Skript zur Vorlesung Stochastik

(nach einem Skript von Prof. Dr. Friedrich Pukelsheim, Universität Augsburg und zu Beginn nach einem Skript von Prof. Dr. Ulrich Wellisch, TH Rosenheim)

Prof. Dr. Wolfgang Bischof

TH Augsburg

11. März 2025

Inhaltsverzeichnis

Einleitung

Stochastik ist die Lehre vom Rechnen mit dem Zufall mit mathematischen Methoden. Die Stochastik umfasst die beiden Fachgebiete Wahrscheinlichkeitstheorie und Statistik.

Im Rahmen dieser Vorlesung werden die beschreibende Statistik und grundlegende Begriffe der Wahrscheinlichkeitstheorie unter Einbezug der Maß- und Integrationstheorie behandelt.

In der Wahrscheinlichkeitstheorie geht es um das Fundament und das Rechengerüst für die mathematische Formulierung des Zufalls. Es werden grundlegende Gesetzmäßigkeiten wie die Gesetze der großen Zahlen oder der zentraler Grenzwertsatz bereitgestellt.

Statistik ist die Kunst, verständliche Zusammenfassungen von Daten mit dem Ziel zu erstellen, Schlüsse daraus zu ziehen. Die Statistik befasst sich mit dem Sammeln von Daten, ihrer anschließenden Beschreibung und ihrer Analyse. Die Grundlage in der angewandten Statistik ist immer ein realer Datensatz, meist eine Stichprobe. In der mathematischen Statistik werden mithilfe der Wahrscheinlichkeitsrechnung die statistischen Verfahren fundiert.

In der angewandten Datenanalyse und statistischen Modellbildung gibt es Überschneidungen und Verbindungen zu den Fachgebieten: Data Analytics, Data Science, Künstliche Intelligenz, Statistisches und Maschinelles Lernen. Diese sowohl in der Forschung als auch in der Anwendung hochaktuellen Themenbereiche bestehen aus der Kombination von Statistik- und Informatik-Inhalten. In Predictive Analytics/Maintenance werden Verfahren des statistischen und maschinellen Lernens verwendet und kombiniert.

Prinzipiell unterscheidet man Statistik in:

(i) *der beschreibenden Statistik, die …*
einfache Zusammenfassungen der beobachteten Daten enthält. Die Zusammenfassungen können entweder aus quantitativen Größen, die aus den Daten berechnet werden können, oder aus einfach zu verstehenden Grafiken bestehen.
Beispiele: Arithmetische Mittel, Median, Häufigkeitstabelle, Säulendiagramm, Histogramm und Boxplot.

(ii) *der schließenden (induktiven, inferentiellen) Statistik, die …*
auf der Wahrscheinlichkeitstheorie basierende Verfahren benutzt, um Schlüsse aus Stichproben zu ziehen.
Beispiele: Parameterschätzung, Hypothesentests, Regression und Anpassungstests.

In dieser Vorlesung werden wir nur die beschreibende Statistik behandeln. Die induktive Statistik ist dann Gegenstand der gleichnamigen Vorlesung im nächsten Semester.

ABC 4

I Beschreibende Statistik

(teilweise nach Skript „Einführung: Stochastik“ von Prof. U. Wellisch, TH Rosenheim)

I.1 Stichproben

Die Grundlage jeder angewandten, statistischen Analyse ist ein Datensatz, in dem Informationen zu einem zufälligen Vorgang enthalten sind. Man nennt die in einer statistischen Untersuchung interessierenden und erfassten Größen „Merkmale“ oder „Variablen“ (Beispiel: Geschlecht und Alter einer Umfrageteilnehmerin). Die Objekte, an denen die Werte der Merkmale erfasst werden, werden „Merkmalsträger“ (statistische Einheiten, Individuen oder Fälle) genannt (Beispiel: Umfrageteilnehmer). Die Gesamtheit der Individuen, die für eine statistische Untersuchung relevant ist, wird als „Grundgesamtheit“ bezeichnet. Die Teilmenge der Grundgesamtheit, an der tatsächlich Werte erfasst wurden, nennt man „Stichprobe“ der Merkmalsträger oder Beobachtungsmenge. Je nach Informationsinhalt und Struktur der Ausprägungsmenge A eines Merkmals unterscheidet man zwischen „kategorialen“ (qualitativen) und „metrischen“ (quantitativen) Merkmalen. Metrische Merkmale mit abzählbarer Ausprägungsmenge nennt man „diskret“, mit überabzählbarer Ausprägungsmenge kontinuierlich (z.B. $A = \mathbb{R}$ oder A ein Intervall in $\mathbb{R}$). Merkmale werden noch nach ihrer Messskala (oder Skalenniveau) unterteilt. Man unterscheidet zwischen

- nominal
- ordinal
- metrisch

skalierten Merkmalen. Ordinal skalierte Merkmale können der Größe nach geordnet werden. Nominal- und ordinal skalierte Merkmale entsprechen zusammengefasst den kategorialen Merkmalen. Im Detail wird in Abschnitt I.3 auf die Unterschiede eingegangen.

Definition I.1
Eine „(univariate) Stichprobe“ x vom Stichprobenumfang $n \in \mathbb{N}$ eines Merkmals X

$$x = (x_1, \ldots, x_n)$$

ist ein n-Tupel mit $x_i \in A$, wobei A die Ausprägungsmenge von X bezeichnet. Ein Merkmal wird für n Merkmalsträger erfasst.

Eine „p-variate Stichprobe“ vom Stichprobenumfang $n \in \mathbb{N}$ von $p \geq 2$ Merkmalen $X_1, \ldots, X_p$ mit Ausprägungsmengen $A_1, \ldots, A_p$ sind die n p-Tupel

$$(x_{11}, \ldots, x_{1p}), \ldots, (x_{n1}, \ldots, x_{np}) \in A_1 \times \ldots \times A_p.$$

An jedem der n Merkmalsträger werden p Merkmale erfasst. Ordnet man die Beobachtungswerte einer ordinal skalierten Stichprobe $(x_1, \ldots, x_n)$ der Größe nach um, nennt man dies die geordnete Stichprobe und man schreibt

$$(x_{\uparrow 1}, \ldots, x_{\uparrow n})$$

wobei gilt, dass $x_{\uparrow 1} \leq x_{\uparrow 2} \leq \ldots \leq x_{\uparrow n}$.

Beispiel I.1 Für die Stichprobe

$$(x_1, \ldots, x_5) = \left(1, 4, 1, \frac{1}{2}, 10\right)$$

erhält man die „geordnete Stichprobe“

$$(x_{\uparrow 1}, \ldots, x_{\uparrow 5}) = \left(\frac{1}{2}, 1, 1, 4, 10\right).$$

I.2 Häufigkeiten

Gegeben sei eine Stichprobe $x = (x_1, \ldots, x_n)$. Im Folgenden bezeichnet $A = \{a_1, \ldots, a_m\}$ die Menge aller in x auftretenden Ausprägungen $a_1, \ldots, a_m, m \leq n$.

Definition I.2.1 (Häufigkeit)
Für $i = 1, \ldots, m$ definiert man die „absolute Häufigkeit“ von a_i als:

$$h(a_i) \equiv h_i := \sum_{j=1}^{n} \mathbb{1}_{\{a_i\}}(x_j)$$

und die „relative Häufigkeit“ von a_i als:

$$f(a_i) \equiv f_i := \frac{1}{n} \sum_{j=1}^{n} \mathbb{1}_{\{a_i\}}(x_j).$$

Dabei ist die „Indikatorfunktion“ $\mathbb{1}_A$ für eine beliebige Teilmenge A einer Grundmenge Ω definiert über:

$$\mathbb{1}_A : \Omega \to \{0, 1\} \qquad \omega \mapsto \begin{cases} 1 & \text{falls} \quad \omega \in A \\ 0 & \text{falls} \quad \omega \notin A \end{cases}$$

Die Zahlen $h_1, \ldots, h_m$ bzw. $f_1, \ldots, f_m$ werden „absolute bzw. relative Häufigkeitsverteilung“ genannt.

Ergänzend können die Häufigkeiten auch für zusätzliche Ausprägungswerte $b_1, \ldots, b_k$, die nicht in der Stichprobe auftreten, als $h(b_i) = f(b_i) := 0$ definiert werden. Man sieht leicht, dass $\sum_{i=1}^{m} h_i = n$ und $\sum_{i=1}^{m} f_i = 1$.

Mit einer Häufigkeitstabelle kann man Beobachtungen, die nur relativ wenig verschiedene Werte annehmen, gut beschreiben. Wir betrachten dabei der Einfachheit halber nur „univariate“ Daten.

I.2 Beispiel: Schulnoten in Mathematik in einer Klasse

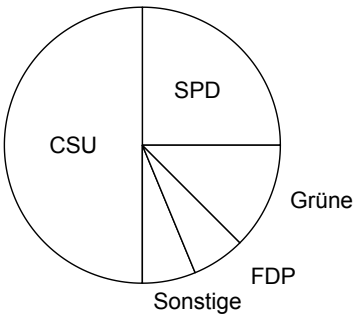
Note	absolute Häufigkeit	relative Häufigkeit
1	2	
2	4	
3	5	
4	3	
5	3	
6	3	

Die Daten einer Häufigkeitstabelle können grafisch in einem „Stab-“ oder „Säulendiagramm“ dargestellt werden. Dabei werden auf der horizontalen Achse die verschiedenen vorkommenden Werte aufgetragen und darüber jeweils ein „Stab“ gezeichnet, dessen Höhe proportional zur zugehörigen (absoluten oder relativen) Häufigkeit ist. Wählt man die relativen Häufigkeiten, kann man das Stabdiagramm besser mit Stabdiagrammen anderer Daten vergleichen, allerdings verliert man dann die Information über die Anzahl der Beobachtungen.

Für den Fall von kategorialen Daten, bei denen es darum geht, Mehrheiten zu erkennen, verwendet man häufig ein „Kuchendiagramm“, um die relativen Häufigkeiten grafisch darzustellen. Dazu zeichnet man einen Kreis und teilt ihn in so viele verschiedene Segmente auf, wie es verschiedene Ausprägungen der Beobachtungen gibt.

Beispiel: Betrachten Sie das folgende Wahlergebnis:

CSU	50.00%
SPD	25.00%
Grüne	12.50%
FDP	6.25%
Sonstige	6.25%



Siehe auch R-Beispiel I.2.

Die Häufigkeitsverteilung einer bivariaten Stichprobe

$$(x_1, y_1), \ldots, (x_n, y_n)$$

zweier Merkmale X und Y vom Umfang n kann mithilfe einer Kontingenztafel (Kontingenztabelle) notiert werden. Häufig sind dabei beide Merkmale nominal skaliert. Der Begriff Kontingenz, also Zusammenhang, deutet bereits an, dass Fragestellungen bzgl. des Zusammenhangs der Merkmale oft im Mittelpunkt stehen. Bezeichne

$$A = \{a_1, \ldots, a_k\}, k \leq n$$

die Menge der Ausprägungen der Stichprobenwerte $x_1, \ldots, x_n$,

$$B = \{b_1, \ldots, b_m\}, m \leq n,$$

die Menge der Ausprägungen der Stichprobenwerte $y_1, \ldots, y_n$, und für $1 \leq i \leq k$, $1 \leq j \leq m$

$$h_{ij} \equiv h\left(a_i, b_j\right) := \sum_{1 \leq r \leq n} \mathbb{1}_{\{(a_i, b_j)\}}\left((x_r, y_r)\right).$$

h_{ij} bezeichnet also die absolute Häufigkeit der Merkmalskombination (a_i, b_j) in der bivariaten Stichprobe. Als Kontingenztafel der absoluten Häufigkeiten bezeichnet man dann die $k \times m$ Matrix (bzw. die entsprechende Tabelle)

$$K := \begin{pmatrix} h_{11} & \ldots & h_{1m} \\ h_{21} & \ldots & h_{2m} \\ \vdots & & \vdots \\ h_{k1} & \ldots & h_{km} \end{pmatrix}$$

oder auch $K^\top$. Ganz analog ist mit den relativen Häufigkeiten $f_{ij} := \frac{h_{ij}}{n}$ die Kontingenztafel der relativen Häufigkeiten definiert. Die Kontingenztafel einer bivariaten Stichprobe wird genauer als 2-dimensionale Kontingenztafel bezeichnet. Entsprechend erhält man für eine p-variate Stichprobe mit $p > 2$ dann eine p-dimensionale Kontingenztafel. Zusätzlich zu den Häufigkeiten h_{ij} bzw. f_{ij} sind die absoluten Randhäufigkeiten

$$h_{i\cdot} := \sum_{j=1}^{m} h_{ij} \text{ und } h_{\cdot j} := \sum_{i=1}^{k} h_{ij}$$

für $i = 1, \ldots, k$ und $j = 1, \ldots, m$ und ganz analog die relativen Randhäufigkeiten von Interesse. Als Tabelle erhält man mit den Randhäufigkeiten eine Kontingenztafel der absoluten Häufigkeiten der Form

h_{11}	$\ldots$	h_{1m}	$h_{1\cdot}$
h_{21}	$\ldots$	h_{2m}	$h_{2\cdot}$
$\vdots$		$\vdots$	$\vdots$
h_{k1}	$\ldots$	h_{km}	$h_{k\cdot}$
$h_{\cdot 1}$	$\ldots$	$h_{\cdot m}$	n

Der Eintrag n rechts unten in der Kontingenztabelle entspricht der Summe der Zeilen oder Spaltenhäufigkeiten, die sich jeweils zum Stichprobenumfang n addieren. Um Hinweise auf einen eventuell vorliegenden Zusammenhang der beiden Merkmale X und Y zu gewinnen, bildet man die bedingten relativen Häufigkeitsverteilungen.

Definition I.2.2 (Bedingte relative Häufigkeitsverteilung)

Die bedingte relative Häufigkeitsverteilung von Y gegeben die Bedingung $X = a_i, i = 1 \ldots, k$, ist durch die relativen Häufigkeiten

$$f_Y\left(b_1 \mid a_i\right) := \frac{h_{i1}}{h_{i\cdot}}, \ldots, f_Y\left(b_m \mid a_i\right) := \frac{h_{im}}{h_{i\cdot}}$$

definiert. Die bedingte relative Häufigkeitsverteilung von X gegeben die Bedingung $Y = b_j, j = 1, \ldots, m$, ist durch die relativen Häufigkeiten

$$f_X\left(a_1 \mid b_j\right) := \frac{h_{1j}}{h_{\cdot j}}, \ldots, f_Y\left(a_k \mid b_j\right) := \frac{h_{kj}}{h_{\cdot j}}$$

definiert.

Beispiel Von 100 Studierenden ist jeweils das Geschlecht Y (Ausprägungen: weiblich und männlich) und die Haarfarbe X (mit den Ausprägungen: blond, hellbraun, braun, rot und schwarz) gegeben. Die bivariate Stichprobe sei durch die folgende Kontingenztabelle der absoluten Häufigkeiten gegeben.

	blond	hellbraun	braun	rot	schwarz	$\sum$
weiblich	20	15	10	5	10	60
männlich	5	10	10	1	14	40
$\sum$	25	25	20	6	24	100

Als bedingte relative Häufigkeitsverteilungen von X unter der Bedingung $Y =$„weiblich“ bzw. unter der Bedingung $Y =$„männlich“ ergibt sich

$$
\begin{array}{lll}
f_X\,(\text{blond} \mid \text{weiblich}) = & \text{bzw.} & f_X\,(\text{blond} \mid \text{männlich}) = \\
f_X\,(\text{hellbraun} \mid \text{weiblich}) = & \text{bzw.} & f_X\,(\text{hellbraun} \mid \text{männlich}) = \\
f_X\,(\text{braun} \mid \text{weiblich}) = & \text{bzw.} & f_X\,(\text{braun} \mid \text{männlich}) = \\
f_X\,(\text{rot} \mid \text{weiblich}) = & \text{bzw.} & f_X\,(\text{rot} \mid \text{männlich}) = \\
f_X\,(\text{schwarz} \mid \text{weiblich}) = & \text{bzw.} & f_X\,(\text{schwarz} \mid \text{männlich}) =
\end{array}
$$

Aufgrund der Werte kann man einen potentiellen Zusammenhang zwischen der Haarfarbe und dem Geschlecht vermuten, da

Eine Kontingenztafel (auch höher-dim.) bzw. die resultierenden bedingten Häufigkeiten können mit einem Mosaik-Plot visualisiert werden. Die folgende Abbildung zeigt einen Mosaik-Plot zu der 2-dimensionalen Kontingenztafel aus obigem Beispiel.

Falls zwischen Geschlecht und Haarfarbe kein Zusammenhang bestünde, würde man lauter (annähernd) durchgezogene Linien im Mosaik-Plot erwarten, was in obigem Mosaik-Plot offenbar nicht der Fall ist.

Besteht zwischen den beiden einer Kontingenztabelle zugrundeliegenden Merkmalen X und Y kein Zusammenhang, würde man erwarten, dass für die bedingten relativen Häufigkeiten gilt, dass

$$f_Y(b_j \mid a_i) \approx f_Y(b_j \mid a_l) \approx \frac{h_{\cdot j}}{n} \; \forall i, l = 1, \ldots, k \text{ und } j = 1, \ldots, m$$

und

$$f_X(a_i \mid b_j) \approx f_X(a_i \mid b_r) \approx \frac{h_{i\cdot}}{n} \; \forall j, r = 1, \ldots, m \text{ und } i = 1, \ldots, k.$$

D.h die relative Häufigkeit einer Ausprägung hängt nicht von der Wahl der Ausprägung ab, bzgl. der man die Häufigkeit bedingt. Diese Überlegung führt zu der folgender Boebachtung.

Unter der Annahme, dass zwischen den Merkmalen X und Y kein Zusammenhang besteht, ist

$$\widetilde{h_{ij}} = \frac{h_{i\cdot} h_{\cdot j}}{n}$$

die erwartete Häufigkeit bei Unabhängigkeit in der entsprechenden Zelle der Kontingenztabelle.
Siehe auch R-Beispiel I.2.

ABC 5, Ü 2, Moodle 1.1-1.2

I.3 Skalenniveaus

Wie schon im ersten Abschnitt erwähnt, unterscheidet man bei Daten grundlegend den Typ der Messskala, in der die Beobachtungswerte angegeben werden. Bei einer

Namensskala (Nominalskala, Kategorien),

unterscheidet man nur zwischen Gleichheit und Ungleichheit, eine

Rangskala (Ordinalskala)

legt auch eine Rangfolge, d.h. vollständige Ordnung, der möglichen Werte fest, eine

metrische Skala (Einheitenskala)

gewichtet darüber hinaus die Abstände zwischen möglichen Beobachtungswerten.

Beispiel: Von den Teilnehmer:innen einer Klausur wurde das Geschlecht (Namensskala), die Platzierung (Rangzahl), die erreichte Note (mindestens Rangskala) sowie die Körpergröße (metrische Skala) beobachtet.

Bemerkung: Metrisch skalierte Beobachtungen können stets angeordnet werden, ordnal skalierte Beobachtungen können stets in Kategorien zusammengefasst werden.

I.4 Empirische Verteilungs- und Quantilfunktion

Definition I.3.1 (Empirische Verteilungsfunktion)

Die empirische Verteilungsfunktion F_n einer Stichprobe $x = (x_1, \ldots, x_n)$ eines metrischen Merkmals ist definiert als die Funktion

$$F_n : \mathbb{R} \to [0,1], F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty,t]}(x_i).$$

Satz I.3

Die empirische Verteilungsfunktion $F_n(t), t \in \mathbb{R}$, einer Stichprobe $x = (x_1, \ldots, x_n)$ mit den Ausprägungen $a_1, \ldots, a_m$ ist eine isotone (d.h. monoton wachsende), rechtsseitig stetige Treppenfunktion mit den Sprungstellen $a_1, \ldots, a_m$ und den entsprechenden relativen Häufigkeiten $f(a_1), \ldots, f(a_m)$ als Sprunghöhen. Für $t < x_{\uparrow 1}$ ist $F_n(t) = 0$ und für $t \geq x_{\uparrow n}$ ist $F_n(t) = 1$.

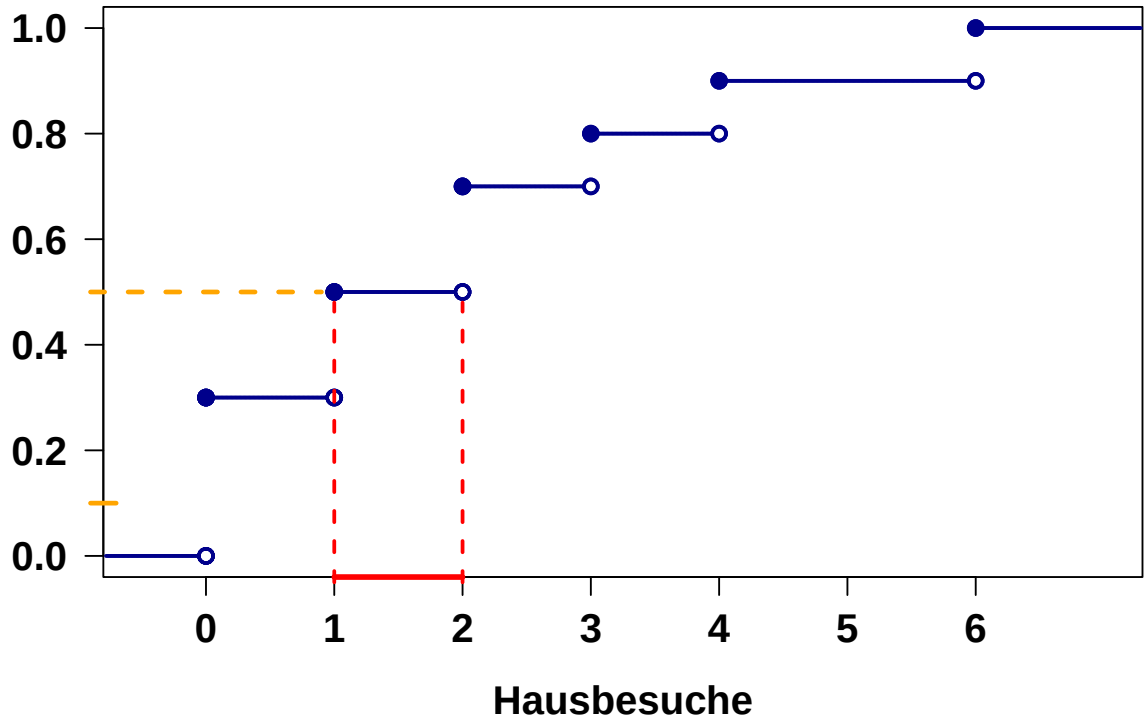
Beweis: selber

Beispiel: Ein Arzt hat zwanzig Tage lang die Anzahl seiner Hausbesuchen erfasst:

Tag	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Hausbesuche	4	0	2	3	2	4	0	3	0	1	1	6	0	2	0	0	1	6	2	1

Daraus ergeben sich die folgenden Häufigkeiten und die dazu gehörige empirische Verteilungsfunktion:

Index	j	1	2	3	4	5	6	7	$\sum$
Ausprägung (sortiert)	a_j	0	1	2	3	4	5	6	
absolut	$h(a_j)$	6	4	4	2	2	0	2	20
kumuliert absolut	$H(a_j)$								
relativ	$f(a_j)$								
kumuliert relativ	$F(a_j)$	$\frac{3}{10}$	$\frac{5}{10}$	$\frac{7}{10}$	$\frac{8}{10}$	$\frac{9}{10}$	$\frac{9}{10}$	1	



Empirische Verteilungsfunktion F_n in Abhängigkeit der Anzahl der täglichen Hausbesuche (orangefarben gestrichelt: zwei Beispiele zur Urbildabbildung)

Für die empirische Verteilungsfunktion F_n einer Stichprobe x vom Umfang n und $p \in (0,1)$ gilt entweder:

$$F_n^{-1}(\{p\}) = \emptyset,$$

oder

$$F_n^{-1}(\{p\}) = \left[x_{\uparrow(np)}, x_{\uparrow(np+1)}\right),$$

wobei $F_n^{-1}(\{p\})$ die Urbilder der empirischen Verteilungsfunktion bezeichnet. Als Beispiele hierfür sind $p = 0.1$ mit $F_n^{-1}(\{0.1\}) = \emptyset$ und $p = 0.5$ mit $F_n^{-1}(\{0.5\}) = [x_{\uparrow(20\cdot 0.5)}, x_{\uparrow(20\cdot 0.5)+1}) = [x_{\uparrow 10}, x_{\uparrow 11}) = [1, 2)$ eingezeichnet.

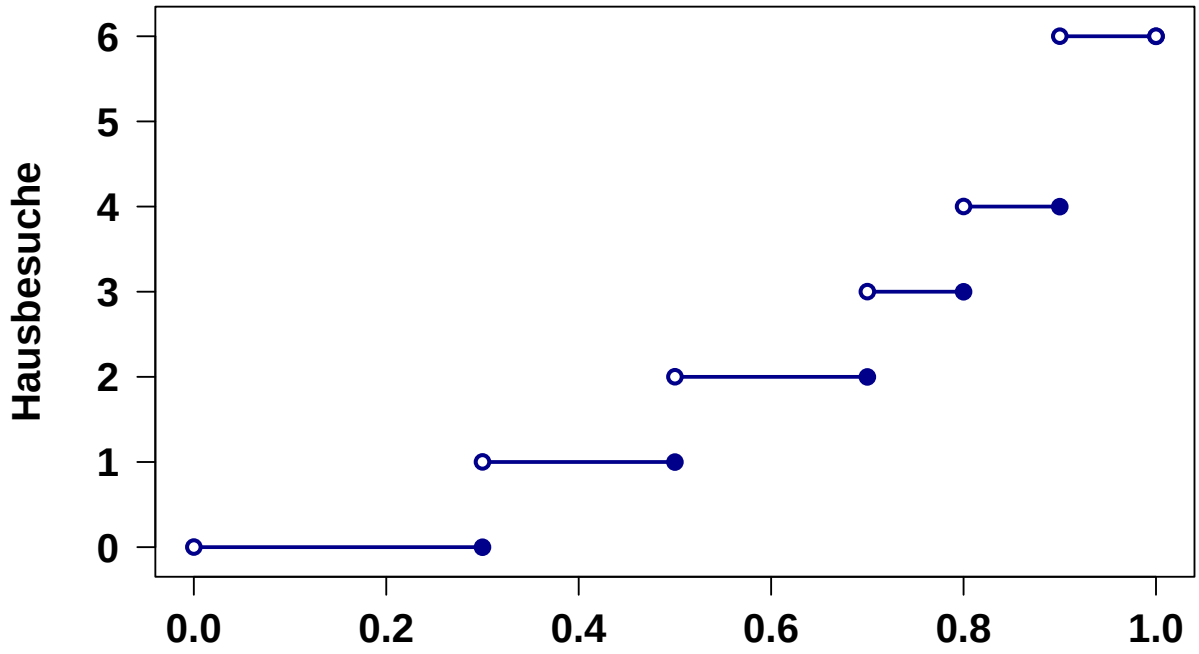
Definition I.3.2 (Empirische Quantilsfunktion)

Die empirische Quantilsfunktion F_n^{-1} einer empirischen Verteilungsfunktion F_n ist definiert als die verallgemeinerte Inverse von F_n, d.h. als

$$F_n^{-1} : \big(0,1\big) \to \mathbb{R};\ F_n^{-1}(p) = \min\left\{t \in \mathbb{R} : F_n(t) \geq p\right\}.$$

In Worten: $F_n^{-1}(p)$ ist die kleinste Zahl t, an deren Stelle die empirische Verteilungsfunktion

einen Wert größer oder gleich p hat.



Empirische Quantilfunktion F_n^{-1} in Abhängigkeit der Wahrscheinlichkeit p

Bemerkungen:

- Die empirische Quantilsfunktion F_n^{-1} ist keine gewöhnliche Inverse, da F_n wegen der Treppenstufen nicht injektiv und damit auch nicht umkehrbar ist.
- Die empirische Quantilsfunktion F_n^{-1} ist isoton und linksstetig.
- Âchtung: Das Symbol F^{-1} steht für drei unterschiedliche Dinge: die gewöhnliche Inverse, die verallgemeinerte Inverse und die Urbildfunktion. Was gemeint ist, erschließt sich der geneigte Leser aus dem Zusammenhang.

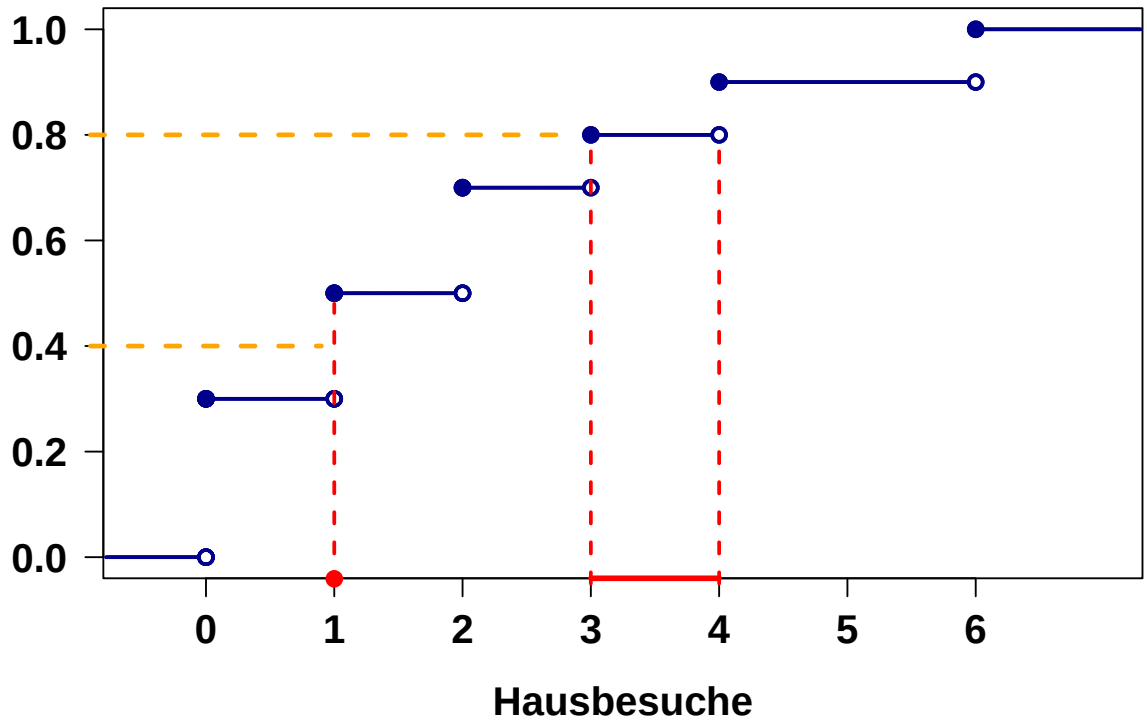
I.5 Empirische Quantile, Quantilintervalle, Quartile

Definition I.5.1 (Empirische Quantile)
Für $p \in (0,1)$ ist das empirische p-Quantil (man sagt auch p-Perzentil oder $(p \cdot 100\%)$-Quantil) x_p einer Stichprobe $x = (x_1, \ldots, x_n)$ eines metrisch skalierten Merkmals definiert als ein Wert, für den gilt, dass mindestens $100p$ Prozent der Beobachtungswerte kleiner oder gleich diesem Wert sind und für den zusätzlich gilt, dass mindestens $100(1-p)$ Prozent größer oder gleich diesem Wert sind.

Bemerkung: Es gilt (siehe auch Graph):

$$x_p := \begin{cases} \left[x_{\uparrow(np)}, x_{\uparrow(np+1)}\right], & \text{falls } np \in \mathbb{N} \text{ und } x_{\uparrow(np)} < x_{\uparrow(np+1)} \\ x_{(\lceil np \rceil + 1)} & \text{falls } np \notin \mathbb{N} \text{ oder } (np \in \mathbb{N} \text{ und } x_{\uparrow(np)} = x_{\uparrow(np+1)}) \end{cases}$$

wobei $\lceil z \rceil$ den ganzzahligen Anteil einer reellen Zahl z bezeichnet.



Empirische Verteilungsfunktion F_n und mögliche empirische Quantile für $p = 0,4$ (eindeutig die Zahl $x_{\uparrow(0,4\cdot 20)} = x_{\uparrow 5} = x_{\uparrow 6}$) und $p = 0,8$ (beliebige Zahl zwischen $x_{\uparrow(0,8\cdot 20)} = x_{\uparrow 16} = 3$ und $x_{\uparrow 17} = 4$)

Wie wir in obiger Bemerkung gesehen haben, sind die empirischen Quantile im Fall $np \in \mathbb{N}$ und $x_{\uparrow(np)} < x_{\uparrow(np+1)}$ nicht eindeutig bestimmt, sondern können aus einem sog. Quantilintervall gewählt werden. Oft will man aber dennoch genau eine Zahl angeben.

In der Literatur und bei statistischer Software finden sich viele Möglichkeiten, das eindeutige Quantil zu bestimmen, so kann beispielsweise beim R-Befehl `quantile` über die Option `type` aus neun verschiedenen Varianten gewählt werden.

Am gebräuchlisten sind die Intervallmitte (`type = 2`) oder die linke Intervallgrenze (`type = 1`):

```
anz_besuche <- c(4,0,2,3,2,4,0,3,0,1,1,6,0,2,0,0,1,6,2,1)

# bei eindeutigen Quantile ist kein Unterschied,
# wohl aber bei Quantilintervallen:
quantile( anz_besuche, probs = c(0.4,0.8), typ=1 )

## 40% 80%
##   1   3

quantile( anz_besuche, probs = c(0.4,0.8), typ=2 )

## 40% 80%
## 1.0 3.5
```

Definition I.5.2

Einige empirische Quantile sind so bedeutend, dass sie eigene Namen bekommen haben: Das

25%-Quantil heißt „(empirisches) erstes oder unteres Quartil“, das 50%-Quantil wird „(empirisches) zweites Quartil oder (empirischer) Median“ genannt und das 75%-Quantil bezeichnen wir als „(empirisches) drittes oder oberes Quartil“. Sobald wir die Begriffe (empirisches) erstes/unteres, zweites und drittes/oberes Quartil oder (empirischer) Median verwenden, meinen wir bei Uneindeutigkeit der entsprechenden Quantile die Intervallmitte.

Darüber hinaus definieren wir das „(empirische) 0-tes Quartil“ als kleinsten Beobachtungswert und das „(empirische) 4-te Quartil“ als größten Beobachtungswert.

Ferner definieren wir den Interquartilsabstand („IQR“ (inter quartile range)) als die Differenz zwischen oberen und unterem Quartil.

Bemerkungen:

- Der empirischen Median der Stichprobe teilt die Stichprobe in zwei (etwa) gleich mächtige Mengen von Stichprobenwerte, die kleiner oder gleich dem empirischen Median bzw. größer oder gleich dem empirischen Median sind.
- Definition I.5.1 lässt sich auch auf ordinale skalierte Merkmale anwenden. Da es bei ordinal skalierten Merkmale im Allgemeinen keine Intervallmitte gibt, verwendet man das kleinste p-Quantil als eindeutiges empirisches p-Quantil. Zum Beispiel ist B der Median des amerikanischen Notenvektors (A, A, B, C, C, D).
- Für das empirische p-Quantil gilt: $x_p \in [F^{-1}(p), F^{-1}(p+)]$, wobei $F^{-1}(p+)$ den rechtsseitigen Grenzwert der Quantilfunktiton an der Stelle p bezeichnet.
- Wenn in einem R-Datenvektor `x` metrische Daten stehen, liefert `quantile( x, type=2 )` genau die fünf Quartile.

Beispiel: Beobachtungswerte: 1,2,3:

(empirischer) Median	= 0.5-Quantil	= 50%-Quantil	=
drittes Quartil	= 0.75-Quantil	= 75%-Quantil	=
$0.\bar{3}$-Quantilintervall			=
$0.\bar{3}$-Quantil			irgendein Wert in

ABC 6

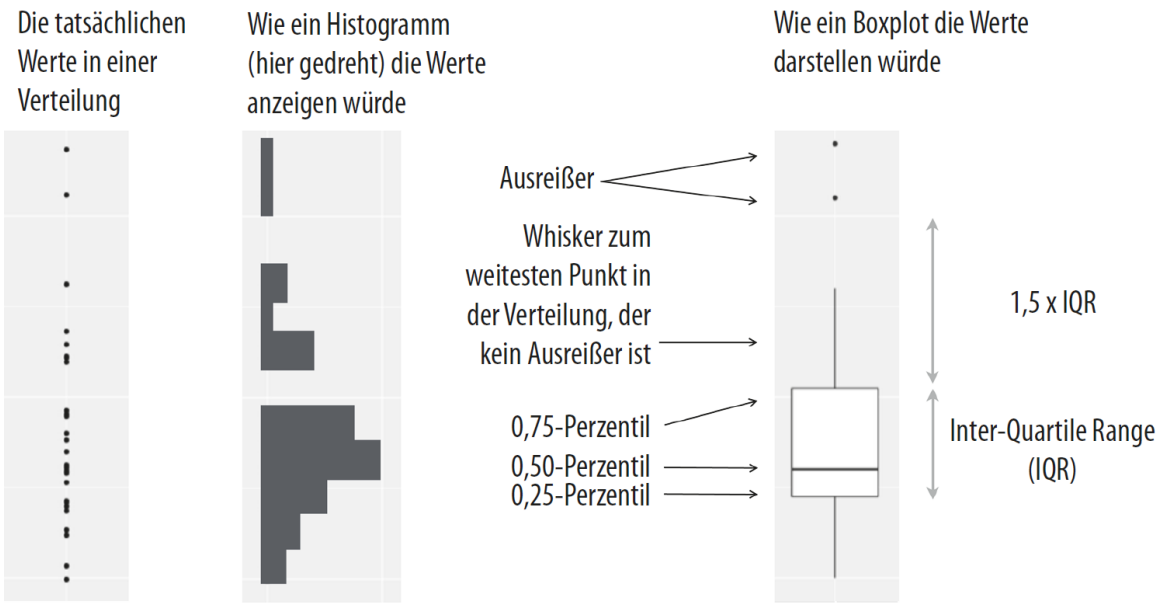
I.6 Boxplots

Metrische Merkmale stellt man am besten mit Boxplots oder Histogrammen (siehe nächster Abschnitt) dar. Boxplots eignen sich besonders gut, um Ausreißer anzuzeigen; Histogramme vermitteln einen guten Eindruck von der Verteilung der Werte.
Ein Boxplot ist wie folgt aufgebaut:

- Die Länge der Box wird bestimmt durch
  - das erste Quartil. Dies ist die untere Kante der Box,
  - das dritte Quartil. Das ist die obere Kante der Box.

  Bemerkung: Die Länge der Box ist also der Interquartilsabstand IQR.
- Der Median (bezeichnet mit x_{med} oder $x_{1/2}$) ist im Boxplot als eine Linie dargestellt, die den Kasten zwischen den Quartilen in zwei Bereiche aufteilt.

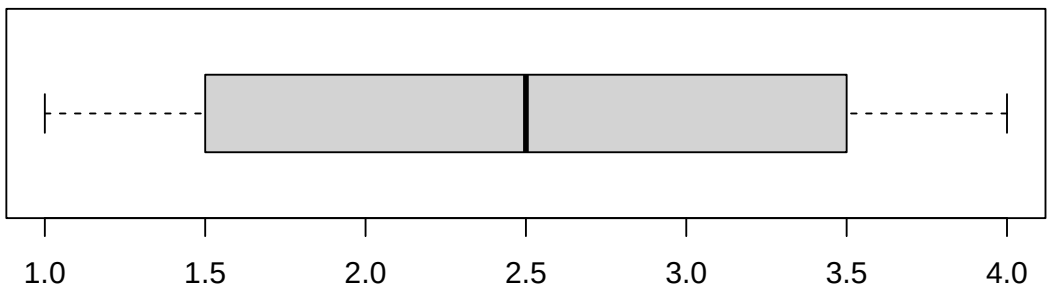
- Mittig an den Kasten werden zwei Linien gezeichnet („whisker“ = Schnurrbart). Diese enden jeweils bei dem äußersten Datenpunkt der maximal den 1,5-fachen Interquartilsabstand von der Box hat.
- Werte, die außerhalb dieser Linien liegen, werden als Ausreißer gekennzeichnet (in R mit einem Kreis oder Stern).



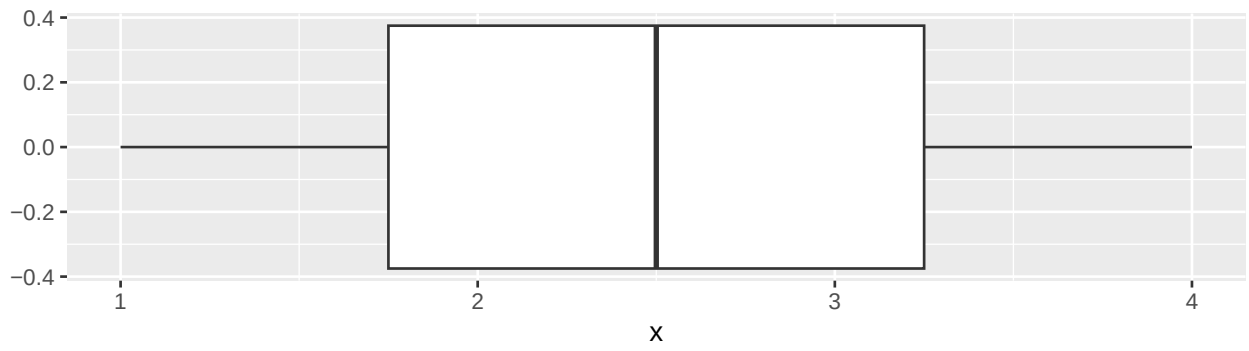
Bestandteile eines Boxplots, aus: Wickham H., Grolemund G., R für Data Science, O'Reilly, 2018, S. 92.

Hinweis: Die Funktion `boxplot` aus der Standard R-Bibliothek berechnet die Quartile - wie bei uns definiert - als Mitte des Quantilintervalls. Die Funktion `geom_boxplot` dagegen verwendet eine andere Definition und zwar die, die dem standardmäßigem `type=7` in der `quantile`-Funktion entspricht. Besonders bei kleinen Stichproben kann dies zu unterschiedlichen Boxplots führen, wie die beiden unteren Boxplots für die Datenpunkte 1,2,3 und 4 zeigen:

```
boxplot( 1:4, horizontal = TRUE)
```



```
ggplot( tibble(x=1:4)) + geom_boxplot(aes(x=x))
```

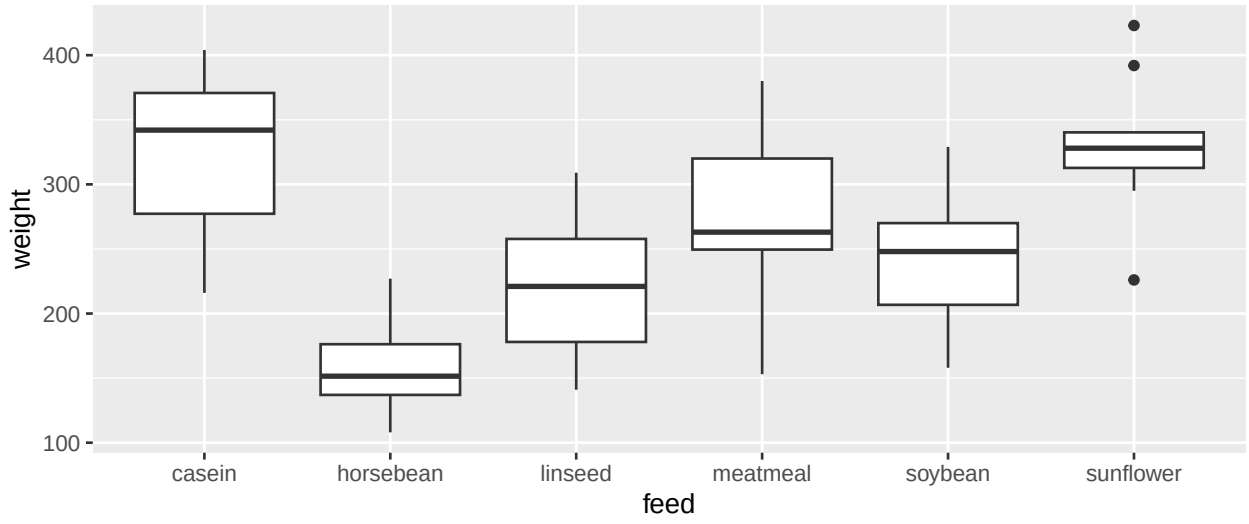


Bemerkung: Die Form der Darstellung des Bereichs zwischen Minimal- und Maximalwert erinnert an einen Schurrbart (engl.: whisker), weswegen das Kastendiagramm auch Box-Whisker-Plot genannt wird. An der Skala lassen sich die Zahlenwerte der dargestellten Größen erkennen.

Box-Whisker-Plots eignen sich sehr gut zum Vergleich mehrerer (Teil-)Stichproben oder meherer anhand eines Merkmals gruppierter Datensätze, wie die folgende beiden Beispiele zeigen:

Im R-internen Datensatz *chickwts* sind die Gewichte von Hühnern erfasst, die mit sechs unterschiedlichen Futtermitteln gefüttern wurden. Ein gruppierter Boxplot zeigt auf einen Blick, dass die Gewichtsverteilungen je Futtermittel sehr unterschiedlich sind:

```
library(tidyverse)
ggplot(chickwts) + geom_boxplot( aes(x=feed, y=weight) )
```



Beispiel: Von den Teilnehmern einer Klausur wurde das Geschlecht (Namensskala) und die erreichte Punktzahl (metrische Skala) beobachtet. Die 12 teilnehmenden Studenten erreichten die Punktzahlen

$$20, 22, 20, 23, 23, 30, 18, 5, 18, 26, 12, 35.$$

Die 10 teilnehmenden Studentinnen erreichten die Werte

$$25, 14.5, 21, 26, 13.5, 23, 20, 10, 26, 16.$$

	12 männl. TN	10 weibl. TN	22 TN insges.
kleinster Beobachtungswert			
erstes Quartil / Intervall für 25%-Quantil			
Median / Intervall für 50%-Quantil			
drittes Quartil / Intervall für 75%-Quantil			
größter Beobachtungswert			

Wie im letzten Abschnitt erörtert, kann man bei der Statistiksoftware *R* über die Option `type` einstellen, welches Quartil geliefert wird (siehe auch R-Beispiel I.6):

```
# Daten in Vektoren einlesen
PkteMaennl <- c(20, 22, 20, 23, 23, 30, 18, 5, 18, 26, 12, 35)
PkteWeibl  <- c(25, 14.5, 21, 26, 13.5, 23, 20, 10, 26, 16)
PkteGesamt <- cbind(PkteMaennl, PkteWeibl)
# minimale Quartile
# Anmerkung: Gibt man im Befehl "quantile" keine Quantile an,
# werden standardmäßig die Quartile ausgegeben. Die Option
# "typ=1" sorgt dafür, dass die minimalen Quantile ausgegeben
# werden
quantile(PkteMaennl, typ=1)

##   0%  25%  50%  75% 100%
##    5   18   20   23   35

# Mitte des Quartilintervalls
quantile(PkteMaennl, typ=2)

##   0%  25%  50%  75% 100%
##  5.0 18.0 21.0 24.5 35.0

quantile(PkteWeibl, typ=2)

##   0%  25%  50%  75% 100%
## 10.0 14.5 20.5 25.0 26.0

quantile(PkteGesamt, typ=2)

##    0%   25%   50%   75%  100%
##  5.00 15.25 20.50 25.00 35.00

# Alle Quantile in 10%-Schritten (Mitte des Quartilintervalls)
quantile(PkteMaennl, p=seq(0,1,by=0.1), typ=2)

##   0%  10%  20%  30%  40%  50%  60%  70%  80%  90% 100%
##    5   12   18   18   20   21   23   23   26   30   35
```

Um diese Kenngrößen zu visualisieren, zeichen wir nun noch die zugehörigen Boxplots:

Man sieht so auf einen Blick, dass sich die Punkteverteilung der männlichen Teilnehmer über einen weiteren Bereich erstrecken als die der weiblichen Teilnehmer, dass aber die „mittlere Hälfte“ bei den männlichen Teilnehmern enger konzentriert ist als bei den weiblichen Teilnehmern. Die mittlere Punktzahl sowohl bei den beiden Teilnehmergruppen wie bei der Gesamtgruppe liegt bei 20.5 bis 21 Punkten.

Einen gruppierten Boxplot erzeugt man in Standard-R durch den Befehl *boxplot* (siehe nochmal R-Beispiel I.6):

```
boxplot(list(PkteMaennl,PkteWeibl,PkteGesamt),
        names=c("12 männl. TN", "10 weibl. TN", "22 TN insges."))
```

I.7 Histogramme

Wenn sehr viele verschiedene metrisch skalierte Beobachtungswerte vorliegen, ist eine Darstellung im Stabdiagramm nicht sinnvoll, wie z.B. für die Durchschnittsnoten der Studierenden in ihrem Abschlusszeugnis:

1.97, 2.33, 3.51, 5.11, 1.12, 2.59, 4.18, 2.81, 3.27, 1.50.

Daher betrachtet man bei vielen verschiedenen numerischen Werten Intervalle, so genannte Klassen, und zählt, wie viele Werte in die einzelnen Klassen fallen. Die Anzahl h_i der Messwerte, die in die i-te Klasse fallen, nennt man die (absolute) Häufigkeit der i-ten Klasse. Dividiert man diese Anzahl durch die Gesamtanzahl n aller Beobachtungswerte, so erhält man die relative Häufigkeit $f_i = \frac{h_i}{n}$ der i-ten Klasse.

Klassifizierte Beobachtungswerte werden in einem „Histogramm “ graphisch dargestellt. Dazu zeichnet man über den einzelnen Klassen Rechtecke, deren **Fläche** ein Maß dafür ist, wie viele Stichprobenwerte in der zugehörigen Klasse liegen. Das wird dadurch erreicht, indem man als Höhe eines Rechtecks einen Wert proportional zu $\frac{h_i}{\Delta x_i}$ bzw. $\frac{f_i}{\Delta x_i}$ wählt (wobei Δx_i die Breite der i-ten Klasse bezeichnet).

Da wenige Klassen einen großen Informationsverlust zur Folge haben, andererseits zu viele Klassen die Übersicht gefährden, muss man bei der Klassenbildung einen guten Kompromiss finden. Dazu gibt es einige Faustregeln:

- Die Klassen sollten gleich breit gewählt werden. Ausnahmen für die erste und letzte Klasse sind jedoch erlaubt (siehe unteres Beispiel).
- Die Anzahl der Klassen sollte etwa zwischen 5 und 20 liegen, jedoch $\sqrt{n}$ nicht wesentlich überschreiten (wobei n der Umfang der Stichprobe ist).

Wir verwenden die „untere-Intervallgrenze-dabei-Regel“, d.h. ein Klassenintervall enthält seine linke Grenze, aber nicht seine rechte Grenze. So enthält das Klassenintervall 1.5-2.5 alle Werte die größer oder gleich 1.5 und kleiner als 2.5 sind.

In unserem Beispiel mit den Durchschnittsnoten ergibt sich:

Klassenintervall	Häufigkeit	
	abs.	rel.
1.0-1.5	1	10%
1.5-2.5	3	30%
2.5-3.5	3	30%
3.5-4.5	2	20%
4.5-5.5	1	10%
5.5-6.0	0	0%

Siehe auch R-Beispiel I.7.

Ü 3; Moodle 1.3-1.7

I.8 Arithmetisches Mittel, Median und Modalwert

Definition: Für metrisch, ordinal oder nominal skalierte Beobachtungswerte $x_1, \ldots, x_n \quad (n \in \mathbb{N})$ kann man unterschiedliche Lagemaße für die Mitte der Beobachtungswerte benutzen.

(i) Für metrisch skalierte Beobachtungswerte ist der „arithmetische Mittelwert“ , i.Z. $\bar{x}$, wie folgt definiert

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

(ii) Seien die metrisch oder ordinal skalierten Beobachtungswerte der Größe nach von klein nach groß geordnet. Wenn n ungerade ist, ist der „Median“ der $\frac{n+1}{2}$-te Wert; wenn n gerade ist, ist der Median der Durchschnitt des $\frac{n}{2}$-ten und des $\frac{n}{2}+1$-ten Werts. Der Median wird mit $x_{1/2}$ bezeichnet.

(iii) Der „Modalwert“ ist der Wert, der am häufigsten auftritt. Er kann für beliebig skalierte Werte berechnet werden. Wenn mehrere Werte am häufigsten auftreten, werden all diese Werte als „Modalwerte“ bezeichnet.

Beispiel:

Note	Häufigkeit
1	5
2	4
3	6
4	3
5	1

$$\begin{aligned} \bar{x} &= \\ x_{1/2} &= \text{Median}(x_1, \ldots, x_{19}) = \\ x_{mod} &= \text{mod}(x_1, \ldots, x_{19}) = \end{aligned}$$

Ein R-Beispiel findet man im nächsten Abschnitt. ABC 7

I.9 Stichprobenvarianz und Stichproben-Standardabweichung

Wir interessieren uns nun für Kenngrößen, die die Streuung oder Variabilität der Beobachtungswerte beschreiben. Wir betrachten in diesem Abschnitt metrisch skalierte Beobachtungswerte: $x_1, \ldots, x_n$.

Definition I.9.1: Die „Stichprobenvarianz“, i.Z. „s^2 “ ist wie folgt definiert:

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

Beispiel: Beobachtungswerte: 3,4,6,7,10

$$\begin{aligned} \bar{x} &= \frac{3+4+6+7+10}{5} = 6 \\ s^2 &= \\ &= \end{aligned}$$

Die folgende Formel ist für die Berechnung der Stichprobenvarianz häufig nützlich:

Satz:

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

Beweis:

In unserem Beispiel:

$$s^2 = \frac{9 + 16 + 36 + 49 + 100 - 5 \cdot 36}{4} = \cdots = 7.5$$

Da die Stichprobenvarianz über die quadrierten Abstände berechnet wird, erhält man ein besseres Gefühl für die Streuung der Daten, wenn man die Wurzel aus der Stichprobenvarianz zieht.

Definition I.9.2: Die Stichproben-Standardabweichung, i.Z. s, wird wie folgt definiert:

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

Die Standardabweichung hat dieselbe Einheit wie die Beobachtungswerte.

Weitere Streuungsmaße, die gelegentlich verwendet werden, sind in folgender Definition zusammengefasst:

Definition I.9.3: Sei $x = (x_1, \ldots, x_n)$ eine Stichprobe eines metrischen Merkmals mit arithmetischen Mittel $\bar{x}$ und Median $x_{\frac{1}{2}}$, dann heißt

$$\begin{aligned}
\delta &:= \frac{1}{n}\sum_{i=1}^{n}|x_i - \bar{x}| \quad \text{„die mittlere absolute Abweichung vom Mittelwert“,} \\
\delta_{Med} &:= \frac{1}{n}\sum_{i=1}^{n}\left|x_i - x_{\frac{1}{2}}\right| \quad \text{„die mittlere absolute Abweichung vom Median“,} \\
\text{MAD} &:= \text{Median der Stichprobe } \left(\left|x_1 - x_{\frac{1}{2}}\right|, \ldots, \left|x_n - x_{\frac{1}{2}}\right|\right) \\
&\quad \text{die „Median-Deviation“ und} \\
R &:= x_{\uparrow n} - x_{\uparrow 1} \text{ die „Spannweite (range)“}
\end{aligned}$$

der Stichprobe x.

Siehe R-Beispiel I.8.9.

I.10 Korrelation

Häufig enthalten die zu untersuchenden Datensätze Paare von mindestens ordinal-skalierten numerischen Werten, die in einer bestimmten Beziehung zueinander stehen .

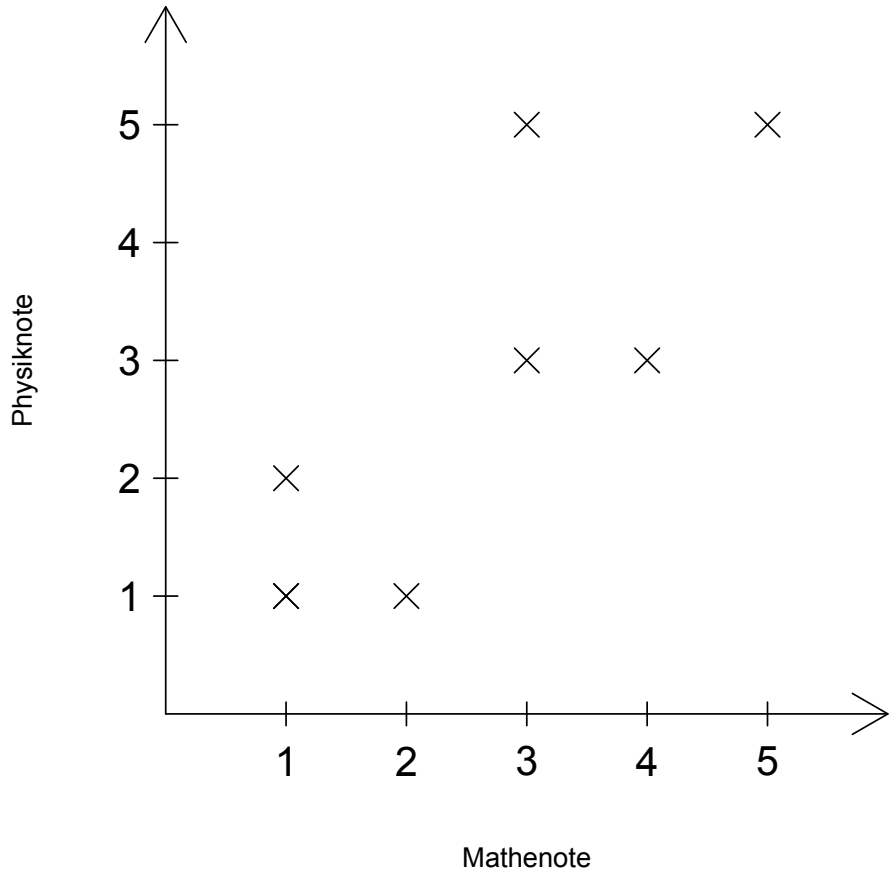
Notation: i-ter Datenpunkt: (x_i, y_i) $\quad i = 1, \ldots, n$

Beispiel: Noten in Mathematik und Physik

Mathe	Physik	Rang Mathe	Rang Physik
1	1		
2	1		
1	2		
3	3		
4	3		
1	1		
5	5		
3	5		

Bei metrisch- oder ordinal-skalierten Werten, wie z.B. den Noten können Ränge berechnet werden. Treten dabei innerhalb eines Merkmals identische Werte auf, ist die Rangvergabe nicht eindeutig. Man spricht in diesem Fall von Bindungen oder Ties und behilft sich mit Durchschnittsrängen, d.h. jedem Beobachtungswert wird das arithmetische Mittel aller möglichen Ränge zugewiesen.

Um einen Überblick über Wertepaare zu erhalten, empfiehlt es sich ein „Streudiagramm“ zu zeichnen:



Um eine Kennzahl zu berechnen, die den Zusammenhang zwischen den einzelnen Werte der Datenpaare angibt, betrachten wir bei metrisch skalierten Werten die Abweichungen $x_i - \bar{x}$ und $y_i - \bar{y}$. Wenn große Werte der x-Variablen gewöhnlich mit großen Werten der y-Variablen und kleine Werte der x-Variablen gewöhnlich mit kleinen Werten der y-Variablen zusammen auftreten, dann sind die Vorzeichen von $x_i - \bar{x}$ und $y_i - \bar{y}$ - egal ob positiv oder negativ - gewöhnlich gleich.

Def.: Seien s_x und s_y die Stichprobenstandardabweichungen der x- bzw. y-Werte. Der „Stichprobenkorrelationskoeffizient“ der Datenpaare $(x_i, y_i), i = 1, \ldots, n$ ist wie folgt definiert:

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

Wenn $r > 0$, nennt man die Datenpaare „positiv korreliert“, wenn $r < 0$, dann nennt man sie „negativ korreliert“.

Satz: (Eigenschaften von r)

(i) $-1 \leq r \leq 1$

(ii) Wenn für zwei Konstanten a und b mit $b > 0$ gilt:

$$y_i = a + bx_i \quad (i = 1, \ldots, n),$$

dann ist $r = 1$.

(iii) Wenn für zwei Konstanten a und b mit $b < 0$ gilt:

$$y_i = a + bx_i, \quad i = 1, \ldots, n,$$

dann ist $r = -1$.

Beweis: Übungsaufgabe 4.10

Interpretation von r:

- Ein Wert von $|r| = 1$ bedeutet, dass es einen vollkommen linearen Zusammenhang gibt.
- Ein Wert von $|r| \approx 0.8$ deutet auf einen relativ starken linearen Zusammenhang hin.
- Ein Wert von $|r|$ um 0.3 besagt, dass der lineare Zusammenhang relativ schwach ist.

In unserem Beispiel erhalten wir:

$$\begin{aligned}
\bar{x}_{math} &= \\
s_{math} &= \\
\bar{x}_{phy} &= \\
s_{phy} &= \\
\Rightarrow r &= \\
&\approx
\end{aligned}$$

Bei metrisch- und auch bei nur ordinal-skalierten Werten kann man „Spearmans Korrelationskoeffinzient“ berechnen, das ist der oben definierte Korrelationskoeffizient für die (Durchschnitts-)Ränge. Man setzt in obigen Formel also nicht die Werte selbst, sondern deren (Durchschnitts-)Ränge ein.

Man kann zeigen, dass Spearmans Korrelationskoeffinzient nicht die Stärke des linearen Zusammenhangs misst, sondern misst, inwieweit große bzw. kleine Werte des einen Merkmals mit großen bzw. kleinen Werten des anderen Merkmals korrespondieren (vgl. Fahrmeir et al. 2016, S. 133f). Die Eigenschaften groß und klein sind dabei relativ auf die anderen Werte des Merkmals zu verstehen.

In unserem Beispiel ergibt sich mit den Stichprobenstandardabweichungen der Ränge $s_{Rg(Mathe)} = 2.375$ und $s_{Rg(Physik)} = 2.36$:[1]

$$\begin{aligned}
r_{Sp} &= \\
&\approx
\end{aligned}$$

Bemerkung: Die Korrelation misst den Zusammenhang, **NICHT** die Kausalität!

Siehe R-Beispiel I.10.

ABC 8; Ü 4; Moodle 1.8-1.10

[1] Das arithmetische Mittel der Ränge für n Beobachtungswerte ergibt sich unabhängig von der Werten wegen der Formel zur arithmetischen Summe zu: $\overline{rg} = \frac{1}{n}\sum_{i=1}^{n} i = \frac{1}{n} \cdot \frac{n(n+1)}{2} = \frac{n+1}{2}$.