DATENANALYSE I

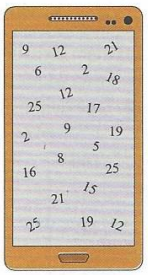
Die Theorie und Daten in diesem Skript stammen nicht aus dem Buch Frommenweiler.
Lesen Sie als Ergänzung dazu im Buch Frommenweiler die Seiten 248 bis 272.

Warum macht man Datenanalyse? Der Mensch ist neugierig, er möchte sein Wissen ständig erweitern. Der Sinn und Zweck der Datananalyse soll uns an den folgenden Beispielen veranschaulicht werden.

1) Drei einführende Beispiele (ohne Kommentare oder Rechnung)

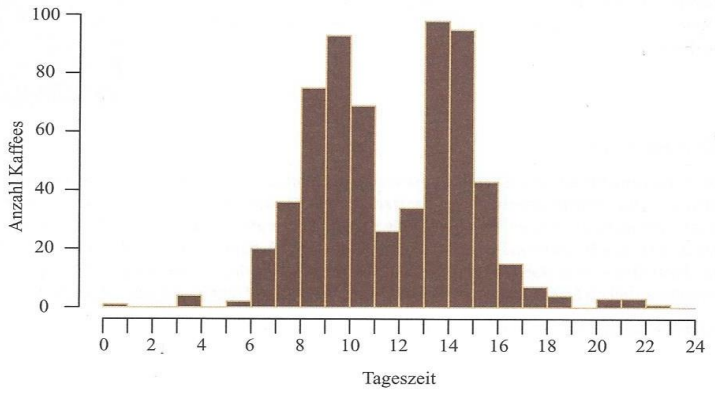
Smartphone

An einer Berufsmaturitätsschule wurden 21 zufällig ausgewählte Lernende befragt, wie viel Zeit sie pro Woche mit ihrem Smartphone telefonieren, Mitteilungen schreiben, chatten oder anderswie den Bildschirm ihres Smartphones aktiv betätigen. Musikhören ab Smartphone sollte nicht eingerechnet werden.
Nebenstehend sind die erfassten Zeiten in der Einheit Stunden völlig ungeordnet im Display abgebildet.



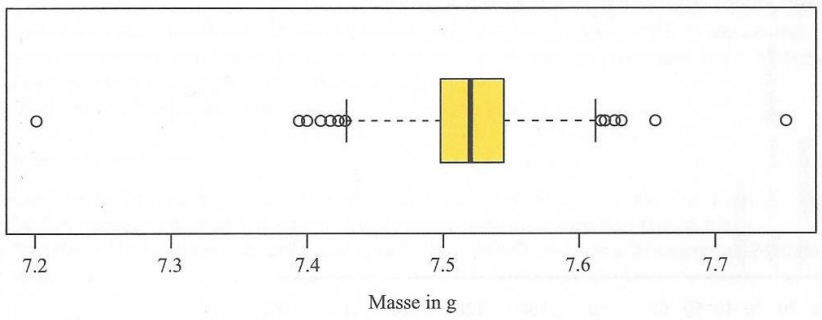
Kaffee

In einem Grossbetrieb wurde an einem Tag der Kaffeekonsum beobachtet. Wann immer ein Kaffee herausgelassen wurde, wurde die Tageszeit erfasst. Insgesamt wurden 629 Tassen Kaffee konsumiert. Stellt man die Tageszeiten in einem Diagramm dar, zeigt sich Folgendes:



1-€-Münze

Gemäss der Europäischen Zentralbank ist die Masse von einer 1-€-Münze 7.5 Gramm. In einer Untersuchung wurde von insgesamt 2000 1-€-Münzen die Masse überprüft.
Dafür wurden von einer Bank acht Packungen zu 250 Münzen zur Verfügung gestellt.
Die Darstellung visualisiert das Ergebnis der Messung:



Täglich werden weltweit grosse Mengen an Daten gesammelt. Man möchte anhand solcher Erhebungen Aussagen machen und Prognosen stellen können. Dazu werden die Daten aufbereitet, neu organisiert, graphisch dargestellt und mit verschiedenen Kennzahlen beschrieben. In der Mathematik bezeichnet man dies als Statistik.

2) Die Methoden der Datengewinnung

Datengewinnung

Um Fragestellungen in Studien zu beantworten, sind Daten notwendig. Die einführenden Beispiele zeigen, dass es verschiedene Möglichkeiten zur **Datengewinnung** gibt. Grundsätzlich unterscheidet man zwischen dem Erfassen von im Prinzip vorhandenen Daten und dem Erfassen von Daten, welche zuerst erzeugt werden müssen. Werden im Prinzip vorhandene Daten erfasst, spricht man von einer **Erhebung**. Bei einer Erhebung werden mit den statistischen Objekten keine zusätzlichen Aktionen durchgeführt. Um Daten zu erzeugen, werden Experimente durchgeführt.

Es gibt: **Experimente, Befragungen, Beobachtungen uns statistische Erhebungen**:

Methoden der Datengewinnung

Experimente

In **geplanten Experimenten** werden geeignete Daten erzeugt. Zu jedem Experiment gehört eine **Versuchsanordnung**. Diese beschreibt, was genau untersucht wird und welche Daten wann, wie oft, wo und wie erfasst werden. Sind an einem Experiment Versuchspersonen oder andere Lebewesen beteiligt, spricht man anstatt von der Versuchsanordnung häufig von einem **Forschungsdesign**.
In Experimenten können **kausale Zusammenhänge** bewiesen werden, weil es möglich ist, an einem System nur einen Faktor zu verändern. Tritt ein Ereignis nur dann auf, ist dieser Faktor kausal relevant. Experimente werden häufig in den Natur- und Ingenieurwissenschaften, aber auch in der Medizin, der Psychologie und der Soziologie durchgeführt.
Ein typisches Experiment ist Beispiel 21.2, Kniearthrose, hier wird die Wirksamkeit einer Therapie untersucht.

Befragungen

Befragungen sind ein klassisches Instrument, mit welchem Daten erhoben werden. Im Gegensatz zum Experiment werden dabei keine zusätzlichen Aktionen vorgenommen, um die Daten zu erzeugen. Befragungen werden vor allem in den Geistes- und Sozialwissenschaften, der Psychologie und den Wirtschaftswissenschaften, aber auch in der Markt- und Meinungsforschung häufig durchgeführt. Es gibt verschiedene Formen: persönlich (Face-to-Face) oder telefonisch durchgeführte **Interviews** oder **standardisierte Fragebogen**, die online oder in Papierform ausgefüllt werden.
Die Umfrage im Beispiel 21.3, Warenhaus, ist ein typisches Beispiel für eine Befragung.

Beobachtungsstudien

Gleich wie bei Befragungen werden bei **Beobachtungsstudien** grundsätzlich vorhandene Daten erhoben. Die Daten werden jedoch nicht durch Befragen, sondern durch Beobachten oder Messen erfasst.
Das Erfassen der Kaffeezeiten im Beispiel 21.4, Kaffee, ist typisch für die Datenerfassung mittels Beobachtung.

Sekundärstatistische Erhebung

Bei einer **sekundärstatistischen Erhebung** werden Daten einer bereits bestehenden **Datensammlung** verwendet. Die für eine Fragestellung relevanten Daten werden aus der Datensammlung herausgefiltert.
Im Beispiel 21.5, Weitsprung, greift der nationale Leichtathletikverband auf eine bestehende Datensammlung zurück.

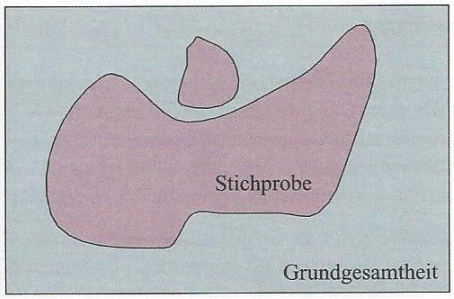
3) Fehler bei der Datengewinnung

Bei der Datengewinnung können **Fehler** passieren, deshalb ist davon auszugehen, dass eine Datensammlung in vielen Fällen fehlerhaft ist. Oft ist es schwierig, zu entscheiden, wo die Fehler liegen und wie mit ihnen umzugehen ist.
Fehler entstehen auf verschiedene Arten. Einige häufige **Fehlertypen** werden hier kurz beschrieben:

Zufällige Fehler	… sind nicht zu verhindern. Im Beispiel 21.6, Übergewicht und Bluthochdruck, wurde der Taillenumfang als Variable erfasst. Stellen wir uns vor, dass bei ein und derselben Person von zehn verschiedenen Ärztinnen der Taillenumfang gemessen wird. Sicher sind wir nicht überrascht, wenn nicht jede Ärztin denselben Umfang misst. Einerseits wird die vermessene Person ihren Bauch kaum zehnmal genau gleich anspannen, andererseits lässt sich die Messung grundsätzlich nicht äusserst präzise durchführen.
Systematische Fehler	… haben ihre Ursache im Messsystem. Die Messwerte sind dann grundsätzlich zu gross oder grundsätzlich zu klein. Beim Beispiel 21.4, Kaffee, kann man sich gut vorstellen, dass die Uhr einer Kaffeemaschine nicht korrekt eingestellt ist, sodass die gemessenen Zeiten systematisch falsch sind.
Übertragungsfehler	… können beim Auf- und Abschreiben oder Übertragen von Werten passieren. Schnell kann es geschehen, dass die Körpertemperatur von 39.6 °C als 36.9 °C notiert wird oder dass beim Aufbereiten und Abschreiben von Strichlisten Zeilen oder Spalten verwechselt werden.
Mutwillige Fehler	… liegen vor, wenn einzelne Daten oder gar ganze Datensätze absichtlich gefälscht werden. Es kommt vor, dass einzelne Messwerte mutwillig angepasst oder gelöscht werden, weil Messfehler vermutet werden. Ganze Datensätze werden manipuliert, um bestimmte Versuchsergebnisse überhaupt oder zumindest klarer und besser zu erreichen.

4) Grundbegriffe der Datenanalyse

4.1 Grundgesamtheit
Die Menge (Anzahl) der für eine Untersuchung relevanten Personen oder Objekte.
4.2 Stichprobe
Ist eine Teilmenge der Grundgesamtheit: Repräsentative Zufallswahl.
4.3 Stichprobenumfang
Die Anzahl der in der Stichprobe untersuchten Personen oder Objekte.



Befragung der gesamten Grundgesamtheit nennt man *Vollerhebung*. Eine solche wird meist aus organisatorischen (Zeit!) oder finanziellen Gründen nicht durchgeführt. Daher nimmt man eine oder mehrere *Stichproben*.

Für eine Stichprobe werden einzelne Personen oder Objekte aus der Grundgesamtheit ausgewählt. Die Auswahl der Stichprobe kann das Ergebnis einer Studie wesentlich beeinflussen (verfälschen) und ist deshalb oft ein Kritikpunkt. Deshalb sollte die Auswahl der Stichproben zufällig sein!

Bei vielen Studien ist eine zufällige Auswahl der Stichproben nicht möglich: Zum Beispiel wenn Probanden freiwillig teilnehmen. Die Resultate solcher Erhebungen müssen dann besonders kritisch hinterfragt werden. In diesem Fall können *systematische Abweichungen* auftreten. Diese systematischen Abweichungen werden *Sampling-Bias* oder kurz *Bias* genannt:

4.4 Bias
Systematische Abweichung, welche bei einer nicht zufälligen Stichprobe auftritt.

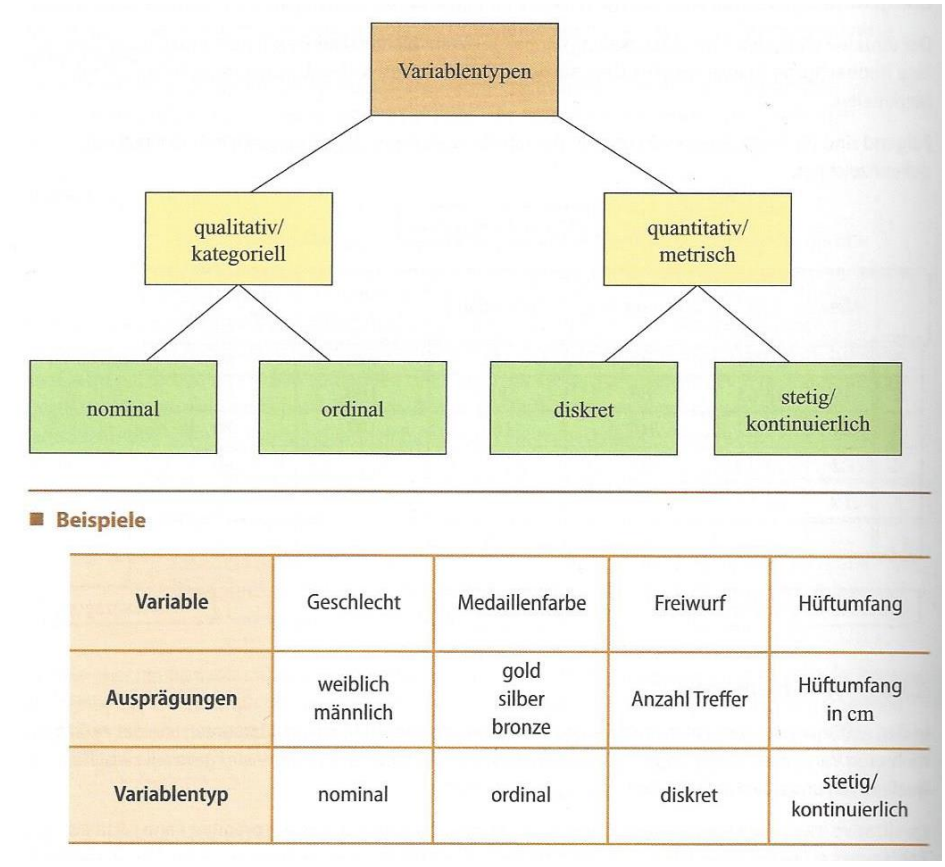
4.5 Repräsentativität
Eine Stichprobe ist dann *repäsentativ*, wenn sie ein unverzerrtes Abbild der Grundgesamtheit wiedergibt. Das heisst, für alle Merkmale entspricht die Verteilung der Häufigkeiten in der Stichprobe jener in der Grundgesamtheit.

4.6 Urliste
Das unmittelbare Ergebnis einer Datengewinnung ist die Urliste. Darin sind die ursprünglichen erhaltenen Werte aufgelistet ohne dass diese auf irgend eine Weise bearbeitet wurden.

4.7 Datensatz
Zusammengefasste Daten einer Studie nennt man Datensatz, egal ob die Daten in Form der Urliste vorhanden sind oder bereits neu geordnet wurden.

Bei einer Datengewinnung werden einzelne Untersuchungseinheiten beobachtet, befragt oder vermessen. Das können Personen oder Objekte sein. Bei jeder Untersuchungseinheit wird mindestens ein *Merkmal* oder eine *Variable* erhoben. Nur ein Merkmal ergeben *univariate* Daten. Die Erhebung mehrer Merkmale pro Untersuchungseinheit ergeben *multivariate* Daten.

Variablentypen (Siehe auch Frommenweiler S. 249)



■ Beispiele

Variable	Geschlecht	Medaillenfarbe	Freiwurf	Hüftumfang
Ausprägungen	weiblich männlich	gold silber bronze	Anzahl Treffer	Hüftumfang in cm
Variablentyp	nominal	ordinal	diskret	stetig/ kontinuierlich

Nach der Datengewinnung wird die Urliste häufig neu organisiert. Nicht selten werden die Beobachtungen nach der Ausprägung einer Variablen (Merkmal) geordnet. Man spricht dann von einer *geordneten Stichprobe.*

Beispiele

(1) Wird eine Stichprobe nach einem **nominalen Merkmal** geordnet, werden die Beobachtungen mit gleicher Ausprägung hintereinander aufgelistet.

Urliste	
	Geschlecht
Alice	w
Bob	m
Marla	w
Paula	w
Felix	m

$\Rightarrow$

Geordnete Stichprobe	
	Geschlecht
Alice	w
Marla	w
Paula	w
Bob	m
Felix	m

(2) Wird eine Stichprobe nach einem **ordinalen Merkmal** geordnet, können die Beobachtungen entsprechend der Ordnung der Merkmalsausprägungen aufgelistet werden.

Urliste	
	Fitness
Alice	schlecht
Bob	mittel
Marla	gut
Paula	gut
Felix	schlecht

$\Rightarrow$

Geordnete Stichprobe	
	Fitness
Alice	schlecht
Felix	schlecht
Bob	mittel
Marla	gut
Paula	gut

(3) Wird eine Stichprobe nach einem **quantitativen Merkmal** geordnet, werden die Beobachtungen nach der Grösse der Ausprägungen aufgelistet.

Urliste	
	Alter in J.
Alice	4
Bob	27
Marla	5
Paula	3
Felix	8

$\Rightarrow$

Geordnete Stichprobe	
	Alter in J.
Paula	3
Alice	4
Marla	5
Felix	8
Bob	27

Wird nach einem quantitativen Merkmal geordnet, entspricht die geordnete Stichprobe einer Rangliste, üblicherweise aufsteigend sortiert. Jeder Beobachtung kann nun ein *Rang*, beziehungsweise eine Ordnungszahl zugeordnet werden.

Rang (Ordnungszahl)

Der Rang einer Beobachtung (einzelne Daten) gibt an, welchen Platz die Beobachtung in der geordneten Stichprobe hat. Bei gleichen Ausprägungen werden die entsprechenden Plätze gemittelt.

Werden die Stichprobenwerte $x_1, x_2, \ldots, x_n$ geordnet, erhält man die geordneten Stichprobenwerte $x_{[1]}, x_{[2]}, \ldots, x_{[n]}$, wobei $x_{[1]} \leq x_{[2]} \leq \ldots \leq x_{[n]}$. Der Stichprobenwert $x_{[1]}$ ist somit der kleinste beobachtete Wert und $x_{[n]}$ der grösste beobachtete Wert. Der Rang $[k]$ der Beobachtung $x_{[k]}$ gibt an, wie viele Beobachtungen kleiner oder gleich dieser Beobachtung sind.

Beispiele

(1) Stichprobe ohne gleiche Ausprägungen:

$$\begin{array}{l} x_1 = 8 \\ x_2 = 27 \\ x_3 = 5 \end{array} \Rightarrow \begin{array}{l} x_{[1]} = 5 \\ x_{[2]} = 8 \\ x_{[3]} = 27 \end{array}$$

(2) Kommen in der Stichprobe Beobachtungen mit gleicher Ausprägung vor, werden die entsprechenden Plätze gemittelt.

$$\begin{array}{l} x_1 = 9 \\ x_2 = 34 \\ x_3 = 5 \\ x_4 = 9 \\ x_5 = 9 \\ x_6 = 2 \\ x_7 = 5 \\ x_8 = 11 \end{array} \Rightarrow \begin{array}{l} x_{[1]} = 2 \\ x_{[2.5]} = 5 \\ x_{[2.5]} = 5 \\ x_{[5]} = 9 \\ x_{[5]} = 9 \\ x_{[5]} = 9 \\ x_{[7]} = 11 \\ x_{[8]} = 34 \end{array}$$

Übungen

1. Zu den erfassten Zeiten aus Beispiel 21.1, Smartphone, sollen einige Überlegungen und Berechnungen angestellt werden.

   a) Wie viele Lernende wurden befragt?
   b) Welches ist die kleinste, welches die grösste erfasste Zeit?
   c) Wie gross ist der Durchschnitt aller erfassten Zeiten?
   d) Welche Zeit liegt so, dass es gleich viele kleinere wie grössere Zeiten gibt?
   e) Wie viele Zeiten liegen im Bereich von 15 bis 20 Stunden?
   f) Wie viel Prozent der Zeiten liegen im Bereich von 15 bis 20 Stunden?

2.

Im Kapitel 22 wurden vier Möglichkeiten zur Datengewinnung beschrieben. Wie wurden in den einführenden Beispielen die Daten gewonnen? Kreuzen Sie an:

	Experiment	Befragung	Beobachtungs-studie	Daten-sammlung
Smartphone	☐	☐	☐	☐
Kaffee	☐	☐	☐	☐
1-€-Münze	☐	☐	☐	☐

3.

Folgend sind Fälle aus der Datengewinnung beschrieben, bei welchen Fehler passiert sind. Um welchen Fehlertyp handelt es sich?

Fall A:
An der Leichtathletik-WM 1987 in Rom sprang der Italiener Giovanni Evangelisti bei seinem letzten Sprung im Weitsprungwettbewerb nur rund $7.80\ \mathrm{m}$. Offenbar wollte der Kampfrichter seinem Landsmann die Bronzemedaille sichern und mass eine Weite von $8.38\ \mathrm{m}$.

Fall B:
Obwohl längst widerlegt, hält sich die Mär vom äusserst eisenreichen Spinat hartnäckig. Der Mythos beruht der Legende nach auf einem Kommafehler eines Lebensmittelanalytikers.

Fall C:
Der niederländische Sozialpsychologe und ehemalige Professor Diederik Stapel hat während seinen Forschungsarbeiten zusätzliche Daten erfunden.

Fall D:
Ein Arzt misst bei der Skifahrerin Lara Gut die Körpergrösse von $1.60\ \mathrm{m}$. Ein Ausrüster hat Guts Körpergrösse als $1.61\ \mathrm{m}$ gemessen.

Fall E:
Die Lehrerin gibt in ihrem Tabellenkalkulationsprogramm eine falsche Formel zur Berechnung der Probennoten ein, weshalb jede Arbeit mit 0.5 Notenpunkten zu hoch bewertet wird.

Fall F:
Im selbst erfassten Datensatz aus Aufgabe 2 hat Meret Muster dreimal das Wasservolumen gemessen, welches in ihrem geschlossenen Mund Platz findet: $7.9\ \mathrm{cl}$, $8.0\ \mathrm{cl}$ und $8.3\ \mathrm{cl}$ waren die drei Messergebnisse.

Kreuzen Sie an:

	zufälliger Fehler	systematischer Fehler	Übertragungs-fehler	mutwilliger Fehler
Fall A	☐	☐	☐	☐
Fall B	☐	☐	☐	☐
Fall C	☐	☐	☐	☐
Fall D	☐	☐	☐	☐
Fall E	☐	☐	☐	☐
Fall F	☐	☐	☐	☐

4.

Bilden Sie die geordnete Stichprobe:

$$
\begin{aligned}
x_1 &= 36.9 \\
x_2 &= 37.4 \\
x_3 &= 39.1 \\
x_4 &= 40.5 \\
x_5 &= 39.1 \\
x_6 &= 36.2
\end{aligned}
\quad \Longrightarrow \quad
\begin{aligned}
x_{[\ldots]} &= \ldots \\
x_{[\ldots]} &= \ldots \\
x_{[\ldots]} &= \ldots \\
x_{[\ldots]} &= \ldots \\
x_{[\ldots]} &= \ldots \\
x_{[\ldots]} &= \ldots
\end{aligned}
$$