2. Deskriptive Statistik

Hajo Holzmann

Philipps-Universität Marburg

- Untersuchungseinheiten: mögliche, statistisch zu erfassende Einheiten
- je Untersuchungseinheit: ein oder mehrere Merkmale oder Variablen beobachten
- mögliche Werte eines Merkmals: Merkmalsausprägungen

- Untersuchungseinheiten: mögliche, statistisch zu erfassende Einheiten
- je Untersuchungseinheit: ein oder mehrere Merkmale oder Variablen beobachten
- mögliche Werte eines Merkmals: Merkmalsausprägungen

Untersuchungseinheit	Merkmal	Merkmalausprägungen
Baum	Baumart	Eiche, Buche, ...
arbeitslose Person	Schulabschluss	keiner, Hauptschule, Realschule, Gymnasium
Person	Familienstand	ledig, verheiratet, geschieden, ...

Stichproben

Grundgesamtheit = Menge der möglichen Untersuchungseinheiten

Stichprobe = zufällig gewonnene, endliche Teilmenge der Grundgesamtheit

Stichprobenumfang = Anzahl der erhobenen Daten

Datentypen

- *Kategorielle (oder nominale) Daten* für jedes Datum welche Kategorie, z.B. Autotypen, Baumart, Nationalität
- *Ordinale Daten* kategorielle Daten mit geordneten Kategorien, z.B. Noten, Erdbebenstärke auf Richter Skala
- *Zähldaten* oder *diskrete Daten*: Zählen bestimmter Merkmale , z.B. Anzahl mit Geigerzähler registrierten Zerfälle einer Probe,
- *Stetige (oder kontinuierliche) Daten* können in Wertebereich – zumindest theoretisch – jeden beliebigen Zahlenwert annehmen, z.B. Größe, Alter, Länge.

Datentypen

- *Kategorielle (oder nominale) Daten* für jedes Datum welche Kategorie, z.B. Autotypen, Baumart, Nationalität
- *Ordinale Daten* kategorielle Daten mit geordneten Kategorien, z.B. Noten, Erdbebenstärke auf Richter Skala
- *Zähldaten* oder *diskrete Daten*: Zählen bestimmter Merkmale , z.B. Anzahl mit Geigerzähler registrierten Zerfälle einer Probe,
- *Stetige (oder kontinuierliche) Daten* können in Wertebereich – zumindest theoretisch – jeden beliebigen Zahlenwert annehmen, z.B. Größe, Alter, Länge.

qualitative Daten: kategorielle und ordinale Daten

quantitative oder metrische Daten: Zähldaten und stetige Daten.

2.2 Beschreibung kategorieller Daten

- absolute Häufigkeiten: Wieviele Daten in jeder Kategorie

  $\rightarrow$ auch Kategorien erwähnen, in die keine Daten fallen.

2.2 Beschreibung kategorieller Daten

- absolute Häufigkeiten: Wieviele Daten in jeder Kategorie
  $\rightarrow$ auch Kategorien erwähnen, in die keine Daten fallen.
- relative Häufigkeiten: Anteil der Daten in jeder Kategorie
  $\rightarrow$ absolute Häufigkeiten / Stichprobenumfang.

  stets zusammen mit Stichprobenumfang angeben.

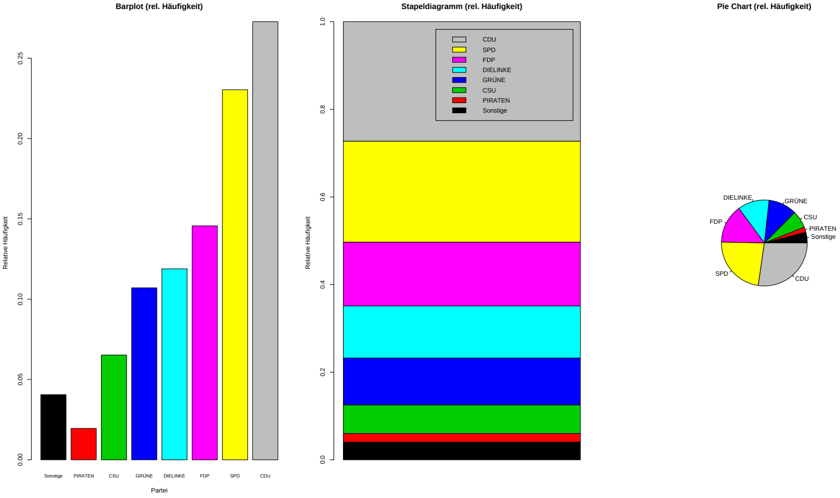
2.2 Beschreibung kategorieller Daten

- absolute Häufigkeiten: Wieviele Daten in jeder Kategorie
  $\rightarrow$ auch Kategorien erwähnen, in die keine Daten fallen.
- relative Häufigkeiten: Anteil der Daten in jeder Kategorie
  $\rightarrow$ absolute Häufigkeiten / Stichprobenumfang.
  stets zusammen mit Stichprobenumfang angeben.

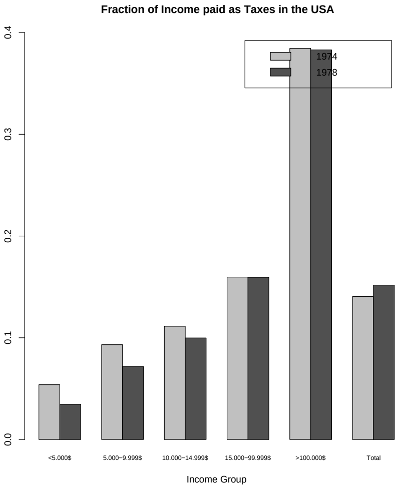
Visualisierung: relative / absolute Häufigkeiten als

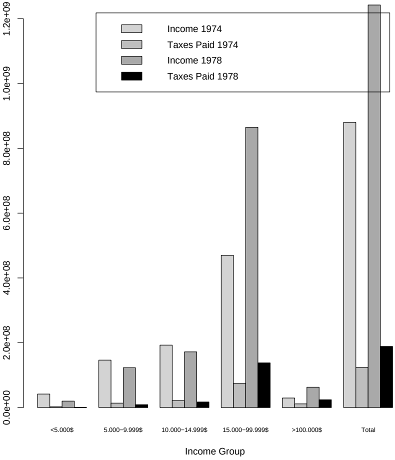
- Balkendiagramme: einzelne Balken
- Stapeldiagramm: übereinander in einem Balken der Größe nach
- Tortendiagramme bzw. Kreisdiagramm: als Kreis / Tortensegmente

Visualisierung (Wahlergebnisse)



Visualisierung (Simpson Paradoxon)





2.3 Zusammenfassung numerischer Daten

Lagemaße: Wo (auf der reellen Achse) befinden sich die Daten?

Streumaße: Wie weit *streuen* die Daten um ein Lagemaß?

2.3 Zusammenfassung numerischer Daten

Lagemaße: Wo (auf der reellen Achse) befinden sich die Daten?

Streumaße: Wie weit *streuen* die Daten um ein Lagemaß?

Weiter:

Maße für Schiefe: Sind die Daten symmetrisch um ihr Lagemaß?

Maße für heavy tails: Gibt es viele Daten, die besonders weit vom Lagemaß entfernt liegen?

Mittelwert: arithmetisches Mittel der Daten.

Daten $x_1, \ldots, x_n \in \mathbb{R}$, dann

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{x_1 + \cdots + x_n}{n},$$

Lagemaße: Mittelwert

Mittelwert: arithmetisches Mittel der Daten.

Daten $x_1, \ldots, x_n \in \mathbb{R}$, dann

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i = \frac{x_1 + \cdots + x_n}{n},$$

gewichtetes Mittel: Gewicht $g_i > 0$ für Beobachtung x_i, dann

$$\frac{\sum_{i=1}^{n} g_i x_i}{\sum_{i=1}^{n} g_i} = \frac{g_1 x_1 + \cdots + g_n x_n}{g_1 + \cdots + g_n},$$

Lagemaße: Median

Ordnungsstatistiken: geordneten Werte $x_{(1)} \leq \ldots \leq x_{(n)}$, d.h. $x_{(1)}$ kleinste, $x_{(n)}$ größte Wert.

Median (lat. *medius:* der mittlere) einfachste Lagemaß.

$$\mathrm{med}(x) = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{für } n \text{ ungerade} \\ \dfrac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{für } n \text{ gerade,} \end{cases}$$

$\rightarrow$ mindestens 50% der Daten $\geq$ und 50% der Daten $\leq \mathrm{med}(x)$.

$x = (x_1, \dots, x_n)$ beobachtete Daten.

Varianz:

$$\mathrm{var}(x) = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{(x_1 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}$$

Standardabweichung (engl. *standard deviation*)

$$\mathrm{sd}(x) = \sqrt{\mathrm{var}(x)}.$$

Variationskoeffizienten: relative Schwankung im Verhältnis zu ihrem Mittelwert

$$\frac{\mathrm{sd}(x)}{|\bar{x}|}$$

Bsp.: Energieumsatzrate

Interquartilsabstand

Quantile: für $0 < \alpha < 1$

$$q_\alpha(x) = \begin{cases} x_{([n\cdot\alpha+1])}, & \text{falls } n \cdot \alpha \text{ keine ganze Zahl ist,} \\ \frac{1}{2}\left(x_{(n\cdot\alpha)} + x_{(n\cdot\alpha+1)}\right), & \text{falls } n \cdot \alpha \text{ eine ganze Zahl ist.} \end{cases}$$

$\rightarrow$ mindestens $\alpha \cdot 100\%$ der Daten $\leq q_\alpha(x)$
und $(1-\alpha) \cdot 100\%$ der Daten $\geq q_\alpha(x)$.

Quantile: für $0 < \alpha < 1$

$$q_\alpha(x) = \begin{cases} x_{([n\cdot\alpha+1])}, & \text{falls } n\cdot\alpha \text{ keine ganze Zahl ist,} \\ \frac{1}{2}\left(x_{(n\cdot\alpha)} + x_{(n\cdot\alpha+1)}\right), & \text{falls } n\cdot\alpha \text{ eine ganze Zahl ist.} \end{cases}$$

$\rightarrow$ mindestens $\alpha \cdot 100\%$ der Daten $\leq q_\alpha(x)$
und $(1-\alpha)\cdot 100\%$ der Daten $\geq q_\alpha(x)$.

unteres Quartil: $q_{0,25}(x)$,

oberes Quartil: $q_{0,75}(x)$,

Interquartilsabstand

$$\mathrm{IQR}(x) = q_{0,75}(x) - q_{0,25}(x).$$

Maß für Schiefe

Schiefe (engl.: skewness) von $x_1, \ldots, x_n$:

$$\operatorname{skew}(x) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\operatorname{sd}(x)} \right)^3 .$$

$\to$ kennzeichnet Abweichung von symmetrischer Lage um $\bar{x}$.

Maß für Schiefe

Schiefe (engl.: skewness) von $x_1, \ldots, x_n$:

$$\text{skew}(x) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\text{sd}(x)} \right)^3 .$$

$\rightarrow$ kennzeichnet Abweichung von symmetrischer Lage um $\bar{x}$.

Ist $\text{skew}(x) < 0$: linksschief

Ist $\text{skew}(x) > 0$: rechtsschief.

Verteilungsschwänze (heavy tails)

Kurtosis von $x_1, \dots, x_n$:

$$\text{kurtosis}(x) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\text{sd}(x)} \right)^4 - 3.$$

$\rightarrow$ kennzeichnet Abweichung von Verteilungsschwänzen der Normalverteilung.

Verteilungsschwänze (heavy tails)

Kurtosis von $x_1, \dots, x_n$:

$$\text{kurtosis}(x) = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\text{sd}(x)} \right)^4 - 3.$$

$\rightarrow$ kennzeichnet Abweichung von Verteilungsschwänzen der Normalverteilung.

Ist $\text{kurtosis}(x) < 0$: low tails

Ist $\text{kurtosis}(x) > 0$: heavy tails.

im Vergleich zur Normalverteilung.

Boxplot

- Graphische Darstellung der 5 Zahlen Median, unteres und oberes Quartil, Max. und Min.
- Box. zwischen $q_{0.25}$ und $q_{0.75}$, darin Median als Strich
- Striche (engl. Whiskers) bis Max. und Min.

Boxplot

- Graphische Darstellung der 5 Zahlen Median, unteres und oberes Quartil, Max. und Min.
- Box. zwischen $q_{0.25}$ und $q_{0.75}$, darin Median als Strich
- Striche (engl. Whiskers) bis Max. und Min.

Histogramm

- Unterteilung des Wertebereichs in disjunkte Intervalle,
- Plotte Rechtecke auf Intervalle, Höhe: Anzahl (Anteil) Daten in dem Intervall

Boxplot

- Graphische Darstellung der 5 Zahlen Median, unteres und oberes Quartil, Max. und Min.
- Box. zwischen $q_{0.25}$ und $q_{0.75}$, darin Median als Strich
- Striche (engl. Whiskers) bis Max. und Min.

Histogramm

- Unterteilung des Wertebereichs in disjunkte Intervalle,
- Plotte Rechtecke auf Intervalle, Höhe: Anzahl (Anteil) Daten in dem Intervall

Rug-Plot

- Ergänzend zu Histogramm,
- Plotte Daten als Striche auf x-Achse

lineare Transformationen: Für $a, b \in \mathbb{R}$, $a \neq 0$,

$$f(x_i) = ax_i + b, \qquad i = 1, \ldots, n.$$

Bsp.: Grad Celsius in Grad Kelvin, Euro in Dollar.

lineare Transformationen: Für $a, b \in \mathbb{R}$, $a \neq 0$,

$$f(x_i) = ax_i + b, \qquad i = 1, \ldots, n.$$

Bsp.: Grad Celsius in Grad Kelvin, Euro in Dollar.

Standardisierung.

$$f(x_i) = \frac{x_i - \bar{x}}{\operatorname{sd} x}.$$

Transformation positiver Daten $x_i > 0$.

- Logarithmieren: $f(x_i) = \log(x_i)$.

  $\rightarrow$ rechtsschiefe Daten symmetrisch machen.

Logarithmus und Potenzen

Transformation positiver Daten $x_i > 0$.

- Logarithmieren: $f(x_i) = \log(x_i)$.

  $\rightarrow$ rechtsschiefe Daten symmetrisch machen.
- Allgemeiner: Box-Cox-Transformationen für $\alpha > 0$:

$$f(x_i) = \frac{x_i^\alpha - 1}{\alpha},$$

  - für $\alpha \to 0$: erhalte Logarithmus.
  - für $0 < \alpha < 1$: rechtsschiefe Daten symmetrisch machen.
  - für $1 < \alpha$: linksschiefe Daten symmetrisch machen.

Visualisierung (Deutschland Daten)