Science Werkstudent soll er für folgende (sortierte) Liste das 20%-Quantil ausrechnen:

–50, –28.2, –17.7, –17.5, –2.8, –2.4, –0.6, 1.9, 6.1, 8.1, 8.7, 13.3, 14.6, 16.8,
18.6, 22.9, 23.4, 24.6, 28.5, 30.1, 30.8, 32.2, 32.3, 33, 34.8, 35.9, 36, 36.5, 39.5,
41.7, 43.8, 45.1, 46.8, 54.4, 55.2, 71.2, 71.6, 74.8, 109.6, 120, 135.3, 334.2

Seine Überlegungen:

- **Python oder Excel:** 6.5
- **Per Hand:** Anzahl der Elemente (42) mal das Quantil: $42 \cdot 0.2 = 8.4$. Da ungerade, nehm ich jetzt den Wert an der 8. Stelle und den an der 9. Stelle, also 1.9 und 6.1 was dann 8 ergibt, teile durch 2 und komme auf 4. Was ja nachvollziehbar ist und mir die 6.1 dann quasi als Grenzwert geben sollte.
- **Problem:** $6.1 < 6.5$, wie kommt die 6.5 zustande??

BEISPIELE ZUM MITDENKEN - KENNWERTE

Alice interessiert sich für die Anzahl von Teilnehmer:innen einer Mastervorlesung ihrer Fakultät. Daher zählt sie diese in $n = 10$ Vorlesungen und erhält die Urliste 14, 24, 22, 19, 18, 36, 15, 29, 41, 17.

Gesucht: Mittelwert, die Quartile sowie das 90% Quantil der Urliste.

BEISPIELE ZUM MITDENKEN - INTERPRETATION

Gesucht: Bedeutung der ausgerechneten Kennwerte für die Urliste

14, 24, 22, 19, 18, 36, 15, 29, 41, 17.

Mögliche Interpretationen (für die untersuchte Stichprobe):

Frage: Gilt das für alle Mastervorlesungen (an dieser Fakultät)?
⇒ Thema der schließenden Statistik (Kapitel 4)!

BEISPIEL: WELCHEN SERVER WÜRDEN SIE NEHMEN? (ZAHLEN VON CLAUDE) I

Sie sind als CIO dafür verantwortlich, die Webseite Ihrer Firma auf einen anderen Server auszulagern. Sie haben drei Angebote vorliegen:

Server A mittlere Response Time: 100 msec, 500 € pro Monat

Server B mittlere Response Time: 100 msec, 200 € pro Monat

Server C mittlere Response Time: 100 msec, 100 € pro Monat

Für welchen Server entscheiden Sie sich?

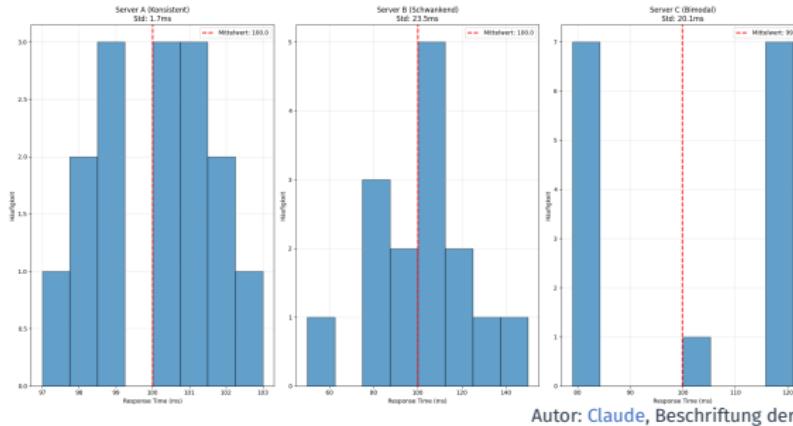
So schnell bitte für gar keinen!

Schritt 1 Rohdaten der Response times ansehen

- server_a = [98, 102, 99, 101, 100, 103, 97, 101, 99, 102, 100, 98, 101, 99, 100]
- server_b = [50, 150, 75, 125, 80, 120, 85, 115, 90, 110, 95, 105, 100, 100, 100]
- server_c = [80, 80, 81, 79, 80, 120, 121, 119, 120, 121, 80, 79, 120, 119, 100]

BEISPIEL: WELCHEN SERVER WÜRDEN SIE NEHMEN? (ZAHLEN VON CLAUDE) II

Schritt 2 (optional) Graphische Auswertung



Autor: Claude, Beschriftung der Achsen ist zu klein

Schritt 3 Standardabweichung (grob: durchschnittliche Abweichung der Antwortzeiten vom Mittelwert) ausrechnen

Server A Mittelwert: 100.0 ms, Standardabweichung: 1.7 ms

Server B Mittelwert: 100.0 ms, Standardabweichung: 23.5 ms

Server C Mittelwert: 100.0 ms, Standardabweichung: 20.1 ms

Schritt 4 Kosten und konsistent schnelle oder schwankende oder bimodale Antwortzeiten abwägen

VARIANZ UND STANDARDABWEICHUNG I

Problem aller Lagewerte: Ungeeignet um **Streuung** (Unterschiedlichkeit, Varianz) der Stichprobenwerte 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 häufig 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. (mehr Details: Kapitel 4)
- 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

Gesucht: Absolute Häufigkeit aller auftretenden Werte, Mittelwert, Median, Varianz und Standardabweichung der Stichprobe

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

AUGEN AUF BEI DER BERECHNUNG DER STANDARDABWEICHUNG!

Achtung: Sie müssen Ihrem Tool angeben, ob Sie die empirische Standardabweichung s (durch $n - 1$ teilen) oder die Standardabweichung einer Zufallsvariable σ (durch n teilen) ausrechnen möchten!!

Daher: Vergleichen Sie Ihr Tool der Wahl mit einem Beispiel aus der Vorlesung.



Quelle: clipartmag.com

Beispiel: $x = [8, 5, 6, 3, 3, 4, 9, 9, 2, 0, 0]$

Doku: MATLAB std, Python numpy.std, R std

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

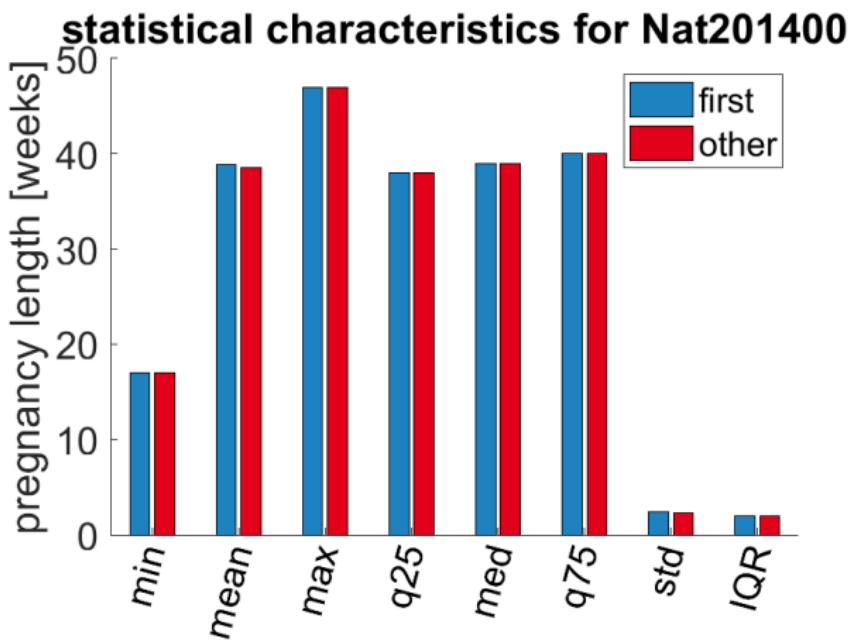
Spannweite (Range) $R = x_{max} - x_{min}$ (Differenz von größtem und kleinstem Stichprobenwert); **Vorteil:** einfach zu berechnen, **Nachteil:** starke Beeinflussung durch Ausreißer.

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

Beispiel: Berechnen Sie Spannweite und IQR der Stichprobe

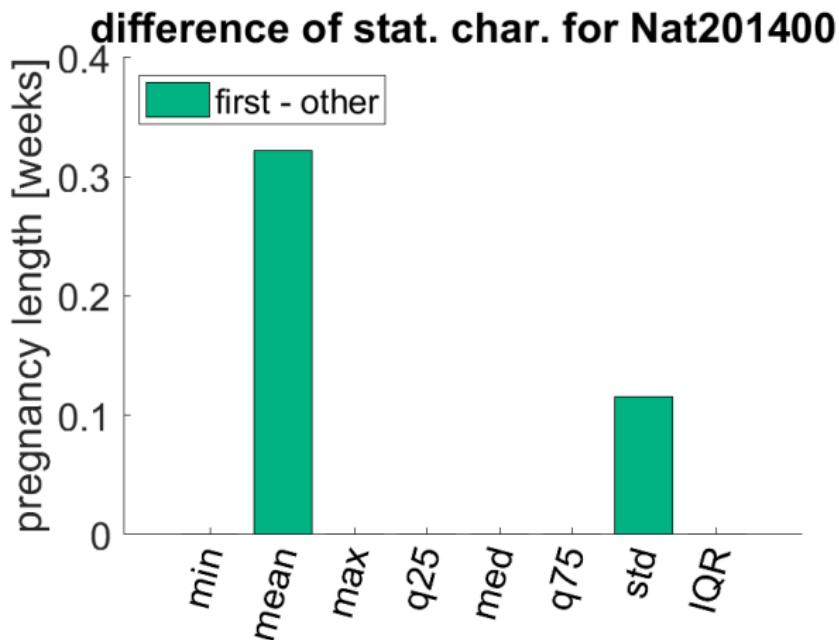
1, 1, 2, 2, 3, 3, 3, 3, 4, 4, 7.

VERGLEICH STATISTISCHER KENNWERTE - BABYS I



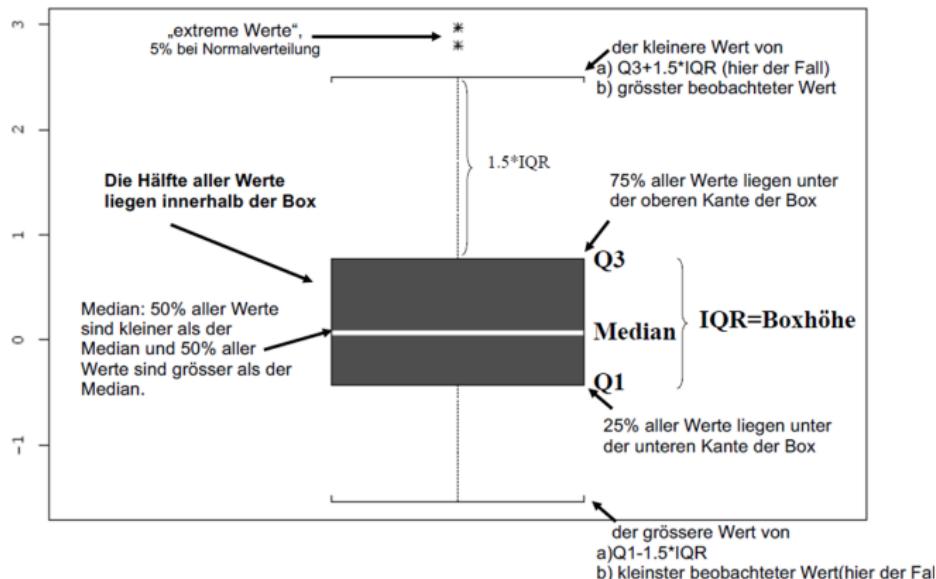
- hinsichtlich der Lage- und Streuungskennwerte scheint kaum ein Unterschied sichtbar

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

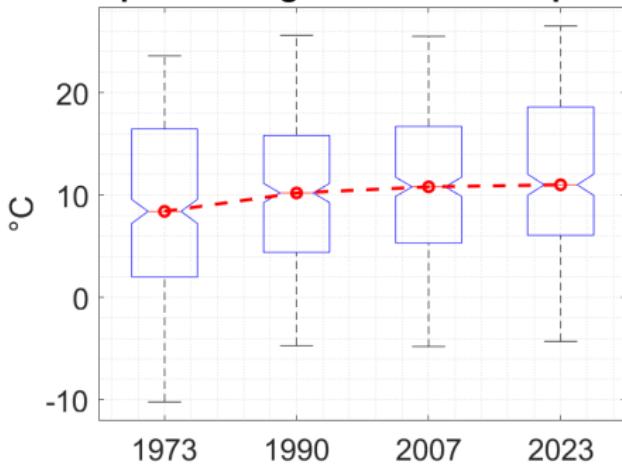


Quelle: Oliver Dürr

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

BOXPLOT - TEMPERATUREN

boxplot for Tagesmittel Lufttemperatur



- 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.
Zusätzlich: Linie für Mediane über der Zeit
- Boxplot für Baby-Problem unübersichtlich aufgrund zu vieler Ausreißer

Fazit: In Konstanz ist es tatsächlich etwas wärmer geworden.

Aber: Gab es nur einen lokalen Höhepunkt? Wird es wieder kälter?

WIRD ES WIEDER KÄLTER?



Bild 4: Hörlepark Konstanz, 15.01.2021, Quelle: www.suedkurier.de

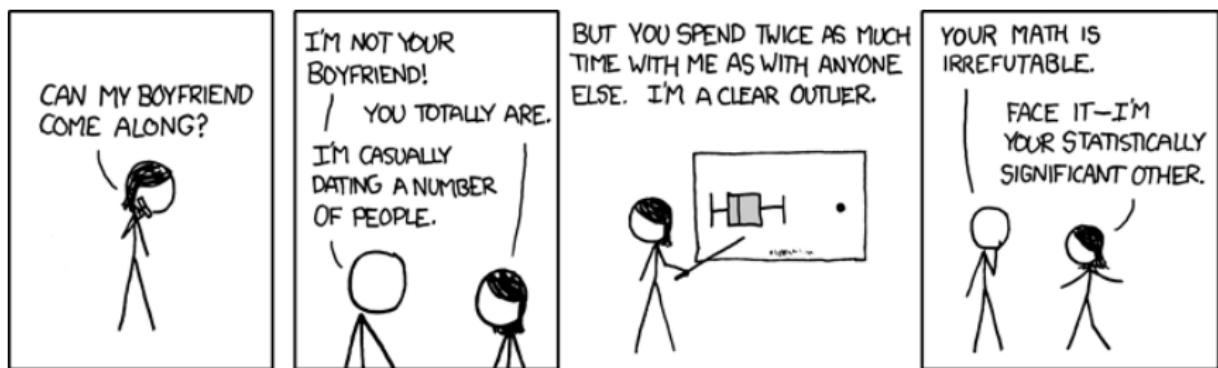
Leider nein: 2021 war ein „Jahrhundertwinter“ - wir hatten in Konstanz 30cm Schnee, der ein oder zwei Wochen liegenblieb gefolgt von einer längeren Frostperiode.

Merke: Wetter \neq Klima (Quelle [Wetterlexikon des DWD](#))

Wetter ist der physikalische Zustand der Atmosphäre zu einem bestimmten Zeitpunkt oder in einem auch kürzeren Zeitraum an einem bestimmten Ort oder in einem Gebiet.

Klima wird repräsentiert durch die statistischen Gesamteigenschaften der Wettererscheinungen über einen genügend langen Zeitraum. Im allgemeinen wird ein Zeitraum von 30 Jahren zugrunde gelegt.

BOXPLOTS IN DER ANWENDUNG I



Quelle: xkcd.com

BOXPLOTS IN DER ANWENDUNG II

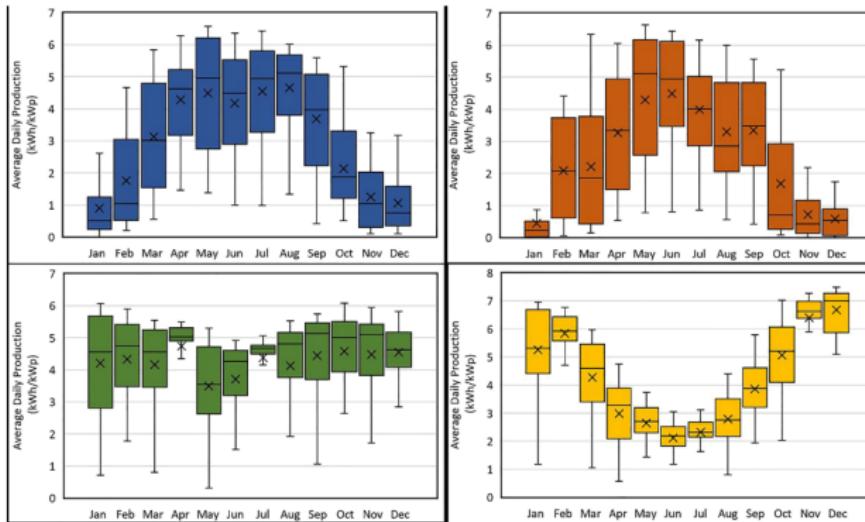
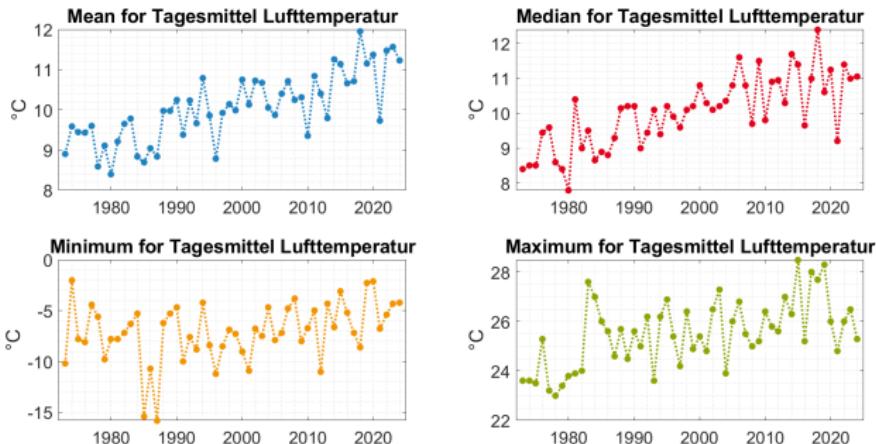


FIGURE 5 Box plot showing the average, minimum, and maximum daily photovoltaic (PV) production in 2016 for each location at an optimized tilt angle (see Table 1). Clockwise from top left: Amsterdam (NL), Oslo (NO), Perth (AU), and São Paulo (BR). The X represents the mean (average), and the line in the 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

AUSBLICK: ZEITREIHEN



Fazit: Zeitreihen zeigen, dass es in Konstanz tatsächlich tendenziell wärmer geworden ist. Kurven “zappeln” aber aufgrund jahresbedingter zufälliger Schwankungen.

Lösung: z.B. gleitende Mittelwerte (moving averages)

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 (2nd edition).

O'Reilly.

Online verfügbar unter www.greenteapress.com.

[Downey, 2025] Downey, A. B. (2025).

Think stats: Exploratory Data Analysis in Python (3rd edition).

O'Reilly.

Online verfügbar unter allendowney.github.io/ThinkStats.

[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.

[Hyndman and Fan, 1996] Hyndman, R. J. and Fan, Y. (1996).

Sample quantiles in statistical packages.

The American Statistician, 50(4).

[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.

VERWENDETE ODER EMPFOHLENE LITERATUR III

[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.