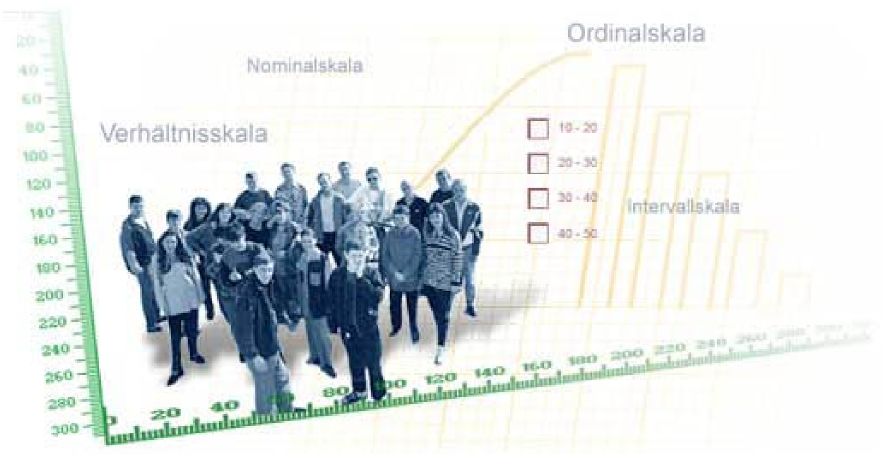
GST - Grundbegriffe der Statistik



Lernziele und Überblick

Diese Lerneinheit hat das Ziel, Sie mit den Grundlagen der deskriptiven Statistik vertraut zu machen. Es werden für die Statistik grundlegende Begriffe wie Datenerhebung, Vollerhebung, Teilerhebung, Einheit und Gesamtheit erläutert sowie die Begriffe Merkmal, Merkmalsausprägung, Skala und Transformation eingeführt.

Lernziele

- Sie können die Begriffe Datenerhebung, Vollerhebung, Teilerhebung, Einheit, die in der deskriptiven Statistik verwendet werden erklären.
- Sie können qualitative und quantitative (diskrete, stetige) Merkmale unterscheiden und beschreiben.
- Sie können die 4 Skalenniveautypen erklären und ihre Unterschiede benennen:
  - Nominalskala,
  - Ordinalskala,
  - Intervallskala,
  - Verhältnisskala
- Sie können die verschiedenen Transformationen beschreiben und mit Beispielen unterlegen:
  - ein-eindeutige Transformation,
  - streng monotone Transformation,
  - eine lineare Transformation (Lineartransformation),
  - eine proportionale Transformation.

Diese Lerneinheit enthält Tabellen zu den Skalen und Merkmalsklassifikationen. Mit der Hilfe dieser Tabellen können Sie besser verstehen, auf welchem Skalenniveau Ihre (gemessenen) Daten vorliegen.

Den Stoff dieser Lerneinheit werden Sie für alle folgenden Lerneinheiten benötigen.

Gliederung der Lerneinheit

1. Einleitung
2. Gegenstand der Statistik
3. Statistische Einheiten
4. Merkmal, Merkmalsträger, Ausprägungen
5. Merkmalsarten und Skalen
6. Qualitative Daten
7. Quantitative Daten
8. Transformationen

Zeitbedarf und Umfang

Für die Durcharbeitung dieser Lerneinheit benötigen Sie ca. 180 Minuten und für die Übungen der Statistiksoftware **R** ca. 150 Minuten.

1 Einleitung

Diese Lerneinheit soll Sie mit wichtigen Grundbegriffen und Definitionen der beschreibenden Statistik bekannt machen.

Sie fragen sich vielleicht, weshalb erst Grundbegriffe, es geht doch um Datenbeschreibung.
Mit der Datenbeschreibung verhält es sich wie mit dem Kochen. Erst wenn wir über die Zutaten Bescheid wissen, sind wir in der Lage, ein hervorragendes Menü zuzubereiten. Milch kann man kaum braten oder backen, aus Tomatenmark lassen sich nicht wieder Tomaten herstellen.

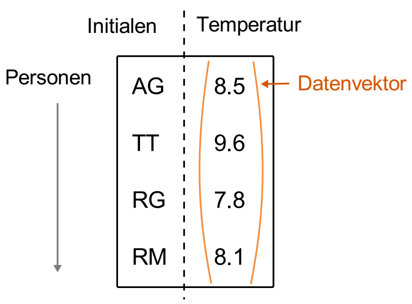
Zu den Zutaten der Statistik gehören Merkmale, Merkmalsträger und vieles mehr.
Jetzt viel Spaß bei der ersten richtigen Lektion, Sie werden viele Übungen finden, die Ihnen das Verständnis erleichtern sollen. Und keine Bange, diese Einheit ist trotz des großen Umfangs gut zu bewältigen.

Was ist eigentlich Statistik?

Wir ziehen für den aktuellen Stand die Begriffsklärung von Wikipedia zurate. (Statistik – Wikipedia) Der Link führt zu einer sehr umfangreichen Einführung in Begriffe der Statistik. Für unsere Lerneinheiten werden wir uns auf das beschränken, was wir behandeln. Definitionen und Grundlagen, die zur beschreibenden Statistik gehören, waren lange Zeit stark vom wissenschaftlichen Anspruch der empirischen Sozialforschung geprägt. In der Schließenden Statistik dominierten streng mathematische Definitionen aus der Wahrscheinlichkeitstheorie und der Entscheidungstheorie. Heute gehören statistische Methoden zum festen Bestandteil vieler Anwendungsgebiete. Viele Anwender und Anwenderinnen haben oft einen pragmatischen Umgang mit den Begriffen und verwenden teilweise auch unterschiedliche Namen für die gleichen Dinge. Ein solcher Wandel im Laufe der Zeit ist durchaus üblich, genau wie bei der Umgangssprache. Der Duden wird ebenfalls regelmäßig aktualisiert. In der Wikipedia Einführung stellt man fest, dass sich gegenüber früheren Definitionen, die sich auch in vielen klassischen Lehrbüchern finden, Daten an sich in den Vordergrund gerückt sind. Das ist sicher ein Ergebnis der immer weiter fortschreitenden Digitalisierung, die es ermöglicht, für fast alle Fragen Daten zu finden. In den Medien und im Internet werden wir von Daten fast erschlagen. Für die meisten Menschen ist es sehr schwierig, die Qualität der Daten und der aus den Daten gewonnenen Informationen zu beurteilen. Kenntnisse in Statistik ermöglichen es uns, zu einem eigenen kritischen Urteil zu gelangen. Wir können dann selbst Daten sammeln, wichtige Kenngrößen ermitteln und dadurch erste Schritte einer Analyse selbst durchführen. Dass das gar nicht so schwer ist, haben sie in unserer Einführung schon erfahren.

Wie kommen Daten zustande?

Wir interessieren uns für eine Eigenschaft eines Untersuchungsobjekts und beobachten, messen und dokumentieren diese Eigenschaft auf geeignete Weise. Nehmen wir zum Beispiel die Temperatur von Leitungswasser – das bei allen etwas unterschiedlich aus dem Hahn kommt (in °C, eigentlich ist eine genauere Messvorschrift erforderlich). Messen vier Personen an einem Stichtag um 08:00 Uhr morgens die Temperatur ihres Wasser, können wir die Messergebnisse zusammentragen und aufschreiben. Um die Daten zuordnen zu können, verwenden wir die Initialen der Personen. Zur Messung haben sich alle auf den Standard „Badethermometer“ geeinigt. Die Ergebnisse wurden in eine Tabelle eingetragen.



Es ist nützlich, die Daten als Vektor aufzuschreiben – **`temp= (8.5, 9.6, 7.8, 8.1)`** - auch wenn es für unsere Zwecke nicht auf die Reihenfolge ankommt. Sie können die Daten dann leicht mit Hilfe der Statistik Software R erfassen und bearbeiten. Schauen Sie sich das folgende Beispiel an.

Quellcode

```
temp         = c(8.5, 9.6, 7.8, 8.1)
names(temp) = c("AG","TT","RG","RM")
temp
AG  TT  RG  RM
8.5 9.6 7.8 8.1
str(temp)
Named num [1:4] 8.5 9.6 7.8 8.1
- attr(*, "names")= chr [1:4] "AG" "TT" "RG" "RM"

```

Mit Daten dieser Art, also mit Vektoren von Messergebnissen oder Beobachtungen, werden wir uns in diesem Kurs hauptsächlich befassen. Ob die Daten dabei als Tabelle oder in der mathematischen Schreibweise dargestellt werden spielt keine Rolle.

Wir könnten neben der Temperatur noch mit einem einfachen Messpapier den PH-Wert des Wassers ermitteln. Dann würden wir unsere Daten einfach ergänzen und hätten folgende Datenstruktur:

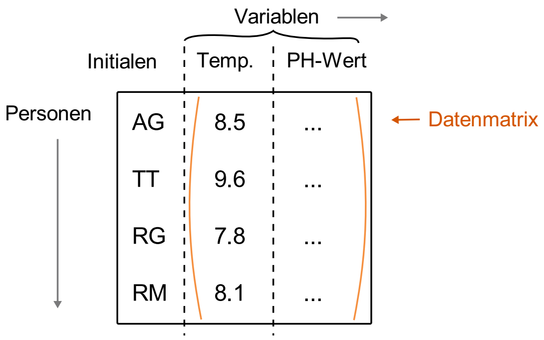


Abb.: Bsp. Datenmatrix mit 2 Spalten

Ergebnis wäre eine Datenmatrix mit 2 Spalten. Die Initiale können wir wieder zur Identifizierung der Messungen verwenden. Auch solche Daten werden in unserem Kurs eine Rolle spielen.

Wenn wir die Wassertemperatur an aufeinander folgenden Tagen erfassen und aufschreiben hätten dann wird der Temperaturverlauf über mehrere Tage bestimmt.

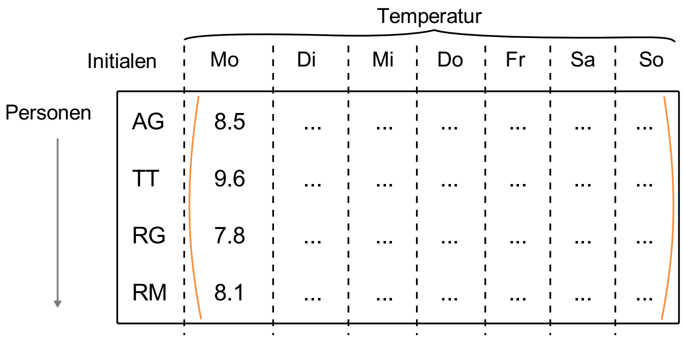


Abb.: Bsp. Temperaturverlauf

Hier messen wir also je Individuum (Person) einen Zeitverlauf. Bei der Analyse müssen wir berücksichtigten, dass die Werte jeweils durch die Wasserversorgung des einzelnen Haushalts beeinflusst werden.

Wie sieht es mit etwas Lehrbuch ferneren Daten aus?

Einige stellen wir Ihnen im Laufe unserer Lerneinheiten vor. Solche Datensätze sind oft sehr umfangreich, und eine Bearbeitung ist ohne geeignete Programme, der Software und ausreichender Rechenkapazität, der Hardware nicht möglich. Umfangreiche Berechnungen wurden noch vor 80 Jahren in großen Rechensälen, Übrigens überwiegend von Frauen, mithilfe einfacher Rechenmaschinen durchgeführt. Die Frauen dieser Rechensäle haben versucht, Genauigkeit und Geschwindigkeit zu erhöhen und haben dadurch entscheidende Beiträge zur Entwicklung der Computerwissenschaft geleistet. Heute können wir auf einfach bedienbare Software und Hardware zurückgreifen und sind von umständlicher Handarbeit befreit. Auch was die Datenstrukturen anbelangt, gibt es neue Entwicklungen. Damit Sie einen Eindruck über die Möglichkeiten bekommen, was moderne statistische Methoden und Ansätze der Data Science liefern, erhalten sie einen kleinen Einblick in die Struktur großer Datensätze: **Big Data**.

Datenstrukturen und Big Data

Viele Daten werden heute von kommerziellen Anbietern wie Google gesammelt. Dabei werden für einige Variablen die Daten von möglichst vielen Nutzern erfasst und gespeichert. Die Daten haben dann folgende Gestalt:

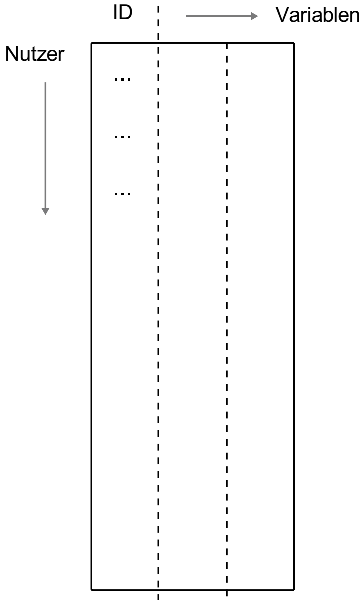


Abb.: Bsp. Viele Nutzer

Die Anzahl der Nutzer und Nutzerinnen ist in der Regel groß bis sehr groß; wir reden von Tausenden, Millionen, sogar Milliarden, wenn wir an die Daten über Impfungen in der Pandemie denken. Die Anzahl der Variablen ist meistens ungefähr 2 - 100.

In einem anderen Zusammenhang werden ungeheuer viele Variablen erfasst. Das können zum Beispiel detaillierte genetische oder molekulare Eigenschaften sein.

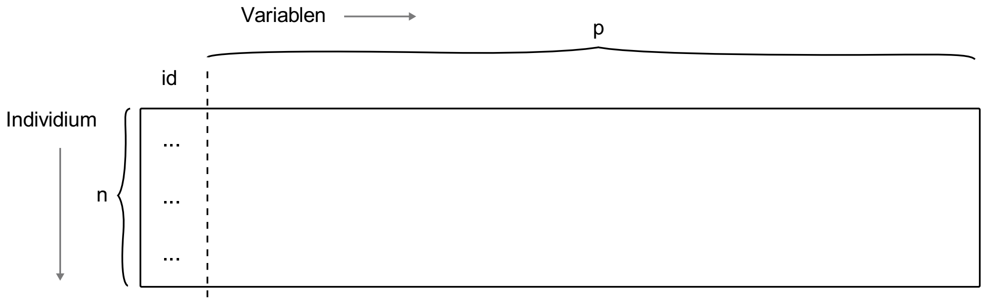


Abb.: Bsp. Viele Variablen

$n \ll p$ die Anzahl der Individuen ist viel kleiner als die Anzahl der Variablen

Hier reden wir von einer mittleren Anzahl von Individuen, die etwa mehreren Hundert bis zu wenigen 1000 reicht. Dafür, können aber sehr viele Variablen auftreten, von über 1000 bis zu mehreren Millionen. Im Folgenden betrachten wir noch deutlich komplexere Datenstrukturen. Häufig gibt es nicht für alle Individuen oder Beobachtungseinheiten die gleichen Informationen. Ein Versicherungsunternehmen erfasst für seine Kunden und Kundinnen jeweils alle Verträge. Zu den Verträgen selbst gibt es unterschiedlich komplexe Angaben, und die Datenstruktur entspricht dann einer Liste.

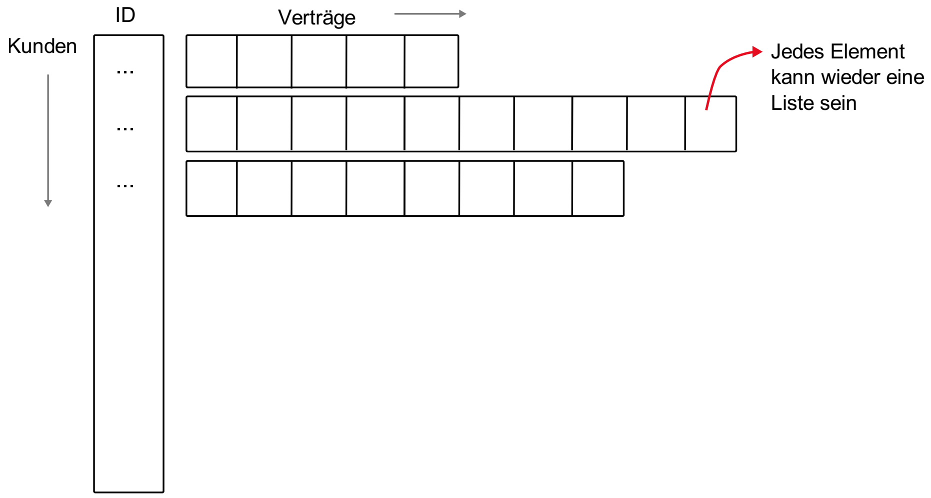


Abb.: Bsp. Viele Variablen

Alle Personen die zum Kundenstamm gehören können eine unterschiedliche Anzahl von Verträgen abgeschlossen haben. Das heißt, die Einträge können von Person zu Person unterschiedlich viele Elemente enthalten. Weiter ist zu beachten, dass jeder Vertrag selbst wieder durch eine Liste charakterisiert ist. Solche Daten treten zum Beispiel bei der Telekom, bei Krankenversicherungen und in vielen anderen Zusammenhängen auf. Die Anzahl der Kunden kann dabei von Tausenden, Hunderttausend bis zu Millionen reichen.

Um hier statistische Auswertungen zu erhalten, werden die Angaben einzelner Personen zunächst ausgewertet und neuen Variablen zugeordnet, wie zum Beispiel Angaben, die das Risiko für eine Versicherung charakterisieren. Die Schufa, bekannt bei allen Wohnungssuchenden, sammelt solche Angaben auch. Eine statistische Analyse von Schufa Anfragen würde dann Anhaltspunkte über die Zahlungsfähigkeit der Mietsuchenden einer Vermittlungsagentur wie z. B. ImmoScout erlauben.

Was für alle diese Datenstrukturen wichtig ist: die Daten müssen so aufbereitet, beschrieben und dargestellt werden, dass auch Nichtfachleute und statistische Laien die Plausibilität von Schlussfolgerungen, die aus der Datenanalyse gezogen werden, nachvollziehen können.Wissenschaftliche Zeitschriften, die großen Wert auf qualitativ hochwertige Inhalte legen, verlangen neben angemessenen Analysen auch leicht verständliche elementare Beschreibungen und Zusammenfassung der verwendeten Daten. Und solche Zusammenfassungen und Datenbeschreibung werden Sie nach dem Durcharbeiten unserer Lerneinheiten selbst erstellen können.

2 Gegenstand der Statistik

Wir wollen Ihnen zunächst eine Definition des Begriffs Statistik geben, damit Sie sagen können, womit Sie sich in unserem Kurs beschäftigen werden.

Definition

Statistik

Statistik ist die Lehre von Verfahren und Methoden zur Gewinnung, Erfassung, Analyse, Charakterisierung, Abbildung, Nachbildung und Beurteilung von beobachtbaren Daten über die Wirklichkeit (Empirie).

Anmerkungen

1. **Datenanalyse**
   Bei der Analyse von Daten werden geeignete statistische Verfahren zum Zwecke der Erkenntnisgewinnung angewendet.
2. **Datencharakterisierung**
   Die grafische und tabellarische Darstellung von Daten sowie die Berechnung von zusammenfassenden, den empirischen Sachverhalt beschreibenden, Kennzahlen wird als Datencharakterisierung bezeichnet. Sie ist Gegenstand der deskriptiven Statistik.
3. **Datenbeurteilung**
   Die Beurteilung von Daten kann einerseits durch Schlüsse auf der Basis unvollständiger Daten, z. B. Schlüsse von der Stichprobe auf ihre Grundgesamtheit, erfolgen, andererseits, allgemeiner, auf der Basis unsicherer Daten, unter Anwendung der Wahrscheinlichkeitsrechnung. Dies ist Gegenstand der induktiven Statistik.
4. **Datenaufbereitung**
   Die Aufbereitung von Daten umfasst die Ordnung, Zusammenfassung und Darstellung des erhobenen statistischen Datenmaterials in Datendateien, Tabellen und/oder geeigneten Grafiken.
5. **Datenmissbrauch**
   Man sieht statistischen Ergebnissen nicht an, ob sie manipuliert wurden. Der Missbrauch von Daten ist kein Problem der Statistik, sondern eines der Personen, die mit Daten umgehen.

Die klassische Herangehensweise der empirischen Sozialforschung zur beschreibenden Statistik wird im folgenden Rolloverbild verdeutlicht.

Rolloverbild

Statistik

2.1 Datenerhebung

In unseren Lerneinheiten werden Sie Methoden zur Darstellung von Daten mit Hilfe von Tabellen, Maßzahlen sowie statistischen Graphiken kennenlernen. Diesen Teil der Statistik werden wir behandeln.

Definition

Datenerhebung

Datenerhebung ist der Vorgang zur Ermittlung und zur Erfassung von Ausprägungen eines statistischen Merkmals.

Anmerkungen

Unter einer Erhebung (Datenerhebung) versteht man jede systematische Gewinnung von statistischen Daten.

Primärerhebung ist eine statistische Untersuchung des Datenmaterials durch eine Erhebung nach speziellen Ausprägungen von sachlichen, räumlichen und zeitlichen Identifikationsmerkmalen.

Vollerhebung ist eine statistische Untersuchung aller statistischen Einheiten einer Gesamtheit.

Teilerhebung ist eine Erhebung, die sich auf einen Teil einer statistischen Gesamtheit beschränkt. Eine Teilerhebung liegt vor, wenn nur n von N Einheiten der Grundgesamtheit erhoben werden ($n < N$).

Hinweis

Wir werden den Begriff Grundgesamtheit im Folgenden möglichst vermeiden. Der Zusammenhang zwischen Daten und Grundgesamtheit lässt sich korrekt nur mit Begriffen der Wahrscheinlichkeitsrechnung darstellen. Deshalb nehmen wir die Daten, die wir verwenden als gegeben hin.

2.2 Deskriptive und induktive Statistik

Die Statistik kann in zwei große Bereiche unterteilt werden:

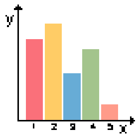
- deskriptive oder beschreibende Statistik,
- induktive oder schließende Statistik.

Die **deskriptive Statistik** (*lat.*: descriptio - Beschreibung) dient der Betrachtung der Daten an sich. Mithilfe der deskriptiven Statistik sollen gewonnene Daten verdichtet bzw. so dargestellt werden, dass das Wesentliche deutlich hervortritt.

Um eine übersichtliche Darstellung zu erreichen, muss das, oft sehr umfangreiche, Material auf geeignete Art und Weise zusammengefasst werden. In der deskriptiven Statistik werden zu diesem Zweck insbesondere die folgenden drei Darstellungsformen benutzt:

- Tabellen
- grafische Darstellungen

Nr.	Name	PLZ
1	Ralf	10999
2	Lena	11471
3	Abu	79115



- charakteristische Maßzahlen

$\overline{x}$

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

Der Schluss vom Teil aufs Ganze, im Sinne der induktiven Statistik, zählt nicht mehr zur deskriptiven Statistik.

Die **induktive Statistik** (*lat.*: inductio - Hineinführen) dient dazu, aus den erhobenen Fakten Schlüsse auf die Ursachenkomplexe zu ziehen, die zu diesen Daten geführt haben. Die induktive Statistik basiert auf der Wahrscheinlichkeitstheorie.

Die Einteilung in deskriptive und induktive Statistik wurde verwendet, um die unterschiedliche Zielsetzung der in diesen beiden Bereichen verwendeten Methoden herauszustellen („Beschreiben“ im Gegensatz zu „geplant Analysieren“).

Lässt man diese strenge Einteilung außer Acht, ergeben sich viele Möglichkeiten, nach verborgener Struktur, nach Ursache-Wirkungsrelationen usw. in Daten zu suchen. In Verbindung mit der auf leistungsfähigen Rechnern heute verfügbaren Software, die auch neuere grafische Verfahren zur Datenbeschreibung umfasst, hat sich als neues Konzept die **explorative Datenanalyse** durchgesetzt.

Synonyme

Synonyme für induktive Statistik: analytische oder inferentielle Statistik.

3 Statistische Einheiten

Bei jeder statistischen Untersuchung müssen wir uns zunächst darüber klar werden, was wir wissen wollen und wie oder wo wir die benötigten Informationen gewinnen können. Wir gehen in unseren Lerneinheiten davon aus, dass die verwendeten Daten an sich geeignet sind.

Beispiel

Standortsuche

Will sich ein Unternehmen in einer Industrieregion ansiedeln, interessiert sich die Unternehmensleitung zum Beispiel für die Wirtschaftsstruktur dieser Region.
Das heißt, sie will wissen, welcher Art die bereits dort vorhandenen Betriebe sind.

Einzelne Unternehmen bilden dabei natürliche Einheiten, die beobachtet werden können. Je nach Fragestellung, wie z. B.:

- Wie stehen die Konkurrenten da?
- Gibt es ausreichend Zulieferer für die geplante Produktion?

ergeben sich unterschiedliche Gruppen von Unternehmen, wir sagen Gesamtheiten, die uns interessieren und über die wir Informationen sammeln wollen.

Merkmalsträger

Die Individuen, Objekte, Versuchseinheiten, Beobachtungseinheiten, an denen wir die Daten erheben, messen, beobachten etc werden in der Sozialforschung **Merkmalsträger** genannt, da es sich bei den Daten meist um Eigenschaften von Personen handelt. Wir verwenden vor allem die Begriffe „Merkmalsträger“ und „Beobachtungseinheit“. Diese sind als Synonyme aufzufassen, allgemein werden sie auch als „statistische Einheiten“ bezeichnet. Kurz, meint alles das selbe.

Definition

Statistische Einheit

Statistische Einheiten sind die kleinsten Elemente in der Statistik. Eine statistische Einheit (Element, Merkmalsträger) ist Träger von Informationen bzw. Eigenschaften, die im Rahmen einer statistischen Untersuchung von Interesse sind.

Definition

Statistische Gesamtheit

Die statistische Gesamtheit ist eine endliche Menge wohl unterschiedener, sachlich, räumlich und zeitlich gleich abgegrenzter statistischer Einheiten.

Synonyme

Synonyme: Masse, Grundgesamtheit, Population, Kollektiv.

4 Merkmal, Merkmalsträger, Ausprägungen

In diesem Kapitel werden die Begriffe:

- Merkmal,
- Merkmalsträger und
- Merkmalsausprägung

vorgestellt und erläutert.

Sie sollten nach dem Durcharbeiten dieses Kapitels die wesentlichen Begriffe voneinander unterscheiden und mit ihnen umgehen können.

4.1 Statistische Merkmale

Beispiel

Merkmale

Für die Beschreibung der Wirtschaftsstruktur einer Industrieregion sind u.a. folgende Eigenschaften der Unternehmen wichtig:

- R = Rechtsform (AG, GmbH, KG, BGB-Gesellschaft...),
- G = Unternehmensgröße (klein, mittel, groß),
- Z = Zahl der Beschäftigten (tatsächliche Anzahl),
- U = Umsatz (in Millionen Euro).

Definition

Merkmal

Ein Merkmal ist eine Eigenschaft einer statistischen Einheit, die Grundlage oder Gegenstand einer statistischen Untersuchung ist.

Anmerkung

In der Regel werden die Merkmale mit großen lateinischen Buchstaben X, Y, Z bezeichnet.

Synonyme

Synonyme: Eigenschaft, Erhebungsmerkmal, Erfassungsmerkmal.

4.2 Merkmalsausprägungen

Beispiel

Merkmalausprägungen, Industrieregion

Die RuckZuck Baumaschinen GmbH, ein in der entsprechenden Industrieregion ansässiges mittelständisches Unternehmen mit 150 Beschäftigten, erwirtschaftete im letzten Jahr einen Umsatz von 20 Mio. €.
Diese Angaben lassen sich leicht als die Merkmalausprägungen der auf der vorherigen Seite genannten Merkmale aufschreiben:

R = GmbH, G = mittel, Z = 150, U = 20

Gewinnen wir diese Werte für viele oder alle Firmen der Region, können wir beurteilen, wie sich unser Unternehmen in die Wirtschaftslandschaft einfügt.

Definition

Merkmalausprägung

Den Wert, den ein Merkmal (eine Variable) für einen Merkmalsträger annimmt nennt man Merkmalsausprägung. Die potentiellen Merkmalsausprägungen bestehen aus allen Werten, die ein Merkmal annehmen kann. Konkrete Werte für eine Beobachtungseinheit nennt man auch Realisierungen.

Anmerkungen

In der Regel werden die Merkmalsausprägungen mit kleinen lateinischen Buchstaben x_1, x_2, x_3, ... bezeichnet.

Ein Merkmal hat endlich (oder unendlich) viele Merkmalsausprägungen. Somit steht ein Merkmal für eine Menge von Merkmalsausprägungen.

Man spricht auch von Variablen, wenn den Merkmalsausprägungen Zahlen zugeordnet werden können. Der Begriff Variable wird also meistens bei quantitativen Merkmalsausprägungen benutzt.

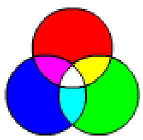
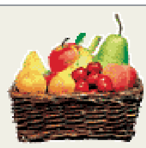
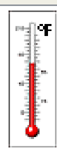
Synonyme

Synonyme: Merkmal, Variable

Synonyme: Merkmalsausprägung, Merkmalswert, Realisation, Beobachtung, Modalität.

Beispiel

Merkmalsausprägungen

Merkmal (untersuchte Größe)	Merkmalsausprägungen (konkreter Wert des Merkmals)	
Farbe	rot, grün, blau, gelb, violett	
Handelsklasse	Extra, I, II, III	
Temperatur (in Grad Fahrenheit)	32° F, 50 °F, 68 °F, 86 °F	

Tab.: Beispiele für Merkmalsausprägungen

Die Merkmalsausprägungen für Farbe und Handelsklassen sind vollständig, für Temperatur nur beispielhaft.

4.3 Merkmalsträger

Merkmale können wir für die unterschiedlichsten Einheiten, wir sagen auch Merkmalsträger, erheben.

Beispiel

Merkmalsträger

Erfassen wir von allen Kommilitonen und Kommilitoninnen Alter, Geschlecht, Körperlänge sowie ihre Zufriedenheit mit dem Studium, so ist die Einheit oder der Merkmalsträger ein Kommilitone bzw. eine Kommilitonin. Die Merkmale sind Alter, Geschlecht usw.

Von Claudia (Merkmalsträgerin) erfahren wir zum Beispiel folgende ihrer Eigenschaften als Merkmale:

Alter	= 29
Geschlecht	= weiblich
Länge	= 173
Zufriedenheit	= 2

Die Angaben im Beispiel können wir erst verstehen, wenn wir erfahren, dass die Ausprägungen des Merkmals Zufriedenheit folgendermaßen verschlüsselt (kodiert) wurden:

1 = sehr zufrieden

2 = zufrieden

3 = indifferent

4 = unzufrieden

5 = sehr unzufrieden.

Schreiben wir die Werte der Merkmale nebeneinander auf und tragen untereinander die verschiedenen Merkmalsträger ein, erhalten wir eine sogenannte Datenmatrix. In ihr finden wir alle Angaben, die wir erhoben haben.

Datenmatrix

In der folgenden Tabelle (Datenmatrix) sind Daten der Studentinnen und Studenen aus Claudias Semester eingetragen.

Beispiel

Index	Name	Alter	Geschlecht	Länge	Zufriedenheit
1	Claudia	29	w	173	2
2	Paula	27	w	165	3
3	Klaus	19	m	179	1
4	Peter	33	m	176	4
5	...	...	...	...	...

Tab.: Beispiel für Daten von Studierenden

4.4 Schreibweisen

Wie bereits bei den Anmerkungen zu den Definitionen „Merkmale" und „Merkmalsausprägung" erwähnt, werden die Merkmale mit großen lateinischen Buchstaben X, Y, Z und die Merkmalsausprägungen mit kleinen lateinischen Buchstaben x , y , ... bezeichnet (vgl. 4.1 Statistische Merkmale f.).

Während das A für das Merkmal „Alter" allgemein steht, steht das a_1 für eine konkrete Merkmalsausprägung, d. h. für das Alter des 1. Merkmalsträgers.

Beispiel

Schreibweisen

Aus den vorigen Beispielen (vgl. 4.3 Merkmalsträger) kennen Sie Claudia bereits.

Bei dieser Merkmalsträgerin gilt $a_1 = 29$, bei Paula $a_2 = 27$ und bei Klaus $a_3 = 19$.
Das Merkmal „Geschlecht" wird mit B bezeichnet. Für Claudia gilt $b_1 = w$.
Das Merkmal Größe wird mit C bezeichnet. Claudias Größe hat den Wert $c_1 = 173$.

Hinweis

Häufig wird statt des Begriffs „Merkmal" den Begriff „Variable" völlig synonym benutzt. In diesem Kurs verwenden wir den Begriff **Merkmal**.

Ausnahmen bestätigen natürlich die Regel ;-)
(siehe auch Anmerkungen zur Definition „Merkmalsausprägung").

5 Merkmalsarten und Skalen

Nicht alle Merkmale tragen vergleichbare Informationen. Um Merkmale systematisch einordnen zu können, hat man sich in der Statistik auf sogenannte Merkmalsskalen (kurz: Skalen) geeinigt.

Wir kommen noch einmal zurück auf unsere Beispiele Merkmalausprägungen bzw. Merkmalsträger und Datenmatrix.

Beispiel

Merkmalsskalen

Bei den Firmendaten Geschäftsform und Unternehmensgröße liegen qualitative Angaben und bei der Umsatz- und Beschäftigtenzahl quantitative Angaben vor. Die Befragung der Kommilitonen und Kommilitoninnen erfolgt über das Alter und die Größe als quantitative, über das Geschlecht und die Zufriedenheit als qualitative Angaben.

Hinweis

Aufgepasst, auch wenn die Ausprägungen von Zufriedenheit als Zahlen kodiert worden sind, bleiben die Angaben dennoch qualitativ!

Anmerkungen

In der Statistik hat sich weitgehend eine differenzierte Einteilung der Skaleneigenschaften von Merkmalen durchgesetzt. Sie beruht im Wesentlichen auf Rechenoperationen, die mit den Ausprägungen eines Merkmals durchgeführt werden können.
Als Rechenoperationen kommen in Frage:

Operation	Operator	Symbol
Vergleichsoperatoren	gleich	=
	ungleich	≠
	kleiner	<
	größer	>
arithmetische Operationen	Addition	+
	Subtraktion	-
	Multiplikation	x
	Division	/

Tab.: Rechenoperationen

Dadurch ergibt sich die folgende Einteilung von Skalen:

Beispiel

Skalierung

Skalenart	zulässige Operationen	Beispiel für Merkmal	Beispiele für Operationen
Nominalskalierung	=, ≠	Geschlecht	Geschlecht von Claudia ≠ Geschlecht von Peter
Ordinalskalierung	=, ≠ , <, >	Konfektionsgröße	XXL > XL > L > M > S > XS
Intervallskalierung	=, ≠, <, >, +, -	Länge	Länge (Klaus) - Länge (Claudia) < Länge (Peter) - Länge (Paula)
Verhältnisskalierung	=, ≠, <, >, +, - , x , /	Umsatz	Der Umsatz ist um 7 % gegen- über dem Vorjahr gestiegen.

Tab.: Skalenarten

5.1 Merkmalsarten

Entsprechend den Merkmalsausprägungen können qualitative und quantitative Merkmale (eine Art der Klassifizierung) unterschieden werden:

- die Ausprägungen qualitativer Merkmale unterscheiden sich durch ihre Art,
- die Ausprägungen quantitativer Merkmale durch ihre Größe.

Qualitative oder kategoriale Merkmale sind Größen, die endlich viele Ausprägungen besitzen und höchstens ordinalskaliert sind. Von Bedeutung ist dabei, dass Ausprägungen eine Qualität und nicht ein Ausmaß widerspiegeln.

Liegen den Ausprägungen eine Messung zugrunde, so spricht man von quantitativen, kardinalen oder metrischen Merkmalen. Damit sind alle Messungen gemeint, deren Werte Zahlen darstellen, also im herkömmlichen Sinn Ausprägungen von quantitativen Merkmalen bzw. Variablen meinen.
Bei quantitativen Merkmalen bzw. Variablen kann hinsichtlich der Anzahl möglicher Merkmalsausprägungen zwischen diskreten und stetigen Merkmalen unterschieden werden.

Multiple Choice

Übung GST-01

Qualitative und quantitative Merkmale

Unterscheiden Sie zwischen qualitativen und quantitativen Merkmalen.

	quali.	quanti.	Auswertung
Familienstand			
Körpergröße			
Art der Medaille			
Farbe			
Temperatur in Celsius			
Nettomiete			
Anzahl von Äpfeln im Korb			
Prüfergebnis			
Temperatur in Kelvin			
Kinderzahl			

5.2 Diskrete Merkmale

Definition

Diskrete Merkmale

Ein Merkmal wird diskretes Merkmal genannt, wenn es in einem endlichen Intervall nur einzelne von vielen Merkmalswerten annehmen kann.

Anmerkungen

Zumeist stellen die Merkmalsausprägungen bei diskreten Merkmalen positive ganze Zahlen dar, wie 0, 1, 2, 3, 4, ...

Ein diskretes Merkmal kann nur endlich viele oder höchstens abzählbar unendlich viele Ausprägungen besitzen.

Das monatliche Bruttogehalt ist ein diskretes Merkmal.
Grund: bei Währungen gibt es eine kleinste Geldeinheit
(z. B. den Euro-Cent: 0,01 €).

Synonyme

Synonyme: ganzzahliges Merkmal, diskontinuierliches Merkmal.

Diskrete Merkmale

Beispiel

Beispiele für diskrete Merkmale und deren jeweilige Merkmalsausprägungen finden Sie in der folgenden Tabelle.

Beobachtungseinheiten (eine endliche Menge)	**Merkmal** (untersuchte Größe)	**Merkmalsausprägungen** (konkreter Wert des Merkmals)
Angestellte eines Betriebes	Monatliches Nettogehalt	2.500,00 €, 3.000,00 €, 3.500,00 €
Krankenhäuser	Anzahl der Geburten (aufeinanderfolgende Tage)	3, 5, 6, 4
Städte	Anzahl der Einwohner	10.899, 209.877, 158.333, 1.789.654

Tab.: Beispiele für diskrete Merkmale

5.3 Stetige Merkmale

Definition

Stetige Merkmale

Ein Merkmal wird als stetiges Merkmal bezeichnet, wenn es in einem endlichen Intervall jeden beliebigen der theoretisch möglichen Merkmalswerte (unendlich viele) annehmen kann.

Synonym

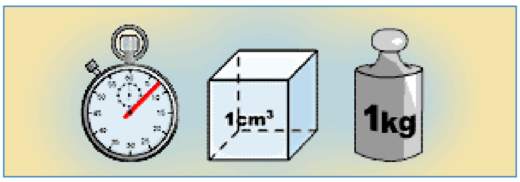
Synonym für stetiges Merkmal: kontinuierliches Merkmal.

Anmerkungen

Im Allgemeinen lassen sich alle Messgrößen als stetige Merkmale auffassen, die

- der Zeit (z. B. Alter, Lebensdauer),
- der Masse (z. B. Gewicht),
- dem Raum (z. B. Länge, Flächeninhalt, Volumeninhalt) oder >
- den Funktionen dieser Größen (z. B. Geschwindigkeit)

zugeordnet sind.



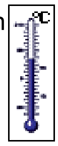
Hinweis

Vorsicht:
Bei stetigen Merkmalen ist jede praktisch durchführbare und durch die jeweilige Grenze der Messgenauigkeit bedingte Messung - diskret.

Beispiel

Messgenauigkeit

Misst man Temperaturen, so werden die Messergebnisse zumeist nur in vollen Gradzahlen oder mit einer Nachkommastelle angegeben. Angenommen, die Messgenauigkeit läge bei 0,1 °C, so wären alle Maßangaben Vielfache von 0,1 °C.



Das Wesen des stetigen Merkmals besteht nicht darin, dass man für jede Beobachtungseinheit genau einen Punkt auf der Zahlengerade angeben kann. Vielmehr besteht es unabhängig von den technischen Möglichkeiten des Messvorganges darin, dass jeder Punkt eines Intervalls von vornherein als Merkmalsausprägung gedacht werden kann.

Beispiel

Stetige Merkmale

Beispiele für stetige Merkmale und deren jeweilige Merkmalsausprägungen finden Sie in der folgenden Tabelle.

Beobachtungseinheit	**Merkmal** (untersuchte Größe)	**Beispiele Merkmalsausprägungen**	
In einer Stadt lebende Personen	Alter	7 Tage, 10 Monate, 20 Jahre, 25 Jahre	
Fußballstadien	Breite der Fußballtore	7,323 m, 7,321 m, 7,318 m, 7,316 m	
Behälter	Volumen	10 l, 5 l, 2 l	

Tab.: Beispiele für stetige Merkmale

In der folgenden Übung geht es um den Unterschied zwischen stetigen und diskreten Merkmalen.

Multiple Choice

Übung GST-02

Stetige und diskrete Merkmale

Unterscheiden Sie zwischen stetigen und diskreten Merkmalen.

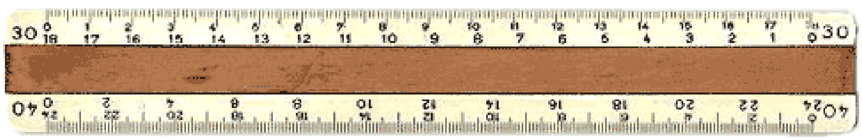
	diskret	stetig	Auswertung
Ton auf dem Klavier			
Körpergröße			
Erfolglose Versuche für eine Prüfung vor dem Bestehen			
Ton auf der Geige			
Temperatur in Celsius			
Nettomiete			
Anzahl von Äpfeln im Korb			
Prüfergebnis			
Temperatur in Kelvin			
Kinderzahl			

5.4 Skalen

Definition

Skala

Eine Skala (*lat.*: scala - Leiter, Treppe) ist sozusagen eine „Messlatte“ für Merkmalsausprägungen eines statistischen Erhebungsmerkmals.



Anmerkungen

Um die Ausprägungen eines Merkmals messen oder erfragen zu können, muss man zunächst eine Skala festlegen, die alle potentiellen Ausprägungen eines Merkmals beinhaltet.

Eine statistische Skala ist eine relationstreue Abbildung von Merkmalsausprägungen eines Erhebungsmerkmals auf eine Zeichen- bzw. Zahlenmenge.

Von der Skala hängt entscheidend die Anwendung statistischer Analyseverfahren ab. Auf der Grundlage der Skala werden die Ausprägungen eines statistischen Merkmals erfasst.

Definition

Skalierung

Die Skalierung ordnet einem Merkmal ein Skalenniveau zu.

5.5 Skalentypen

In der deskriptiven Statistik kommt vor allem den vier bereits genannten Skalentypen (siehe Skalierung) eine besondere praktische und theoretische Bedeutung zu. Diese vier Skalentypen werden hierarchisch (*griech.*: hieros - heilig, arhe - herrschen) folgendermaßen geordnet. Deshalb sprechen wir auch von **Skalenniveaus**. Der Begriff enthält eine Bewertung des Informationsgehalts der Daten.

Tab.: Skalentypen

Skalenniveau		Erklärung
Skala	Nominalskala	Gleichwertigkeit
	Ordinalskala	Anordnung
	Intervallskala	Differenzenbildung
	Verhältnisskala	Quotientenbildung

Definitionen

Kategorialskala

Die Nominal- und Ordinalskala werden häufig zur sogenannten Kategorialskala zusammengefasst.

Die einzelnen in der Tabelle genannten Skalentypen werden Ihnen in den nächsten beiden Abschnitten vorgestellt.

- 6 Qualitative Daten
- 7 Quantitative Daten

6 Skalen für qualitative Daten

Betrachten wir zunächst die Skalen, die für die Darstellung von qualitativen Merkmalen (Daten) von Bedeutung sind. Die Nominal- und die Ordinalskala werden häufig zur sogenannten Kategorialskala zusammengefasst.

In diesem Kapitel werden die beiden Skalen und deren Merkmalsarten vorgestellt.

6.1 Nominalskala

Definition

Nominalskala

Die Nominalskala (*lat.*: nominalis - zum Namen gehörig, begrifflich) stellt die Skala mit dem niedrigsten Niveau dar, weil sie

- den niedrigsten Informationsgehalt und
- die geringste Fehlerempfindlichkeit

aufweist.

Anmerkungen

Die Nominalskala ist eine Skala, mit der lediglich die Gleichartigkeit oder die Verschiedenartigkeit von Merkmalsausprägungen zum Ausdruck gebracht werden kann.

Die Werte einer Nominalskala unterliegen keiner Rangfolge und sind nicht direkt vergleichbar.

6.2 Nominales Merkmal

Definition

Nominales Merkmal

Merkmale, deren Ausprägungen nur mit einer nominalskalierten Skala wiedergegeben werden können, werden als nominale Merkmale oder nominalskalierte Merkmale bezeichnet.

Anmerkung

Ausprägungen von nominalen Merkmalen unterliegen keiner Rangfolge.

Beispiel

Nominale Merkmale

Farbangaben, Familienstand oder Beruf. Man kann nicht sagen, dass Ärztinnen besser sind als Mathematiker und umgekehrt, oder dass ledige Personen schlechter sind als verheiratete und umgekehrt.

Anmerkung

Nominale Merkmale sind in mindestens zwei diskrete Merkmalsausprägungen unterteilt.

Definition

Dichotom

Ein Dichotom (*griech.*: dicha - zweifach, tome - Schnitt) ist ein nominales Merkmal, das nur zwei mögliche Ausprägungen besitzt.

Anmerkungen

Die Ausprägungen eines nominalen Merkmals werden auch als Kategorien oder Attribute bezeichnet. Die Ausprägungen können nicht der Größe nach geordnet werden.

Die Kategorien bzw. Merkmalsausprägungen sind diskret, weil Zwischenstufen nicht zugelassen sind, d. h. eine Ausprägung fällt entweder in die eine oder in die andere Kategorie, liegt aber nicht „zwischen“ zwei Kategorien.

Beispiele für nominale Merkmale und deren Ausprägungen

Beispiel

Merkmal (untersuchte Größe)	**Merkmalsausprägungen** (konkreter Wert des Merkmals)
Familienstand	ledig, verheiratet, geschieden, verwitwet
Geschlecht	männlich, weiblich, divers
Postleitzahl	13353, 10707, 10820
Staatsangehörigkeit	deutsch, französisch, chinesisch

Tab.: Beispiele für nominale Merkmale

6.3 Ordinalskala

Definition

Ordinalskala

Die Ordinalskala (*lat.*: ordinare - ordnen), auch Rangskala genannt, ist eine Skala mit der eine natürliche Rangfolge von Merkmalsausprägungen zum Ausdruck gebracht werden kann. Sie hat ein etwas höheres Niveau als die Nominalskala.

Anmerkung

Die Rangfolge wird in einer Ordinalskala festgehalten. Die Werte der Ordinalskala lassen sich in ihrer Intensität unterscheiden und nach der Stärke dieser Intensität ordnen.

Bei einem auf einer Ordinalskala gemessenen Merkmal, dessen Ausprägungen sich nach der Intensität unterscheiden, ist das Ordnungsprinzip die Stärke bzw. der Grad der Intensität.

Die Ordinalskala findet in der statistischen Analyse wissenschaftlicher Sachverhalte eine breite Anwendung.

Beispiel

Ordinalskala

Leistungsmerkmale sind beispielsweise die Plätze der Finalergebnisse bei einer Olympiade also 1, 2, 3, 4, ... bzw. Gold, Silber, Bronze, 4. Platz, 5. Platz, ...

Qualitätsmerkmale sind beispielsweise die Handelsklassen von Obstsorten - Extra, I, II, III

Mangel der Ordinalskala

Bei einem ordinalskalierten Merkmal lassen sich die Abstände zwischen verschiedenen Ausprägungen – also den Werten der Skala – nicht interpretieren. Man weiß zwar, dass eine Merkmalsausprägung größer als eine andere Ausprägung ist, über das „wie viel größer“ wird jedoch nichts gesagt.

6.4 Ordinales Merkmal

Definition

Ordinales Merkmal

Ein Merkmal heißt ordinales oder ordinalskaliertes Merkmal, wenn seine (meist begrifflichen) Ausprägungen auf einer Ordinalskala erfasst werden können.
Die Merkmalsausprägungen eines ordinalen Merkmals werden auch als Kategorien oder Attribute bezeichnet.

Anmerkungen

Die ordinalen Merkmale sind in mindestens zwei diskrete Kategorien (Merkmalsausprägungen) unterteilt.

Die Merkmalsausprägungen treten in vergleichbaren, diskreten Kategorien auf und lassen sich nach Größe, Stärke oder Intensität anordnen. Zum Zählen kommt zusätzlich ein ordnendes Vergleichen hinzu. Somit wird mehr Information als bei einer Nominalskala verarbeitet.

Die Rangwerte bezeichnen die Platzziffer der Werte in der geordneten Reihenfolge. Rangwerte bilden die Grundlage des Rangkorrelationskoeffizienten von Spearman, der in der empirischen Wirtschaftsforschung eine breite Anwendung erfährt.
Den Rangkorrelationskoeffizienten nach Spearman werden Sie in Lerneinheit RPK kennen lernen.

Beispiel

Ordinale Merkmale, Zensuren

Bei den Zensuren weiß man, dass die Note „sehr gut“ (1) besser als „gut“ (2) ist. Der Abstand zwischen „sehr gut“ und „gut“ (2 - 1 = 1) lässt sich aber nicht sinnvoll interpretieren oder etwa mit demjenigen zwischen „befriedigend“ (3) und „ausreichend“ (4) vergleichen.

Das liegt daran, dass die Noten 1 und 2 willkürlich gewählte Zahlen sind und gewissen Leistungen zugeordnet werden. Man könnte die Leistung „gut“ auch mit 13 und die Leistung „sehr gut“ mit 16 beziffern; der Abstand wäre dann:

„sehr gut“ (16) - „gut“ (13) = 3.

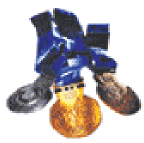
Hinweis

Wir wollen nicht übertreiben:
Es stimmt zwar, dass Zensuren eigentlich ordinale Merkmale sind. Bei der Berechnung von Abschluss- oder Gesamtnoten wird das in der Regel ignoriert. Das ist zwar nicht die reine Lehre, aber auch nicht ganz schlimm.

Beispiel

Ordinale Merkmale, Olympiamedaillen

Aus der olympischen Medaillenvergabe weiß man, dass die Goldmedaille besser ist als Silber und Silber besser als Bronze.



Dabei kann aber durchaus der Unterschied Gold und Silber groß sein, während der Unterschied zwischen Silber und Bronze kleiner ist.

Diese Information geht verloren, wenn nur die Rangplätze zählen.

Olympische Spiele 2020: Schwimmen 400 Meter Lagen der Frauen
(Wikipedia)

Gold: Yui Ohaschi
(Japan) 4:32,08 min

Silber: Emma Weyant
(Vereinigte Staaten) 4:32,76 min

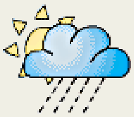
Bronze: Hali Flickinger
(Vereinigte Staaten) 4:34,90 min

Differenz zwischen Gold und Silber: 0,68 s
Differenz zwischen Silber und Bronze: 2,14 s

Ordinale Merkmale und deren Ausprägungen

Beispiele für ordinale Merkmale und deren jeweilige Merkmalsausprägungen finden Sie in der folgenden Tabelle:

Beispiel

Beobachtungseinheiten	**Merkmal** (untersuchte Größe)	**Merkmalsausprägungen** (konkreter Wert des Merkmals)
Gruppe von Studierenden	Prüfungsergebnis in der Statistik-Klausur	sehr gut, gut, befriedigend, ausreichend...
Wetterstationen 	Windstärke (nach Beaufort)	Windstille (0), leiser Zug (1), leichte Brise (2), schwache Brise (3), mäßige Brise (4), frische Brise (5), starker Wind (6), steifer Wind (7), stürmischer Wind (8), Sturm (9), schwerer Sturm (10), orkanartiger Sturm (11), Orkan (12)
Personen	Alter	Kind, Jugendliche, Erwachsene
Warenbestand eines Kaufhauses (Hemden) 	Konfektionsgröße	XS extra small S small M medium L large XL extra large XXL extra extra large

Tab.: Beispiele für ordinale Merkmale

Hinweis

Die hier angegebenen Ausprägungen kommen dadurch zustande, dass eigentlich kontinuierliche Werte in Klassen eingeteilt werden. Das ist natürlich sehr praxisnah. Wir müssen uns aber darüber im Klaren sein, dass jede Klasseneinteilung willkürlich ist, und, das wird meist vergessen, zu einem **Informationsverlust** führt. Mehr dazu finden Sie in den folgenden Lerneinheiten.

7 Skalen für quantitative Daten

Betrachten wir jetzt die Skalen, die für die Darstellung von quantitativen Merkmalen (Daten) von Bedeutung sind. In diesem Kapitel werden diese Skalen und Merkmalsarten vorgestellt. Die Intervall- und die Verhältnisskala werden häufig zur sogenannten Kardinalskala zusammengefasst.

7.1 Intervallskala

Definition

Intervallskala

Die Intervallskala ist eine Skala, die keinen natürlichen Nullpunkt besitzt. Die Intervallskala ist die niedrigstwertige Kardinalskala.

Anmerkungen

Im Vergleich zur Ordinalskala erlaubt die Intervallskala zusätzlich zur Anordnung der Merkmalsausprägungen auch den Vergleich der Abstände zwischen den Ausprägungen.

Die Intervallskala ist in der Regel nicht mehr diskret, sondern kontinuierlich.

7.2 Intervallskaliertes Merkmal

Definition

Intervallskaliertes Merkmal

Ein Merkmal, dessen Ausprägungen auf einer Intervallskala gemessen werden, heißt intervallskaliert.

Beispiel

Intervallskaliertes Merkmal

Als typisches Beispiel für ein intervallskaliertes Merkmal wird häufig die Temperatur in Grad Celsius oder Fahrenheit angeführt.

Anders gesagt, das Merkmal „Temperatur“ mit den Ausprägungen in „Grad Celsius“ (°C) ist intervallskaliert.

Die Temperaturskala wurde nach dem schwedischen Naturforscher Anders Celsius (1701-1744) benannt und geht vom Nullpunkt gleich 0 °C aus, den Celsius „willkürlich“ festgelegt hat.

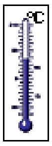
Temperatur-Schwankungen von -1 °C bis +9 °C im Februar und +18 °C bis +28 °C im Mai sind gleich groß, da die Differenz zwischen maximaler und minimaler Temperatur in beiden Monaten 10 °C beträgt.

Eine sinnvolle Aussage lautet: 18 °C sind wärmer als 9 °C.

Keine sinnvolle Aussage ist hingegen: 18 °C sind doppelt so warm wie 9 °C.

Beispiel

Intervallskaliertes Merkmal

Beobachtungseinheiten		**Merkmal** (untersuchte Größe)	**Merkmalsausprägungen** (konkreter Wert des Merkmals)
Wetterstationen in Deutschland		Abendtemperatur (in °C)	0 °C, 10 °C, 20 °C, 30 °C
Gruppe von Kindern		Geburtstag	04.03.2010, 09.03.2011, 01.03.2012

Tab.: Beispiele für intervallskalierte Merkmale

7.3 Verhältnisskala

Definition

Verhältnisskala

Die Verhältnisskala besitzt einen natürlichem Nullpunkt.

Anmerkungen

Werte einer Verhältnisskala sind also immer größer oder gleich Null.
Während die Intervallskala nur den Vergleich von Differenzen gemessener Werte erlaubt, können bei der Verhältnisskala auch die Quotienten (das Verhältnis) gemessener Werte verglichen werden.

Die sinnvolle Berechnung von Quotienten ist möglich, weil Verhältnisskalen einen eindeutig festgelegten und nicht willkürlich gesetzten Nullpunkt haben.
Man darf auch hier nicht durch Null dividieren!

Beispiel

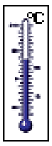
Verhältnisskala

Die in °C gemessene Temperatur entspricht nicht den Anforderungen einer Verhältnisskala, denn bei einer Temperatur von +18 °C ist nicht die doppelte Wärmemenge wie bei +9 °C vorhanden.

Bei der Festlegung der Celsiusskala wurde von Anders Celsius der Gefrierpunkt des Wassers willkürlich zum Nullpunkt der Skala erklärt.

Bei Kelvin ist dagegen der niedrigste theoretisch erreichbare Temperaturzustand als Nullpunkt bestimmt worden. -273,15 °C entsprechen 0 K.

Hier wurde also kein willkürlicher Nullpunkt gewählt, sondern der absolute Temperatur-Nullpunkt.
Kälter wird's nicht!



Die Kelvinskala ist somit eine Verhältnisskala und es ist physikalisch sinnvoll, eine Temperatur von 18 K als doppelt so warm wie 9 K zu bezeichnen.
Und das ist übrigens beides ziemlich kalt!

7.4 Verhältnisskaliertes Merkmal

Definition

Verhältnisskaliertes Merkmal

Ein Merkmal, das auf einer Verhältnisskala gemessen wurde, heißt verhältnisskaliert.

Beispiel

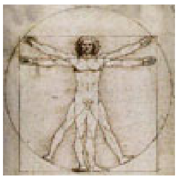
Verhältnisskaliertes Merkmal, physikalische Einheiten

Physikalische Größen wie Längen in Metern (m), Gewichte in Kilogramm (kg), Zeit in Sekunden (s) sind Beispiele für verhältnisskalierte stetige Merkmale.

Beispiel

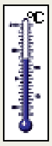
Verhältnisskaliertes Merkmal, Länge

Wir verwenden häufig Verhältnisse. So charakterisiert z. B. das Verhältnis von Länge und Breite typische Formen von Rechtecken. Berühmt ist das Verhältnis des Goldenen Schnittes, das als besonders ausgewogen gilt.



Beispiel

Verhältnisskalierte und stetige Merkmale

Beobachtungseinheiten	Merkmal	Merkmalsausprägungen
Personen einer Stadt 	Alter	7 Tage, 10 Monate, 20 Jahre, 25 Jahre
Personen eines Hauses	Körpergewicht	30 kg, 62 kg 300 g, 77 kg 850 g
Kinder einer Schulklasse	Körpergröße	1,32 m, 1,35 m, 1,39 m
Kraftwerkskessel 	Temperatur	273 K, 283 K, 293 K, 303 K

Tab.: Beispiele für verhältnisskalierte und stetige Merkmale

Wir haben Ihnen verschiedenen Skalen und deren Merkmale vorgestellt. Versuchen Sie in der folgenden Übung die Merkmale zuzuordnen.

Multiple Choice

Übung GST-03

Beispiele für unterschiedliche Skalierungen

Welchen Skalenniveautypen würden Sie die folgenden Merkmale zuordnen?
Klicken Sie die entsprechenden Felder an.

	Nominal	Ordinal	Intervall	Verhältnis
Familienstand	☐	☐	☐	☐
Körpergröße	☐	☐	☐	☐
Art der Medaille	☐	☐	☐	☐
Farbe	☐	☐	☐	☐
Temperatur in Celsius	☐	☐	☐	☐
Nettomiete	☐	☐	☐	☐
Anzahl von Äpfeln im Korb	☐	☐	☐	☐
Prüfungsergebnis	☐	☐	☐	☐
Temperatur in Kelvin	☐	☐	☐	☐
Kinderzahl	☐	☐	☐	☐

8 Transformationen

Beim Bearbeiten von Daten kann es nützlich oder erforderlich sein, die Werte umzurechnen. Um solche Umformungen geht es - Transformationen sind einfach mathematische Funktionen.

Definition

Transformationen

Eine Transformation g des Merkmals X in das Merkmal Y ist eine Funktion, die jedem Wert von X einen und nur einen Wert von Y zuordnet: Y = g(X).

Anmerkungen

Durch Transformationen kann das Skalenniveau abgesenkt werden.

Durch Transformationen kann das Skalenniveau nicht erhöht werden.

Durch Transformationen kann das Skalenniveau auf gleichem Niveau bleiben.

Auf diese Transformationsmöglichkeiten werden wir im Folgenden detaillierter eingehen.

8.1 Arten von Transformationen

Man kann verschiedene Arten von Transformationen unterscheiden:

- **Ein-eindeutige Transformation:**
  Es gibt für alle Werte von X genau einen zugehörigen Wert von Y und umgekehrt für jeden Wert von Y genau einen Wert von X.
  (Die Funktion g ist umkehrbar eindeutig.)
- **Streng monotone Transformation:**
  Eine Transformation heißt streng monoton steigend, falls aus
  $x_1 < x_2$ stets $g(x_1) < g(x_2)$ folgt.
  Bei monoton fallenden Transformationen dreht sich die Anordnung der transformierten Werte um:
  $x_1 < x_2$ stets $g(x_1) > g(x_2)$.
- **Lineare Transformation:**
  Eine Lineartransformation liegt vor, wenn mit reellen Konstanten a und b gilt:
  $g(X) = a + bX$.
- **Proportionale Transformation:**
  Eine Proportionaltransformation liegt dann vor, wenn man in der Gleichung
  $g(X) = a + bX$ den Wert $a = 0$ setzt. Somit bleibt der Nullpunkt bei einer proportionalen Transformation stets erhalten.

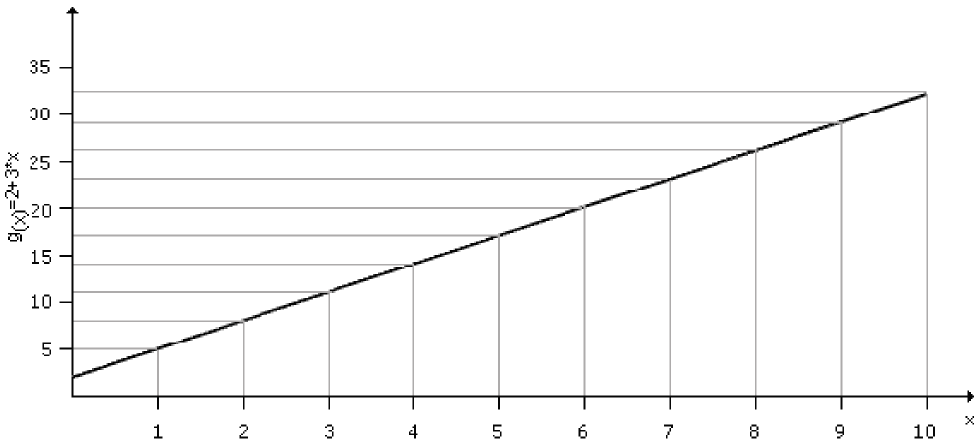


Abb.:
Eine lineare Transformation
$g(x) = 2 + 3x$

Transformationen werden aus verschiedenen Gründen verwendet. Wie bei der Umrechnung von Währungen bei der Urlaubsreise sind die transformierten Werte oft sinnvoller interpretierbar, manchmal gibt es auch innerstatistische Gründe für Transformationen, die uns an dieser Stelle aber noch nicht weiter interessieren sollen.

8.2 Transformationen und Skalen

Die vier genannten Transformationen (ein-eindeutige, streng monotone, lineare und proportionale) entsprechen genau den vier genannten Skalen: Nominal-, Ordinal-, Intervall- und Verhältnisskala.

Bei den gewählten Transformationen bleibt der eigentliche Informationsgehalt erhalten. So ändern sich Farben nicht, wenn wir für „Rot" das französische „rouge" verwenden. In USA schwitzen wir bei Fahrenheit, zuhause bei gleicher Hitze in Celsius.

Tab.: Transformationen und Skalen

Transformation	Skala
ein-eindeutige Transformation	Nominalskala
streng monotone Transformation	Ordinalskala
lineare Transformation	Intervallskala
proportionale Transformation	Verhältnisskala

Noch einmal: Die Skalierung eines Merkmals ändert sich nicht, wenn man bei:

- nominalen Merkmalen eine ein-eindeutige Transformation,
- ordinalen Merkmalen eine monotone Transformation,
- intervallskalierten Merkmalen eine lineare Transformation und
- verhältnisskalierten Merkmalen eine proportionale Transformation

durchführt.

Regeln

Einfache Transformationsregeln

1. Abwärts geht die Transformation **immer**.
2. Aufwärts geht die Transformation **nie**.
3. Lineare Transformation: $Y = a + b X$
4. Proportionale Transformation: $Y = k X$
   (z. B. Umrechnung der Währungen $ in €)
5. Einfach-logarithmische Transformation: $Y = \ln(X)$

Die hier kurz vorgestellten Transformationsregeln werden im Folgenden mit Beispielen erläutert.

8.3 Abwärts immer

Regel

Abwärts-Transformation

Abwärts geht die Transformation immer.

Beispiel

Abwärts-Transformation, Briefporto

Das Merkmal Länge oder Gewicht kann, aber muss nicht auf dem Verhältnisskalen-Niveau gemessen werden. Für manche Fälle genügt aber eine Nominalskala, so z. B., wenn sich die Post bei Standardbriefen für die Fragen interessiert:

- Ist die Dicke des Briefes über 0,5 cm (Porto-Zuschlag) oder darunter (normales Porto)?

- Ist das Gewicht des Briefes über 20 g (Porto-Zuschlag) oder darunter (normales Porto)?

In diesem Fall wird die Längenmessung oder die Gewichtsmessung durch eine Nominalskala mit zwei Kategorien ersetzt.

Beispiel

Abwärts-Transformation, Sekt

Einen Preisvergleich für zwei vergleichbare Artikel anzustellen ist sinnvoll. Z. B. sind die Sektflaschen im Supermarkt so ausgepreist:

- $x_1 =$ 6,13 € (Schneewittchen) und
- $x_2 =$ 10,22 € (Cremant).

Zum Vergleich gibt es viele Möglichkeiten. Diese zeigen Ihnen das verwendete Skalenniveau:

- Verhältnisskala:
  Der Preis des Cremant macht das 1,667-fache des Preises des deutschen Sektes aus. Der dimensionslose Quotient $\frac{x_2}{x_1}$= 1,667 wird als Preismesszahl bezeichnet.
- Intervallskala:
  Der Preis des Cremants liegt um 4,09 € über dem Preis des deutschen Sektes. Preisdifferenz: $x_2 - x_1$.
- Ordinalskala:
  Crement ist teurer als deutscher Sekt ($x_2 > x_1$).
- Nominalskala:
  Der Preis des deutschen Sektes unterscheidet sich vom Preis des Cremants Sektes ($x_1 \neq x_2$).

8.4 Aufwärts nie

Regel

Aufwärts-Transformation

Aufwärts geht die Transformation nie.

Beispiel

Aufwärts-Transformation

Die Merkmale wie „Familienstand“, „Geschlecht“ oder „Postleitzahl“ sind nur nominalskaliert und lassen sich auf keinem höheren Skalenniveau messen.
Das Addieren oder Subtrahieren von Postleitzahlen ist also nicht sinnvoll, wie das folgende Beispiel zeigt:

München-Flughafen (PLZ 85356),
Berlin-Kreuzberg (PLZ 10967)

85356 - 10967 = 74389 (Cleebronn)
Alles klar??

8.5 Lineare Transformation

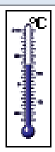
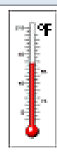
Regel

Lineare Transformationen erfolgen über die Umrechnung nach der Formel $y = a + bx$.

Beispiel

Umrechnung von Celsius in Fahrenheit

Umrechnung einer Temperaturmessung von „Grad Celsius“ (x_i) in „Grad Fahrenheit“ (y_i). Diese Umrechnung erfolgt nach der Formel $y_i = 32 + 1{,}8\ x_i$.
Wir erhalten:

	°C	°F	
	0	32	
	10	50	
	20	68	
	30	86	
	40	104	

Tab.: Umrechnung von Grad Celsius in Grad Fahrenheit

Die Abstände zwischen 10 °C und 20 °C usw. sind jeweils 10 °C und die entsprechenden Abstände zwischen 50 °F und 68 °F bzw. 68 °F und 86 °F sind ebenfalls gleich. Sie betragen 18 °F (wobei $18 = 1{,}8 \times 10 = 1{,}8\ x_i$).

Die Größe a bewirkt eine Verschiebung des Nullpunkts, der bei °C und °F willkürlich festgelegt wurde.

Bei K („Grad Kelvin“, absolute Temperaturmessung) dagegen liegt der absolute Nullpunkt unveränderlich bei -273,15 °C.

Die Größe b bewirkt eine Streckung (bei $b > 1$) bzw. eine Stauchung (bei $b < 1$) des Maßstabs.

Die Intervallskala bei °C wird durch die Transformation in °F nicht zu einer Skala anderen Typs.

8.6 Proportionale Transformation

Die Umrechnung von Währungseinheiten erfolgt proportional, d. h. mit Hilfe eines (Umrechnungs-) Faktors, etwa von Britischen Pfund in Euro.

Beispiel

Proportionale Transformation

Tab.: Proportionale Transformation von Euro in verschiedene Währungen

Faktor (X in €)	Faktor (€ in X)	100000 € entsprechen
1,14112	0,87616	87.616 Pfund (GBP)
1,0001	0,99974	99.974 Dollar (USD)
1,03655	0,96453	96.453 Franken (CHF)
0,21159	4,72047	472.047 Zloty (PLN)
19.152,7	0,00005	5,20544 Bitcoin (BTC)

Währungskurse Stand: 21.09.2022

8.7 Einfach-logarithmische Transformation

Bei vielen Merkmalen wird anstatt der physikalischen Originalwerte der Logarithmus `Y = ln (X)`, verwendet. Beispiele, wo von vornherein Logarithmen verwendet werden, sind der PH-Wert, der Schalldruckpegel oder die Stärke von Erdbeben auf der Richterskala.

Wenn gleichzeitig sehr große und sehr kleine Werte auftreten, kann die Log-Transformation u. a. für die grafische Darstellung der Daten sehr nützlich sein. Dann können die eher dicht beieinander liegenden kleinen Werte besser auseinandergehalten werden und große Werte werden herangezogen und liegen nicht mehr so weit entfernt.

Hinweis

Beachte

log = dekadischer Logarithmus, Logarithmus zur Basis 10 (z. B. log 10^3 = 3)

`log(x)` ist nur für `x > 0` definiert, d. h. die Daten müssen positiv sein, um `log` anwenden zu können. Der Einfachheit halber nehmen wir den dekadischen Logarithmus für unser Beispiel.

`log(x) < 0 für x < 1` (Werte kleiner als 1 werden gespreizt.)
`log10(0.1)= -1,`
`log10(0.01)= -2,`
`log10(0.001)= -3,` …

Die Logarithmen der Datenpunkte 0,1, 0,01, 0,001 scheinen symmetrisch zu streuen.

`log(x) > 0 für x > 1` (Werte größer als 1 werden zusammengezogen.)
`log10(1) = 0,`
`log10(10) = 1,`
`log10(100) = 2`

Die Logarithmen der Datenpunkte 1, 10, 100 scheinen symmetrisch um 1 zu streuen.

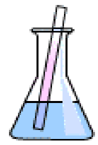
Beispiel

Einfach-logarithmische Transformation, pH-Wert

Der pH-Wert charakterisiert die Wasserstoffionenkonzentration. Er ist definiert als der negative dekadische Logarithmus der molaren H+-Ionen Konzentration: pH = -log [H+].

Ein pH-Wert von 7,4 entspricht einer Wasserstoffionenkonzentration von

[H+] = 10^{-pH} = $10^{-7,4}$ = 39 nmol/L.



Originalwerte	Transformierte Werte
0,01 mol/l	2 pH
0,00001 mol/l	5 pH
0,000001 mol/l	6 pH

Beispiel

Einfach-logarithmische Transformation, Schalldruckpegel

Verschiedene Schallquellen werden häufig als Schalldruckpegel (L in dB) dargestellt:

L = 20 x log (p/p0).

Der Schalldruckpegel L ist eine dimensionslose Zahl und wird in Dezibel (dB) angegeben. Zur Bestimmung des Schalldruckpegels einer bestimmten Schallquelle wird der aktuelle Schalldruck p der Schallquelle zu dem Schalldruck p_o in Beziehung gesetzt, der etwas unterhalb der Hörschwelle liegt (p_o = 2 x 10^{-5} $\frac{N}{m^2}$).

Ein Ton, ein Klang oder ein Geräusch besitzt einen Schalldruck von 2 * 10^{-2} N/m^2. Setzt man diesen Wert in obige Gleichung ein, ergibt sich einen Schalldruckpegel von:

$$L = 20 \cdot \log(\frac{2 \cdot 10^{-2}}{2 \cdot 10^{-5}}) = 20 \cdot \log 10^3$$

Mit log 10^3 = 3 folgt: L= 60 dB.

Typische Werte einiger Geräuschquellen

Beispiel

Schallquelle	dB (ca.)	Schallquelle	dB (ca.)
Laub im Wind	10	durchschnittliches Straßengeräusch	80
Schlafzimmer	15	laute Werkstatt	> 85
öffentliche Bibliothek	38	Presslufthammer	100
Wohnzimmer	40	Rockband	> 110
Gespräch	58	Silvesterkracher	125
Büroraum	> 65	Startender Düsenjet	140

Tab.: Schallquellen und ihre dB-Werte

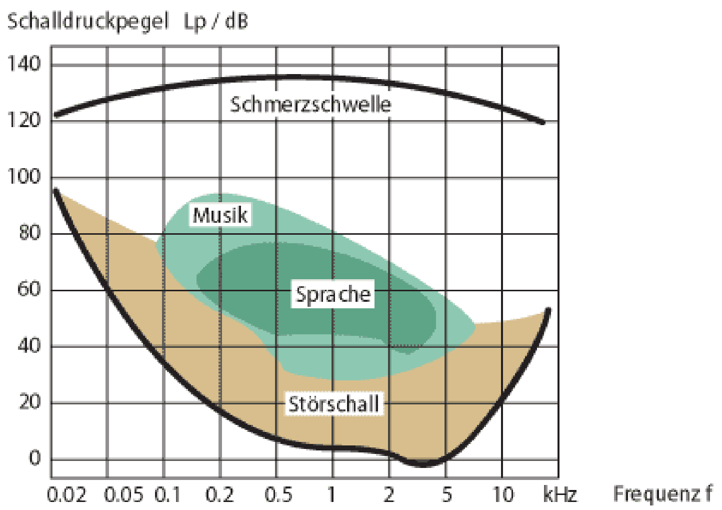


Abb.: Schalldruckpegel und Frequenz

Quelle: www.meta.ch/de/labors/9/Schalldruck.pdf bzw. 91.html, Stand: 29.01.2002

In dem folgenden YouTube-Video wird noch einmal am Beispiel der Richter-Skala gezeigt, wie Logarithmen zur Basis 10 bestimmt werden können.

Filmbeispiel

Dekadischer Logarithmus (in knapp 5 Minuten)

Video: „Dekadischer Logarithmus“

 Datenschutzhinweis
Quelle: https://youtu.be/66p5ZTrXLKw

Neben der reinen Systematik des Informationsgehalts der Merkmalsausprägungen hat die Einteilung in Skalentypen von Merkmalen vor allem die Aufgabe, eine sinnvolle Auswahl von statistischen Beschreibungen der Werte eines Merkmals zu ermöglichen.

Wie das geschieht, erfahren Sie in den nächsten Lerneinheiten.

Zusammenfassung

- Die Gewinnung von Daten setzt die Erfassung von Objekten, den statistischen Einheiten, voraus, an denen die interessierenden Merkmale beobachtet werden können.
- In der Regel werden wir dabei nicht alle möglichen, für die Fragestellung relevanten statistischen Einheiten erheben, also keine Vollerhebung durchführen, sondern uns auf eine Teilerhebung beschränken.
- Für die Qualität eines Datensatzes ist neben dem Ziehungsverfahren (und dem Studiendesign) entscheidend, in welcher Weise die Daten gemessen wurden.
- In der Statistik wird zwischen qualitativen und quantitativen, bei quantitativen auch zwischen diskreten und stetigen Merkmalen unterschieden.
- Die qualitativen Merkmale können nominal oder ordinal skaliert sein.
- Die quantitativen Merkmale können intervall- oder verhältnisskaliert sein.
- Die Skalentypen lassen sich leicht, durch für die Daten zulässigen Rechenoperationen, charakterisieren.
- Transformationen sind mathematische Funktionen.
- Es können verschiedene Arten von Transformationen unterschieden werden: Ein-eindeutige, streng monotone, lineare und proportionale.
- Bei vielen Merkmalen wird anstatt der physikalischen Originalwerte der Logarithmus verwendet.

Sie sind am Ende dieser Lerneinheit angelangt. Auf den folgenden Seiten finden Sie noch die Übungen für R.

Übungen mit der Statistiksoftware R

Die Übungen zu dieser Lerneinheit enthalten noch keine Rechenaufgaben. Sie sollten sich jedoch jetzt mit der Benutzungsoberfläche und den Elementen der Statistiksoftware **R** vertraut machen.

Außerdem werden Sie sehen, wie Sie mit unseren Musterlösungen umgehen können. Sie können die R-Datei aus der Lösung auf Ihrem Rechner speichern und direkt mit **R** öffnen oder den Code kopieren.

Um die Übungen zu bearbeiten, muss die Software „**R**“ auf Ihrem Rechner installiert sein.

Installationshinweise [Manuals | **R** Installation and Administration]

Statistiksoftware R

Übung GST04a

Data Frame für Kursliste erstellen

Sie möchten eine Kursliste der Studierenden eines Kurses zu erstellen. Es sollen folgende Daten erfasst werden.

Vorname	Nachname	Alter	Größe (cm)	Gewicht (kg)

Aufgabe

Erstellen Sie in R einen Data Frame mit Ihren Ergebnissen.

Lösung mit R (Siehe Anhang)

Bearbeitungszeit: 10 Minuten

Statistiksoftware R

Übung GST04b

Unfallklinik

Im folgenden Datensatz ist ein Auszug einer Krankenhausstatistik zu sehen. Erfasst wurde von den Patienten, die wegen eines Verkehrsunfalls eingeliefert wurden, das Alter, das Geschlecht, der Grad der Verletzung, die Aufenthaltsdauer in Tagen und die Art des Fahrzeuges.

i	Alter	Geschlecht	Verletzungsgrad	Verweildauer	Verkehrsmittel
1	35	m	leicht	20	Fahrrad
2	23	m	leicht	11	zu Fuss
3	40	w	mittel	28	zu Fuss
4	29	m	schwer	23	Motorrad
5	24	w	mittel	7	Auto
6	18	w	schwer	15	Auto
7	28	w	leicht	18	Auto
8	20	w	mittel	8	Fahrrad

Aufgabe

Stellen Sie die Häufgkeitsverteilung der an den Unfällen beteiligten Verkehrsmittel als Säulendiagramm dar. (Diese Aufgabe ist ein Vorgriff auf die Lerneinheit 03 DHV)

Lösung mit R (Siehe Anhang)

Bearbeitungszeit: 10 Minuten

Appendix

Lösung für Übung GST-04a

Data Frame für Kursliste erstellen

Lösung mit R

kurs_loesung.R

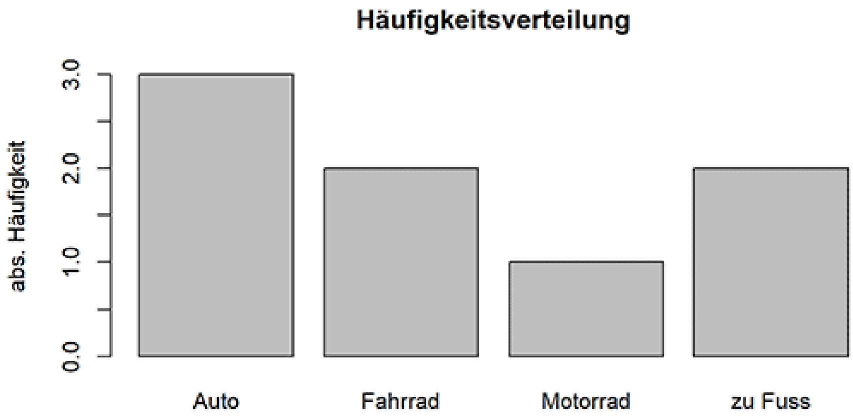
```r
# Erstellen des Data Frames mit Beispieldaten.

kursliste <- data.frame(
  Vorname = c("Max", "Maxi"),
  Nachname = c("Mustermann", "Musterfrau"),
  Alter = c(20, 24),
  Groesse_in_m = c(1.83, 1.68),
  Gewicht_in_kg = c(75, 63)
)
```

Lösung für Übung GST-04b

Unfallklinik

Häufgkeitsverteilung der an den Unfällen beteiligten Verkehrsmittel als Säulendiagramm.



Lösung mit R

unfallklinik_loesung.R

```r
#  Einlesen der Daten in einen Data Frame

patienten <- data.frame(
  Alter = c(35, 23, 40, 29, 24, 18, 28, 20),
  Geschlecht = c("m", "m", "w", "m", "w", "w", "w", "w"),
  Verletzungsgrad = c(
    "leicht",
    "leicht",
    "mittel",
    "schwer",
    "mittel",
    "schwer",
    "leicht",
    "mittel"
  ),
  Verweildauer = c(20, 11, 28, 23, 7, 15, 18, 8),
  Verkehrsmittel = c(
    "Fahrrad",
    "zu Fuss",
    "zu Fuss",
    "Motorrad",
    "Auto",
    "Auto",
    "Auto",
    "Fahrrad"
  )
)

#  Erzeugen des Säulendiagramms mit den Häufigkeiten des Merkmals Verkehrsmittel

barplot(table(patienten$Verkehrsmittel),
        main="Häufigkeitsverteilung",
        ylab="abs. Häufigkeit")
```