Statistik

Vorlesung 13 - Deskriptive Statistik

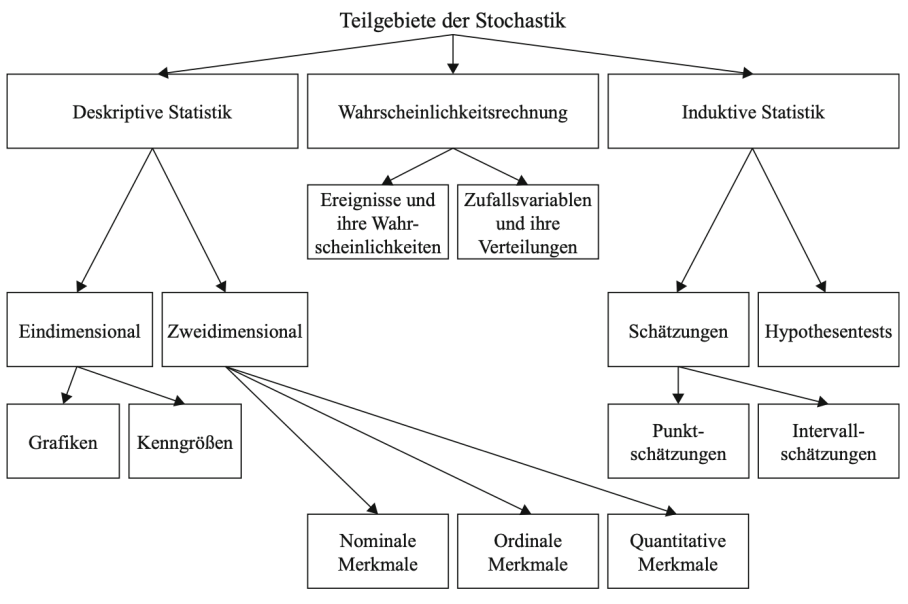
Prof. Dr. Sandra Eisenreich

Hochschule Landshut

Agenda

1. Untersuchungseinheiten und Merkmale
2. Empirische Häufigkeitsverteilung
3. Lage- und Streuungsmaße
4. Box-Plots

Zur Erinnerung: Teilgebiete der Stochastik



Motivation

- Statistik wird in beschreibende (deskriptive) und in beurteilende (schließende) Statistik eingeteilt
- eine Hauptaufgabe der deskriptiven Statistik: übersichtliche grafische und / oder tabellarische Darstellung der für die jeweilige Fragestellung wesentlichen Aspekte vorliegender Daten
  - Achtung: das bedeutet nicht, dass die deskriptive Statistik frei von Beurteilungen ist
  - oft werden grafische / tabellarische Darstellungen verwendet, um zu beeinflussen

Untersuchungseinheiten und Merkmale

Statistische Einheiten, Merkmale, Gesamtheiten

Statistische Einheiten: Objekte, an denen interessierende Größen erfasst werden

Grundgesamtheit: Menge aller für die Fragestellung relevanten statistischen Einheiten

Teilgesamtheit: Teilmenge der Grundgesamtheit

Stichprobe: tatsächlich untersuchte Teilmenge der Grundgesamtheit

Merkmal: interessierende Größe, *Variable*

Merkmalsausprägung: konkreter Wert des Merkmals für eine bestimmte statistische Einheit

Merkmalstypen

diskret: endlich oder abzählbar unendlich viele Ausprägungen

stetig: alle Werte eines Intervalls sind mögliche Ausprägungen

nominalskaliert: Ausprägungen sind Namen, keine Ordnung möglich

ordinalskaliert: Ausprägungen können geordnet, aber Abstände nicht interpretiert werden

intervallskaliert: Ausprägungen sind Zahlen, Interpretation der Abstände möglich

verhältnisskaliert: Ausprägungen besitzen sinnvollen absoluten Nullpunkt

qualitativ: endlich viele Ausprägungen, höchstens Ordinalskala

quantitativ: Ausprägungen geben Intensität wieder

Merkmalstypen

	sinnvoll interpretierbare Berechnungen			
Skalenart	auszählen	ordnen	Differenzen bilden	Quotienten bilden
nominal	ja	nein	nein	nein
ordinal	ja	ja	nein	nein
Intervall	ja	ja	ja	nein
Verhältnis	ja	ja	ja	ja

Empirische Häufigkeitsverteilung

Absolute und relative Häufigkeiten

Besitzt ein Merkmal X genau s mögliche verschiedene Ausprägungen $a_1, a_2, \ldots, a_s$, so erhalten wir die absoluten Häufigkeiten als

$$h_j := \sum_{i=1}^{n} 1_{\{x_i = a_j\}} \quad (j = 1, \ldots, s,\ h_1 + \cdots + h_s = n).$$

Diese führen uns zur empirischen Häufigkeitsverteilung des Merkmals X in der Stichprobe $x_1, \ldots, x_n$.

Oft verwendet man auch relative Häufigkeiten

$$r_j := \frac{h_j}{n} = \frac{1}{n} \sum_{i=1}^{n} 1_{\{x_i = a_j\}} \quad (j = 1, \ldots, s,\ r_1 + \cdots + r_s = 1)$$

Können wir aus den relativen Häufigkeiten immer die absoluten Häufigkeiten rekonstruieren?

Darstellungen empirischer Häufigkeitsverteilungen

Empirische Häufigkeitsverteilungen können in tabellarischer Form oder grafisch als Stab-, Säulen- oder Kreisdiagramme dargestellt werden.

Partei	Zweitstimmen	in Prozent
CDU	12 447 656	26.8
SPD	9 539 381	20.5
Die Linke	4 297 270	9.2
Grüne	4 158 400	8.9
CSU	2 869 688	6.2
FDP	4 999 449	10.7
AfD	5 878 115	12.6
Sonstige	2 325 573	5.0

Abbildung 1: Tabelle: Bundestagswahl 2017

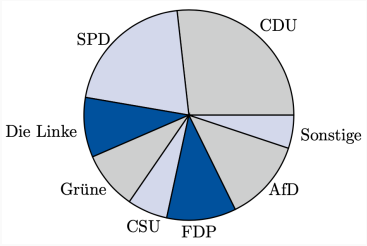


Abbildung 2: Kreisdiagramm und Stabdiagramm: Bundestagswahl 2017

Darstellungen empirischer Häufigkeitsverteilungen

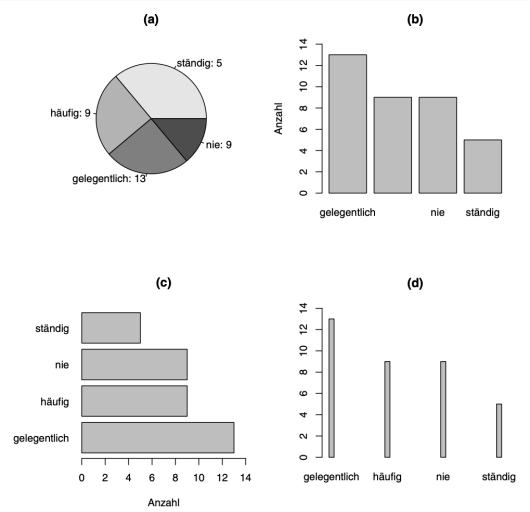


Abbildung 3: (a) Kreisdiagramm, (b) Säulendiagramm, (c) Balkendiagramm, (d) Stabdiagramm für das Merkmal "Brauchen Sie Statistik bei Ihrer aktuellen Stelle?" (bei 36 Absolventen)

Darstellungen empirischer Häufigkeitsverteilungen

Stabdiagramm: Trage über $a_1, \ldots, a_s$ jeweils einen zur x-Achse senkrechten Strich (Stab) mit Höhe $h_1, \ldots, h_s$ (oder $r_1, \ldots, r_s$) ab.

Säulendiagramm: wie Stabdiagramm, aber mit Rechtecken statt Strichen.

Balkendiagramm: wie Säulendiagramm, aber mit vertikal statt horizontal gelegter x-Achse

Kreisdiagramm: Flächen der Kreissektoren proportional zu den Häufigkeiten: Winkel des Kreissektors $j = r_j \cdot 360°$

Histogramme

- ist der Stichprobenumfang n wesentlich kleiner als die Anzahl s der möglichen Merkmalsausprägungen, so erhalten wir bei der Angabe der absoluten Häufigkeiten sehr viele Nulleinträge (durch nicht beobachtete Merkmalswerte)
- wir können aber die Stichprobenwerte $x_1, \ldots, x_n$ in Klassen einteilen
- eine Klasse ist ein *halboffenes Intervall* der Form $[a, b)$
- wählen wir nun $s + 1$ Zahlen $a_1 < a_2 < \cdots < a_s < a_{s+1}$ und damit s disjunkte Klassen

$$[a_1, a_2), [a_2, a_3), \cdots, [a_s, a_{s+1})$$

die alle Werte $x_1, \ldots, x_n$ enthalten, so erhalten wir eine Darstellung der Stichprobe in Gestalt eines Histogramms zur oberen Klasseneinteilung

Histogramme

- wir zeichnen über jedem Teilintervall $[a_j, a_{j+1})$ ein Rechteck
- die Fläche des Rechtecks über $[a_j, a_{j+1})$ entspricht dabei der relativen Klassenhäufigkeit

$$k_j := \frac{1}{n} \sum_{i=1}^{n} 1_{\{a_j \leq x_i < a_{j+1}\}}, \quad j = 1, \ldots, s$$

- die Höhe des Rechtecks über dem Intervall $[a_j, a_{j+1})$ berechnet sich also aus der Gleichung

$$d_j(a_{j+1} - a_j) = k_j, \quad j = 1, \ldots, s$$

Beispiel: Histogramm für jährliche Milchleistung von Kühen

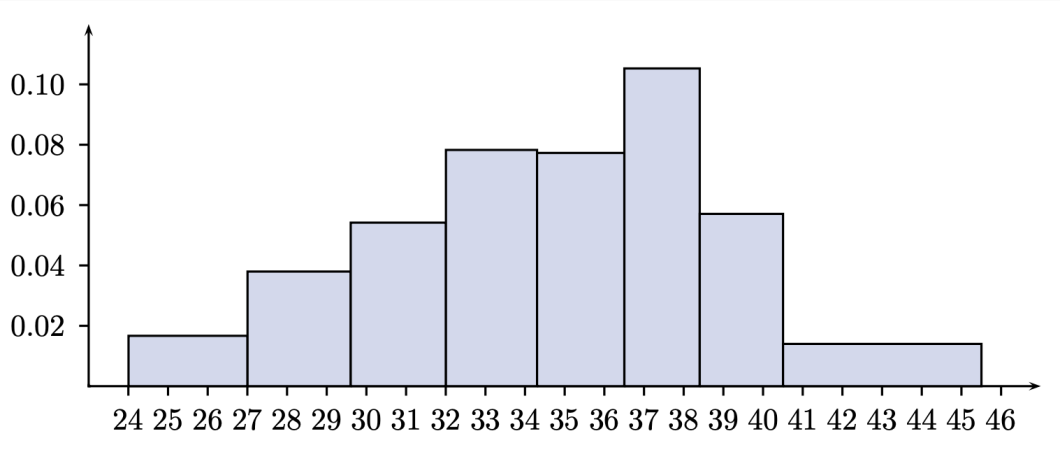
Stichprobe vom Umfang $n = 100$ (jährliche Milchleistung von Kühen, in Vielfachen von 100 Litern)

37.4	37.8	29.0	35.1	30.9	28.5	38.4	34.7	36.3	30.4
39.1	37.3	45.3	32.2	27.4	37.0	25.1	30.7	37.1	37.7
26.4	39.7	33.0	32.5	24.7	35.1	33.2	42.4	37.4	37.2
37.5	44.2	39.2	39.4	43.6	28.0	30.6	38.5	31.4	29.9
34.5	34.3	35.0	35.5	32.6	33.7	37.7	35.3	37.0	37.8
32.5	32.9	38.0	36.0	35.3	31.3	39.3	34.4	37.2	39.0
41.8	32.7	33.6	43.4	30.4	25.8	28.7	31.1	33.0	39.0
37.1	36.2	28.4	37.1	37.4	30.8	41.6	33.8	35.0	37.4
33.7	33.8	30.4	37.4	39.3	30.7	30.6	35.1	33.7	32.9
35.7	32.9	39.2	37.5	26.1	29.2	34.8	33.3	28.8	38.9

Wir wählen $s = 8$ Klassen: $a_1 = 24$, $a_2 = 27$, $a_3 = 29.6$, $a_4 = 32$, $a_5 = 34.3$, $a_6 = 36.5$, $a_7 = 38.4$, $a_8 = 40.5$, $a_9 = 45.5$.

Dann gilt z.B. $k_1 = 5/100$ und $d_1 = k_1/(a_2 - a_1) = 0.0166$.

Beispiel: Histogramm für jährliche Milchleistung von Kühen



Lage- und Streuungsmaße

Lagemaße

- es seien $x_1, \ldots, x_n$ Zahlen, die wir als Stichprobe eines quantitativen Merkmals auffassen
- wir wollen der Stichprobe eine Zahl $\gamma(x_1, \ldots, x_n)$ zuweisen, die deren grobe Lage auf der Zahlengeraden beschreibt
- einzige Bedingung:

$$\gamma(x_1 + a, \ldots, x_n + a) = \gamma(x_1, \ldots, x_n) + a$$

für jede Wahl von Zahlen $x_1, \ldots, x_n$ und a

Arithmetisches und gewichtetes Mittel

Arithmetisches Mittel (Mittelwert; Durchschnitt)

$$\overline{x}_n := \frac{1}{n}(x_1 + \cdots + x_n) = \frac{1}{n}\sum_{j=1}^{n} x_j$$

- die Summe der Quadrate $\sum_{j=1}^{n}(x_j - t)^2$ wird für $t = \overline{x}_n$ minimal

Gewichtetes Mittel

Tritt in der Stichprobe $x_1, \ldots, x_n$ der Wert a_i genau h_i-mal auf ($i = 1, 2, \ldots, s$, $h_1 + \cdots + h_s = n$), so erhalten wir

$$\overline{x}_n = \sum_{i=1}^{s} g_i a_i$$

als gewichtetes Mittel von $a_1, \ldots, a_s$ mit den Gewichten $g_i := h_i/n$, $i = 1, \ldots, s$.

Median

- sortiere die Daten $x_1, \ldots, x_n$ der Größe nach; dabei bezeichne $x_{(j)}$ den j-kleinsten Wert
- wir erhalten also die der Größe nach sortierte Reihe

$$x_{(1)} \leq x_{(2)} \leq \cdots \leq x_{(n-1)} \leq x_{(n)},$$

die sogenannte geordnete Stichprobe

Empirischer Median

Der empirische Median (Zentralwert) ist definiert als

$$x_{1/2} := \begin{cases} x_{\left(\frac{n+1}{2}\right)}, & \text{falls } n \text{ eine ungerade Zahl ist} \\ \frac{1}{2}\left(x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}\right), & \text{falls } n \text{ eine gerade Zahl ist.} \end{cases}$$

Median

- im Gegensatz zum arithmetischen Mittel $\overline{x}_n$, das $\sum_{j=1}^{n}(x_j - t)^2$ minimiert, minimiert $x_{1/2}$ die Summe

$$s(t) := \sum_{j=1}^{n} |x_j - t|$$

  der Abstände als Funktion von t
- da zur Bildung von $\overline{x}_n$ alle Stichprobenwerte mit gleichem Gewicht $1/n$ eingehen, ist das arithmetische Mittel $\overline{x}_n$ extrem ausreißeranfällig
- im Gegensatz dazu ist der Median $x_{1/2}$ robust gegenüber dem Auftreten von Ausreißern

Empirisches p-Quantil

p-Quantil

Für eine Zahl p mit $0 < p < 1$ heißt

$$x_p := \begin{cases} x_{(\lfloor np+1 \rfloor)}, & \text{falls } np \notin \mathbb{N}, \\ \frac{1}{2}\left(x_{(np)} + x_{(np+1)}\right), & \text{falls } np \in \mathbb{N}. \end{cases}$$

das (empirische) p-Quantil von $x_1, \ldots, x_n$.

Dabei bezeichnet $\lfloor y \rfloor := \max\{k \in \mathbb{Z} : k \leq y\}$.

Damit heißen $x_{0.25}$ und $x_{0.75}$ das untere bzw. obere Quartil.

Streuungsmaße

- im Gegensatz zu einem Lagemaß ändert sich der Wert eines Streuungsmaßes $\sigma(x_1, \ldots, x_n)$ bei Verschiebungen der Daten nicht:

$$\sigma(x_1 + a, \ldots, x_n + a) = \sigma(x_1, \ldots, x_n)$$

für jede Wahl von $x_1, \ldots, x_n$ und a.

(empirische) Varianz / Stichprobenvarianz

$$s^2 := \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \overline{x}_n)^2$$

(empirische) Standardabweichung / Stichprobenstandardabweichung

$$s := \sqrt{s^2} = \sqrt{\frac{1}{n-1} \sum_{j=1}^{n} (x_j - \overline{x}_n)^2}$$

- ein Nachteil von s und s^2 ist, wie bereits beim arithmetischen Mittel, die Empfindlichkeit gegenüber Ausreißern

weitere Streuungsmaße

- mittlere absolute Abweichung

$$\frac{1}{n}\sum_{j=1}^{n}|x_j - \overline{x}_n|,$$

- Stichprobenspannweite

$$x_{(n)} - x_{(1)} = \max_{1\le j\le n} x_j - \min_{1\le j\le n} x_j,$$

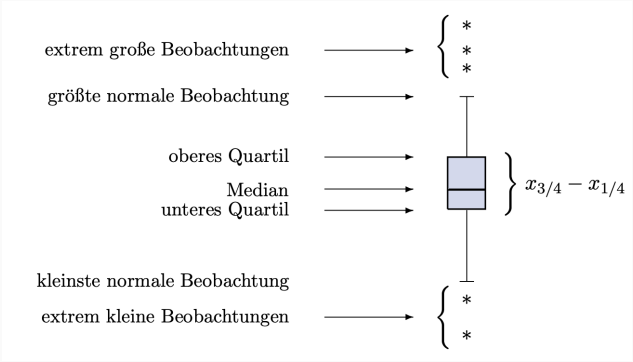
- Quartilsabstand $x_{3/4} - x_{1/4}$
- und die als empirischer Median von $|x_1 - x_{1/2}|, |x_2 - x_{1/2}| \ldots, |x_n - x_{1/2}|$, definierte Median-Abweichung von $x_1, \ldots, x_n$.

- Quartilsabstand und Median-Abweichung sind robuste Streuungsmaße

Box-Plots

Box-Plots

- der Box-Plot dient dem schnellen Vergleich verschiedener Stichproben
- Quantile werden zur Darstellung von Lage und Streuung benutzt und potentielle Ausreißer hervorgehoben



- man zeichnet eine beim Median unterteilte Box vom unteren zum oberen Quartil
- der nach oben aufgesetzte Stab endet bei der größten Beobachtung, die kleiner als $x_{3/4} + 1.5(x_{3/4} - x_{1/4})$ ist
- unten: größer als $x_{1/4} - 1.5(x_{3/4} - x_{1/4})$

Übungsaufgabe

Die unten stehenden Werte sind Druckfestigkeiten (in 0.1 N/mm^2), die an 30 Betonwürfeln ermittelt wurden.

374	358	341	355	342	334	353	346	355	344	349	330	352	328	336
359	361	345	324	386	335	371	358	328	353	352	366	354	378	324

Bestimmen Sie

(a) das arithmetische Mittel

(b) die empirische Varianz und die Standardabweichung der Stichprobe

(c) das untere Quartil und das 90%-Quantil

(d) die Stichprobenspannweite und den Quartilsabstand

(e) die Median-Abweichung

Literatur

- Henze, Norbert; Stochastik für Einsteiger; Springer; 10. Auflage; 2013
- Fahrmeir, Ludwig; Heumann, Christian; Künstler, Rita; Pigeot, Iris; Tutz, Gerhard; Statistik; Springer; 8. Auflage; 2016