Inhalt

1 Grundlagen

In diesem Kapitel besprechen wir die wichtigsten Begriffe und Unterscheidungen von Merkmalen und ihren Skalen.

1.1 Wichtige Begriffe

Grundbegriffe

Beginnen wir mit den grundlegenden Begriffen der beschreibenden Statistik.

- **Merkmalsträger** sind statistische Einheiten, an welchen Größen bzw. Merkmale beobachtet werden.
- Bei der **Grundgesamtheit** handelt es sich um die Menge aller für die Fragestellung relevanten Merkmalsträger.
- Die **Stichprobe** stellt die untersuchte Teilmenge der Grundgesamtheit dar.
- **Merkmale** sind Größen oder Variablen. Ein Merkmal nimmt für gewöhnlich verschiedene Merkmalsausprägungen an.
- Bei **Merkmalsausprägungen** handelt es sich um die Werte, welche ein Merkmal annehmen kann.
- Bei dem **Wertebereich** handelt es sich um die Menge der Merkmalsausprägungen.
- Eine **Beobachtung** ist eine Ausprägung, welche direkt an einer statistischen Einheit ermittelt wurde.

Merkmal, Merkmalsträger, Merkmalsausprägung

Beispiel

Wir machen uns das anhand eines einfachen Beispiels deutlich:

- **Grundgesamtheit:** Studierende eines Studienganges
- **Merkmalsträger:** Student
- **Merkmal:** Geschlecht

- **Merkmalsausprägungen:** männlich, weiblich, divers

Merkmale sind für uns von besonderer Bedeutung, da diese bei einer Erhebung untersucht werden. Wir können Merkmale wie folgt einteilen:

- **Diskrete** und **stetige** Merkmale
- **Qualitative** und **quantitative** Merkmale

Diskrete und stetige Merkmale

Diskrete und stetige Verteilung

- Ein Merkmal, welches nur eine abzählbare oder endliche Menge von Ausprägungen annehmen kann, heißt **diskret**.
  Beispielhaft können die Anzahl von Studierenden in unserem Studiengang oder Anzahl von Fachsemestern genannt werden.
- Ein Merkmal, welches zumindest theoretisch alle reellen Zahlen annehmen kann, nennen wir **stetig**.
  Gewicht, Größe, Preise oder Temperaturen sind Beispiele für Variablen, die wir als stetig bezeichnen können.

Qualitative und quantitative Merkmale

Quantitative und qualitative Merkmale

- Um ein **qualitatives** Merkmal handelt es sich, wenn die Merkmalsausprägungen durch verbale Ausdrücke benannt werden können.
  Dazu gehören z. B. Farbe, Geschlecht und Wochentag.
- Alternativ dazu ist ein Merkmal **quantitativ**, wenn es sich bei den verschiedenen Merkmalsausprägungen um Zahlen handelt.
  Wir können z. B. das Alter, Gewicht oder Einkommen als quantitative Merkmale bezeichnen.

1.2 Skalen

Die verschieden Ausprägungen der Merkmale werden in Skalen erfasst. Dabei unterscheiden wir drei Hauptskalen:

- **Nominalskala**
- **Ordinalskala**
- **Kardinalskala** (metrische Skala)
  - Intervallskala
  - Verhältnisskala
  - Absolutskala

2 Häufigkeiten

In diesem Kapitel betrachten wir die absoluten sowie relativen Häufigkeiten und beschäftigen uns mit der Fragestellung, wie sich diese in klassierten und unklassierten Daten berechnen lassen, Außerdem werden wir verschiedene Möglichkeiten der grafischen Darstellung kennenlernen.

2.1 Absolute und relative Häufigkeiten

Bei den Häufigkeiten unterscheiden wir grundsätzlich:

- **Absolute Häufigkeiten**
  Sie geben an, wie oft eine bestimmte Merkmalsausprägung oder ein bestimmtes Ereignis auftritt.
- **Relative Häufigkeiten**
  Sie geben an, welchen Anteil eine bestimmte Merkmalsausprägung von der Gesamtmenge n annimmt.

Es liegt ein Merkmal X mit den Ausprägungen $a_1, \dots a_l$ in einer Stichprobe im Umfang von $x_1, \dots x_n$ vor, dann können wir mit

$$h(a_j) = |\{i \in \{1, \dots, n\} : \ x_i{=}a_j\}|$$

die Anzahl angeben, in welcher die Ausprägungen $a_j, j \in \{1, \dots, l\}$ vorkommen.

Die **absolute Häufigkeit** eines einzelnen Merkmals a_j bezeichnen wir dabei mit:

$$h(a_j)$$

Die **relative Häufigkeit** eines einzelnen Merkmals a_j bezeichnen wir dabei mit:

$$f(a_j) = \frac{h(a_j)}{n}$$

Absolute und relative Häufigkeit

Beispiel

Es wurden zehn Kinder nach ihrer Lieblingsfarbe gefragt. Anhand dieser nominalen Beobachtungen machen wir uns deutlich, wie wir die Häufigkeiten bestimmen.
Folgende Beobachtungen liegen vor: rot, rot, grün, rot, blau, blau, grün, blau, grün, grün

3 Lagemaße

Im Allgemeinen stehen verschiedene Lagemaße zur Verfügung, um die Lage von Daten und deren Mitte zu beschreiben. Für alle Lagemaße gilt dabei gleichermaßen, dass die vorhandenen Daten auf eine einzige Größe reduziert werden. Diese Größe gibt einen Wert an, der für die Daten typisch ist und um den herum die Daten liegen.

3.1 Modus / Modalwert

> Der **Modus** oder **Modalwert** x_{mod} ist die Ausprägung, welche am häufigsten in einer Stichprobe vorkommt. Jede Ausprägung a_j, deren absolute Häufigkeit
>
> $$h(a_j) \geq h(a_{j^*}) \quad \forall j^* \in \{1, \ldots, l\}$$
>
> erfüllt, wird als Modus x_{mod} bezeichnet.
>
> Sollten mehrere Ausprägungen gleich häufig vorkommen, so haben wir mehrere Modi.

Modalwert, Median (Zentralwert)

Der Modus...

- kann für alle Skalenarten (nominal, ordinal und metrisch) bestimmt werden.
- kommt mindestens einmal vor. Es gibt keine Stichprobe, in welcher nicht mindestens ein Modus vorhanden ist.
- ist robust gegenüber Ausreißern.
- kann nur Werte annehmen, die beobachtet wurden.
- unterscheidet sich zu den anderen Lagemaßen in der Hinsicht, dass auch mehrere Modi vorhanden sein können und nicht immer nur ein Wert.

Ungruppierte Daten

Wir machen uns nun mit Hilfe von drei verschiedenen Beispielen mit unterschiedlichem Skalenniveau deutlich, wie sich jeweils der Modalwert bestimmen lässt.

Beispiel 1: ein Modalwert - Nominalskala

10 Schulkinder haben ihre Lieblingsfarben genannt:
rot, rot, grün, grün, blau, rot, grün, gelb, grün, blau
Wir bestimmen für jede Ausprägung die absolute Häufigkeit.

rot	grün	blau	gelb
3	4	2	1

Die Ausprägung, welche am häufigsten vorhanden ist, ist unser Modalwert: x_{mod} : grün

Beispiel 2: mehrere Modalwerte - Ordinalskala

Es liegen zehn Schulnoten einer Klassenarbeit vor:
sehr gut, gut, gut, ausreichend, mangelhaft, sehr gut, sehr gut, gut, ungenügend, ungenügend

sehr gut	gut	ausreichend	mangelhaft	ungenügend
3	3	1	1	2

Tauchen mehrere Ausprägungen gleich oft auf, so sind alle entsprechenden Ausprägungen Modalwerte. x_{mod} : sehr gut, gut

Beispiel 3: mehrere Modalwerte - metrische Skala

Es liegen die Monatsmieten für Studentenwohnungen und WG-Zimmer in Euro vor:
210; 225; 234; 266; 310; 340; 352; 370; 442; 462; 467; 494; 522; 563; 566; 582

x_{mod}: 210; 225; 234; 266; 310; 340; 352; 370; 442; 462; 467; 494; 522; 563; 566; 582

Klassierte Daten

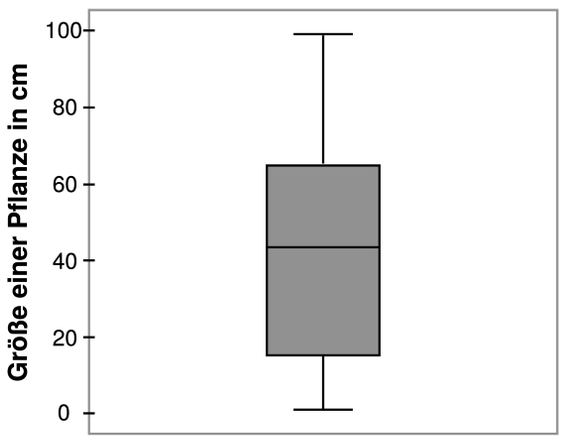
Für den Fall, dass wir den Modus von einem klassierten Merkmal bestimmen wollen, wird als Modus häufig die Klassenmitte der Modalklasse verwendet.

- Bei **äquidistanten** Klassen liegt der Modus in der am häufigsten vorkommenden Klasse. In der Grafik ist dies der flächengrößte Balken. Da alle Balken gleich breit sind, ist es der höchste Balken.

- Bei **nicht-äquidistanten** Klassen, also Klassen mit unterschiedlichen Klassenbreiten, liegt der Modalwert in der Klasse mit der größten Häufigkeitsdichte. Durch die unterschiedlichen Breiten ist dies nicht so einfach zu erkennen.

> Die Mitte der Klasse, welche den größten Histogrammwert aufweist, ist als **Modus** definiert. Die entsprechende Klasse nennt sich **Modalklasse**.

- Die Box wird durch die Whiskers oder Schnurrbarthaare erweitert. Sie stellen die Verbindungslinie vom oberen bzw. unteren Quartil zu den Extremwerten, d. h. dem größten und kleinsten Wert dar.



Zum Erstellen eines **einfachen Boxplot** gehen wir nach dem folgenden Prinzip vor:

1. Ordnen der in der Stichprobe vorliegenden Beobachtungen $x_1, \ldots, x_n$, sodass $x_{(1)}, \ldots, x_{(n)}$ der Größe nach geordnet vorliegen.
2. Berechnung des Medians $\tilde{x}_{0,5}$, des unteren Quartils $\tilde{x}_{0,25}$ und des oberen Quartils $\tilde{x}_{0,75}$.
3. Bestimmen der Extremwerte.
4. Zeichnen des Boxplots.

Beispiel

Wir möchten die Gewichtsverteilung in Kilogramm von 15 zufällig ausgewählten Kindern aus der ersten Klasse in einem Boxplot darstellen. Gewicht der Kinder in kg:

15,50 ; 19,30 ; 22,90 ; 17,80 ; 18,70 ; 27,40 ; 19,90 ; 24,30 ; 22,60 ; 18,90 ; 29,70 ;
17,90 ; 22,30 ; 18,40 ; 23,60

1. Sortieren der Beobachtungen:

   15,50 ; 17,80 ; 17,90 ; 18,40 ; 18,70 ; 18,90 ; 19,30 ; 19,90 ; 22,30 ; 22,60 ; 22,90 ;
   23,60 ; 24,30 ; 31,70 ; 40,00

2. Berechnen von $\tilde{x}_{0,5}$, $\tilde{x}_{0,25}$ und $\tilde{x}_{0,75}$;
   - $\tilde{x}_{0,5} = x_{(7,5)} = x_{(8)} = 19{,}90$
   - $\tilde{x}_{0,25} = x_{(3,75)} = x_{(4)} = 18{,}40$
   - $\tilde{x}_{0,75} = x_{(11,25)} = x_{(12)} = 23{,}60$
3. Bestimmen von $x_{\max} = 40{,}00$, sowie $x_{\min} = 15{,}50$.
4. Zeichnen des Boxplots.

Beispiel

Wie die Lorenzkurve berechnet wird, machen wir uns anhand eines Beispiels deutlich. Dafür betrachten wir die erzielten Treffer beim Dosenwerfen von drei Gruppen.

- Gruppe A - Merkmal A: 2,6,8,8,16 - Gesamtsumme:40
- Gruppe B - Merkmal B: 4,4,8,16 - Gesamtsumme: 32
- Gruppe C - Merkmal A: 1,1,4,16 - Gesamtsumme: 22

Als Beispiel berechnen wir den dritten Punkt der Lorenzkurve für die Gruppe A. Wir haben $n = 5$ Beobachtungen (Wert) vorliegen: $u_3 = \frac{3}{5}$ und $v_3 = \frac{2+6+8}{40}$. Auf die gleiche Art lassen sich alle anderen Punkte berechnen. Die Nummer der Beobachtungen kennzeichnen wir mit #.

Koordinatenpaare für Gr. A:

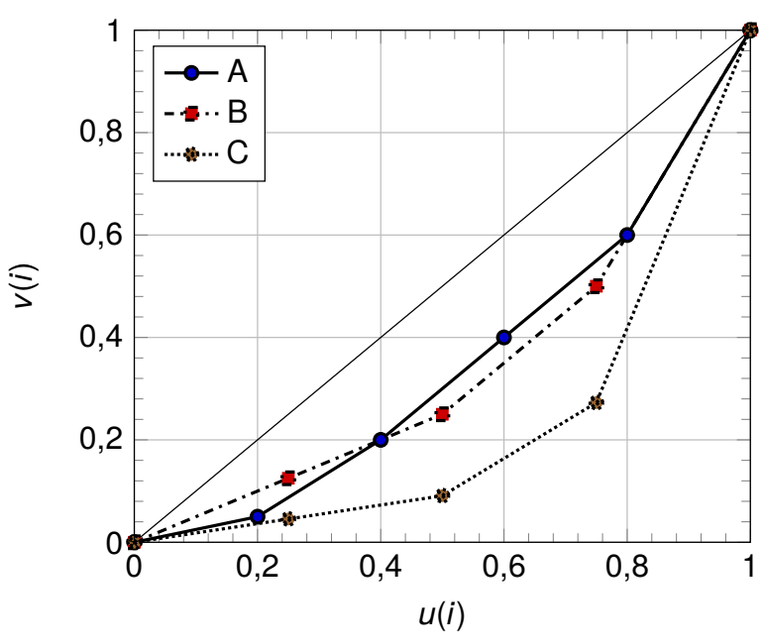
#	Wert	u_i	v_i
1	2	$\frac{1}{5}$	$\frac{2}{40}$
2	6	$\frac{2}{5}$	$\frac{8}{40}$
3	8	$\frac{3}{5}$	$\frac{16}{40}$
4	8	$\frac{4}{5}$	$\frac{24}{40}$
5	16	1	1

Koordinatenpaare für Gr. B:

#	Wert	u_i	v_i
1	4	$\frac{1}{4}$	$\frac{4}{32}$
2	4	$\frac{2}{4}$	$\frac{8}{32}$
3	8	$\frac{3}{4}$	$\frac{16}{32}$
4	16	1	1

Koordinatenpaare für Gr. C:

#	Wert	u_i	v_i
1	1	$\frac{1}{4}$	$\frac{1}{22}$
2	1	$\frac{2}{4}$	$\frac{2}{22}$
3	4	$\frac{3}{4}$	$\frac{6}{22}$
4	16	1	1

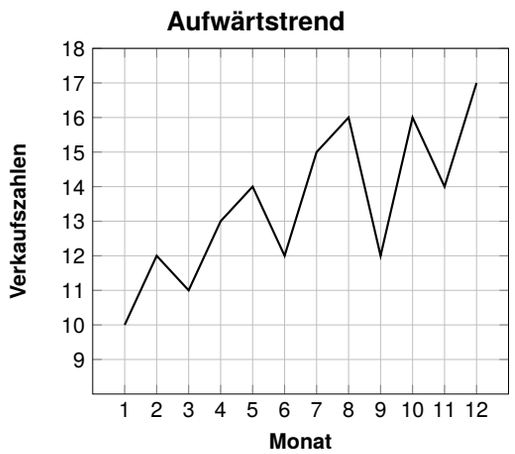


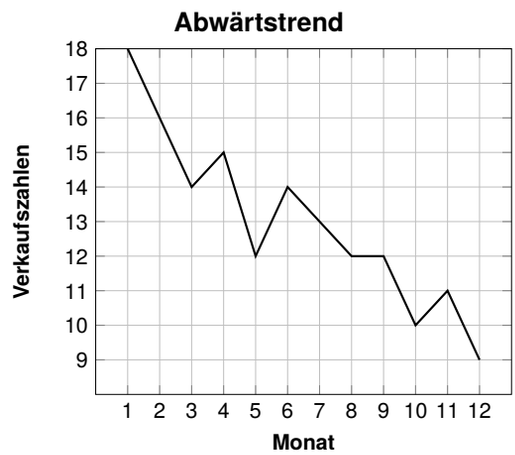
Grafische Beispiele für Zeitreihen

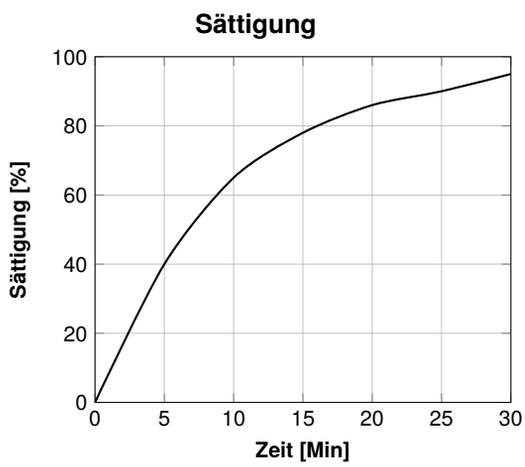
Mit Zeitreihen können wir unterschiedlichen Fragestellungen beantworten, wie z. B.:

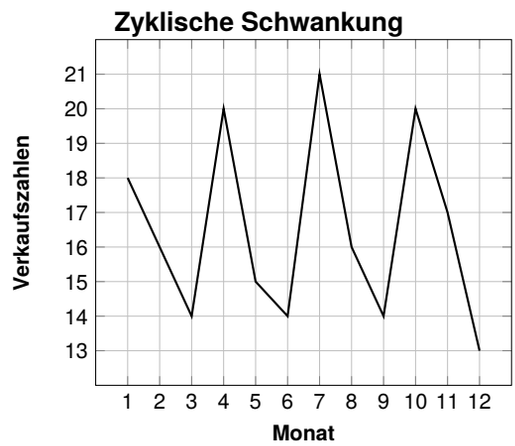
- Lässt sich ein Trend erkennen? Wenn ja, um welchen Verlauf handelt es sich?
- Lassen sich zyklische Schwankungen erkennen?
- Lassen sich mit Hilfe der vorhandenen Zeitreihe Prognosen für die Zukunft geben?

Um die Fragestellungen anschließend zu klären, werden im ersten Schritt die grafischen Verläufe angeguckt, da diese Methode am einfachsten ist. Damit lassen sich erste Rückschlüsse auf das Vorliegen von Trends oder Saisonverhalten schließen.









In vielen Fällen reicht eine Grafik aber alleine nicht aus, um die Zielfragestellung ausführlich zu beantworten, sodass Berechnungen durchgeführt werden müssen.

11.1.1 Besondere Ereignisse

Bei besonderen Ereignissen handelt es sich um Ereignisse, welche sicher auftreten. Dabei können zwei verschiedene Fälle eintreten:

- **sicheres Ereignis**
  Es gilt: $\Omega \subset \Omega$. Es handelt sich um ein sicheres Ereignis, weil alle Ereignisse in Ω enthalten sind.
- **unmögliches Ereignis**
  Es gilt: $\varnothing \subset \Omega$. Es handelt sich um ein unmögliches Ereignis, weil kein Ereignis in Ω enthalten ist, z. B. bei einem Münzwurf das „Werfen der Farbe blau“.

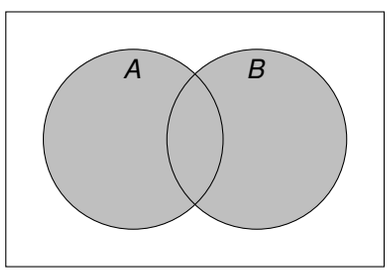
11.1.2 Venn-Diagramm

Venn-Diagramme

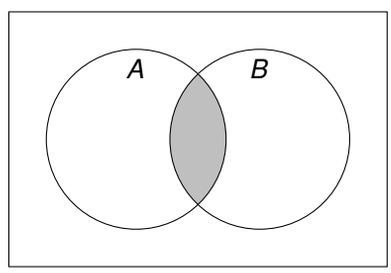
In **Venn-Diagrammen** werden Ereignisse und ihre mögliche Kombination, wie z. B. ihre Schnittmenge veranschaulicht.

Ereignisse lassen sich unterschiedlich miteinander kombinieren, die häufigsten Kombinationen werden nachfolgend an zwei Ereignissen A und B dargestellt. Dies lässt sich auf noch mehr Ereignisse erweitern.

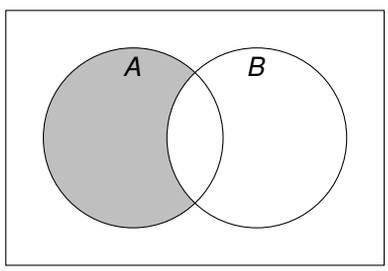
Vereinigung
A oder B: $A \cup B$



Durchschnitt
A und B: $A \cap B$



Differenz
A, aber nicht B: $A \backslash B$



Komplementär
Nicht in A und B: $\overline{A \cap B}$