

WIRTSCHAFTSSTATISTIK

EINLEITUNG

WS 2021/22

DR. E. MERINS

FACH „WIRTSCHAFTSSTATISTIK“

- **Pflichtfach**

- Teilnahmevoraussetzungen = KEINE

- **Aus welchem Bereich kommt „Statistik“?**

- Teilgebiet der Mathematik

- **Niveau**

- Grundkenntnisse der beschreibenden Statistik

- Hochschule ≠ Schule: der Stoff wird schneller vermittelt

- **Lernziele:**

- Verstehen – Wissen – Anwenden

Daten erheben, aufbereiten, verdichten und anhand deskriptiver und induktiver statischer Methoden beurteilen;

Zusammenhänge erkennen;

Daten und Zusammenhänge anhand von Tabellen und Grafen darstellen, analysieren und interpretieren.

LERNFORM

- **Online-Studiengang** → selbständiges Lernen
- **zeitlich parallelaufende Online-Betreuung (E-Mail, Webkonferenz) sowie Präsenzphasen**
- **Vor- und Nachbereitung unbedingt notwendig**
- **kontinuierliche Arbeit unbedingt notwendig** → nicht erst kurz vor der Klausur mit dem Lernen beginnen!
- **kontinuierliches Üben zielführend (bzgl. Klausurerfolg)**

UNTERSTÜTZUNG BEIM LERNEN ...

- **Ca. jede vierte Woche ein Webseminar 90 Minuten (bitte Termine beachten!) → Vorstellung neuer Themen mit Rechenbeispielen**
 - Zum Ablauf: 1. Präsentation vom Betreuer, 2. Fragen per Chat / Mail von Teilnehmern
 - Dokumentation (Folien) und Aufzeichnungen vom Webseminar sowie Übungsaufgaben erscheinen auf **moodle.oncampus.de** direkt nach dem Webseminar
 - Lösungen zu den Übungsaufgaben erscheinen auf **moodle.oncampus.de** zeitlich versetzt
- **Nach Bedarf weitere Webseminare** (Termine werden separat mitgeteilt)
→ Beantwortung vorher zugesandten Fragen
- **Online-Studienmodul mit...**
 - Lernmaterial zum Selbststudium (s. auch Kursmaterialien)
 - Literaturliste (dem Modulhandbuch zu entnehmen)
- **Tipp:** „Formelsammlung“ → kann selbstständig erstellt werden, hilft beim Üben

PRÄSENZVERANSTALTUNGEN

- **Präsenzinhalte:**
 - Kennenlernen
 - Klärung inhaltlicher Fragen
 - gemeinsame Bearbeitung von Übungsaufgaben
 - Prüfungsvorbereitung
- **Anwesenheit**
 - Anwesenheit in den Präsenzveranstaltungen sowie Teilnahme an den Webseminaren ist nicht verpflichtend und ist keine Voraussetzung für die Prüfung

PRÜFUNG

■ Prüfungsform:

- Schriftliche Klausur 120 Minuten in Präsenz
- Termine werden separat mitgeteilt

■ Anwesenheit

- Anwesenheit in den Präsenzveranstaltungen sowie Teilnahme an den Webseminaren ist nicht verpflichtend und ist keine Voraussetzung für die Prüfung

■ Zulässige Hilfsmittel

- Zwei DIN A4 beschriebene Blätter (vier Seiten), selbstbeschrieben oder maschinell.

Tipp: bitte lesbare Schriftgröße wählen! Weitere Hilfsmittel, wie z.B. Lupe, sind nicht erlaubt

- Taschenrechner

■ Leistungsbewertung entsprechend der Skala (s. Datei in dem Ordner 4_Klausur)

THEMEN

■ **Modul 1: Einführung in die Statistik**

- Etwas über die Statistik: Definitionen und Grundbegriffe
- Statistik in Tabellen und Grafiken
- Stichprobe
- Ablauf empirischer Untersuchungen und Operationalisierung
- Datenanalyse mit Statistik-Software

■ **Modul 2: Skalen und Klassierung**

- Merkmale und Skalenniveau
- Klassierung

THEMEN

■ **Modul 3: Häufigkeiten und Häufigkeitsverteilungen**

- Datendokumentation
- absolute und relative Häufigkeiten und Summenhäufigkeiten
- Häufigkeitsverteilungen in Tabellen und Grafiken

■ **Modul 4: Lageparameter**

- Modus, Median, arithmetisches Mittel, Quantile
- Berechnung und Interpretation

■ **Modul 5: Streuungsparameter**

- Quartile, Quartilsabstand, Spannweite, Varianz, Standardabweichung
- Berechnung, Interpretation und grafische Darstellung (Boxplot)

THEMEN

■ **Modul 6: Korrelation und Regression**

- Analyse mehrerer Merkmale: Zusammenhangsanalyse (Korrelation) und Abhängigkeitsanalyse (lineare Regression)
- Bestimmtheitsmaß
- Berechnung, Anwendung und Interpretation

■ **Modul 7: Wahrscheinlichkeitsrechnung**

- Definition und Grundbegriffe
- Zufallsexperimente
- Mengenlehre
- Kombinatorik
- Bedingte Wahrscheinlichkeit

Sehr geehrte Studierende,

herzlich Willkommen zum Kurs Wirtschaftsstatistik im WS2021/22.

Ich freue mich, dass nach einem anstrengenden Corona bedingtem WS2020/21 dieser Kurs wieder zum Teil in Präsenz dargeboten werden kann.

Die erste geplante Präsenzveranstaltung am Freitag den 15.10.2021 findet noch im gewohnten online-Format als Webkonferenz in Adobe Connect in der Zeit von 20:00 – 21:00 statt. Wir werden uns kennenlernen, organisatorische Fragen und den Kursablauf besprechen und gleich in die Statistik einsteigen.

Die folgenden Präsenzveranstaltungen werden wir für die Übungen nutzen, das theoretische Teil ist für die Webkonferenzen vorgesehen.

In Moodle im Abschnitt "Kurs" finden Sie vier Ordner, wo die Unterlagen zum Kurs kontinuierlich abgelegt werden.

Für die weiteren Übungsstunden in Präsenz empfehle ich die Präsentationen (PDF) zu jedem Thema dabei zu haben (entweder in Papierform (ausgedrückt) oder digital (auf Ihrem Laptop/Tablett)) oder gleich eine Formelsammlung zu erstellen, damit Sie die Formeln und ggf. die Rechenbeispiele vor den Augen haben. Außerdem benötigen Sie Schreibutensilien, ein Lineal und einen Taschenrechner.

Noch ein kleines Info für die Studierende, die am 15.10.2021 an der Webkonferenz nicht teilnehmen können. Die Teilnahme an den Präsenz- und Webveranstaltungen ist keine Pflicht. Die organisatorischen Hinweise und Termine finden Sie in dem Ordner "1 Plan und Hinweise". Alle Webveranstaltungen werden aufgenommen und gespeichert.

Alle Fragen beantworte ich gerne per E-Mail. Nutzen Sie auch gerne den „Fachforum zum Modul“ im Moodle.

Viele Grüße,

Dr. E. Merins

Plan und Hinweise

Modulhandbuch:

http://www.beuth-hochschule.de/fileadmin/studiengang/modulhandbuch/b-winf-o/2017_Modulhandbuch_zur_studien- und pruefungsordnung_2017.pdf

ggf. neue Version unter moodle

Wirtschaftsstatistik (Economic Statistics) Auszug aus dem Modulhandbuch Wirtschaftsinformatik Bachelor Online (modifiziert)	
Credit Points	5
Pflicht/ Wahlpflicht	Pflicht
Modulverantwortliche(r)	Prof. Dr. Ulrike Grömping, Beuth Hochschule für Technik Berlin Fachbereich II
Teilnahmevoraussetzungen	wünschenswert „Grundlagen der Mathematik“
Lerninhalte	<p>Das Modul vermittelt Grundkenntnisse der beschreibenden Statistik. Im Mittelpunkt stehen die Beschreibung, Erklärung und Beurteilung der Daten anhand deskriptiver und induktiver statischer Methoden.</p> <p>Nach dem erfolgreichen Studium des Moduls sollen die Studierenden in der Lage sein, Fragestellungen der beschreibenden Statistik selbstständig zu erfassen und zu lösen und darüber hinaus sich in anspruchsvolle Anwendungen statistischer Methoden einzuarbeiten.</p>
Lernziele	Verstehen – Wissen – Anwenden Daten erheben, aufbereiten und verdichten; Zusammenhänge erkennen; Daten und Zusammenhänge anhand von Tabellen und Grafen darstellen, analysieren und interpretieren.
Prüfungsvorleistung	-
Medien-/ Lernform	Multimedial aufbereitetes Online-Studienmodul zum Selbststudium mit zeitlich parallel laufender Online-Betreuung (E-Mail, Webkonferenz) sowie Präsenzphasen. E-Mail: elena@merins.de

Arbeitsaufwand	Präsenzphasen ca. 8 LE (4 Mal je 90 Min. → 6 Stunden) Webseminar ca. 8 LE (4 Mal je 90 Min. → 6 Stunden) Selbststudium ca. 140 LE (105 Stunden)
Präsenzphasen	<i>Bitte Termine beachten!</i> Präsenzinhalte: Kennenlernen, Klärung inhaltlicher Fragen, gemeinsame Bearbeitung von Übungsaufgaben, Prüfungsvorbereitung
Webseminare	<i>Bitte Termine beachten!</i> Webseminarinhalte: Vorstellung neuer Themen anhand Power-Point-Folien, ggf. Beantwortung der Fragen
Anwesenheit	Die Anwesenheit in den Präsenzveranstaltungen in der Beuth-Hochschule sowie Teilnahme an den Webseminaren ist <u>nicht</u> verpflichtend und ist <u>keine</u> Voraussetzung für die Prüfung.
Prüfungsform	Klausur (schriftlich) 120 Minuten (erfordert physische Anwesenheit) Termine werden separat mitgeteilt
Literatur	dem Modulhandbuch zu entnehmen (Link s. oben)

WIRTSCHAFTSSTATISTIK

MODUL 1: EINFÜHRUNG IN DIE STATISTIK

WS 2021/22

DR. E. MERINS

ZUR GESCHICHTE DER STATISTIK

- Die „praktische Statistik“ ist 4.000 – 5.000 Jahre alt
- Der Ursprung ist nicht die Mathematik. Die Mathematik kam erst vor rund 300 Jahren dazu über die Wahrscheinlichkeitsrechnung
- Ausgangspunkt der Statistik: (Staats-)Verwaltung/Management von großen Projekten
- Das Wort Statistik stammt von lateinisch *statisticum* „den Staat betreffend“ und italienisch *statista* Staatsmann oder Politiker, was wiederum aus dem griechischen στατιζω (einordnen) kommt
- Die deutsche Statistik, eingeführt von Gottfried Achenwall 1749, bezeichnete ursprünglich die „Lehre von den Daten über den Staat“. Im 19. Jahrhundert hatte der Schotte John Sinclair das Wort erstmals in seiner heutigen Bedeutung des allgemeinen Sammelns und Auswertens von Daten benutzt.

WO BRAUCHT MAN STATISTIK?

In welchen Gebieten (Wissenschaften) braucht man Statistik?

- **In den empirischen Wissenschaften (auch Realwissenschaften bzw. Erfahrungswissenschaften genannt)**
 - Naturwissenschaften
 - Sozialwissenschaften
 - Biologie / Medizin (Biometrie)
 - Ingenieurwissenschaften (Technometrie)
 - Verhaltenswissenschaften (Psychometrie)

→ man will neue Erkenntnisse gewinnen über einen Ausschnitt der Realität. Dazu werden empirische Untersuchungen durchgeführt; hierbei fallen Daten an, die mit statistischen Methoden ausgewertet werden
- **In allen Bereichen, in denen große Datenmengen anfallen, aus denen man Erkenntnisse gewinnen will**

WAS IST EINE „STATISTIK“?

- **Was ist eine „Statistik“?**

- eine systematische Zusammenstellung von Zahlen und Daten

- **Wozu?**

- zur Beschreibung bestimmter Zustände, Entwicklungen und Phänomene

- **Ziel:**

- Gewinnung von Information aus unübersichtlichen und/oder unstrukturierten und/oder großen Datenmengen



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

GEGENSTAND DER STATISTIK

■ **Datengewinnung**

Es gibt verschiedene Möglichkeiten, wie man Daten erhalten kann. Für die Wirtschaftsstatistik werden neben amtlichen Erhebungen vor allem Berichte, Umfragen und betriebliche Quellen verwendet

- **Datenerhebung** = jede systematische Datengewinnung → Vorgang zur Ermittlung und zur Erfassung von Ausprägungen eines statistischen Merkmals
- **Primärerhebung** → Erhebung neuer Daten nach Vorgaben
- **Sekundärerhebung** → aus bereits vorhandenem Datenmaterial
- **Vollerhebung** → Untersuchung aller statistischen Einheiten einer Gesamtheit
- **Teilerhebung** → $n < N$

GEGENSTAND DER STATISTIK

- **Datenanalysen**

Anwendung statistischer Verfahren zum Zweck der Erkenntnisgewinn

- **Datencharakterisierung**

Beschreibung, Visualisierung, Kennzahlen: 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

GEGENSTAND DER STATISTIK

■ Datenbeurteilung

Die Beurteilung von Daten erfolgt durch:

- Schlüsse auf der Basis unvollständiger Daten, z. B. Schlüsse von der **Stichprobe** auf ihre **Grundgesamtheit**
- Allgemeiner: auf der Basis unsicherer Daten, unter Anwendung der **Wahrscheinlichkeitsrechnung**. Dies ist Gegenstand der induktiven (schließenden) Statistik.

GEGENSTAND DER STATISTIK

- **Datenaufbereitung**

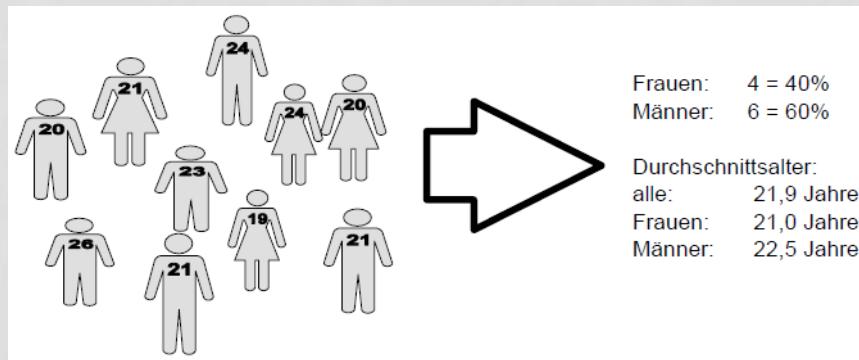
Ordnung, Zusammenfassung und Darstellung des erhobenen statistischen Datenmaterials in Datendateien, Tabellen und/oder geeigneten Grafiken.

- **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

BEREICHE DER STATISTIK

■ **deskriptive oder beschreibende Statistik**



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

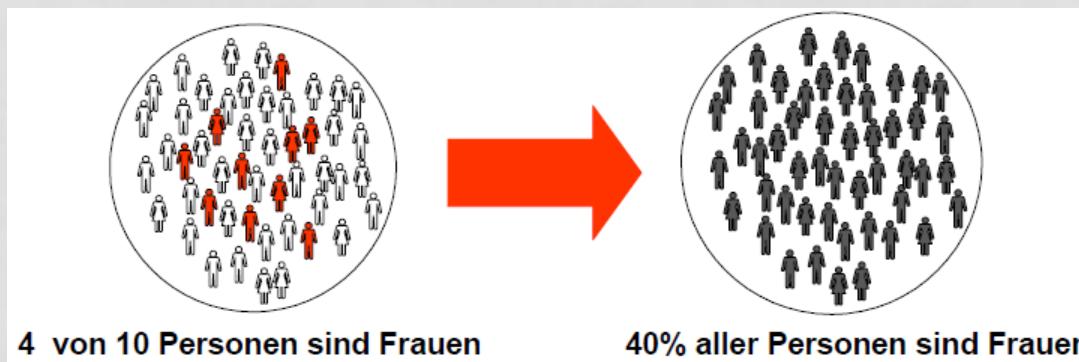
Für eine übersichtliche Darstellung muss das, oft sehr umfangreiche, Material auf geeignete Art und Weise zusammengefasst werden.

Dazu werden insbesondere die drei Darstellungsformen benutzt:

- **Tabellen**
- **grafische Darstellungen**
- **charakteristische Maßzahlen**

BEREICHE DER STATISTIK

- **induktive oder schließende Statistik → Der Schluss vom Teil aufs Ganze**



Probleme der Stichprobe:
Stichprobenfehler
“Repräsentativität”

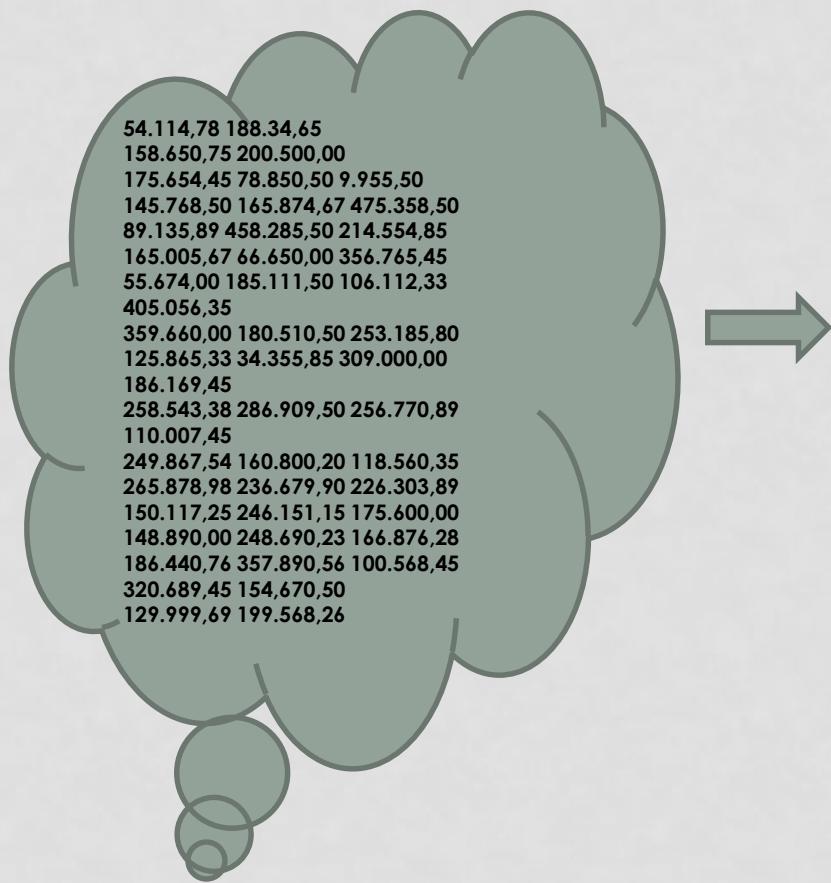
Die induktive Statistik (lat.: *inductio* - Hineinführen) dient dazu, aus den erhobenen Fakten Schlüsse auf die Ursachenkomplexe zu ziehen, die zu diesen Daten geführt haben. Die induktive Statistik basiert auf der Wahrscheinlichkeitstheorie.

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“).

Weitere Synonyme für induktive Statistik: analytische oder inferentielle Statistik. 10

STATISTIK IN TABELLEN UND GRAFIKEN

INPUT



OUTPUT

Umsätze der Meyer AG über die Großhändler
in NRW im Jahr 2008

Umsatzklasse in Tsd. €	Anzahl Großhändler (absolute Häufigkeit)	Anteil Großhändler von Gesamt in % (relative Häufigkeit)
0 bis unter 100	7	14%
100 bis unter 200	23	46%
200 bis unter 300	12	24%
300 bis unter 400	5	10%
400 bis unter 500	3	6%
Summe	50	100%

(Quelle: Umsatzstatistiken der Vertriebsabteilung, 2008)
Tabelle 1

STATISTIK IN TABELLEN UND GRAFIKEN

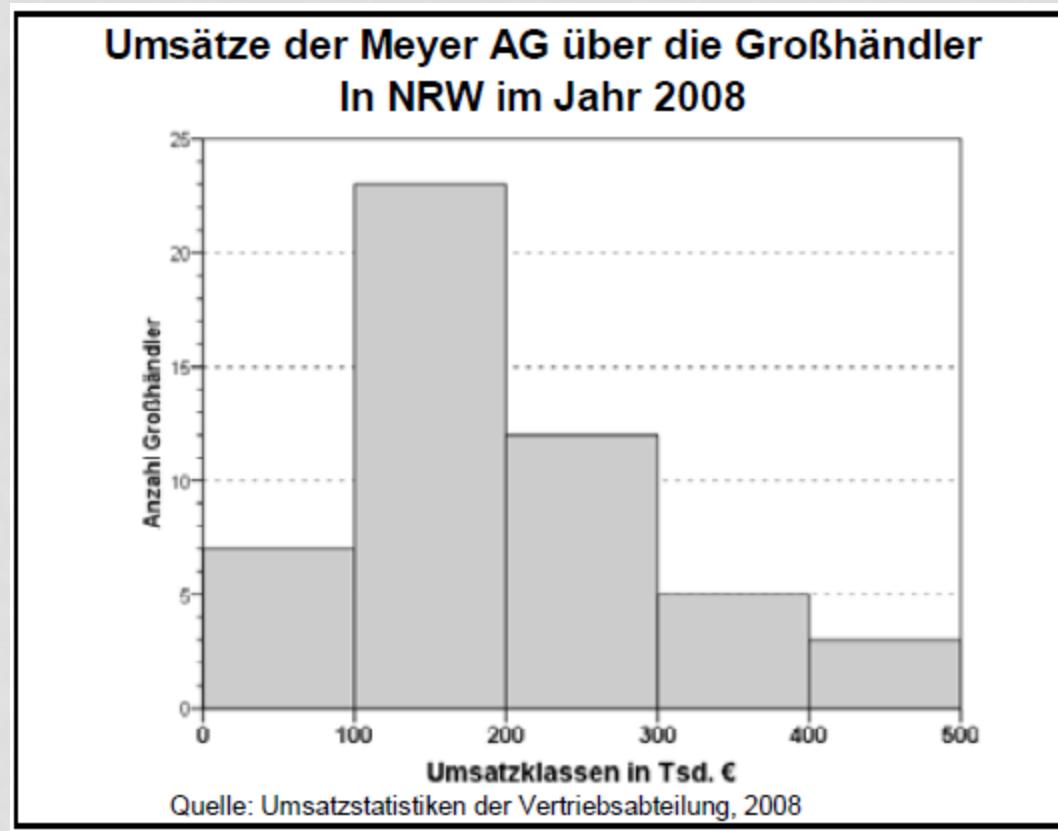


Abbildung 1

TABELLEN VS. GRAFIKEN

Vor- und Nachteil einer „Statistik“ in tabellarischer Darstellung und einer „Statistik“ in graphischer Darstellung:

■ **Tabellarische Darstellung:**

- **Vorteil:** liefert detailliertere Informationen, man kennt die genauen Werte → das ist insbesondere bei Planungsaufgaben wichtig.
- **Nachteil:** Tabellen sind schwerer zu lesen, man braucht Zeit, um die Information zu verarbeiten. Tabellen sind „langweilig“.

■ **Graphische Darstellung:**

- **Vorteil:** Man kann sich sehr schnell ein Bild von den quantitativen Verhältnissen machen, man erkennt sehr schnell die wesentlichen Informationen (wenn das Diagramm gut gestaltet ist ...).
- **Nachteil:** Nur mit Mühe lassen sich genaue Werte ablesen.

FACHTERMINOLOGIE

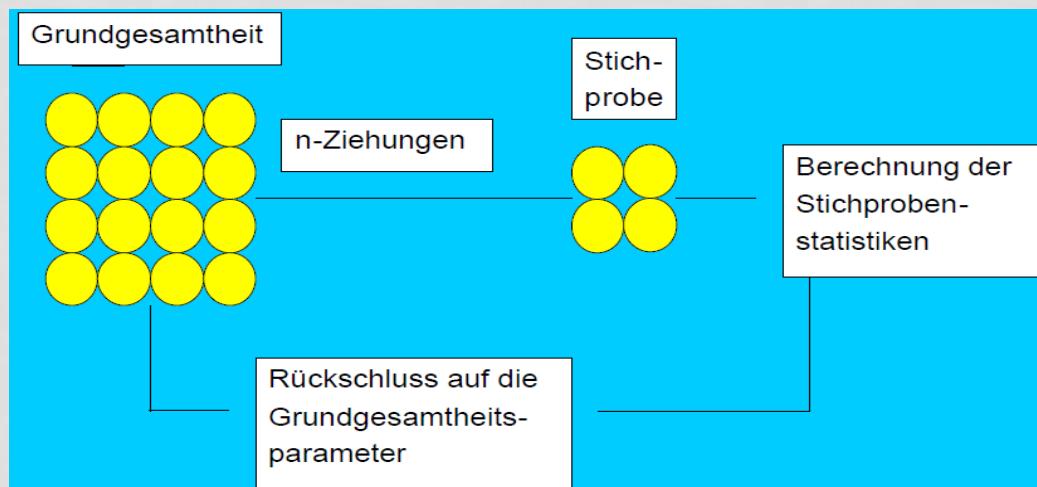
Fachterminologie

Untersuchungseinheiten = statistische Einheiten (Träger der Information)	einzelne Großhändler
Grundgesamtheit = statistische Masse	Alle Großhändler der Meyer AG in NRW im Jahr 2008
Umfang einer Gesamtheit = Anzahl ihrer Einheiten (Elemente)	Anzahl der Großhändler
Merkmal = Variable	Umsatz im Jahr 2008
Information der Tabelle	(klassierte) Häufigkeitsverteilung mit absoluten und relativen (Klassen-)Häufigkeiten

STICHPROBE

Grundgesamtheit → die Menge aller möglichen Erhebungseinheiten

Stichprobe → eine n-elementige Teilmenge der Grundgesamtheit mit N Elementen (Merkmalsträgern)



Ein **Auswahlverfahren** ist die Art und Weise, wie die Elemente der Stichprobe möglichst zweckmäßig ausgewählt werden.

ZUFALLSSTICHPROBE

- **Einfache Zufallsstichproben**

jede mögliche Stichprobe und auch jedes Element besitzen dieselbe Chance ausgewählt zu werden. Dies ist dann eine echte Zufallsstichprobe (meist unrealistisch), der Idealfall einer Stichprobe. Sie ist ein genaues Abbild der Grundgesamtheit, so dass der Schluss von der Stichprobe auf die Grundgesamtheit gewährleistet ist.

- **Geschichtete Zufallsstichproben**

Die Elemente der Grundgesamtheit werden so in Gruppen (Schichten, strata) eingeteilt, dass jedes Element der Grundgesamtheit zu einer – und nur zu einer – Schicht gehört. Danach werden einfache Zufallsstichproben aus jeder Schicht gezogen.

KLUMPENSTICHPROBE

■ Klumpenstichprobe

eine einfache Zufallsauswahl, bei der die Auswahlregeln nicht auf die Elemente der Grundgesamtheit, sondern auf zusammengefasste Elemente (Klumpen, Cluster) angewendet werden und dann jeweils die Daten aller Elemente des ausgewählten Clusters erhoben werden. Ein Nachteil dieses Verfahrens: es kann kein Stichprobenumfang n vorgegeben werden.

Beispiel:

Es soll ein Leistungstest an deutschen Schulkindern durchgeführt werden. Im ersten Schritt werden 'Gemeinden' als Klumpen ausgewählt. Als 'Liste' kann das Telefonvorwahlverzeichnis benutzt werden. Darin sind ca. 8.000 Gemeinden zu finden, aus denen eine Stichprobe gezogen werden kann. Einige der Gemeinden werden über keine Schulen verfügen. Eine Liste der Schulen ist ebenfalls als 'Liste' (Über das verantwortliche Schulamt) vorhanden. Aus den zur Verfügung stehenden Schulen wird dann eine Stichprobe gezogen, anschließend aus den dort existierenden Klassen. Schließlich nehmen Kinder der ausgewählten Klassen an dem Test teil.

WILLKÜRLICHE UND BEWUSSTE AUSWAHLEN

■ **Willkürliche Auswahlen (Auswählen aufs Geratewohl)**

unkontrollierte Aufnahme eines Elementes der Grundgesamtheit in die Stichprobe

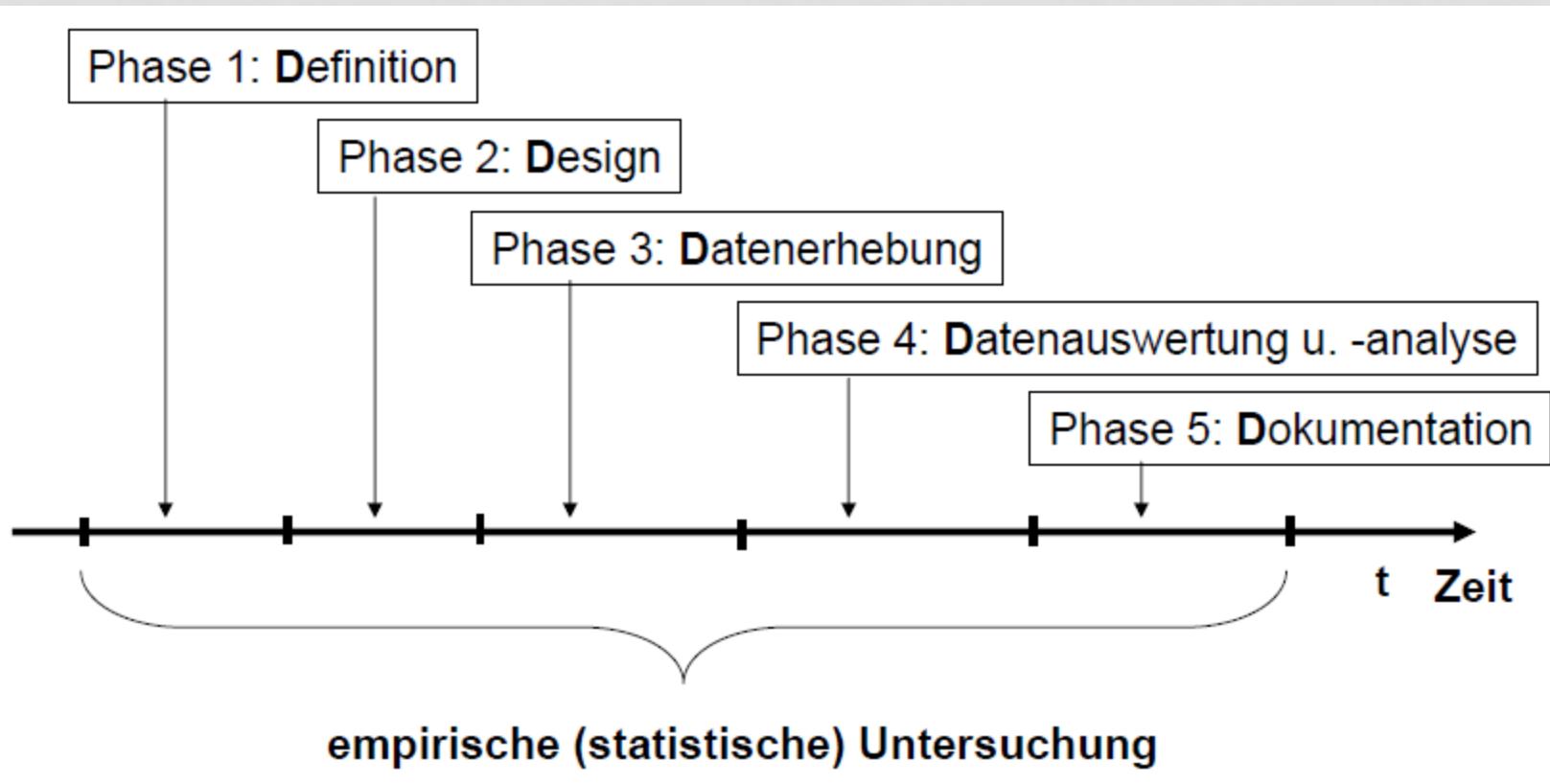
■ **Bewusste Auswahlen (Auswählen nach Gutdünken)**

nach einem Auswahlplan (anhand von Listen und festgelegten Regeln) und diesem Plan zugrunde liegenden angebbaren Kriterien. Es gibt viele verschiedene Arten bewusster Auswahlen:

- **Auswahl extremer Fälle**
- **Auswahl typischer Fälle**
- **Konzentrationsprinzip**
- **Schneeball-Verfahren**
- **Quotaverfahren (bestimmte Merkmale in der Stichprobe sollen exakt in derselben Häufigkeit (in %) vorkommen wie in der Grundgesamtheit)**

ABLAUF EINER EMPIRISCHEN UNTERSUCHUNG

Die 5 D's



DIE 5 D'S

- **Definition (Phase 1)**
 - **Definition des Informationsbedarfs, der Hypothesen, der Begriffe, der Untersuchungseinheiten, über die man Information haben will**

Nur durch eindeutige und verständliche Formulierung der **Zielsetzung** kann gewährleistet werden, dass wirklich das erforscht wird, was erforscht werden soll!

DIE 5 D'S

■ Design-Entscheidungen (Phase 2)

- **Abgrenzung der Grundgesamtheit, evtl. Stichprobenumfang**
- **Erhebungsart:**
 - Querschnitt- oder Längsschnittuntersuchung
 - Primärerhebung oder Sekundärerhebung
 - Vollerhebung oder Teilerhebung
- **Erhebungstechnik:**
 - Befragung (persönlich, telefonisch, schriftlich oder online)
 - Beobachtung (offen oder verdeckt)
 - Dokumentenanalyse

DIE 5 D'S

■ Design-Entscheidungen (Phase 2)

■ „Konstruktion“ von Messinstrumenten (Pretest der z.B. Fragebögen)

Ziel: das Risiko des Misserfolgs zu reduzieren und vorab Gründe für ein eventuelles Versagen zu finden. Außerdem können nach den Pretests eventuell noch Verbesserungen vorgenommen werden.

■ Auswertungsdesign

■ Methodischer Zugang zur Auswertung

■ Entscheidungsspielraum wird eingeschränkt bzw. beeinflusst durch:

■ Budget

■ Zeit

■ Thema/Aufgabenstellung

■ Interdependenzen (gegenseitige Abhängigkeit und Beeinflussung) zwischen den Design-Entscheidungen

DIE 5 D'S

- **Datenerhebung (Phase 3)**
 - entsprechend der getroffenen Entscheidungen
- **Datenauswertung und –analyse (Phase 4)**
 - **Vorbereitung der „maschinellen“ Datenauswertung / –analyse (mit Software)**
 - Dateiaufbau festlegen (Variablendefinition und –codierung), Datenimport
 - Datenbereinigung
 - Datenqualitätssicherung (Kontrolle auf Vollständigkeit und Plausibilität)
 - Datenaufbereitung (Sortierung der Daten, Klassenbildung, ...)
 - Datenauswertung und Datenanalyse (univariate und multivariate Datenanalysen mit Anwendung geeigneter statistischer Methoden)
 - Einsatz von Statistik-Software

DIE 5 D'S

- **Dokumentation (Phase 5)**

Dokumentation in Tabellen und Schaubildern und Interpretation der Ergebnisse

Beispiel für die Gliederung einer Ergebnisstudie:

- Problemstellung
- Vorgehensweise, Beschreibung und Begründung aller Design-Entscheidungen
- Hauptteil: Ergebnisse der empirischen Untersuchung
- Folgerungen, Empfehlungen, Wertungen
- Anhang: Fragebogen, Literatur-, Abbildungs- und Tabellenverzeichnis

Mögliche Reaktionen auf die Ergebnisse der empirischen Untersuchung

„Na klar!“

→ Vermutungen bestätigt

„Aha!!!“

→ Ergebnisse überraschen

„OPERATIONALISIERUNG“ EINES BEGRIFFS

„**Operationalisierung**“ eines Begriffs ist die Angabe derjenigen Vorgehensweisen und **Forschungsoperationen**, mit deren Hilfe zu entscheiden ist, ob und in welchem Ausmaß der mit dem Begriff bezeichnete Sachverhalt in der Realität vorliegt, was bedeutet, dass man beobachtbare Kriterien dafür anzugeben hat, wann ein Sachverhalt vorliegt bzw. je nach Skalenniveau auch in welcher Ausprägung er auftritt.

Etwas weniger abstrakt:

„Operationalisierung“ definiert, wie man den Begriff konkret misst. Die Operationalisierung ist besonders wichtig bei Begriffen ohne direkten empirischen Bezug (so genannte „latente Variable“), z.B. Kundenzufriedenheit, Teamfähigkeit, Intelligenz, Werbewirkung u.a.

Der Ausdruck Operationalisierung bezeichnet im weitesten Sinne die Entwicklung eines Forschungsdesigns für eine konkrete Fragestellung, während es im engeren Sinne um die Formulierung von Messvorschriften geht, d.h., um die Bestimmung von Indikatoren, mit deren Hilfe ein Konstrukt gemessen werden kann.

- die Festlegung der Vorgehensweise (Operation) bei der Definition der Untersuchungsvariablen in einer Untersuchung.

Beispiel: Intelligenz kann operational durch die Anzahl der Lösungen von Intelligenzaufgaben in einem konkreten Intelligenztest definiert werden.

„OPERATIONALISIERUNG“ EINES BEGRIFFS

Beispiel: (Quelle der Geschichte: Becker, B.: Statistik, S. 40 und 79)

Ein Mitarbeiter des Statistischen Bundesamtes wird eines Tages von seiner kleinen Tochter gefragt, warum er so viel Geld verdient, wo er doch nur Zahlen addiert. Der Vater lächelt und sagt: „Mein Liebes, geh’ hinaus in den Garten und zähle die Bäume“. Nach einigen Minuten kommt die Tochter zurück, aufgelöst in Tränen: „Papi, ich kann die Bäume nicht zählen, denn da sind große Bäume, kleine Bäume, Büsche, einige haben Nadeln, andere Blätter, einige haben einen Stamm, andere haben 2 Stämme...“. Ihr Vater lächelt wieder und nimmt sie sanft auf seinen Arm: „Siehst du, das ist es, wofür ich bezahlt werde, ich zähle nicht einfach nur, ich mache Entscheidungen, **was ein Baum ist, wie man Bäume zählt** usw.“

Wenn ein Statistiker mit der Arbeit beginnt, muss er geklärt haben, was er überhaupt quantifizieren will. Will er z.B. Bäume zählen, so muss er wissen, was mit einem Baum gemeint ist, wie man also von einem mehr oder weniger abstrakten Begriff zu einer „operationalisierbaren“ bzw. beobachtbaren Kategorie kommt. Es muss also gesagt werden, ob man z.B. nur Laubbäume meint oder auch Nadelbäume eingeschlossen sein sollen, ob man jeden Baum unabhängig von seiner Größe zählen soll usw..

DATENANALYSE MIT STATISTIK-SOFTWARE

Zur Datenanalyse verwendet man in der Praxis unterschiedliche Statistik-Software.

Marktführer sind:

- **EXCEL (Tabellenkalkulation, einfache statistische Methoden, Grafiken, etc.)**

Nicht für anspruchsvolle statistische Aufgaben geeignet!

DATENANALYSE MIT STATISTIK-SOFTWARE

■ R (**die mächtige Open Source-Lösung, kostenfrei**)

www.r-project.org

Eine populäre Open Source-Statistik-Umgebung, die durch Pakete nahezu beliebig erweiterbar ist und sich zunehmender Beliebtheit erfreut. Mit RStudio existiert eine komfortable Entwicklungsumgebung, die lokal oder in einer Client-Server-Installation über den Webbrower genutzt werden kann. R-Applikationen lassen sich über Shiny auch direkt interaktiv im Web nutzen.

R kann insbesondere Viel-Nutzern, die die Bereitschaft mitbringen, sich intensiver mit Statistik auseinanderzusetzen, uneingeschränkt empfohlen werden.

DATENANALYSE MIT STATISTIK-SOFTWARE

- **SAS (kommerzielle Statistik-Software, der Mercedes unter den Statistik-Programmen)**

SAS ist ein mächtiges und sehr stabiles Tool, welches insbesondere in größeren Organisationen eingesetzt wird und sich im Pharma-Bereich zum Quasi-Standard für viele Analysen entwickelt hat. Die Software besteht aus unterschiedlichen Modulen, die z.T. völlig verschiedene Bedienkonzepte verfolgen. Entsprechend aufwändig ist die Einarbeitung. Im Vergleich zur kommerziellen Konkurrenz gehört SAS (auch aufgrund der Ausrichtung auf größere Unternehmen/Organisationen) zu den teuersten Lösungen. Eine professionelle Statistiksoftware, welche insbesondere in der Biometrie, der klinischen Forschung und im Banken-Sektor Anwendung findet.

DATENANALYSE MIT STATISTIK-SOFTWARE

■ **SPSS (Statistik für Dummies)**

SPSS gilt als besonders einfach zu bedienen, da die Software in den jüngeren Versionen stark in Richtung eines Tools entwickelt wurde, welches Auswertungen weitgehend automatisiert durchführt, ohne dass dem Benutzer besondere Methodenkenntnisse abverlangt werden. Die Stabilität hat gelitten. Während SPSS einige speziellere Module (z.B. für das Direktmarketing) mitbringt, ist das Spektrum gut unterstützter Methoden insgesamt geringer als z.B. bei R oder SAS. Insbesondere in den Sozialwissenschaften und der Psychologie war SPSS auch im universitären Bereich fest verankert.

Der ursprünglich eigenständige Anbieter wurde mittlerweile von IBM übernommen.

DATENANALYSE MIT STATISTIK-SOFTWARE

■ **STATA (Mehr als nur Panel-Analysen)**

Obwohl STATA eine ausgereifte, sehr stabile und leistungsstarke Software ist, ist die Verbreitung - gerade in Unternehmen - gering. Dabei ist STATA für Anwender, die Wert auf ein breites Methodenspektrum, Stabilität, ein ausgereiftes Bedienkonzept inkl. Skriptsprache und einen fairen Preis legen, der teureren kommerziellen Konkurrenz überlegen.

STATA ist eine kommerzielle Statistiksoftware und wird insbesondere in der Ökonometrie angewendet.

DATENANALYSE MIT STATISTIK-SOFTWARE

■ Weitere Programme

Daneben existieren etliche Programme, die sich auf bestimmte Methoden spezialisiert haben. Einige dieser Programme seien in dieser unvollständigen Übersicht zumindest kurz erwähnt:

- Eviews (Ökonometrie, Zeitreihenanalyse)
- SPSS Amos (Modellierung und Schätzung von Strukturgleichungsmodellen)
- WinBUGS und OpenBUGS (speziell für Bayes'sche Statistik). Mit RBUGs und R2OpenBUGS existieren Pakete, die die Funktionalität in R integrieren.
- Mathematica und Matlab (numerische Problemstellungen)
- Etc.

STATISTIK-SOFTWARE IM VERGLEICH

	Stärken	Schwächen
R	<ul style="list-style-type: none"> • Sehr großer Funktionsumfang (weit über 2000 Pakete) • Sehr gut automatisier- und integrierbar (z.B. LaTeX, ODBC, MS ...) • Sehr guter Community-Support sowie kostenpflichtiger Support über Drittanbieter • Umfangreiche Hilfe-Ressourcen frei verfügbar (Manuals, Tutorials) • Alle gängigen Plattformen werden unterstützt (Windows, Linux...) • Zukunftssicher durch große, aktive Entwickler-Community 	<ul style="list-style-type: none"> • Einarbeitung in die R-Syntax kann eine Einstiegshürde darstellen • Stabilität/Qualität wenig genutzter Pakete z.T. nicht auf dem hohen Niveau • Bei Verwendung sehr großer Datensätze wird leistungsfähige Hardware benötigt
SAS	<ul style="list-style-type: none"> • Schnelle Integration neuer statistischer Verfahren, sehr stabile und zuverlässige Routinen • Sehr gute Dokumentation und professioneller Support • Vielzahl von (kostenpflichtigen) Modulen und Schnittstellen, eigene Business Intelligence Software • Gut geeignet für Umgang mit großen Datensätzen • Umfangreiches hauseigenes Schulungsangebot 	<ul style="list-style-type: none"> • Verschiedene, teils komplizierte (aber mächtige) Programmiersprachen • Lizenzmodell verbunden mit hohen Kosten
SPSS	<ul style="list-style-type: none"> • Leicht erlernbar, Bedienung jedoch nicht immer intuitiv • Erweiterbar über kommerzielle Module • Umfangreiche Literatur vorhanden 	<ul style="list-style-type: none"> • Versionen für Windows und MacOS • kurzes Update-Zyklus (1 Jahr) • schwierig automatisier- und integrierbar
STATA	<ul style="list-style-type: none"> • Großer Funktionsumfang - nahezu jede statistische Methode • Einfacher Einstieg durch GUI • Automatisierbar & mit alten Versionen kompatibel • Guter Support durch die STATA-Community, umfangr. Literatur • Lauffähig unter Windows, Mac, Unix • Im Vgl. zur kommerz. Konkurrenz vergleichsweise preiswert • Investitionssicherheit durch 3-jährigen Release-Zyklus 	<ul style="list-style-type: none"> • Eher träge bei der Einarbeitung neuer Methoden (Versionsupdates) • Integration von und in andere Software ist umständlich • Beschränkung auf einen gleichzeitig geöffneten Datensatz

WIRTSCHAFTSSTATISTIK

MODUL 2: SKALEN UND KLAISIERUNG

WS 2021/22

DR. E. MERINS

MESSBARKEIT

- **Informationsbedarf → empirische (statistische) Untersuchung**

Bei einer empirischen Untersuchung messen wir Merkmale bei ausgewählten Untersuchungseinheiten mit einem Messinstrument auf einer Skala.

Ergebnis: Messwerte = Merkmalswerte = Beobachtungswerte

Wir messen bei Kind und seiner Mutter das Merkmal Körpergröße mit einem cm-Maß auf einer cm-Skala.

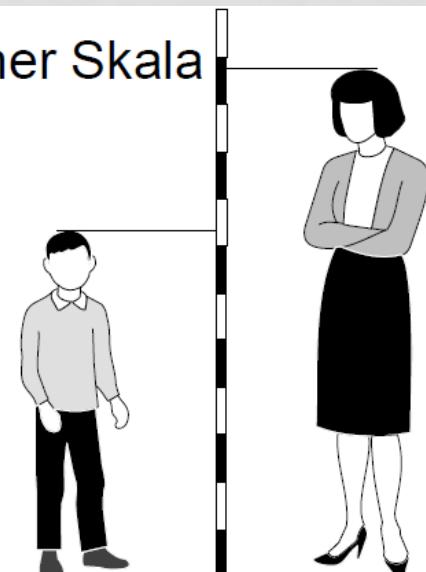
Messergebnisse:

Kind: 121 cm, Mutter: 168 cm.

Messen ... auf einer Skala



**Kind: 121 cm
Mutter: 168 cm**

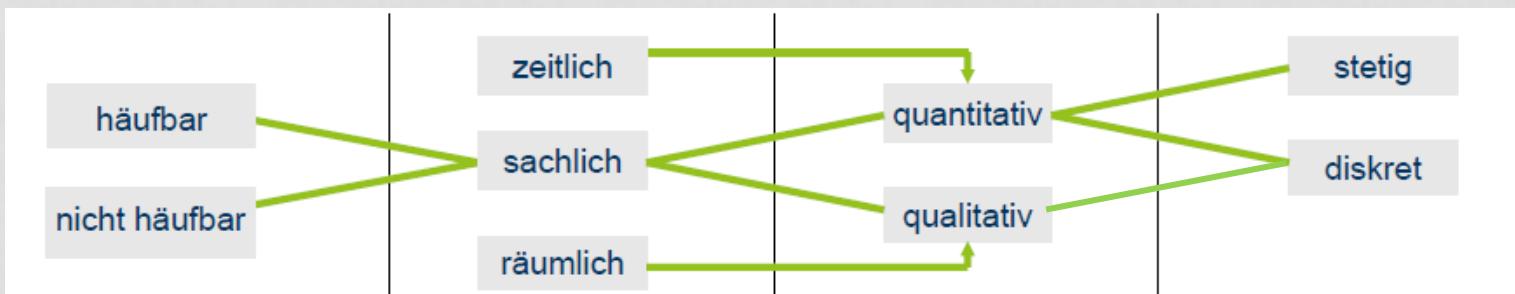
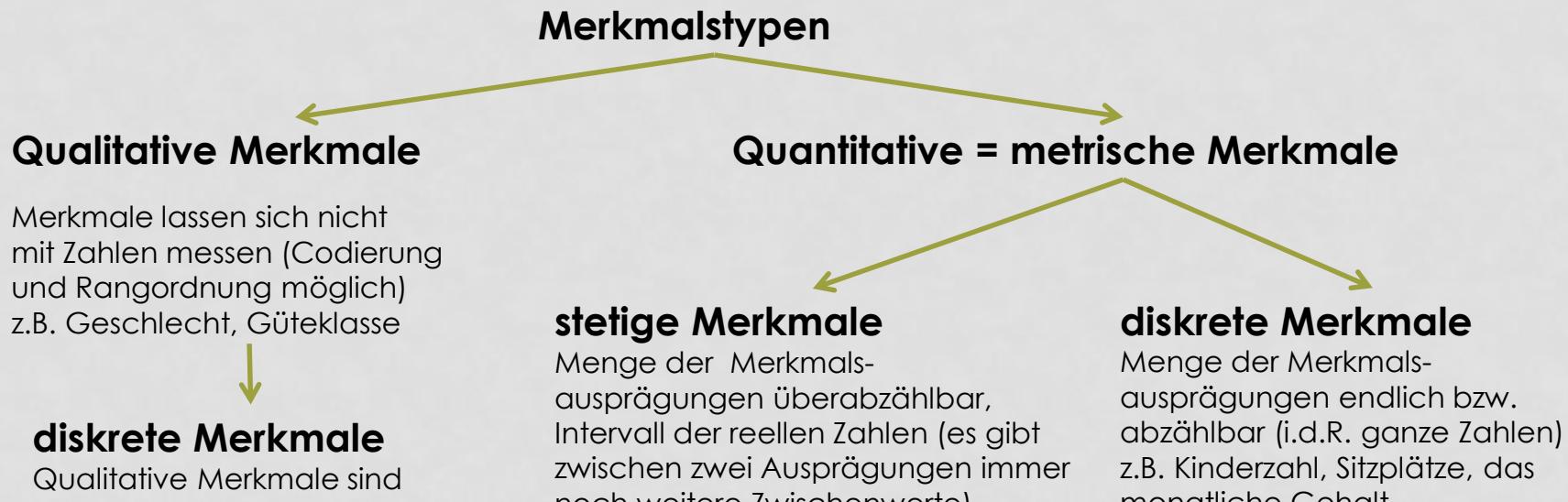


GRUNDBEGRIFFE DER STATISTIK

Grundbegriffe der Statistik

Merkmalsträger	Einzelnes Objekt einer statistischen Untersuchung, Träger der Informationen, für die man sich interessiert. → Untersuchungseinheit → Erhebungseinheit → Unit
Statistische Masse	Menge aller Merkmalsträger, die <ul style="list-style-type: none">• mit dem Untersuchungsziel in Verbindung stehen,• unter sich mindestens eine übereinstimmende Eigenschaft haben,• sich exakt abgrenzen lassen, und zwar<ul style="list-style-type: none">- sachlich- räumlich- zeitlich → Kollektiv, Grundgesamtheit, Population Beispiele: Bevölkerung des Landes, Automobilproduktion
Merkmal	Im Rahmen der statistischen Erhebung relevante Eigenschaften der Merkmalsträger → Statistische Variable
Merkmalsausprägung	Grundsätzlich mögliche Ausformungen eines Merkmals → Wert der Variable, Beobachtungswert

MERKMALE



SKALENNIVEAU

Nach der Art des Merkmals richtet sich, auf welche Weise die Beobachtungswerte bei der statistischen Untersuchung gemessen werden können (Messung = Eindeutige Zuordnung einer Beobachtung zu einem Punkt auf einer Messskala)

Vom **Skalenniveau** hängt auch ab, welche Rechenoperationen mit den Beobachtungswerten und welche statistischen Auswertungsmethoden zulässig sind.

Man unterscheidet folgende **Skalenniveaus**:

I. Nicht metrische Skalen → Anwendung bei qualitativen Merkmalen. Keine Rechenoperationen mit den Merkmalsausprägungen zulässig:

- **Nominalskala**
- **Ordinalskala**

II. Metrische Skalen (Kardinalskalen) → Anwendung bei quantitativen Merkmalen.
Skala hat Nullpunkt und Maßeinheit. Rechenoperationen sind zulässig:

- **Intervallskala**
- **Verhältnisskala** (Ratioskala)
- **Absolutskala**

SKALIERUNG

Skalenart	Besonderheiten	zulässige Operationen	Beispiel für Merkmale	Beispiel für Operationen
Nominalskala	Daten haben nur eine endliche Menge von Ausprägungen, unterliegen keiner Rangfolge und sind nicht vergleichbar. Zuordnung von Zahlen ist lediglich eine Kodierung der Merkmalsausprägungen	=, ≠	Geschlecht, Familienstand, Steuerklasse, PLZ	Geschlecht von Claudia ≠ Geschlecht von Peter
Ordinalskala = Rangskala	Daten haben nur eine endliche Menge von Ausprägungen, können in eine natürliche Rangfolge gebracht werden. Ordnungsprinzip ist die Stärke bzw. der Grad der Intensität, man kann hier allerdings keine Abstände zwischen den einzelnen Ausprägungen interpretieren	=, ≠, <, >	Konfektionsgröße, Schulnoten, Windstärke	XXL > XL > L > M > S > XS
Intervallskala	Besitzt <u>keinen</u> natürlichen Nullpunkt, keine Verhältnisse können gebildet werden. Daten können alle (unendlich viele) Ausprägungen innerhalb eines Intervalls annehmen.	=, ≠, <, >, +, -	Längendifferenzen, IQ, Temperatur in Celsius	morgen wird es 10 Grad kälter als heute
Verhältnisskala = Ratioskala	Besitzt natürlichen Nullpunkt Quotienten (das Verhältnis) gemessener Werte werden verglichen	=, ≠, <, >, +, -, x, /	Umsatz, Körpergröße, Einkommen, Temperatur in Kelvin	Der Umsatz ist um 7% gegenüber dem Vorjahr gestiegen oder doppelt so hoch wie...
Absolutskala	Ausprägungen absolut skalierter Merkmale sind Anzahlen und Stückzahlen. Allgemein: Häufigkeiten oder alles, was man zählen kann	=, ≠, <, >, +, -, x, /	Zahl der Beschäftigten	150 Beschäftigte sind 3 mal so viel wie 50 Beschäftigte

SKALIERUNG, BEISPIELE

Merkmal	Menge der Merkmalsausprägungen	Messinstrument	Skala	Merkealstyp
Familienstand	{ledig, verheiratet, verwitwet, geschieden}	Frage	Nominalskala	qualitatives Merkmal
Hotelgüteklaasse	{***** , **** , *** , ** , * , }	Fragebogen	Rangskala = Ordinalskala	qualitative Merkmale = Rangmerkmale
Klausurnote	{1,0 1,3 1,7 2,0 2,3 2,7 3,0 3,3 3,7 4,0 5,0}	Klausur		
Temperatur ($^{\circ}\text{C}$)	\mathbb{R}	Thermometer	Metrische Skala = Intervallskala	Quantitative Merkmale
Körpergröße	$\{x \mid x \in \mathbb{R} \text{ und } x > 0\}$	cm-Maß	Metrische Skala = Verhältnisskala	= metrische Merkmale
Kinderzahl	$\mathbb{N} \cup \{0\}$	Frage	Metrische Skala = Absolutskala	

KLASSIERUNG BEI QUALITATIVEN MERKMALEN

Beispiel:

Merkmal: Beruf

Merkmalsausprägung:

→ Berufsgruppe: Handwerker = Klasse von z.B.

- Maurer
- Dachdecker
- Schreiner
- Fliesenleger

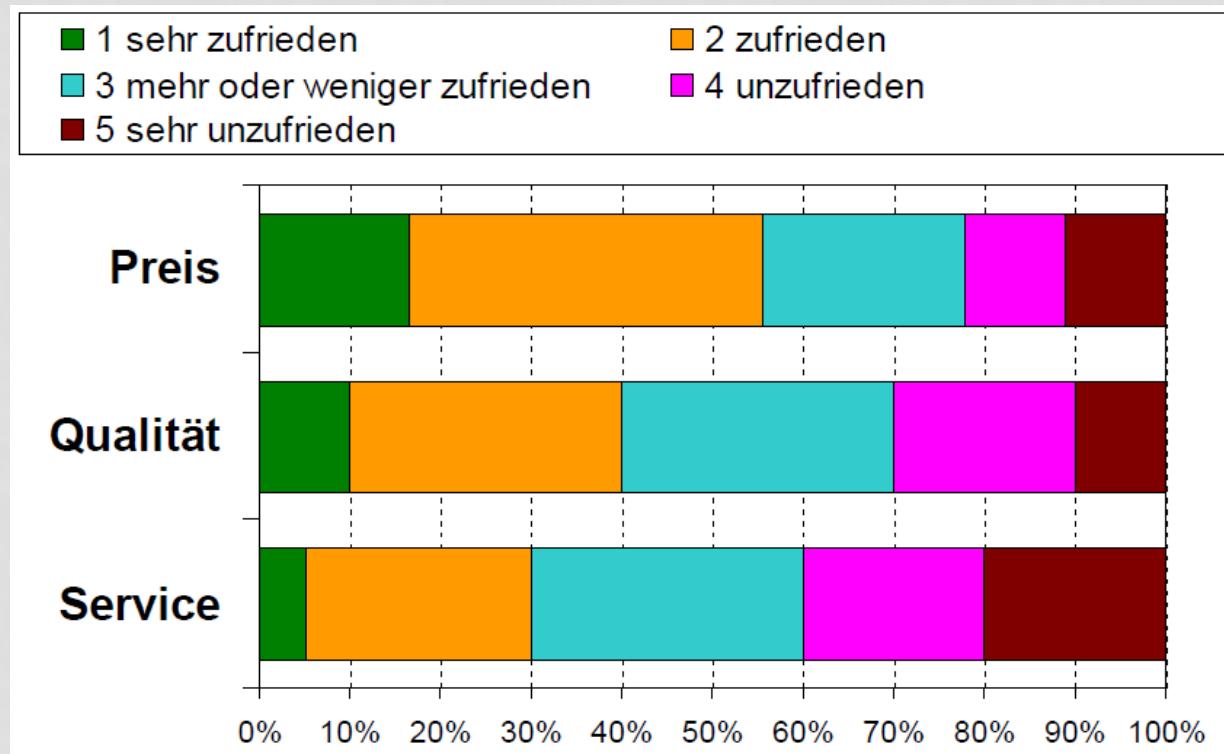
Zielkonflikt: Übersichtlichkeit versus Informationsverlust

KLASSIERUNG BEI RANGMERKMALEN

Beispiel:

Frage: wie zufrieden sind Sie mit einem Produkt bzgl. Preis/Qualität/Service?

Antworten:

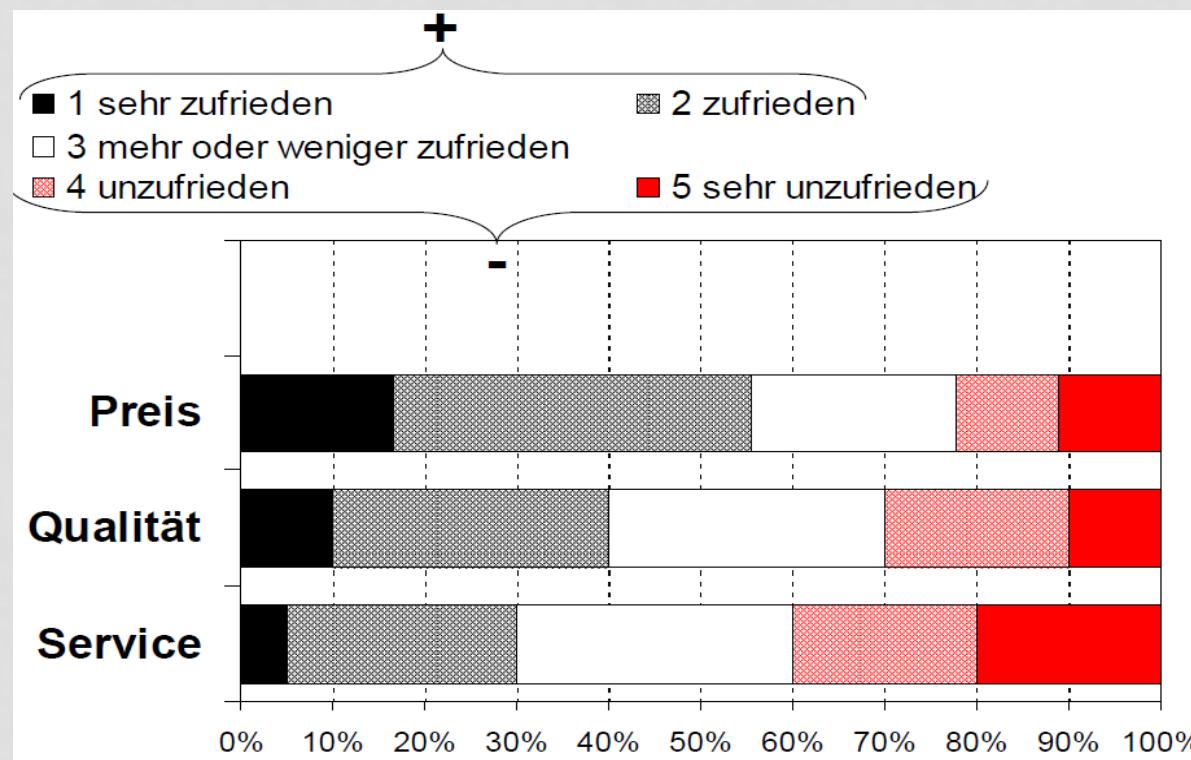


KLASSIERUNG BEI RANGMERKMALEN

Beispiel:

Frage: wie zufrieden sind Sie mit einem Produkt bzgl. Preis/Qualität/Service?

Antworten:

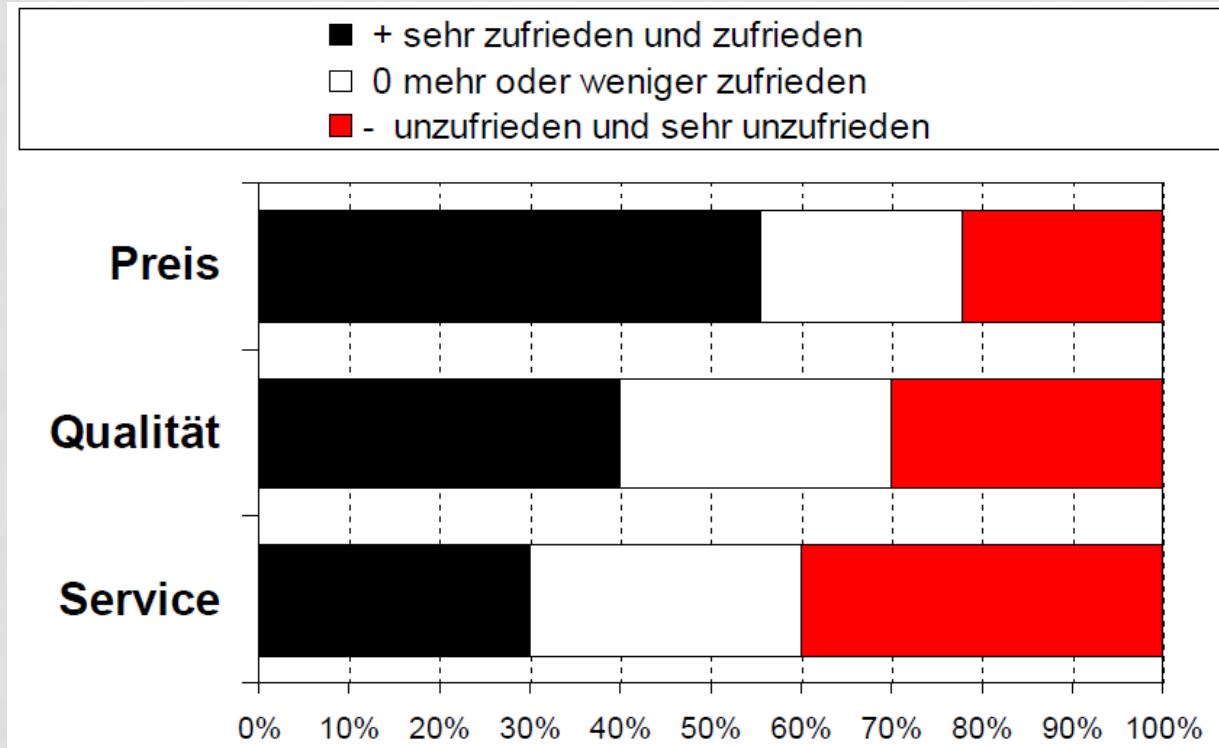


KLASSIERUNG BEI RANGMERKMALEN

Beispiel:

Frage: wie zufrieden sind Sie mit einem Produkt bzgl. Preis/Qualität/Service?

Antworten:



KLASSIERUNG BEI METRISCHEN MERKMALEN

Metrische Merkmale

(vgl. Folie 4)

diskret

z.B. Einwohnerzahl

stetig

z.B. Körpergröße

Klassierung

1. 0 – 19.999
 2. 20.000 – 49.999
 3. 50.000 – 99.999
 4. 100.000 – 249.999
- usw.

Klassierung

- 0 – 99 cm
 - 100 – 139 cm
 - 140 – 159 cm
 - 160 – 169 cm
- usw.

Klassierung

- 0 – 100 cm
 - 100 – 140 cm
 - 140 – 160 cm
 - 160 – 170 cm
- usw.

Klassierung

- 0 bis unter 100 cm
 - 100 b.u. 140 cm
 - 140 b.u. 160 cm
 - 160 b.u. 170 cm
- usw.

wg. Lücken

wg. Über-schneidungen

richtig !!!

$$100 \text{ b.u. } 140 \text{ cm} = \{x \mid x \in \mathbb{R}, 100 \leq x < 140 \text{ cm}\}$$

ENTSCHEIDUNGEN BEI KLASSIERUNG

- **Anzahl der Klassen**
- **Klassenbreite(n)**
 - alle gleich oder unterschiedlich
- **Klassengrenzen (Klassen definieren)**
 - untere Klassengrenzen, obere Klassengrenzen
- **untere/obere offene Randklasse?**
 - „bis unter 50 kg“ bzw. „120 kg und schwerer“

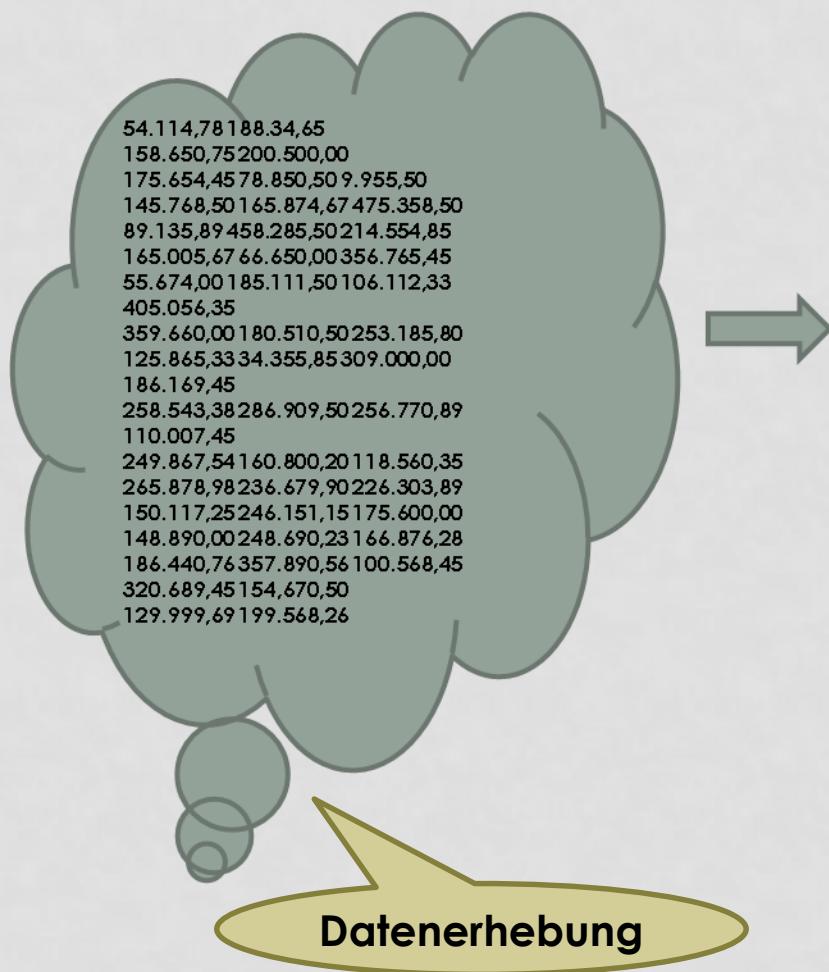
WIRTSCHAFTSSTATISTIK
MODUL 3: HÄUFIGKEITEN UND
HÄUFIGKEITSVERTEILUNGEN

WS 2021/22

DR. E. MERINS

DATEN

INPUT



OUTPUT

Umsätze der Meyer AG über die Großhändler in NRW im Jahr 2008

Umsatzklasse in Tsd. €	Anzahl Großhändler (absolute Häufigkeit)	Anteil Großhändler von Gesamt in % (relative Häufigkeit)
0 bis unter 100	7	14%
100 bis unter 200	23	46%
200 bis unter 300	12	24%
300 bis unter 400	5	10%
400 bis unter 500	3	6%
Summe	50	100%

(Quelle: Umsatzstatistiken der Vertriebsabteilung, 2008)
Tabelle 1

Datenaufbereitung

DATENDOKUMENTATION

Formen als Dokumentation der Daten:

Einzelwerte (Einzelbeobachtungen) → ungeordnete Reihe (**Urliste**,

Rohdaten, Primärdaten) → **INPUT**-Blase auf der Folie 2

→ Die Urliste ist im Bereich der Statistik das direkte Ergebnis einer
Datenerhebung

Vorteile:

Die Urliste enthält alle Beobachtungswerte und damit: keine Auslassungen, keine Übertragungsfehler und keine verlorene Information

Nachteile:

Urlisten können in der Praxis tausende oder Millionen von Datensätze enthalten, die für sich genommen unübersichtlich und nicht auswertbar sind; außerdem können bei einer unkorrigierten Urliste noch offensichtliche Fehler, wie Zahlendreher oder unplausible Daten enthalten sein

HÄUFIGKEITSVERTEILUNGEN

Die Daten einer Urliste müssen in der Praxis also aufbereitet werden, um ihren Zweck zu erfüllen.
Das geschieht meist durch das Bilden von Häufigkeitsverteilungen:

Schritt 1: Sortieren der Daten → **geordnete Reihe** nach irgendeiner Ordnung, z. B. alphabetische Ordnung der Merkmalsträger oder Größenordnung der Merkmalsausprägung

Schritt 2: Verdichten der sortierten Daten auf Merkmalsausprägungen und zählen wie oft diese vorkommen → geordnete Menge von Wertepaaren (Merkmalsausprägung und zugehörige Häufigkeit) heißt **Häufigkeitsverteilung**

Schritt 3: Darstellen tabellarisch von nach Merkmalsausprägungen sortierten Häufigkeitsverteilungen → die **Häufigkeitstabelle**

Für klassierte Daten:

Schritt 1: Einteilung der Werte in Klassen → **klassierte Daten** (Sortierung nicht nötig)

Schritt 2: Verdichten der klassierten Daten → **Häufigkeitsverteilung** für klassierte Daten (klassierte Verteilung)

Schritt 3: Darstellen der klassierten Daten → **Häufigkeitstabelle** für klassierte Daten

ABSOLUTE UND RELATIVE HÄUFIGKEITEN

- **Merkmalsausprägung und zugehörige Häufigkeit**

$$x_j \longrightarrow h(x_j) \quad j = 1, \dots, m$$

absolute Häufigkeit

$$x_j \longrightarrow f(x_j) = \frac{h(x_j)}{n} \quad j = 1, \dots, m$$

$$f(x_j) = \frac{h(x_j)}{n} \cdot 100 \text{ (%)}$$

relative Häufigkeit

Bezug zur
Grundgesamtheit

- **absolute Häufigkeit** → die Anzahl des Auftretens einer bestimmten Merkmalsausprägung
- **relative Häufigkeit** → das Verhältnis der absoluten Häufigkeit und der Summe der Einzelhäufigkeiten

EINDIMENSIONALE HÄUFIGKEITSVERTEILUNG

- **Beispiel 1:**

n = 20 Personen wurden gefragt nach dem

Merkmal X: Familienstand

mit den **j=4 Merkmalsausprägungen:**

x_1 = ledig, x_2 = verheiratet, x_3 = geschieden, x_4 = verwitwet

Primärdaten:

ledig, verheiratet, geschieden, ledig, verheiratet, verwitwet, verheiratet,
ledig, verheiratet, verwitwet, verheiratet, ledig, verheiratet, geschieden,
ledig, verheiratet, verwitwet, verheiratet, ledig, verheiratet

Was können Sie über die Daten sagen? Charakterisieren Sie die Daten

EINDIMENSIONALE HÄUFIGKEITSVERTEILUNG

- Beispiel 1:

Schritt 1: Sortieren

Schritt 2: Verdichten in eine Häufigkeitsverteilung

Schritt 3: Darstellen als eine Häufigkeitstabelle

j	x_j	Anzahl $h(x_j)$	Anteil $f(x_j)$	Anteil in % $f(x_j) (%)$
1	ledig	6	0,30	30
2	verheiratet	9	0,45	45
3	geschieden	2	0,10	10
4	verwitwet	3	0,15	15
	Summe	20	1,00	100

EINDIMENSIONALE HÄUFIGKEITSVERTEILUNG

▪ Beispiel 2:

Frage: Wo wohnen Sie?

Antworten: B C A B C B B A A D k.A. A B B A k.A. A B B (k.A. = keine Antwort)

→ Verdichten in eine **Häufigkeitsverteilung** und **Darstellen** als eine **Häufigkeitstabelle**

i	Wohnort x_i	Anzahl $h(x_i)$	Anteil $f(x_i) (%)$ (bezogen auf alle Antworten)	Anteil $f(x_i) (%)$ (bezogen auf die gültigen Antworten)
1	A	6	30,0%	33,3%
2	B	9	45,0%	50,0%
3	C	2	10,0%	11,1%
4	D	1	5,0%	5,6%
5	k. A.	2	10,0%	-
Summe		20	100,0%	100,0%

SUMMENHÄUFIGKEITEN

→ sinnvoll nur für Rangmerkmale und metrische Merkmale

absolute Summenhäufigkeiten

(absolute kumulierte Häufigkeit)

$$H(x_1) = h(x_1)$$

$$H(x_2) = h(x_1) + h(x_2)$$

$$H(x_3) = h(x_1) + h(x_2) + h(x_3)$$

...

$$H(x_j) = h(x_1) + h(x_2) + \dots + h(x_j)$$

...

$$H(x_i) = h(x_1) + h(x_2) + \dots + h(x_i) = n$$

relative Summenhäufigkeiten

(relative kumulierte Häufigkeit)

$$F(x_1) = f(x_1)$$

$$F(x_2) = f(x_1) + f(x_2)$$

$$F(x_3) = f(x_1) + f(x_2) + f(x_3)$$

...

$$F(x_j) = f(x_1) + f(x_2) + \dots + f(x_j)$$

...

$$F(x_i) = f(x_1) + f(x_2) + \dots + f(x_i) = 1 \text{ (100\%)}$$

EINDIMENSIONALE HÄUFIGKEITSVERTEILUNG MIT SUMMENHÄUFIGKEITEN

→ sinnvoll nur für Rangmerkmale und metrische Merkmale

Index	Merkmals-ausprägungen	absolute Häufigkeit	relative Häufigkeit	relative Häufigkeit in %	absolute Summenhäufigkeit	relative Summenhäufigkeit	relative Summenhäufigkeit
i	x_i	$h(x_i)$	$f(x_i)$	$f(x_i) (%)$	$H(x_i)$	$F(x_i)$	$F(x_i) (%)$
1	a	28	0,11	11,0%	28	0,11	11,0%
2	b	102	0,402	40,2%	$28 + 102 = 130$	$0,110 + 0,402 = 0,512$	51,2%
3	c	61	0,24	24,0%	$130 + 61 = 191$	$0,512 + 0,240 = 0,752$	75,2%
4	d	39	0,154	15,4%	$191 + 39 = 230$	$0,752 + 0,154 = 0,906$	90,6%
5	e	11	0,043	4,3%	$230 + 11 = 241$	$0,906 + 0,043 = 0,949$	94,9%
6	f	9	0,035	3,5%	$241 + 9 = 250$	$0,949 + 0,035 = 0,984$	98,4%
7	h	3	0,012	1,2%	$250 + 3 = 253$	$0,984 + 0,012 = 0,996$	99,6%
8	g	1	0,004	0,4%	$253 + 1 = 254$	$0,996 + 0,004 = 1$	100,0%
Summe		n=254	1	100%	-	-	-

EINDIMENSIONALE KLASSIERTE HÄUFIGKEITSVERTEILUNG MIT SUMMENHÄUFIGKEITEN

Klasse Nr.	Klasse	absolute Häufigkeit	relative Häufigkeit in %	absolute Summenhäufigkeit	relative Summenhäufigkeit	<u>Klassenbreite</u>	<u>Klassenmitte</u>
i		h_i	$f_i (\%)$	H_i	$F_i (\%)$	b_i	m_i
1	0 b.u. 20	30	15%	30	15%	$20 - 0 = 20$	$(20 + 0) / 2 = 10$
2	20 b.u. 50	60	30%	$30 + 60 = 90$	$15 + 30 = 45\%$	$50 - 20 = 30$	$(50 + 20) / 2 = 35$
3	50 b.u. 100	80	40%	$90 + 80 = 170$	$45 + 40 = 85\%$	$100 - 50 = 50$	$(100 + 50) / 2 = 75$
4	100 b.u. 200	30	15%	$170 + 30 = \underline{\underline{200}}$	$85 + 15 = \underline{\underline{100\%}}$	$200 - 100 = 100$	$(200 + 100) / 2 = 150$
Summe		n=200	100%	-	-	-	-

Klassenbreite b_i :

Die Differenz aus der oberen und der unteren Klassengrenze heißt Klassenbreite **b** der Klasse i

$$\rightarrow b_i = x_{k-1} - x_k$$

Klassenmitte m_i :

Das arithmetische Mittel aus der unteren und der oberen Klassengrenze heißt Klassenmitte **m**

$$\text{der Klasse } i \rightarrow m_i = 1/2 (x_{k-1} + x_k)$$

ZWEIDIMENSIONALE HÄUFIGKEITSVERTEILUNG

Zweidimensionale Häufigkeitsverteilung: $G \rightarrow M$ Kreuztabelle

**Randverteilung:
eindim.
Häufigkeits-
verteilung von M**

**Absolute Häufigkeiten
der Merkmals-
ausprägungskombinationen**

**Relative
Spaltenhäufigkeiten
(bedingte relative
Häufigkeiten)**

**Relative
Zeilenhäufigkeiten
(bedingte relative
Häufigkeiten)**

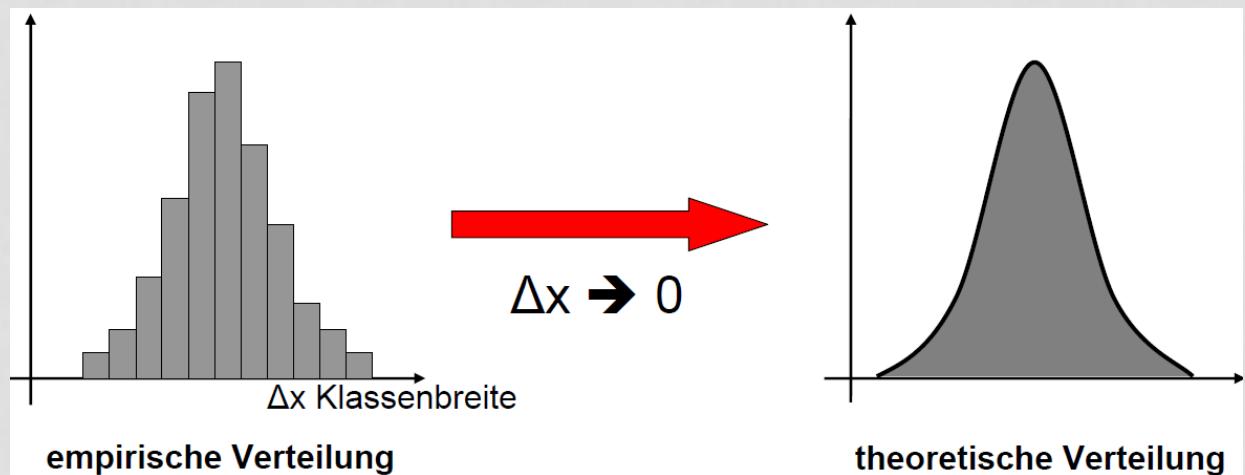
**Relative Häufigkeiten
der Merkmals-
ausprägungskombinationen**

**Randverteilung:
eindim.
Häufigkeits-
verteilung von G**

$M/G \rightarrow$	w	m	Σ
A	400 40,0% 33,3% 20,0%	800 80,0% 66,7% 40,0%	1.200 60%
B	600 60,0% 75,0% 30,0%	200 20,0% 25,0% 10,0%	800 40%
Σ	1.000 50%	1.000 50%	2.000 100%

EMPIRISCHE VERTEILUNGSFUNKTION

Die Verteilungsfunktion enthält die gesamte Information, die in den Daten steckt, nur die ursprüngliche Reihenfolge geht verloren



In der mathematischen Statistik klingt das dann in etwa wie folgt:

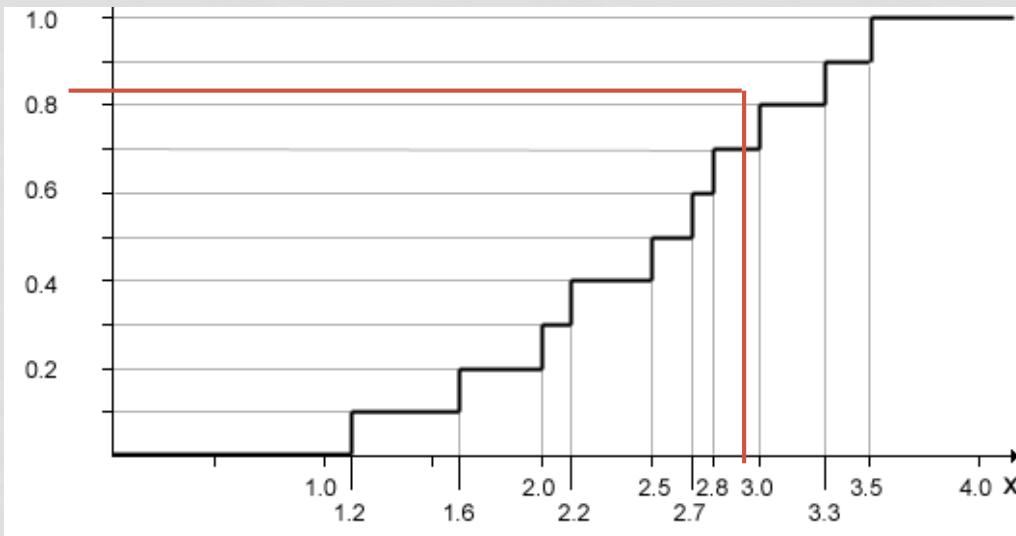
„Das Maximum der Abweichungen der empirischen Verteilungsfunktion von der theoretisch zugrunde liegenden konvergiert mit Wahrscheinlichkeit Eins gegen Null.“

EMPIRISCHE VERTEILUNGSFUNKTION

Eigenschaften:

- Die empirische Verteilungsfunktion $F(x)$ ist (relative) Summenhäufigkeitskurve, relative Summenfunktion
- Die empirische Verteilungsfunktion $F(x)$ gibt für jede beliebige reelle Zahl x den Anteil der Merkmalsträger an, für die das Merkmal X einen Wert x_i annimmt, der kleiner oder gleich x ist
- Wertebereich: $0 \leq F(x) \leq 1$
- $F(x)$ ist monoton nichtfallend (steigt oder ist konstant)
- $F(x)$ ist eine Treppenfunktion mit Sprungstellen bei x_1, x_2, \dots, x_i
- Die Größe der Sprünge beträgt $f_i = F(x_i) - F(x_{i-1})$

EMPIRISCHE VERTEILUNGSFUNKTION



Treppenfunktion

Die Abbildung zeigt die **empirische Verteilungsfunktion** für das Merkmal Abiturnoten.

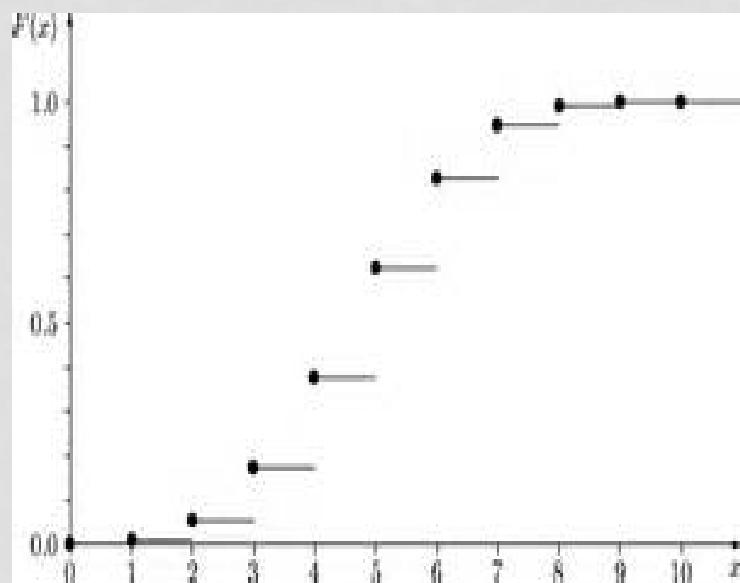
Greift man auf der x-Achse den Wert 3 heraus, so lässt sich der dazugehörige y-Wert 0.8 wie folgt interpretieren: 80 % der Abiturienten haben im schlechtesten Fall den Notendurchschnitt 3 bekommen.

Anders formuliert:

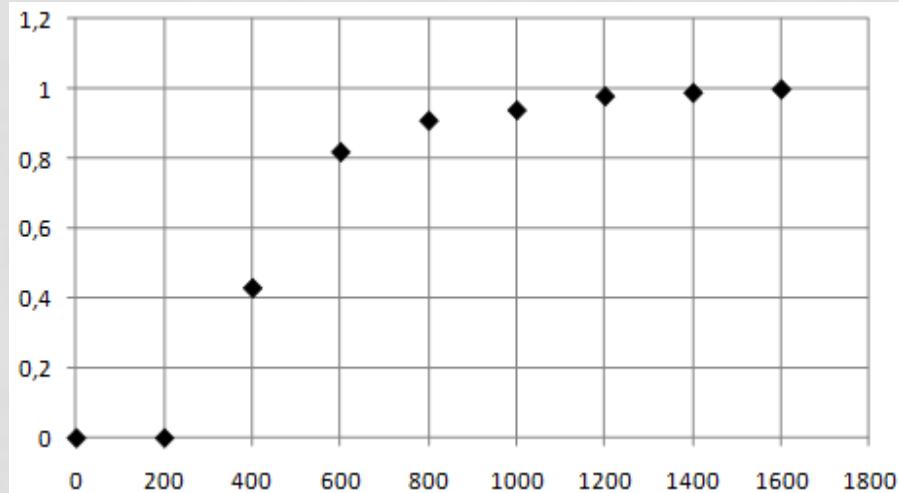
Der Notendurchschnitt ist bei 80 % der Schüler kleiner oder gleich 3.

EMPIRISCHE VERTEILUNGSFUNKTION

Beispiel für eine andere Darstellung der Treppenfunktion:

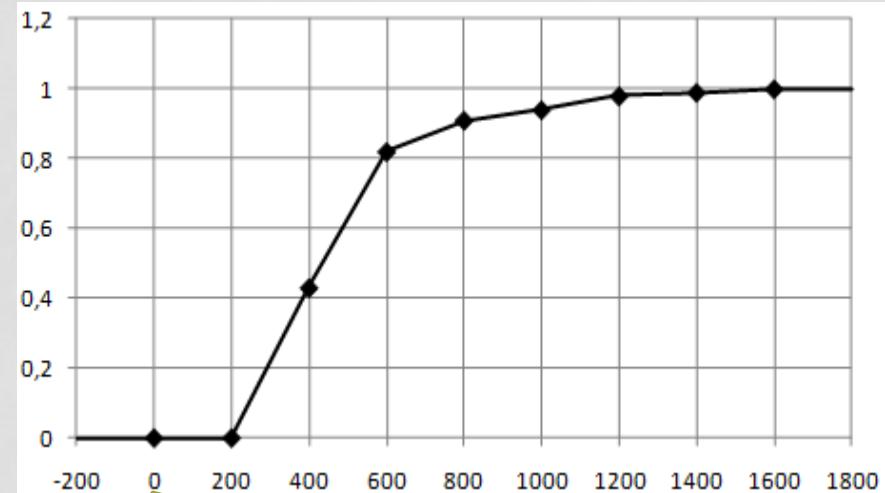


EMPIRISCHE VERTEILUNGSFUNKTION BEI KLASSIERTEN DATEN



↑

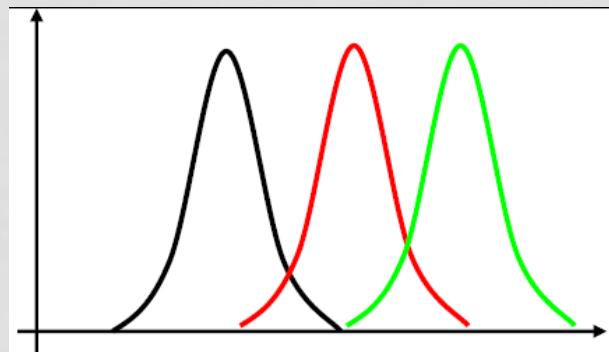
Obere Klassengrenze



↑

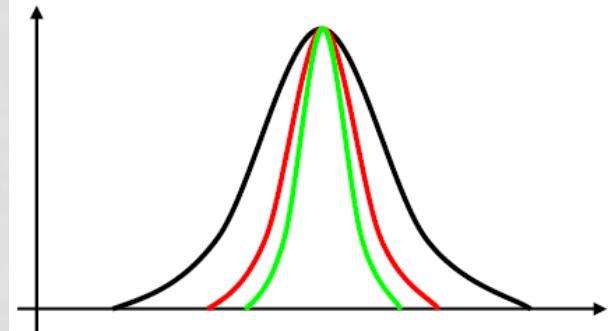
EIGENSCHAFTEN DER HÄUFIGKEITSVERTEILUNGEN

Lage

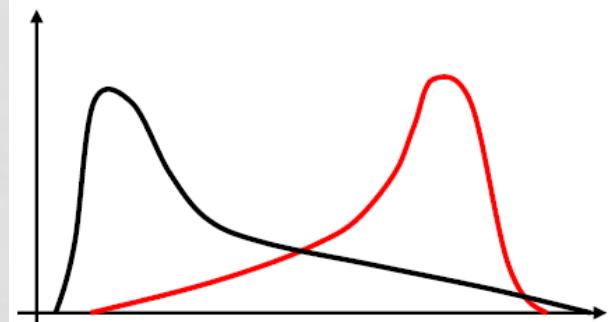


Streuung

= Wölbung, Form



Schiefe



GRAFISCHE DARSTELLUNG DER HÄUFIGKEITSVERTEILUNG

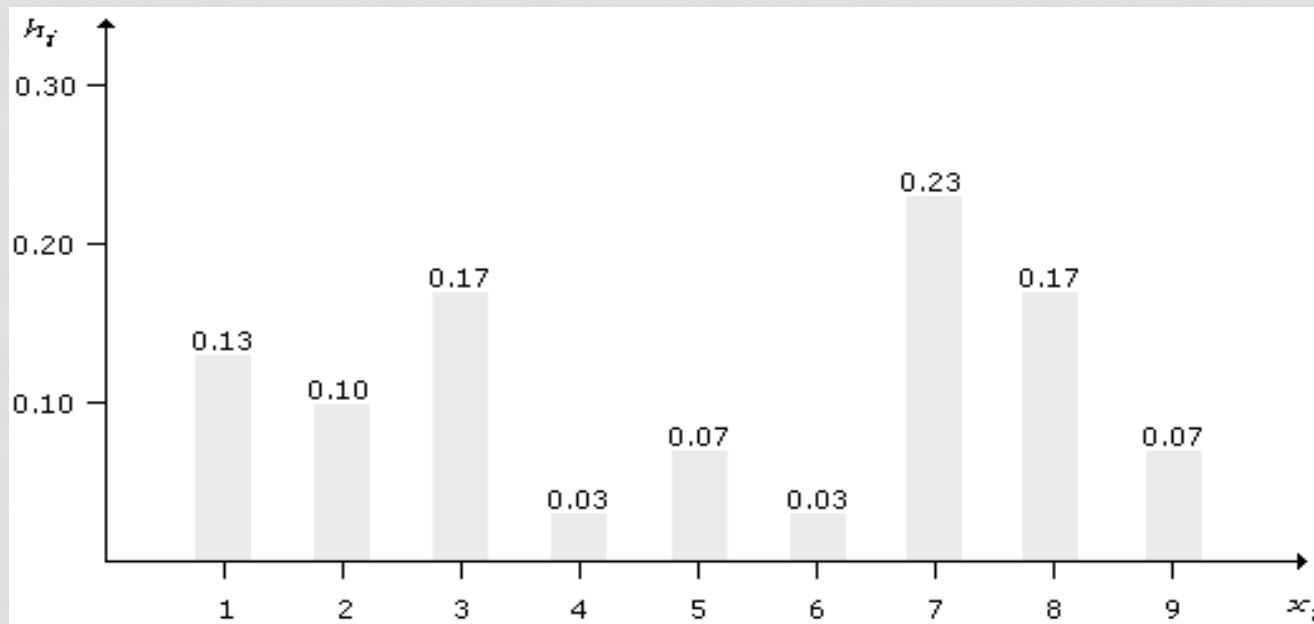
- **Ziel:**
 - ein anschauliches Bild der Daten
 - das Wesentliche der Verteilung aufzuzeigen
- **Wahlentscheidung:**
 - Form der grafischen Darstellung
 - Achsenmaßstab
 - Evtl. Ausschnitt darstellen

→ Manipulationen sind denkbar (optische Täuschung!)
- **Die am weitesten verbreiteten grafischen Darstellungsformen:**
 - Säulendiagramm
 - Stabdiagramm
 - Balkendiagramm
 - Kreisdiagramm
 - Histogramm (bei klassierten Daten)

SÄULENDIAGRAMM

Säulendiagramm

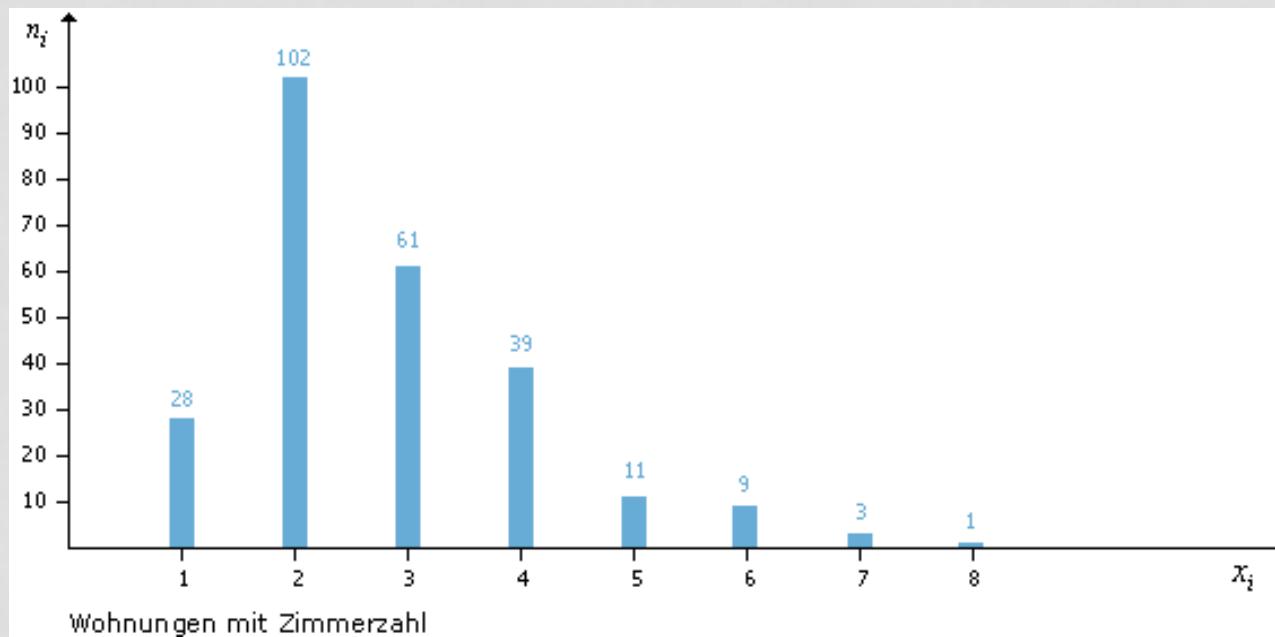
- höhenproportionale Darstellungsform einer Häufigkeitsverteilung durch auf der **x-Achse senkrecht stehende, nicht aneinander grenzende Säulen** (mit beliebiger Breite)
- eignet sich besonders, um wenige Ausprägungen zu veranschaulichen. Bei mehr als 15 Kategorien leidet die Anschaulichkeit und es sind Liniendiagramme zu bevorzugen.



STABDIAGRAMM

Stabdiagramm / Liniendiagramm

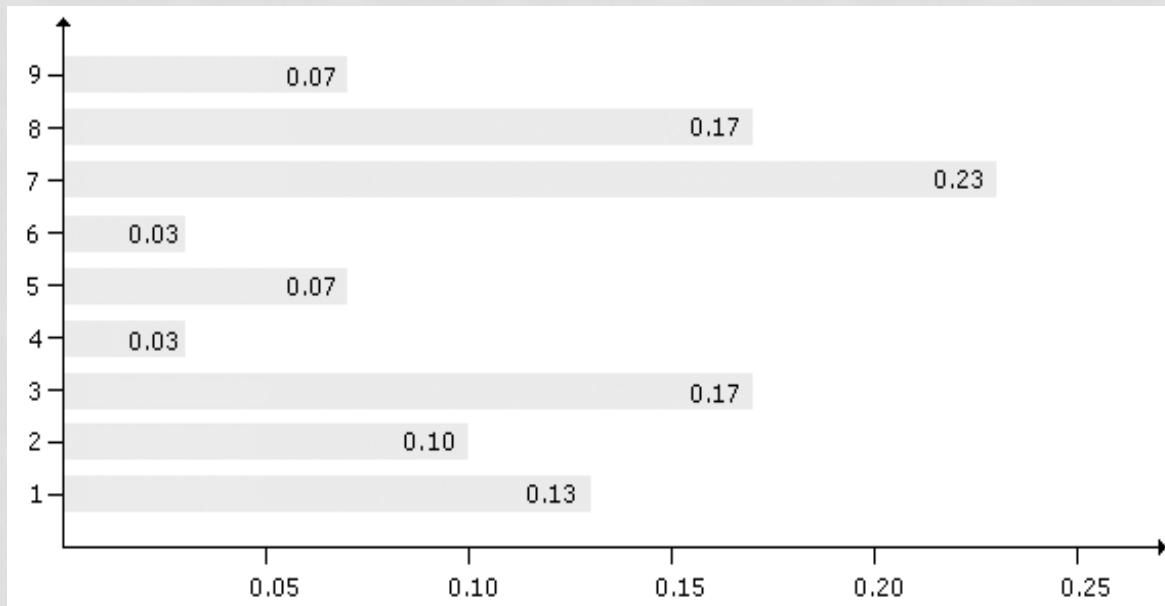
- Stabdiagramm = Säulendiagramm mit schmalen Säulen
- Liniendiagramm = Säulendiagramm mit sehr schmalen Säulen in Breite einer Linie



BALKENDIAGRAMM

Balkendiagramm

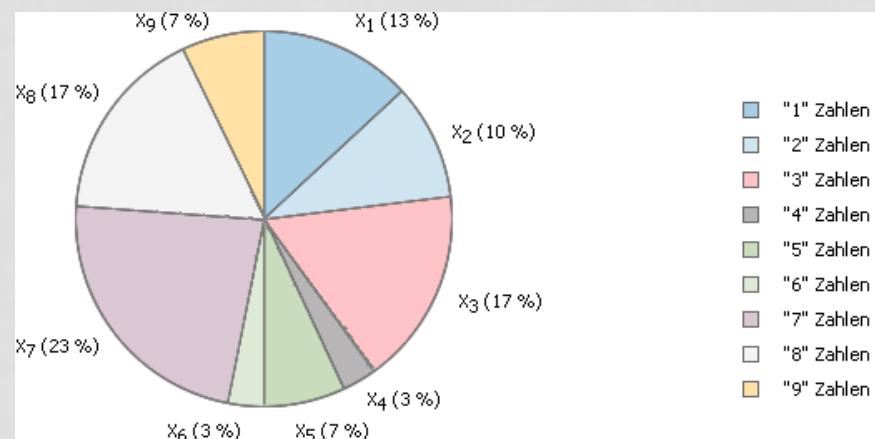
- einer der am häufigsten verwendeten Diagrammtypen
- Balkendiagramm = Säulendiagramm mit **horizontalen Balken**
- eignet sich sehr gut zur Darstellung von Rangfolgen (= Reihenfolge mehrerer vergleichbarer Objekte, deren Sortierung eine Bewertung festlegt, z.B. hier Weltrangliste in Musik)



KREISDIAGRAMM

Kreisdiagramm (Kuchen- oder Tortendiagramm)

- **Kreisförmig, in mehrere Sektoren eingeteilt**, wobei jeder Kreissektor einen Teilwert und der Kreis somit die Summe der Teilwerte (das Ganze) darstellt
- Faustregel: max. 7 Teilwerte, sonst unübersichtlich. Zur besseren Übersichtlichkeit die Teilwerte im Uhrzeigersinn der Größe nach sortieren
- eignet sich zur Darstellung von diskreten Daten, besonders für das Nominal- und das Ordinalskalenniveau zu empfehlen.
- Verwenden wenn:
 - nur eine Datenreihe wird dargestellt
 - keine negativen Werte auftreten
 - keine Nullwerte vorhanden sind
 - die Kategorien Teile des gesamten Kreisdiagramms repräsentieren



HISTOGRAMM

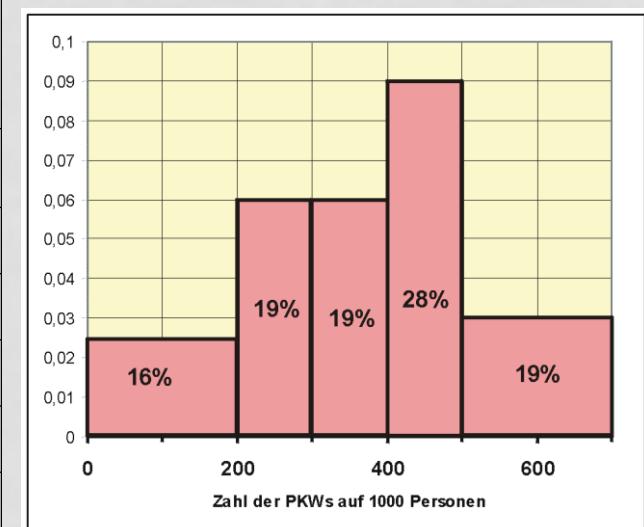
Histogramm

- grafische **flächenproportionale Darstellung der Häufigkeiten von klassierten Daten**
- Im Unterschied zum Säulendiagramm muss bei einem Histogramm die x-Achse immer eine Skala sein, deren Werte geordnet sind und gleiche Abstände haben
- direkt **nebeneinanderliegende Rechtecke** (keine Abstände dazwischen) **der Breite der jeweiligen Klasse**
- Absolute oder relative Häufigkeiten der Klassen werden durch die Flächen der Rechtecke dargestellt: **Fläche = Breite x Höhe**
 - Die Breite der Rechtecke entspricht der Breite der Klasse
 - Die Höhe der Rechtecke entspricht den Klassenhäufigkeiten
 - Die Fläche eines Rechtecks = $c \cdot f(x_j)$, wobei $f(x_j)$ die relative Klassenhäufigkeit der Klasse j und c ein Proportionalitätsfaktor ist. Ist c gleich dem Stichprobenumfang ($c = n$), so ist die Fläche eines jeden Rechtecks gleich der absoluten Klassenhäufigkeit. Das Histogramm wird absolut genannt wenn Summe der Flächeninhalte aller Rechtecke = n. Verwendet das Histogramm die relativen Klassenhäufigkeiten ($c = 1$), wird das Histogramm relativ oder normiert genannt (Summe der Flächeninhalte aller Rechtecke ist 1).

HISTOGRAMM

Histogramm

Klasse	Zahl der PKW pro 1.000 Personen	absolute Häufigkeit	Klassenbreite	Rechteckhöhe
i		Anzahl der Länder (h_i)	b_i	$r_i = h_i/b_i$
1	0 b. 200	5	200-0=200	$5/200=0,025$
2	Ü. 200 b. 300	6	300-200=100	$6/100=0,06$
3	Ü. 300 b. 400	6	400-300=100	$6/100=0,06$
4	Ü. 400 b. 500	9	500-400=100	$9/100=0,09$
5	Ü. 500 b. 700	6	700-500=200	$6/200=0,03$
Summe		32	-	-

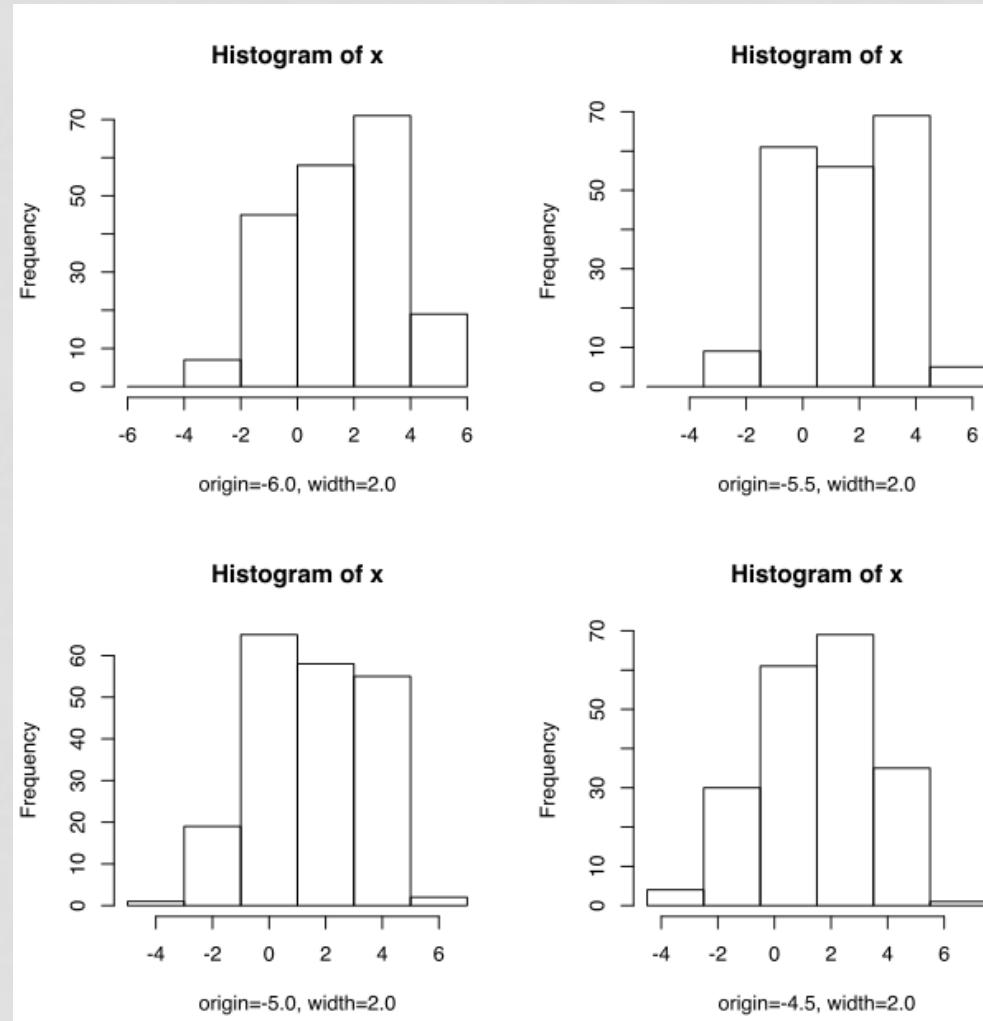


HISTOGRAMM

Beispiel:

Vier Histogramme für den gleichen Datensatz: die Klassenbreiten sind in jedem Histogramm gleich 2.0, aber der Beginn der ersten Klasse verschiebt sich von -6,0 über -5,5 und -5,0 auf -4,5.

Fazit: Neben dem Problem der Klassenanzahl bzw. Klassenbreite spielt also auch die Wahl der (linken) Klassengrenzen eine Rolle



UND NOCH PAAR GRAPHEN...

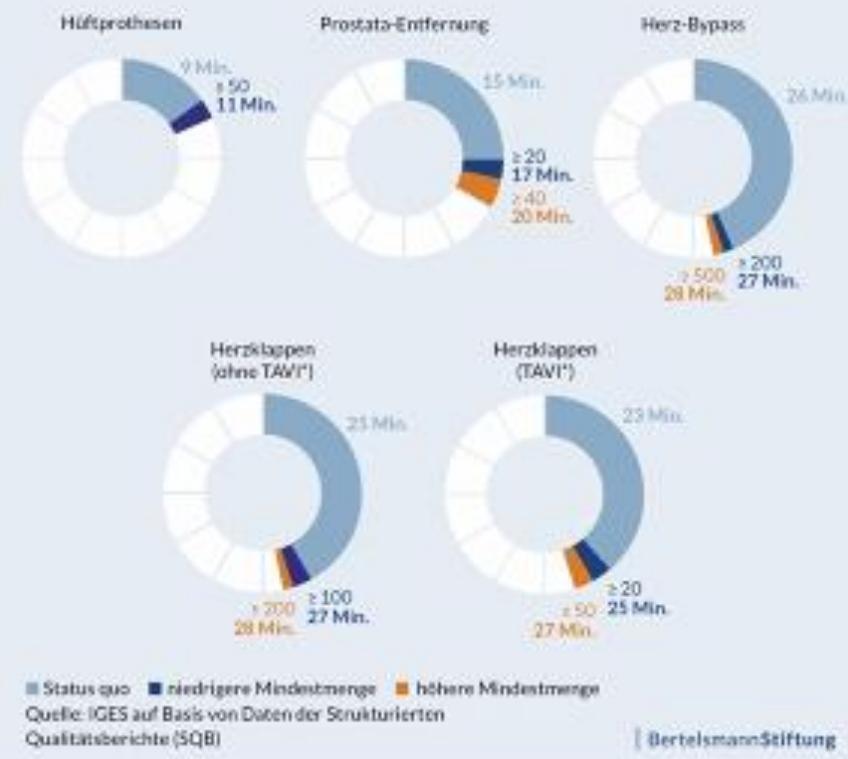
Beispiel:

Segmente

Thema:

Erreichbarkeitsanalyse in der medizinischen Versorgung

Simulation: Wie wirken sich Mindestmengen auf die durchschnittlichen Fahrzeiten zur nächstgelegenen Klinik aus?



UND NOCH PAAR GRAPHEN...

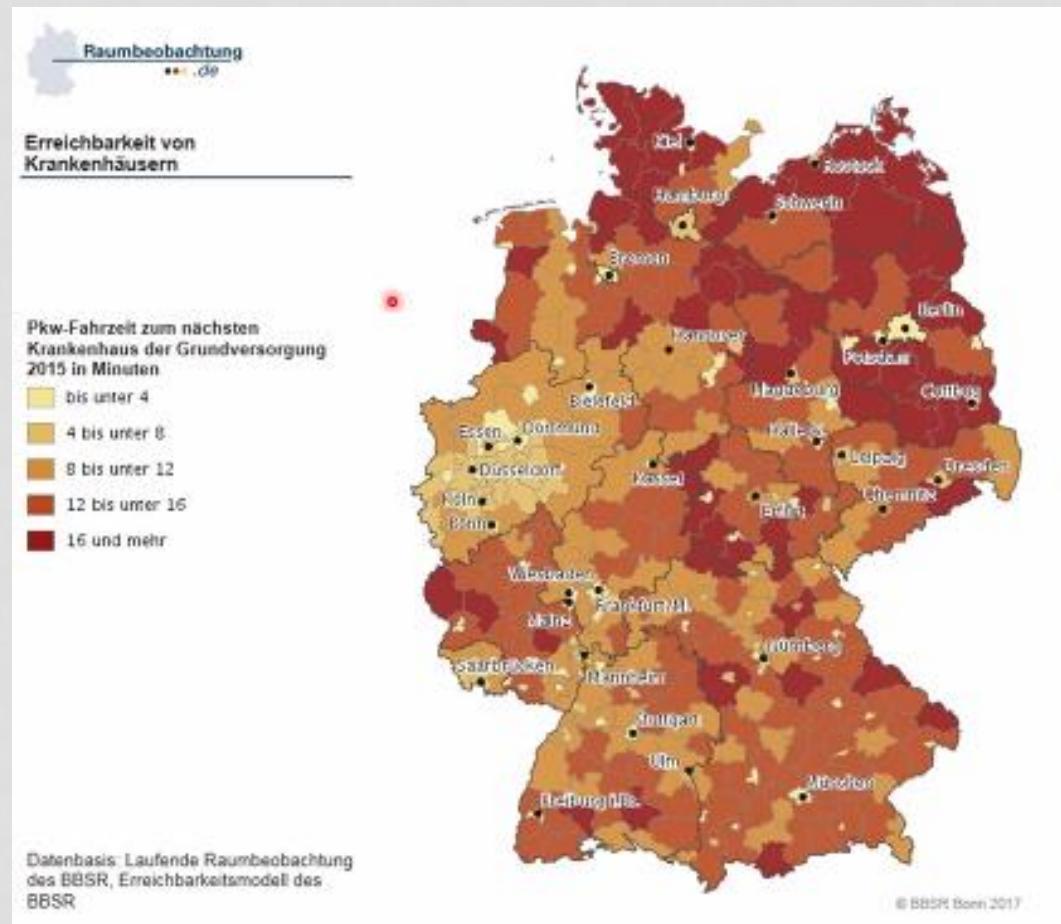
Beispiel:

Wärmekarte

(erzeugt mit dem Programm GIS)

Thema:

Erreichbarkeitsanalyse in der medizinischen Versorgung



UND NOCH PAAR GRAPHEN...

Beispiel:

Punkte-Wärmekarte
(erzeugt mit dem
Programm GIS)

Thema:

Erreichbarkeitsanalyse in der
medizinischen Versorgung



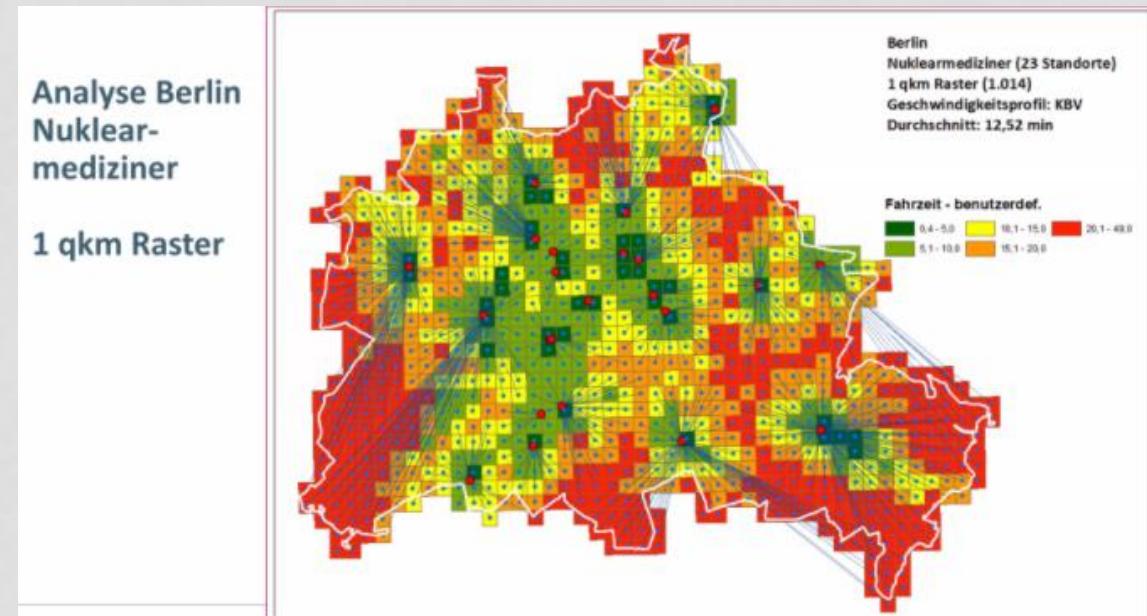
UND NOCH PAAR GRAPHEN...

Beispiel:

Clusteranalyse mit Zentroiden
(die schnellste Strecke zum
roten Punkt wird genommen)
(erzeugt mit dem Programm
GIS)

Thema:

Erreichbarkeitsanalyse in der
medizinischen Versorgung,
Erreichbarkeit in Zeitlicher
Abhängigkeit



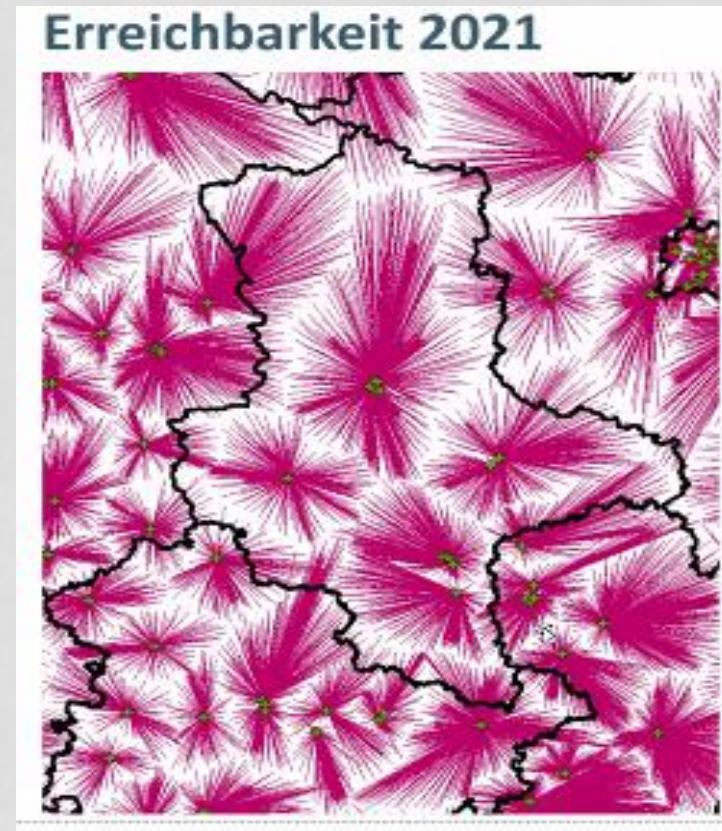
UND NOCH PAAR GRAPHEN...

Beispiel:

Netzwerkanalyse in einer
geographischen Karte
(erzeugt mit dem
Programm GIS)

Thema:

Erreichbarkeitsanalyse in der
medizinischen Versorgung.
Ermittlung der Fahrzeiten
anhand der realen
Straßenprofile.



WIRTSCHAFTSSTATISTIK

MODUL 4: LAGEPARAMETER

WS 2021/22

DR. E. MERINS

LAGEPARAMETER

- Lageparameter beschreiben die “Lage” der Elemente der Grundgesamtheit bzw. der Stichprobe in Bezug auf die Messskala
- noch Lokationsmaße genannt

ALLGEMEINE LAGEPARAMETER

Allgemeine Mittelwerte:

- **Modus** \bar{x}_D
- **Median** \bar{x}_z
- **arithmetisches Mittel** \bar{x}
- **Quantil** \tilde{x}_p

Spezielle Mittelwerte:

- **geometrisches Mittel** \bar{x}_G
- **harmonisches Mittel** \bar{x}_H

Spezielle Mittelwerte werden wir in diesem Kurs nicht berechnen.
Sie müssen aber wissen, dass solche existieren

MODUS

- **Modus (oder Modalwert)** \bar{x}_D

Der **Modus** oder **Modalwert** ist die am häufigsten auftretende Merkmalsausprägung (**maximale Häufigkeit**). Er wird hauptsächlich für nominale Merkmale verwendet, ist aber auch für alle anderen (diskreten) Merkmalstypen sinnvoll.

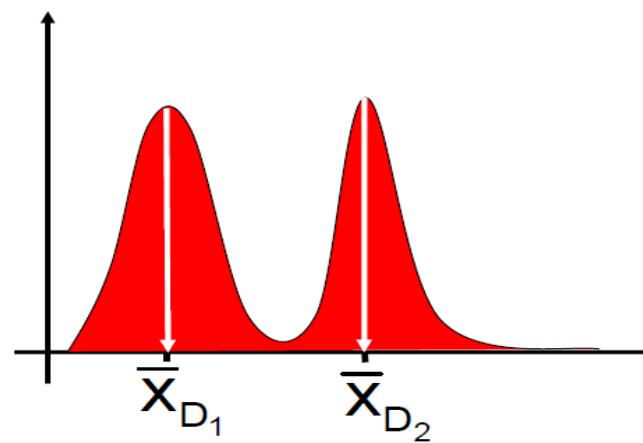
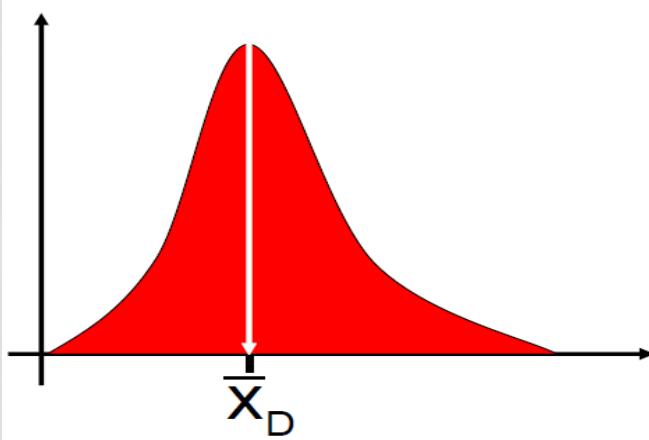
Bei klassierten Daten ist der Modalwert die **Mitte der Klasse mit den größten Häufigkeiten**. Diese Klasse nennt man die **Modalklasse**.

Bemerkung:

Gibt es mehrere Merkmalsausprägungen mit der gleichen maximalen Häufigkeit, so existieren mehrere Modalwerte → **Multimodale Verteilungen** (bimodale Verteilung: zwei Modalwerte; trimodale Verteilung: drei Modalwerte; usw.)

MODUS

- **Modus (oder Modalwert) \bar{x}_D :** Merkmalsausprägung, die am häufigsten vorkommt



unimodale Verteilung

Dichtekurve hat nur
ein lokales Maximum

multimodale Verteilung

Dichtekurve hat mehrere lokale
Maxima (bimodale Verteilung,
Trimodale Verteilung usw.)

MEDIAN

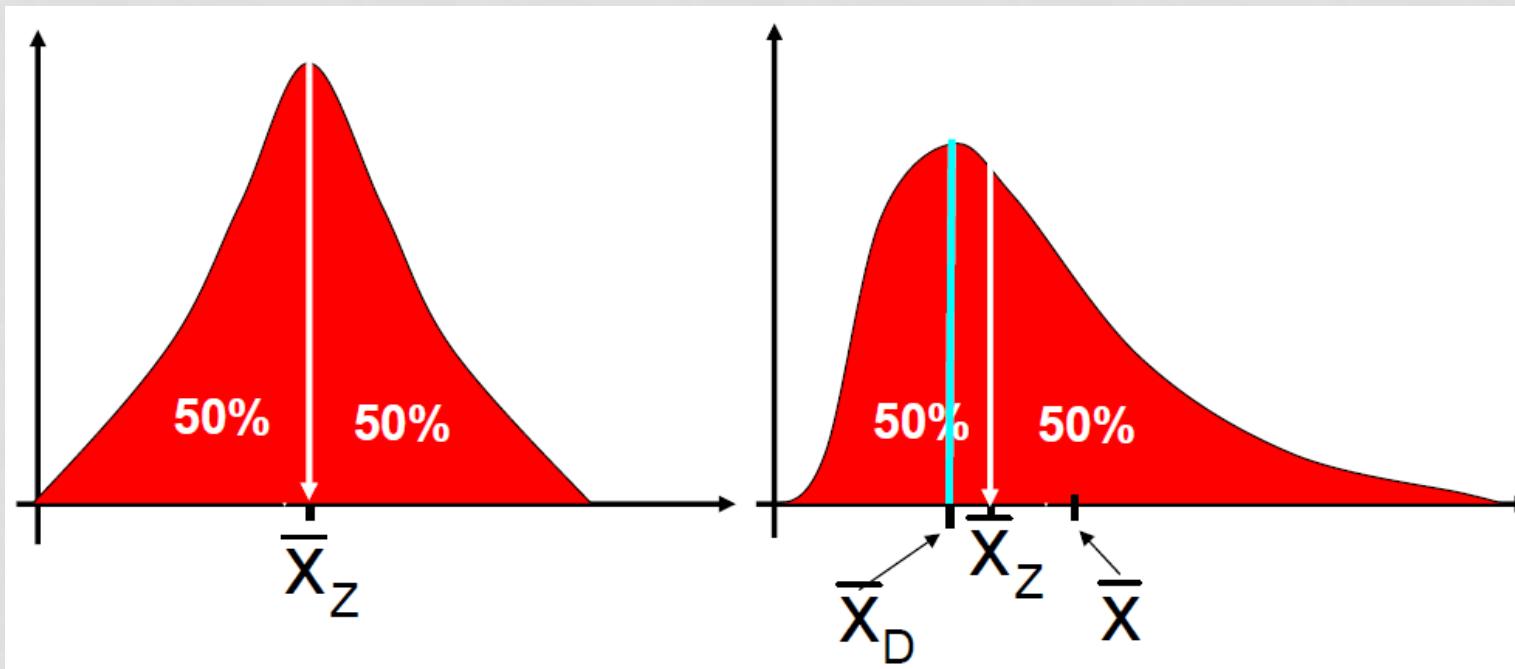
- **Median (oder Zentralwert)** \bar{x}_z

Mindestens 50% der Werte liegen links und mindestens 50% rechts des Medians (den Median selbst ggf. mit eingerechnet).

Median ist ein sehr robustes Lokationsmaß. Robuste statistische Kenngrößen sind wenig anfällig gegen Datenausreißer. Man muss die Hälfte der Daten gegen $+\infty$ oder $-\infty$ verschieben, um den Median selbst gegen $\pm\infty$ wandern zu lassen.

MEDIAN

- **Median (oder Zentralwert) \bar{x}_z** : Wert in der Mitte der geordneten Reihe.
50% der Beobachtungswerte liegen
unter dem Median, 50% darüber.



MEDIAN

- **Median (oder Zentralwert)** \bar{x}_z

Für ordinale und metrische Merkmale ist der empirische Median (oder Zentralwert) definiert als:

$$\bar{x}_z := x_{0,5} := \begin{cases} x_{\frac{n+1}{2}}, & \text{falls } n \text{ ungerade} \\ \frac{1}{2} * (x_{\frac{n}{2}} + x_{\frac{n}{2}+1}), & \text{falls } n \text{ gerade} \end{cases}$$

MEDIAN

- Median (oder Zentralwert) \bar{x}_z

Beispiel 1:

Die Anzahl n der Merkmalsausprägungen ist ungerade, z.B. das Alter von 7 Lehrern
(n = 7)

28	31	40	45	52	53	62
x_1	x_2	x_3	x_4	x_5	x_6	x_7
3 Werte			\bar{x}_z	3 Werte		

$$\bar{x}_z = x_{\frac{n+1}{2}} = x_{\frac{7+1}{2}} = x_4 = 45$$

In der Tabelle stehen links und rechts neben dem Median gleich viele Werte.

MEDIAN

- Median (oder Zentralwert) \bar{x}_z

Beispiel 2:

Die Anzahl n der Merkmalsausprägungen ist gerade, z.B. das Alter von 8 Lehrern
(n = 8)

28	31	40	45	52	53	58	62
x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
3 Werte			\bar{x}_z		3 Werte		

$$\bar{x}_z = \frac{1}{2} * (x_{\frac{n}{2}} + x_{\frac{n}{2}+1}) = \frac{1}{2} * (x_4 + x_5) = \frac{1}{2} * (45 + 52) = 48,5$$

Bei einer geraden Anzahl von Werten berechnet man den Median aus den beiden mittleren Werten.

MEDIAN

- **Median (oder Zentralwert)** \bar{x}_z

Bemerkungen:

Falls das betrachtete Merkmal nur ordinal skaliert ist (z.B. Zeugnisnoten), so ist bei geradem n zu beachten, dass der Median nur dann existiert, wenn beide infrage kommenden Merkmalsausprägungen gleich sind.

Beispiel:

bei den Zeugnisnoten 1 2 3 4 5 6 existiert kein Median, denn 3,5 als Zeugnisnote ist nicht üblich.

Aber: 1 2 3 3 4 5 hat den Median 3.

MEDIAN BEI KLASSIERTEN DATEN

- **Median (oder Zentralwert)** \bar{x}_z

Für metrische Daten in Klassen, kann die exakte Merkmalsausprägung des Medians nicht bestimmt werden → **Näherungswerte für Median**

$$\bar{x}_z := x_{k-1} + (x_k - x_{k-1}) * \frac{0,5 - F_{k-1}}{f_k}$$

wobei $k = \text{Einfallsklasse}$ (Klasse mit $F(x)=50\%$)

MEDIAN BEI KLASSIERTEN DATEN

Schritt 1: Bestimmung der Einfallsklasse k

→ Klasse mit $F(x)=50\%$

Schritt 2: Berechnung des Näherungswertes für Median

→ Näherungswert, weil die Verteilung in den Klassen nicht bekannt ist. Es wird angenommen, dass die Beobachtungswerte in den Klassen gleich verteilt sind.

$$\bar{x}_z = \underline{x}_{k-1}^* + (\overline{x}_k^* - \underline{x}_{k-1}^*) \cdot \frac{(0,5 - F_{k-1})}{f_k}$$

Untergrenze der Einfallsklasse

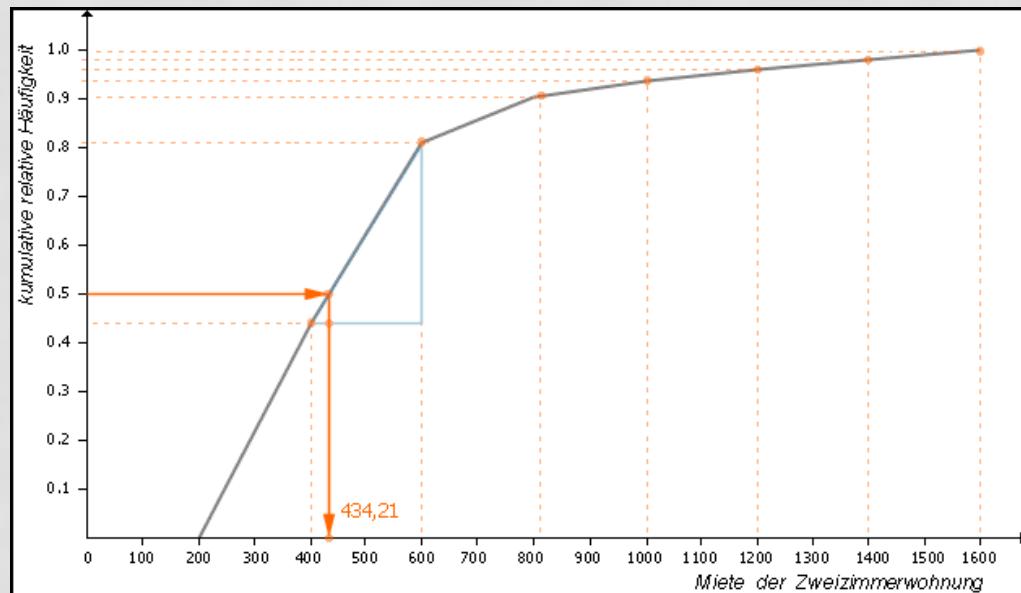
Obergrenze der Einfallsklasse

relative Summenhäufigkeit der Klasse
Unterhalb der Einfallsklasse (Anteilswert)

relative Häufigkeit der Einfallsklasse
Anteilswert

MEDIAN BEI KLASSIERTEN DATEN

Beispiel 1:



Das Ergebnis:

Näherungswert für Median aus klassierten Daten $\bar{x}_z = 434,21 \text{ €}$.

Der tatsächliche Median der Daten ist 423 €.

Achtung!! Es existiert Fehler der Näherung.

MEDIAN BEI KLASSIERTEN DATEN

Beispiel 2:

Klassen-Nr. i	Größenklassen (cm)	h_i	f_i (%)	H_i	F_i (%)
1	100 b.u. 150	40	40%	40	40%
2	150 b.u. 170	40	40%	80	80%
3	170 b.u. 200	20	20%	100	100%
Summe		100	100%	-	-

Einfallsklasse: $k = 2$

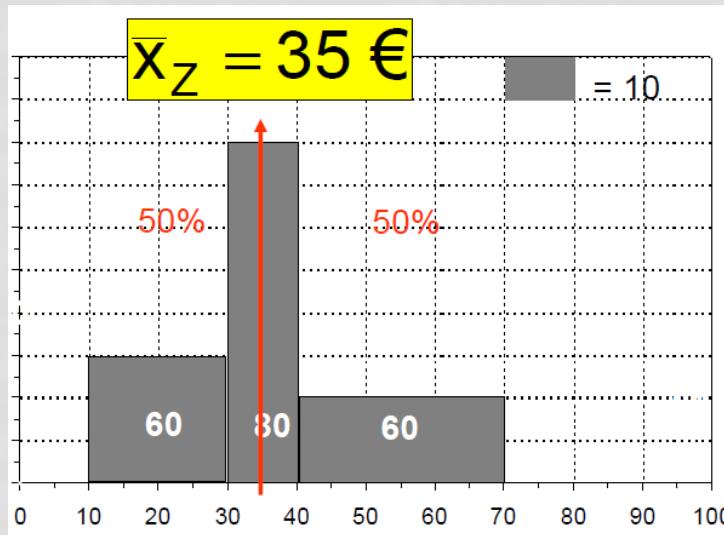
$$\bar{x}_z = x_{k-1}^* + (x_k^* - x_{k-1}^*) \cdot \frac{(0,5 - F_{k-1})}{f_k}$$

$$\bar{x}_z = 150 + (170 - 150) \cdot \frac{(0,5 - 0,4)}{0,4} = 150 + 20 \cdot \left(\frac{0,1}{0,4}\right) = 155 \text{ (cm)}$$

MEDIAN BEI KLASSIERTEN DATEN

Beispiel 3:

Umsatzklassen (€)	Anzahl h_i	Anteil in % f_i	H_i	F_i (%)
10 b.u. 30	60	30%	60	30%
30 b.u. 40	80	40%	140	70%
40 b.u. 70	60	30%	200	100%
Summe	200	100%	-	-



$$\bar{x}_z = x_{k-1}^* + (x_k^* - x_{k-1}^*) \cdot \frac{(0,5 - F_{k-1})}{f_k}$$

Einfallsklasse $k=2$ (30 b.u. 40)

$$\bar{x}_z = 30 + (40 - 30) \frac{0,5 - 0,3}{0,4} = 35 \text{ (€)}$$

ARITHMETISCHES MITTEL

- **Arithmetisches Mittel** \bar{x}

Das **arithmetische Mittel** (oder **Mittelwert**, oder **Durchschnitt** genannt) ist sinnvoll für beliebige metrische Merkmale.

$$\bar{x} = \bar{x}_n = \frac{x_1 + x_2 + \dots + x_n}{n} = \frac{1}{n} \sum_{i=1}^n x_i$$

Statistiker-Witz:

Steht jemand mit einem Fuß auf der Herdplatte und mit dem anderen im Eiskasten, dann sagt der Statistiker: im Durchschnitt ist ihm angenehm warm.

ARITHMETISCHES MITTEL

- **Arithmetisches Mittel** \bar{x}

Eigenschaften:

- Die Summe der Abweichungen der Einzelwerte vom arithmetischen Mittel ist stets gleich null $\sum(x_i - \bar{x}) = 0$
- bekanntester Mittelwert
- nur für quantitative Merkmale sinnvoll
- empfindlich gegen Ausreißer (Vorsicht bei schießen Verteilungen!)

ARITHMETISCHES MITTEL AUS HÄUFIGKEITSTABELLEN

- Arithmetisches Mittel \bar{x}

$$\bar{x} = \frac{1}{n} \sum_{i=1}^j x_i * h(x_i) = \sum_{i=1}^j x_i * f(x_i)$$

x_1, \dots, x_j	<i>Merkmalsausprägungen</i>
$h(x_1), \dots, h(x_j)$	<i>absolute Häufigkeiten</i>
$f(x_1), \dots, f(x_j)$	<i>relative Häufigkeiten</i>

ARITHMETISCHES MITTEL AUS HÄUFIGKEITSTABELLEN

- Arithmetisches Mittel \bar{x}

Fall 1: Absolute Häufigkeit $h(x_i)$

$$\bar{x} = \frac{1}{n} \sum_{i=1}^j x_i * h(x_i) = \frac{1}{n} * (x_1 h(x_1) + x_2 h(x_2) + \dots + x_j h(x_j))$$

$$n = \sum_{i=1}^j h(x_i) = h(x_1) + h(x_2) + \dots + h(x_j)$$

$h(x_i)$ absolute Häufigkeit der Merkmalsausprägung x_i

n Summe der absoluten Häufigkeiten

j Anzahl der Merkmalsausprägung x_i

ARITHMETISCHES MITTEL AUS HÄUFIGKEITSTABELLEN

- Arithmetisches Mittel \bar{x}

Beispiel:

Berechnung des arithmetischen Mittels über die absoluten Häufigkeiten:

Note x_i	1	2	3	4	5	6
Anzahl Schüler $h(x_i)$	5	8	14	16	5	2

Schüler insgesamt:

$$n = \sum_{i=1}^6 h(x_i) = 5 + 8 + 14 + 16 + 5 + 2 = 50$$

Durchschnittsnote:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^j x_i * h(x_i) = \frac{1}{50} * (1 * 5 + 2 * 8 + 3 * 14 + 4 * 16 + 5 * 5 + 6 * 2) = \frac{164}{50} = 3,3$$

ARITHMETISCHES MITTEL AUS HÄUFIGKEITSTABELLEN

- Arithmetisches Mittel \bar{x}

Fall 2: Relative Häufigkeit $f(x_i) = \frac{h(x_i)}{n}$

$$\bar{x} = \sum_{i=1}^j x_i * f(x_i) = (x_1 f(x_1) + x_2 f(x_2) + \dots + x_j f(x_j))$$

$f(x_i)$ relative Häufigkeit der Merkmalsausprägung x_i

n Summe der absoluten Häufigkeiten

j Anzahl der Merkmalsausprägung x_i

ARITHMETISCHES MITTEL AUS HÄUFIGKEITSTABELLEN

- Arithmetisches Mittel \bar{x}

Beispiel:

Berechnung des arithmetischen Mittels über die relativen Häufigkeiten:

Note x_i	1	2	3	4	5	6
Anzahl Schüler $h(x_i)$	5	8	14	16	5	2
Relative Häufigkeit $f(x_i) = h(x_i)/n$	0,1	0,16	0,28	0,32	0,1	0,04

Schüler insgesamt:

$$n = \sum_1^6 h(x_i) = 5 + 8 + 14 + 16 + 5 + 2 = 50$$

Durchschnittsnote:

$$\bar{x} = \sum_{i=1}^j x_i * f(x_i) = 1 * 0,1 + 2 * 0,16 + 3 * 0,28 + 4 * 0,32 + 5 * 0,1 + 6 * 0,04 = 3,3$$

ARITHMETISCHES MITTEL BEI KLASSIERTEN DATEN

Näherungswert für Arithmetisches Mittel

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k m_i * h_i = \sum_{i=1}^k m_i * f_i$$

m_1, \dots, m_k

Klassenmitten (!!)

h_1, \dots, h_k

absolute Klassenhäufigkeiten

f_1, \dots, f_k

relative Klassenhäufigkeiten

→ Näherungswert, weil die Verteilung in den Klassen nicht bekannt ist. Es wird angenommen, dass die Beobachtungswerte jeweils in den Klassenmitten liegen.

ARITHMETISCHES MITTEL BEI KLASSIERTEN DATEN

- Arithmetisches Mittel \bar{x}

Fall 1: Absolute Häufigkeit h_i

$$\bar{x} = \frac{1}{n} \sum_{i=1}^k m_i * h_i = \frac{1}{n} * (m_1 h_1 + m_2 h_2 + \dots + m_k h_k)$$

$$n = \sum_{i=1}^k h_i = h_1 + h_2 + \dots + h_k$$

h_i absolute Häufigkeit der i -ten Klasse

n Summe der absoluten Häufigkeiten

k Anzahl der Klassen

m_i Klassenmitte der i -ten Klasse

ARITHMETISCHES MITTEL BEI KLASSIERTEN DATEN

▪ Arithmetisches Mittel \bar{x}

Beispiel:

klassierte Häufigkeitstabelle für das Körpergewicht, Berechnung über die absoluten Häufigkeiten:

Klasse x_i	41 bis 50	51 bis 60	61 bis 70	71 bis 80	81 bis 90
Häufigkeit h_i	20	15	10	4	1
Klassenmitte m_i	45,5	55,5	65,5	75,5	85,5

Der Häufigkeit wird die Klassenmitte zugeordnet. Man unterstellt, dass alle 15 Schüler z. B. der Klasse x_2 das Körpergewicht 55,5 kg haben.

Schüler insgesamt:

$$n = \sum_{i=1}^5 h_i = 20 + 15 + 10 + 4 + 1 = 50$$

Durchschnittsgewicht:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^j h_i * m_i = \frac{1}{50} * (20 * 45,5 + 15 * 55,5 + 10 * 65,5 + 4 * 75,5 + 1 * 85,5) = \frac{2785}{50} = 55,7$$

ARITHMETISCHES MITTEL BEI KLASSIERTEN DATEN

- Arithmetisches Mittel \bar{x}

Fall 2: Relative Häufigkeit $f_i = \frac{h_i}{n}$

$$\bar{x} = \sum_{i=1}^k m_i * f_i = m_1 f_1 + m_2 f_2 + \dots + m_k f_k$$

f_i relative Häufigkeit der i -ten Klasse

n Summe der absoluten Häufigkeiten

k Anzahl der Klassen

m_i Klassenmitte der i -ten Klasse

ARITHMETISCHES MITTEL BEI KLASSIERTEN DATEN

- **Arithmetisches Mittel** \bar{x}

Beispiel:

klassierte Häufigkeitstabelle für das Körpergewicht, Berechnung über die relativen Häufigkeiten:

Klasse x_i	41 bis 50	51 bis 60	61 bis 70	71 bis 80	81 bis 90
Häufigkeit h_i	20	15	10	4	1
Relative Häufigkeit $f_i = h_i/n$	0,40	0,30	0,20	0,08	0,02
Klassenmitte m_i	45,5	55,5	65,5	75,5	85,5

Schüler insgesamt:

$$n = \sum_1^5 h_i = 20 + 15 + 10 + 4 + 1 = 50$$

Durchschnittsgewicht:

$$\bar{x} = \sum_{i=1}^k m_i * f_i = 45,5 * 0,40 + 55,5 * 0,30 + 65,5 * 0,20 + 75,5 * 0,08 + 85,5 * 0,02 = 55,7$$

MEDIAN VS. MITTELWERT

Beispiel:

Abteilung mit 9 Personen hat folgende Einkünfte in €:

1.200	1.050	950	1.100	900	1.800	6.600	1.150	1.000
-------	-------	-----	-------	-----	-------	-------	-------	-------

$$\bar{x} = 1.660$$

Dieser Durchschnitt liefert ein falsches Bild, weil die Mehrzahl (7 von 9 Personen) höchstens 1.200 € verdient. Der Wert 6.600 € zieht den Mittelwert nach oben.

Man sucht nach einem Wert, der die Verteilung der Einkünfte besser charakterisiert. Dazu werden die Verdienste der Größe nach sortiert:

900	950	1.000	1.050	1.100	1.150	1.200	1.800	6.600
-----	-----	-------	-------	-------	-------	-------	-------	-------

$$\bar{x}_z = x_{\frac{n+1}{2}} = x_{\frac{9+1}{2}} = x_5 = 1.100$$

Der Median beschreibt die Verteilung besser als der Mittelwert, Ausreißer haben auf den Median keinen Einfluss.

NEIGUNG / SCHIEFE

Folgende Faustregel setzt Modus, Median und arithmetisches Mittel in Beziehung:

rechtsschiefe (linkssteile) Häufigkeitsverteilung:

Modus < Median < arithmetisches Mittel

$$\bar{x}_D < \bar{x}_z < \bar{x}$$

linksschiefe (rechtssteile) Häufigkeitsverteilung:

Modus > Median > arithmetisches Mittel

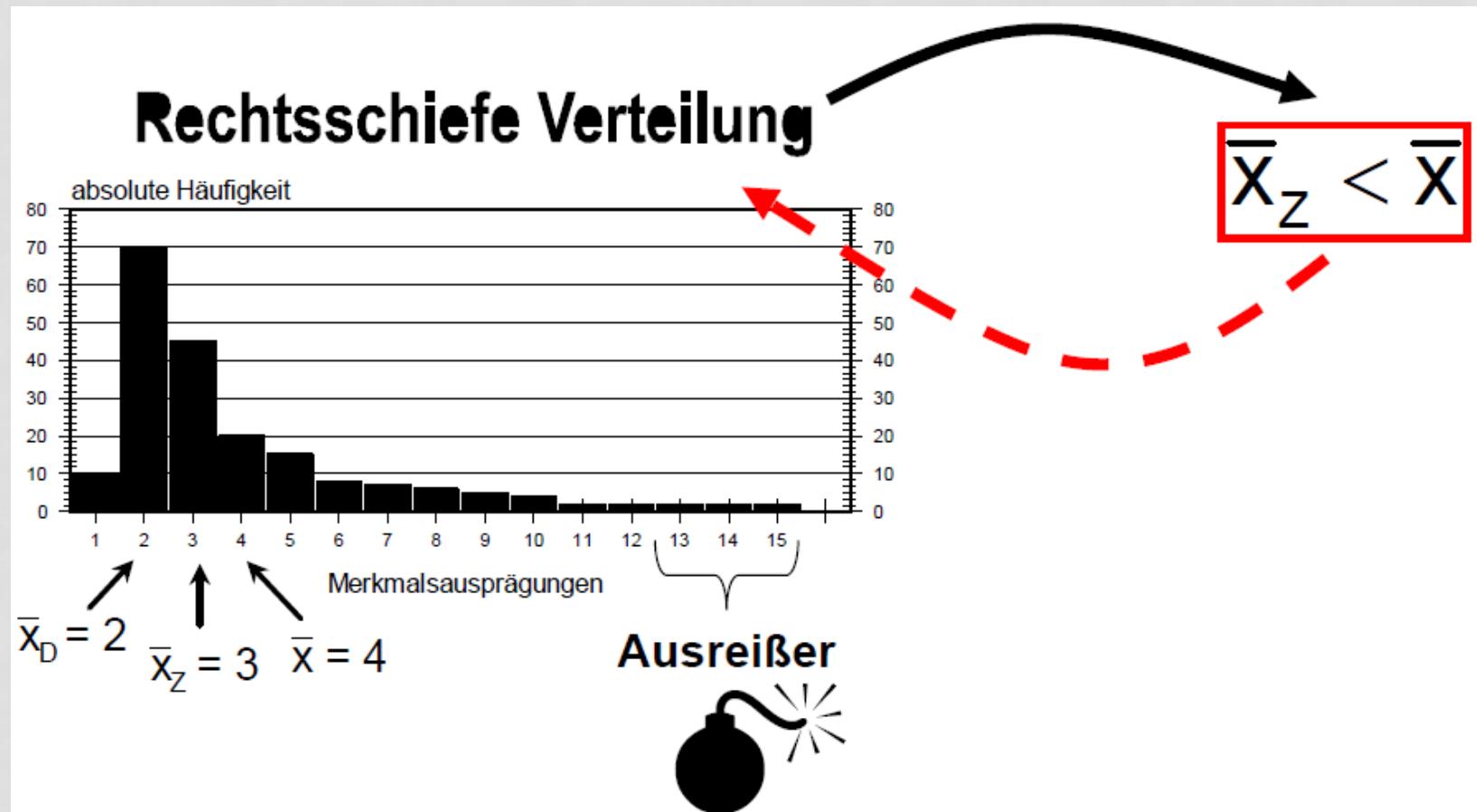
$$\bar{x}_D > \bar{x}_z > \bar{x}$$

unimodale symmetrische Häufigkeitsverteilung:

Modus ≈ Median ≈ arithmetisches Mittel

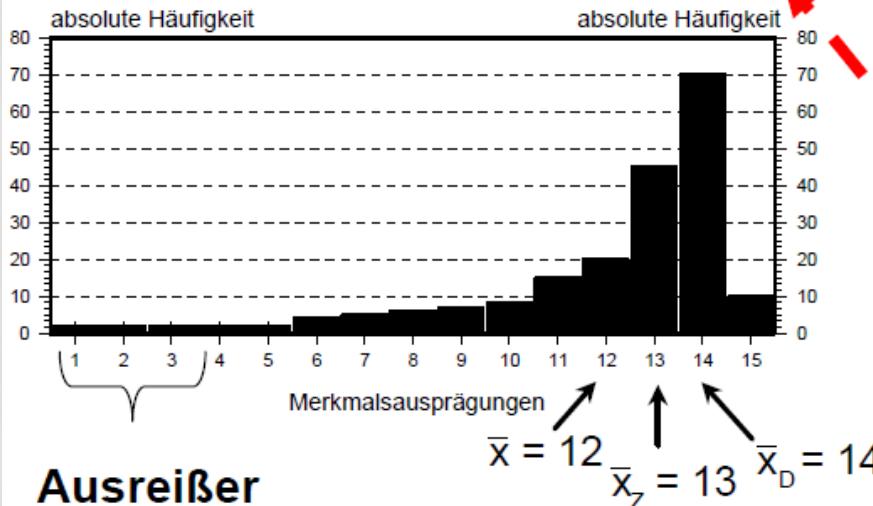
$$\bar{x}_D \approx \bar{x}_z \approx \bar{x}$$

NEIGUNG / SCHIEFE



NEIGUNG / SCHIEFE

Linksschiefe Verteilung



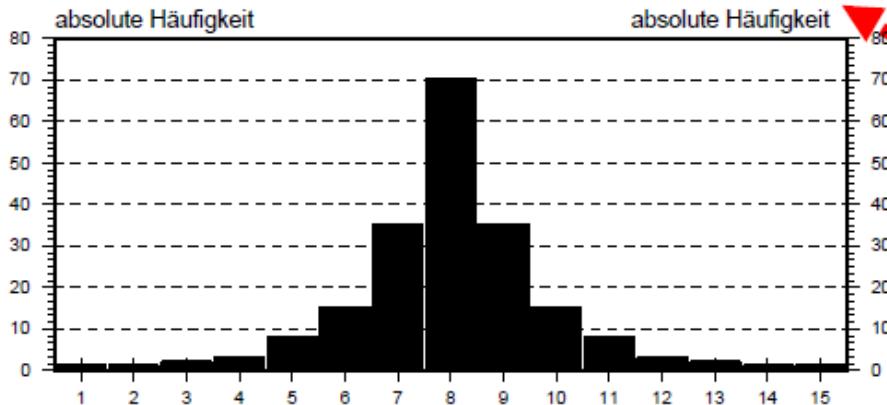
$$\bar{x}_z > \bar{x}$$

Ausreißer



NEIGUNG / SCHIEFE

Symmetrische Verteilung



$\overline{X}_Z = \overline{X}$

$$\overline{X}_D = 8 \quad \overline{X}_Z = 8 \quad \overline{X} = 8$$

BEISPIELÜBUNG

Modus, Median, arithmetisches Mittel, Schiefe

x_i	$h(x_i)$	$f(x_i) (%)$	$H(x_i)$	$F(x_i) (%)$
1	1	8,3%	1	8,3%
2	6	50,0%	7	58,3%
3	2	16,7%	9	75,0%
4	2	16,7%	11	91,7%
9	1	8,3%	12	100,0%
Summe	12	100,0%	-	-

$$\bar{x}_D = 2$$

$$\bar{x}_Z = 2$$

$$\bar{x} = \frac{1}{12}(1 \cdot 1 + 2 \cdot 6 + 3 \cdot 2 + 4 \cdot 2 + 9 \cdot 1) = \frac{36}{12} = 3$$

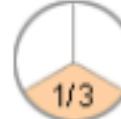
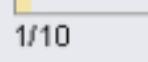
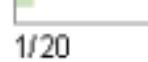
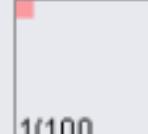
wegen $\bar{x}_Z < \bar{x}$ rechtsschiefe Verteilung

QUANTILE

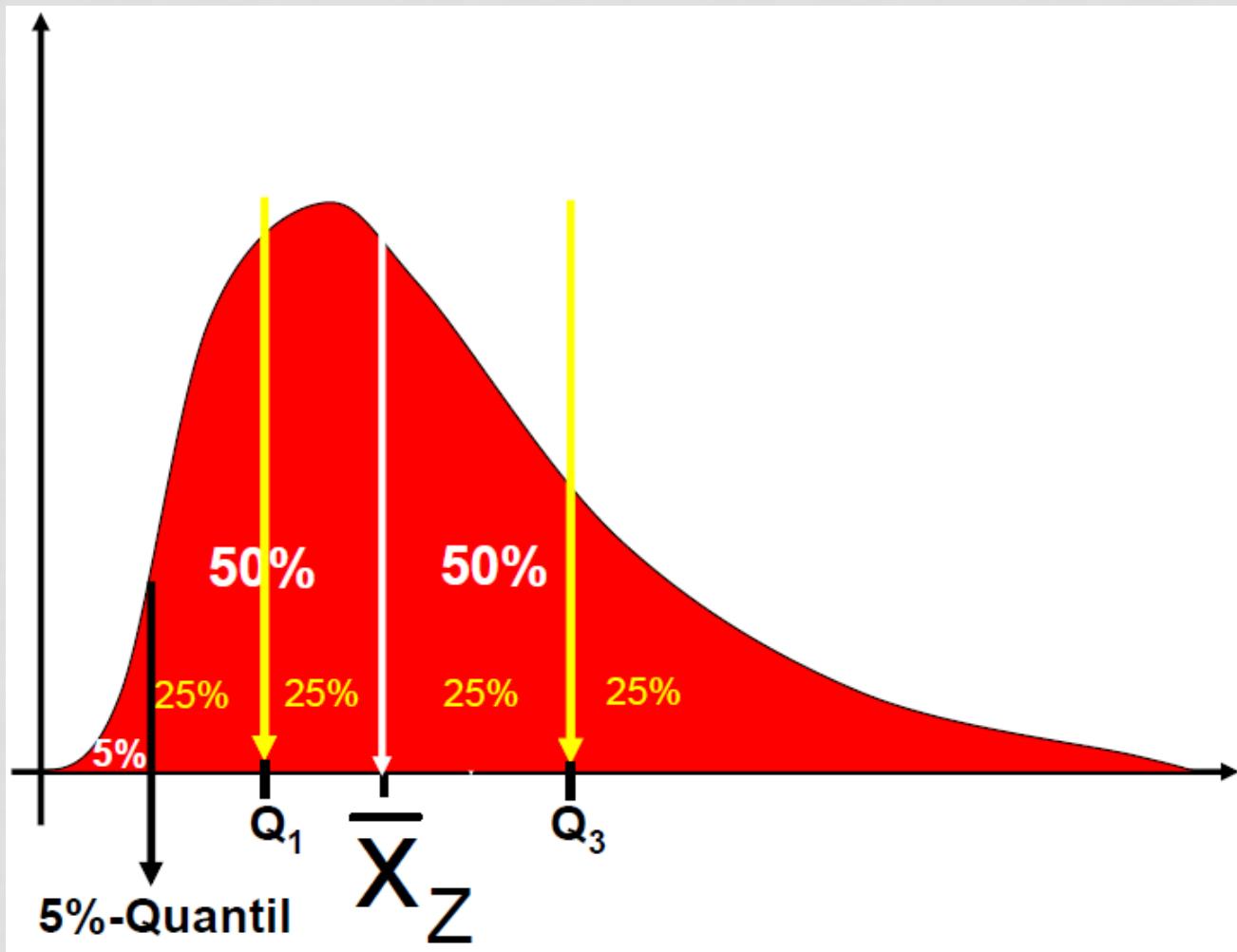
Ein **Quantil** ist ein Lagemaß in der Statistik.

Quantile teilen eine Verteilung in Abschnitte gleicher Häufigkeit.

Spezielle Quantile:

Benennung der Quantile \tilde{x}_p	Anzahl der Intervalle	
Terzile	3	
Quartile	4	
Quintile	5	
Dezile	10	
Vigintile	20	
Perzentile (Zentile)	100	

QUANTILE



WIRTSCHAFTSSTATISTIK

MODUL 5: STREUUNGSPARAMETER

WS 2021/22

DR. E. MERINS

EINLEITUNG

Problem der Lageparameter:

Die Lageparameter schweigen sich aus über die Streuung der Daten.
Das arithmetische Mittel (der Durchschnitt) und auch der Median verdecken oft eine große Ungleichheit.

Statistiker-Witz (frei nach Franz Josef Strauß):

Zwei Männer sitzen im Wirtshaus.

Der eine verdrückt eine ganze Kalbshaxe, der andere trinkt zwei Maß Bier.

Statistisch (im Mittelwert) gesehen ist das für jeden eine Maß Bier und eine halbe Haxe.

Aber in Wirklichkeit der eine hat sich überfressen, und der andere ist besoffen.

→ **die Berechnung des Durchschnitts ist nicht immer sinnvoll**

→ **der Durchschnitt kann offensichtlich nicht immer alles beschreiben**

STREUUNG UM DEN MITTELWERT

Beispiel:

In der folgenden Häufigkeitstabelle und den darauf folgenden Säulendiagrammen ist die Notenverteilung zweier Schülergruppen (Mädchen und Jungen) dargestellt, deren Mittelwert gleich ist.

Schüler Nr.	1	2	3	4	5	6	7	8	9	10	
Note Mädchen	3,2	3,5	2,9	3,3	3,4	2,5	2,7	2,8	3,1	2,6	$\bar{x}=3,0$
Note Jungs	1,0	1,0	2,0	2,5	3,2	2,8	3,5	2,0	6,0	6,0	$\bar{x}=3,0$

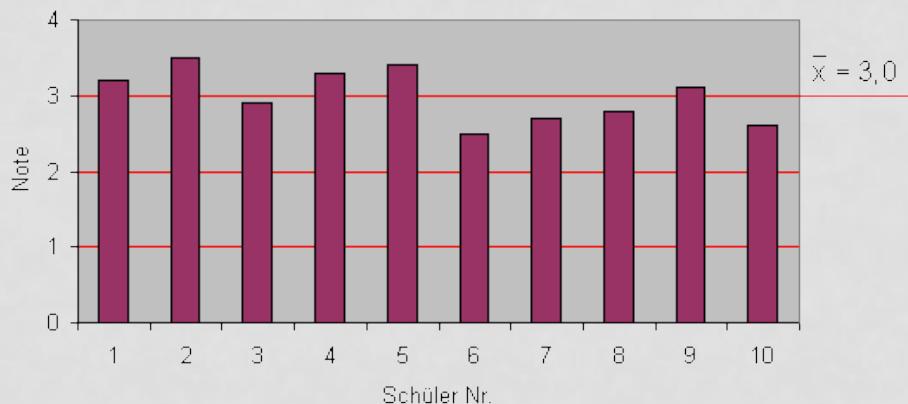
$$\bar{x}_{\text{Mädchen}} = 1/10 * (3,2 + 3,5 + 2,9 + 3,3 + 3,4 + 2,5 + 2,7 + 2,8 + 3,1 + 2,6) = 3,0$$

$$\bar{x}_{\text{Jungs}} = 1/10 * (1,0 + 1,0 + 2,0 + 2,5 + 3,2 + 2,8 + 3,5 + 2,0 + 6,0 + 6,0) = 3,0$$

STREUUNG UM DEN MITTELWERT

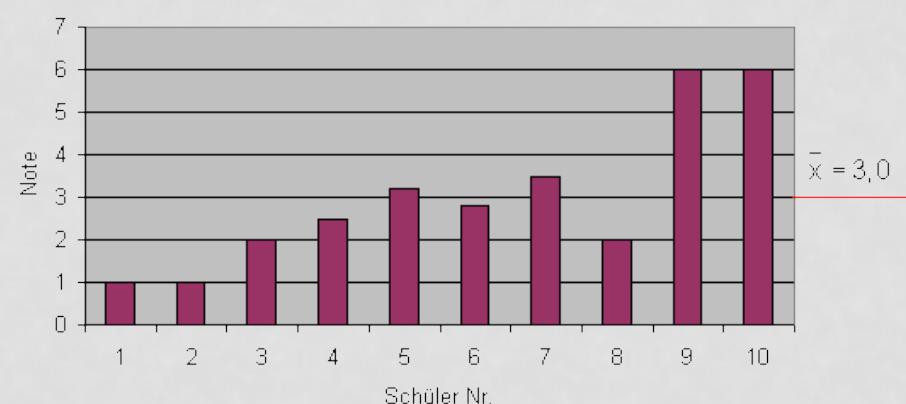
Beispiel:

Notenverteilung Mädchen:



Die Noten liegen alle sehr nahe am Mittelwert
→ Sie streuen wenig um den Mittelwert

Notenverteilung Jungen:



Die Abweichungen vom Mittelwert sind groß
→ Sie streuen stark um den Mittelwert

Die Statistik bietet Möglichkeiten, die **Streuung** näher zu untersuchen und mit Hilfe der **Streuungsparametern** die Streuung zu beschreiben.

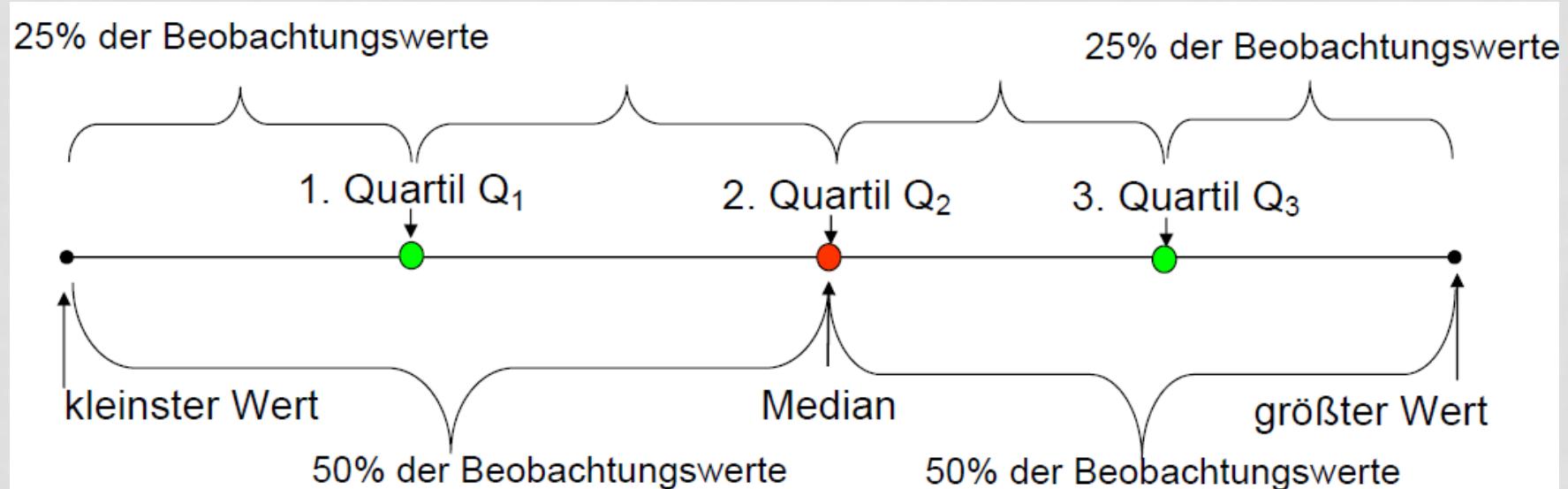
STREUUNGSPARAMETER

Forderungen an eine „gute“ Kennzahl zur Messung der Streuung:

- Bezugspunkt, um den die Werte streuen (\rightarrow Lageparameter)
- alle Beobachtungswerte werden berücksichtigt
- Streuung = 0 (alle Werte sind gleich) \rightarrow Streuungsparameter = 0
- je größer die Streuung, umso größer der Streuungsparameter
- der Streuungsparameter ist unabhängig von der Anzahl der Beobachtungswerte n

QUARTIL

5-Punkte-Zusammenfassung der geordneten statistischen Reihe:



Der (Inter-)Quartilsabstand (engl.: *interquartile range*, **IQR**) bezeichnet die Differenz zwischen dem oberen und dem unteren Quartil **Q₃-Q₁** und umfasst daher 50% der Verteilung.

Der Quartilsabstand wird als **Streuungsmaß** verwendet.

QUARTILSABSTAND

Zusammenfassung:

Der Median teilt einen nach Größe sortierten Datensatz in der Mitte

→ links und rechts vom Median liegen gleich viele Beobachtungswerte. Unterteilt man die linke und die rechte Hälfte nach gleicher Vorschrift, wie man den Median bestimmt, so erhält man 4 gleich große Bereiche, die durch drei Quartils aufgeteilt werden.

25% aller geordneten Beobachtungswerte sind kleiner als das 1. Quartil.

50% aller geordneten Beobachtungswerte sind kleiner als das 2. Quartil.

75% aller geordneten Beobachtungswerte sind kleiner als das 3. Quartil.

Zwischen dem 1. und 3. Quartil liegen 50% aller Beobachtungswerte.

Dieser Bereich wird auch **Quartilsabstand** genannt.

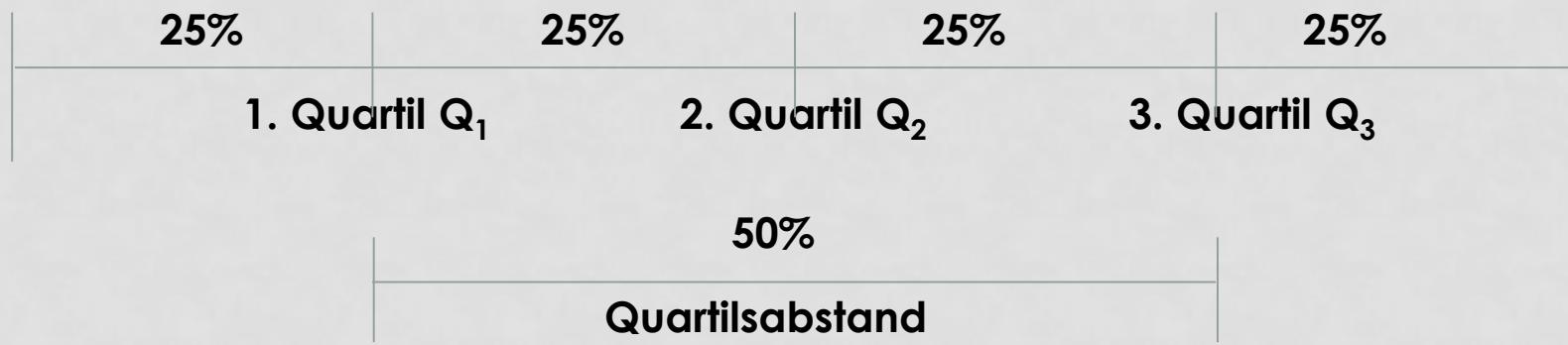
QUARTILSABSTAND

Beispiel:

Die Liste enthält von 13 Schülern die Körpergröße.

Die Merkmalsausprägungen (Beobachtungswerte) wurden nach der Größe geordnet.

Schüler Nr.	1	2	3	4	5	6	7	8	9	10	11	12	13
Größe in m	1,60	1,67	1,67	1,68	1,68	1,70	1,70	1,72	1,73	1,75	1,76	1,78	1,84



QUARTILSABSTAND

Beispiel:



$$\bar{x}_z = Q_2 = x_{\frac{n+1}{2}} = x_{\frac{13+1}{2}} = x_7 = 1,70$$

1. Quartil: $Q_1 = \frac{1}{2}(x_3 + x_4) = \frac{1}{2}(1,67 + 1,68) = 1,675$

3. Quartil: $Q_3 = \frac{1}{2}(x_{10} + x_{11}) = \frac{1}{2}(1,75 + 1,76) = 1,755$

$$Q_A = IQR = Q_3 - Q_1 = 1,755 - 1,675 = 0,08$$

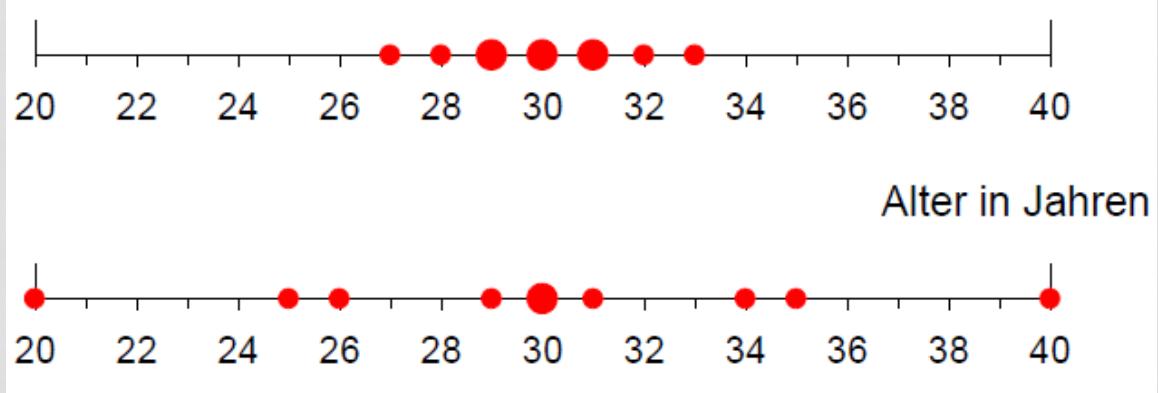
SPANNWEITE

Spannweite (oder Variationsbreite) w : Ausdehnung der Werte (Maß für die Breite des Streubereichs einer Häufigkeitsverteilung)

Für ordinale und metrische Merkmale gilt:

$$w = x_{\max} - x_{\min}$$

Fall 1: $w = 33 - 27 = 6$



Fall 2: $w = 40 - 20 = 20$

SPANNWEITE

$$w = x_{\max} - x_{\min}$$

Beispiel:

Schüler Nr.	1	2	3	4	5	6	7	8	9	10
Note Mädchen	3,2	3,5	2,9	3,3	3,4	2,5	2,7	2,8	3,1	2,6
Note Jungs	1,0	1,0	2,0	2,5	3,2	2,8	3,5	2,0	6,0	6,0

$$w_{Mädchen} = 3,5 - 2,5 = 1$$

$$w_{Jungs} = 6,0 - 1,0 = 5,0$$

QUARTILSABSTAND VS. SPANNWEITE

Vergleich zwischen Quartilsabstand und Spannweite:

Quartilsabstand

Von Ausreißern unabhängig

Gibt die Breite des mittleren Bereichs an, in dem ca. 50% aller Werte liegen

Spannweite

Vom kleinsten und größten Wert abhängig

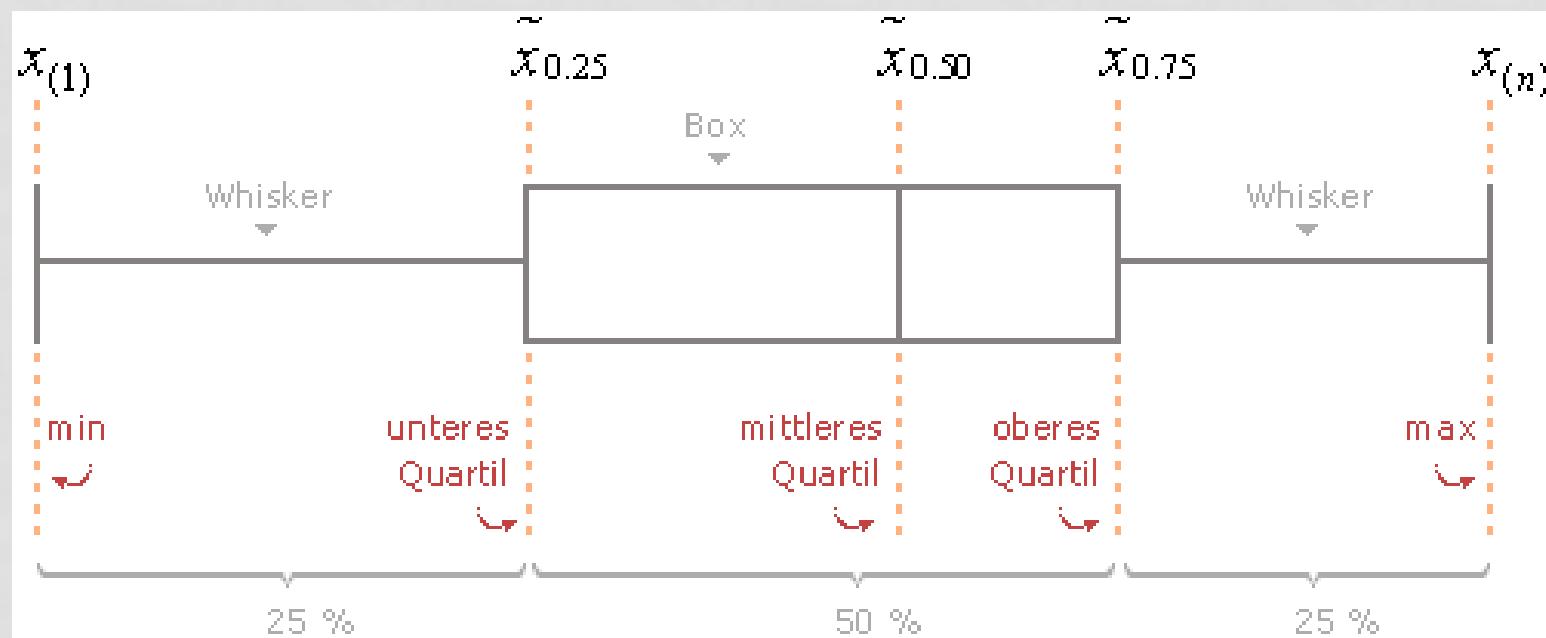
Gibt die Gesamtbreite an, in dem alle Werte liegen

BOXPLOT

Die grafische Darstellung der 5-Punkte-Zusammenfassung heißt
Box-and-Whisker-Plot

Die 5-Punkte-Zusammenfassung besteht aus:

Minimum, Q1, Median, Q3, Maximum



BOXPLOT

Aus einem **Boxplot** lassen sich Informationen über die:

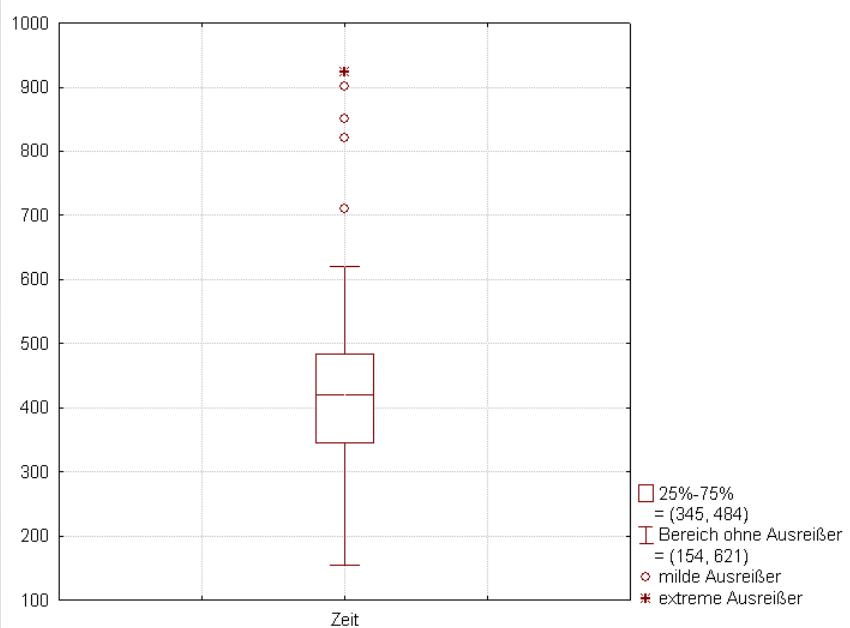
- Lokalisation (Lage des Median)
- Streuungsmaße:
 - **Spannweite** → Ausdehnung eines Boxplots (Differenz $w = x_{\max} - x_{\min}$)
 - **Quartilsabstand** → Ausdehnung der Box (Differenz $IQR = Q_3 - Q_1$)
- Schiefe (Vergleich der beiden Hälften der Box oder der Längen der Whisker)

eines Datensatzes sowie über den evtl. vorliegenden Ausreißer gewinnen.

Eine der Definitionen der Whisker besteht darin, die Länge der Whisker auf maximal das 1,5-Fache des **Interquartilsabstands** ($1,5 \times IQR$) zu beschränken. Der Whisker endet nicht genau nach dieser Länge, sondern bei dem Wert aus den Daten, der noch innerhalb dieser Grenze liegt. Die Länge der Whisker wird also durch die Datenwerte und nicht allein durch den IQR bestimmt. Dies ist auch der Grund, warum die Whisker nicht auf beiden Seiten gleich lang sein müssen. Gibt es keine Werte außerhalb der Grenze von $1,5 \times IQR$, wird die Länge des Whiskers durch den maximalen und minimalen Wert festgelegt. Andernfalls werden die Werte außerhalb der Whisker separat in das Diagramm eingetragen.

BOXPLOT

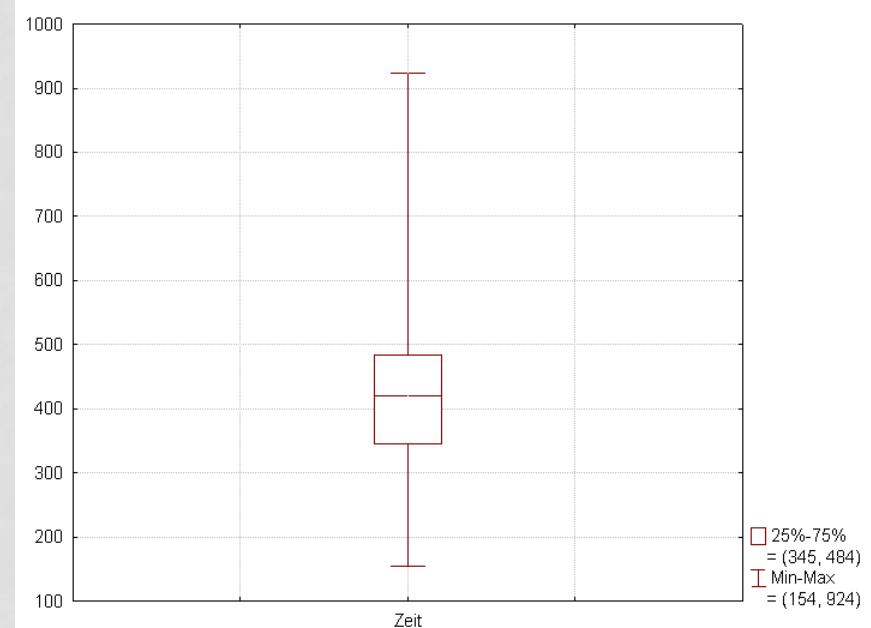
Beispiel:



$$\text{Quartilsabstand: } 484 - 345 = 139$$

$$\text{Spannweite: } 621 - 154 = 467$$

Häufig werden Ausreißer, die zwischen $1,5 \times \text{IQR}$ und $3 \times \text{IQR}$ liegen, als „milde“ Ausreißer bezeichnet und Werte, die über $3 \times \text{IQR}$ liegen, als „extreme“ Ausreißer.



$$\text{Quartilsabstand: } 484 - 345 = 139$$

$$\text{Spannweite: } 924 - 154 = 770$$

VARIANZ

In der beschreibenden Statistik nennt man das arithmetische Mittel der Abweichungsquadrate die **Varianz**.

Eigenschaften:

- wichtiger Streuungsparameter
- Voraussetzung: metrisches Merkmal
- Ausgangswert für weitere folgende Streuungsparameter:
 - **Standardabweichung**
 - **Variationskoeffizient**

→ **Mittelwert und Varianz bzw. Standardabweichung hängen eng zusammen.**

VARIANZ

Konstruktion der Varianz:

Bezugspunkt:

$$\bar{x}$$

Einzelstreuung/Einzelabweichung:

$$(x_i - \bar{x})$$

Summe der Einzelabweichungen:

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

Summe der quadratischen Abweichungen:

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

Varianz:

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

$$s^2 = \frac{1}{n} ((x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = (\underbrace{\frac{1}{n} \sum_{i=1}^n x_i^2}_{\text{Formel (1)}}) - \bar{x}^2$$

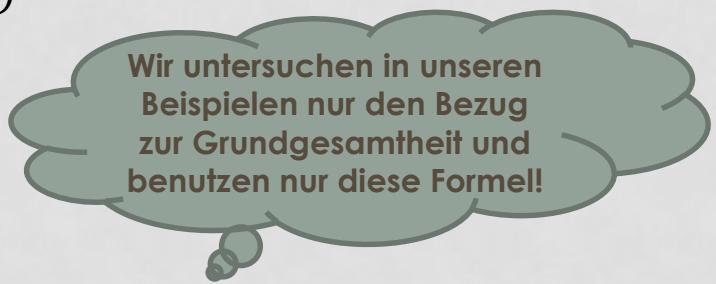
VARIANZ

Konstruktion der Varianz:

Bemerkung:

Handelt es sich bei den zu untersuchenden Daten um die Grundgesamtheit (Population), dann wird mit $1/n$ gewichtet:

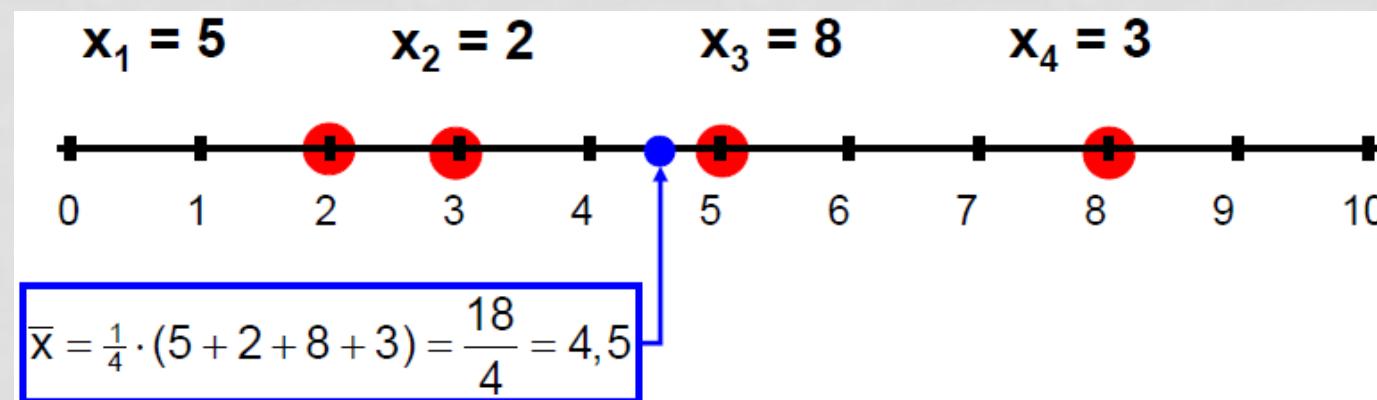
$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$



Wir untersuchen in unseren Beispielen nur den Bezug zur Grundgesamtheit und benutzen nur diese Formel!

VARIANZ

Beispiel:



Berechnung der Varianz

$$s^2 = \frac{1}{4} \cdot ((5 - 4,5)^2 + (2 - 4,5)^2 + (8 - 4,5)^2 + (3 - 4,5)^2) = \\ \frac{1}{4} \cdot (0,25 + 6,25 + 12,25 + 2,25) = \frac{21}{4} = 5,25$$

VARIANZ AUS HÄUFIGKEITSTABELLEN

Formel (1):

$$s^2 = \frac{1}{n} \sum_{i=1}^j (x_i - \bar{x})^2 * h(x_i) = \sum_{i=1}^j (x_i - \bar{x})^2 * f(x_i)$$

Formel (2):

$$s^2 = \left(\frac{1}{n} \sum_{i=1}^j x_i^2 * h(x_i) \right) - \bar{x}^2 = \left(\sum_{i=1}^j x_i^2 * f(x_i) \right) - \bar{x}^2$$

x_1, \dots, x_j **Merkmalsausprägungen**

$h(x_1), \dots, h(x_j)$ **absolute Häufigkeiten**

$f(x_1), \dots, f(x_j)$ **relative Häufigkeiten**

j **Anzahl der Merkmalsausprägung x_i**

VARIANZ AUS HÄUFIGKEITSTABELLEN

Berechnung der Varianz aus einer Häufigkeitstabelle nach **Formel (1)**:

Fall 1: Absolute Häufigkeit h_i

$$n = \sum_{i=1}^j h_i = h_1 + h_2 + \dots + h_j$$

$$s^2 = \frac{1}{n} \sum_{i=1}^j (x_i - \bar{x})^2 * h_i = \frac{1}{n} * ((x_1 - \bar{x})^2 h_1 + (x_2 - \bar{x})^2 h_2 + \dots + (x_j - \bar{x})^2 h_j)$$

h_i absolute Häufigkeit der Merkmalsausprägung x_i

n Summe der absoluten Häufigkeiten

j Anzahl der Merkmalsausprägung x_i

VARIANZ AUS HÄUFIGKEITSTABELLEN

Berechnung der Varianz aus einer Häufigkeitstabelle nach **Formel (1)**:

Fall 2: Relative Häufigkeit f_i

$$s^2 = \sum_{i=1}^j (x_i - \bar{x})^2 * f_i = ((x_1 - \bar{x})^2 f_1 + (x_2 - \bar{x})^2 f_2 + \dots + (x_j - \bar{x})^2 f_j)$$

f_i relative Häufigkeit der Merkmalsausprägung x_i

n Summe der absoluten Häufigkeiten

j Anzahl der Merkmalsausprägung x_i

VARIANZ AUS HÄUFIGKEITSTABELLEN

Beispiel:

Häufigkeitstabelle

Note x_i	1	2	3	4	5	6
Anzahl Schüler h_i	5	8	14	16	5	2
Relative Häufigkeit $f_i = h_i/n$	0,1	0,16	0,28	0,32	0,1	0,04

Schüler insgesamt:

$$n = \sum_1^6 h_i = 5 + 8 + 14 + 16 + 5 + 2 = 50$$

VARIANZ AUS HÄUFIGKEITSTABELLEN

Beispiel:

Berechnung der Varianz über die absolute Häufigkeit:

i	x_i	h_i	$x_i \cdot h_i$	\bar{x}	$x_i - \bar{x}$	$(x_i - \bar{x})^2 h_i$
1	1	5	5	3,28	-2,28	25,992
2	2	8	16	3,28	-1,28	13,107
3	3	14	42	3,28	-0,28	1,098
4	4	16	64	3,28	0,72	8,294
5	5	5	25	3,28	1,72	14,792
6	6	2	12	3,28	2,72	14,797
Σ		50	164	$\bar{x} = 164/50 = 3,28$		78,08

$$s^2 = \frac{1}{50} \sum_{i=1}^{6} (x_i - \bar{x})^2 * h_i = \frac{78,08}{50} = 1,562$$

VARIANZ AUS HÄUFIGKEITSTABELLEN

Beispiel:

Berechnung der Varianz über die relative Häufigkeit:

i	x_i	h_i	f_i	$x_i f_i$	\bar{x}	$x_i - \bar{x}$	$(x_i - \bar{x})^2 f_i$
1	1	5	0,1	0,1	3,28	-2,28	0,520
2	2	8	0,16	0,32	3,28	-1,28	0,262
3	3	14	0,28	0,84	3,28	-0,28	0,022
4	4	16	0,32	1,28	3,28	0,72	0,166
5	5	5	0,1	0,50	3,28	1,72	0,296
6	6	2	0,04	0,24	3,28	2,72	0,296
Σ		50	1	$\bar{x}=3,28$			$s^2=1,562$

$$s^2 = \sum_{i=1}^{6} (x_i - \bar{x})^2 * f_i = 1,562$$

VARIANZ AUS HÄUFIGKEITSTABELLEN

Berechnung der Varianz aus einer klassierten Häufigkeitstabelle nach **Formel (1)**:

Fall 1: Absolute Häufigkeit h_i

$$n = \sum_{i=1}^k h_i = h_1 + h_2 + \cdots + h_k$$

$$s^2 = \frac{1}{n} \sum_{i=1}^k (m_i - \bar{x})^2 * h_i = \frac{1}{n} * ((m_1 - \bar{x})^2 h_1 + (m_2 - \bar{x})^2 h_2 + \cdots + (m_k - \bar{x})^2 h_k)$$

h_i absolute Häufigkeit der i -ten Klasse

n Summe der absoluten Häufigkeiten

k Anzahl der Klassen

m_i Klassenmitte der i -ten Klasse

VARIANZ AUS HÄUFIGKEITSTABELLEN

Berechnung der Varianz aus einer klassierten Häufigkeitstabelle nach **Formel (1)**:

Fall 2: Relative Häufigkeit f_i

$$s^2 = \sum_{i=1}^k (m_i - \bar{x})^2 * f_i = ((m_1 - \bar{x})^2 f_1 + (m_2 - \bar{x})^2 f_2 + \dots + (m_k - \bar{x})^2 f_k)$$

f_i relative Häufigkeit der i -ten Klasse

n Summe der absoluten Häufigkeiten

k Anzahl der Klassen

m_i Klassenmitte der i -ten Klasse

VARIANZ AUS HÄUFIGKEITSTABELLEN

Beispiel:

klassierte Häufigkeitstabelle für die Körpergröße:

Klasse x_i	150 b. u. 160	160 b. u. 170	170 b. u. 180	180 b. u. 190
Häufigkeit h_i	9	12	7	2
Klassenmitte m_i	155	165	175	185
Relative Häufigkeit $f_i = h_i/n$	0,3	0,4	0,23	0,07

Schüler insgesamt:

$$n = \sum_1^4 h_i = 9 + 12 + 7 + 2 = 30$$

VARIANZ AUS HÄUFIGKEITSTABELLEN

Beispiel:

Berechnung der Varianz über die absolute Häufigkeit:

i	Klasse x_i	m_i	h_i	$m_i \cdot h_i$	\bar{x}	$m_i - \bar{x}$	$(m_i - \bar{x})^2 h_i$
1	150 b. u. 160	155	9	1.392	165,67	-10,67	1.024,64
2	160 b. u. 170	165	12	1.980	165,67	-0,67	5,39
3	170 b. u. 180	175	7	1.225	165,67	9,33	609,34
4	180 b. u. 190	185	2	370	165,67	19,33	747,30
Σ			30	4.970	$\bar{x}=4.970/30=165,67$		2.386,67

$$s^2 = \frac{1}{30} \sum_{i=1}^6 (m_i - \bar{x})^2 * h_i = \frac{2.386,67}{30} \approx 80$$

VARIANZ AUS HÄUFIGKEITSTABELLEN

Beispiel:

Berechnung der Varianz über die relative Häufigkeit:

i	Klasse x_i	m_i	h_i	f_i	$m_i f_i$	\bar{x}	$m_i - \bar{x}$	$(m_i - \bar{x})^2 f_i$
1	150 b. u. 160	155	9	0,3	46,5	165,67	-10,67	34,1547
2	160 b. u. 170	165	12	0,4	66,0	165,67	-0,67	0,1796
3	170 b. u. 180	175	7	0,23	40,25	165,67	9,33	20,0212
4	180 b. u. 190	185	2	0,07	12,95	165,67	19,33	26,1554
Σ			30	1	$\bar{x} = 165,6$ 7			80,51

$$s^2 = \sum_{i=1}^{6} (m_i - \bar{x})^2 * f_i \approx 80$$

STANDARDABWEICHUNG

Standardabweichung:

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

Die Standardabweichung ist ein Maß dafür, wie hoch die Aussagekraft des Mittelwertes ist. Eine kleine Standardabweichung bedeutet, alle Beobachtungswerte liegen nahe am Mittelwert (kleine Streuung).

Eine große Standardabweichung bedeutet, die Beobachtungswerte sind weit um den Mittelwert gestreut.

bei normalverteilten Daten liegen ca. 95% der Beobachtungswerte im Intervall $[\bar{x} - 2s, \bar{x} + 2s]$.

STREUUNGSPARAMETER

Spannweite w :

$$w = x_{max} - x_{min}$$

(Inter)Quartilsabstand:

$$Q_A = IQR = Q_3 - Q_1$$

Varianz:

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

Standardabweichung:

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

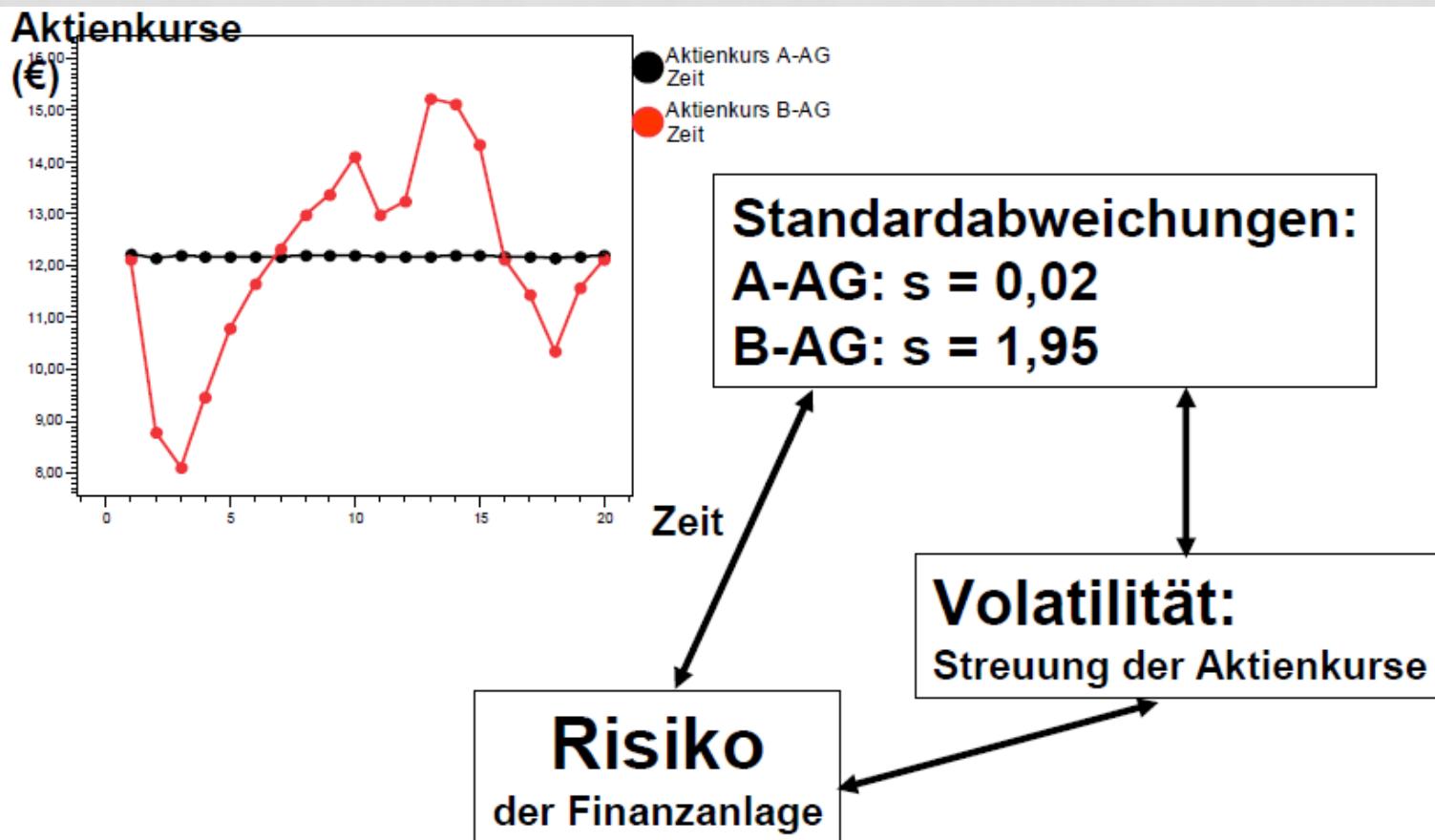
Variationskoeffizient:

(dimensionslose Größe)

$$v = \frac{s}{\bar{x}}$$

STREUUNGSPARAMETER

Beispiel für die Anwendung:



„EINIGE STATISTIKBEGRIFFE IN ENGLISCH“

deutsch

Grundgesamtheit

Stichprobe

arithmetisches Mittel

Modus

Spannweite

Varianz

Standardabweichung

englisch

population

sample

mean

mode

range

variance

standard deviation

(std dev)

WIRTSCHAFTSSTATISTIK

MODUL 6: KORRELATION UND REGRESSION

WS 2021/22

DR. E. MERINS

ZUR GEMEINSAMEN ANALYSE MEHRERER MERKMALE

- Bei einer empirischen Untersuchung werden in der Regel **mehrere Merkmale** gemessen, z.B. X, Y, Z , usw.
- Merkmale werden einzeln ausgewertet (Charakterisierung und Bereinigung) → **univariate Analyse**
- Mehrere Merkmale werden gemeinsam ausgewertet → **multivariate Analyse**

$$X \leftrightarrow Y$$

Zusammenhang (z. B. : Lohnniveau \leftrightarrow Preisniveau)

$$X \rightarrow Y$$

Abhängigkeit (z. B. : Preis \rightarrow Absatzmenge)

ZUR GEMEINSAMEN ANALYSE MEHRERER MERKMALE

- Die **Analyse von Zusammenhängen** ist ein Teilgebiet der **multivariaten** (lat.: multus - vielfach, varia - Allerlei) Statistik. Dabei erfassen statistische Untersuchungen mehrere Merkmale eines Merkmalsträgers gleichzeitig (sogenannte multivariate Datensätze).
- Bei der Analyse von Zusammenhängen können folgende Fragestellungen auftreten:
Besteht überhaupt ein Zusammenhang zwischen Merkmalen (oder Variablen)?
Und wenn ja, dann:
 - Wie stark ist dieser Zusammenhang?
 - Wie lässt sich die Stärke (der Grad, die Intensität) des Zusammenhangs bzw. die Abhängigkeit zwischen zwei (oder mehreren) Merkmalen messen?
 - Lässt sich der Zusammenhang in einer bestimmten Form (Typ, Art) darstellen?
 - Lassen sich die beobachteten Werte einer Variable X durch die Werte einer oder mehrerer anderen Variablen Y (Y_1, Y_2, \dots) näherungsweise bestimmen?

ZUSAMMENHANGSANALYSE

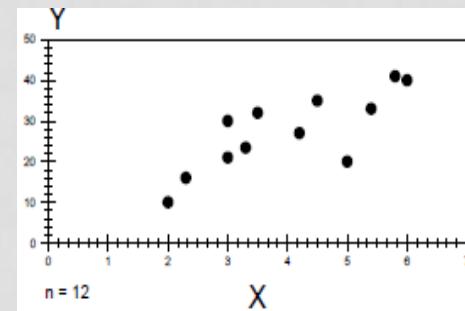
Zusammenhangsanalyse (Interdependenzanalyse)

Es wird eine Wechselwirkung der Variablen untereinander untersucht. Ein

Zusammenhangsmaß, auch Assoziationsmaß genannt, gibt in der Statistik die Stärke und ggf. die Richtung eines Zusammenhangs

Zusammenhangsanalyse zwischen zwei metrischen Merkmalen X und Y

- Zusammenhangsanalyse → Korrelationsanalyse
- Zusammenhangsmaß → Korrelationskoeffizient $-1 \leq r_{xy} \leq +1$
- Grafische Darstellung → Streudiagramm



ZUSAMMENHANGSANALYSE

- **Korrelationsanalyse** (oder Maßkorrelationsanalyse) → wird geprüft, ob zwei Variablen X und Y **linear zusammenhängen** und **wie stark** dieser **Zusammenhang** ist
- Korrelationsanalyse mit dem Spezialfall der **Rangkorrelationsanalyse** → **Zusammenhang zweier ordinalskalierter Merkmale** mit Hilfe von Rangzahlen. Der **Rangkorrelationskoeffizient nach SPEARMAN** hat eine besondere praktische Bedeutung wegen seiner einfachen Berechnung
- **Kontingenzanalyse** (oder Assoziationsanalyse, lat.: contingentia - Zufälligkeit) → **Zusammenhangsanalyse** auf der Basis einer Kontingenztabelle (=Häufigkeitstabelle, s. Modul 03). Je größer der Unterschied zwischen den Häufigkeiten in den Tabellenfeldern ist, umso stärker ist der Zusammenhang bzw. die Abhängigkeit zwischen den Merkmalen

Hinweis: Typen und Arten des Zusammenhangs sind in den Kurs-Materialien „Abschnitt IV. Multivariate Daten Teil 10: ZHA-Zusammenhänge“ gut beschrieben

ZUSAMMENHANGSANALYSE BEI NICHT METRISCHEN MERKMALEN

Rangkorrelationskoeffizient

- ein Maß für die Stärke des Zusammenhangs zweier ordinalskalierter Merkmale
- der Spearmansche Rangkorrelationskoeffizient nutzt Ränge statt der Beobachtungswerte → ein Spezialfall von Pearsons Korrelationskoeffizient, bei dem die Daten in Ränge konvertiert werden, bevor der Korrelationskoeffizient berechnet wird
- benötigt keine Annahme, dass die Beziehung zwischen den Variablen linear ist
- robust gegenüber Ausreißern

Kontingenzkoeffizient

- ein Maß für die Stärke des Zusammenhangs zweier (oder mehrerer) nominaler oder ordinaler Merkmale. Er basiert auf dem Vergleich von tatsächlich ermittelten Häufigkeiten zweier Merkmale mit den Häufigkeiten, die man bei Unabhängigkeit dieser Merkmale erwartet hätte
- kann bei beliebig großen Kreuztabellen angewendet werden
- der Kontingenzkoeffizient C liegt zwischen 0 und +1, d.h., $0 \leq C \leq 1$.

Phi-Koeffizient (auch Vierfelder-Korrelationskoeffizient)

- ein Maß für die Stärke des Zusammenhangs zweier dichotomer Merkmale
- basiert auf einer Kontingenztafel, die die gemeinsame Häufigkeitsverteilung der Merkmale enthält

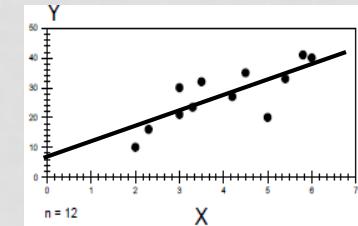
ABHÄNGIGKEITSANALYSE

Abhängigkeitsanalyse (Dependenzanalyse)

Es wird zwischen unabhängigen und abhängigen Merkmalen unterschieden. Es geht um einen gerichteten Zusammenhang. Man hat vorab eine sachlogisch begründete Vorstellung über den Zusammenhang zwischen den Merkmalen, d.h. man weiß oder vermutet, welche der Merkmale auf andere Merkmale einwirken (können).

Abhängigkeitsanalyse zwischen zwei metrischen Merkmalen X und Y

- Abhängigkeitsanalyse → **Regressionsanalyse**
- Abhängigkeitsmaß → **Regressionsfunktion** $\hat{y} = a + b \cdot x$
- **Grafische** Darstellung → **Streudiagramm**
+ **Regressionsgerade**



MULTIVARIATE ANALYSEMETHODEN

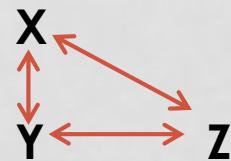
X, Y, Z Merkmale

Beispiel:

Zusammenhangsanalyse (Interdependenzanalyse)

Mitarbeiterzufriedenheit

Kundenzufriedenheit

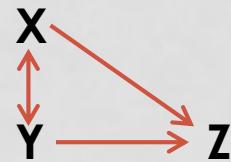


Motivation der Mitarbeiter

Abhängigkeitsanalyse (Dependenzanalyse)

Verkaufsfläche

Anzahl Personal



Filialumsatz

STREUDIAGRAMM

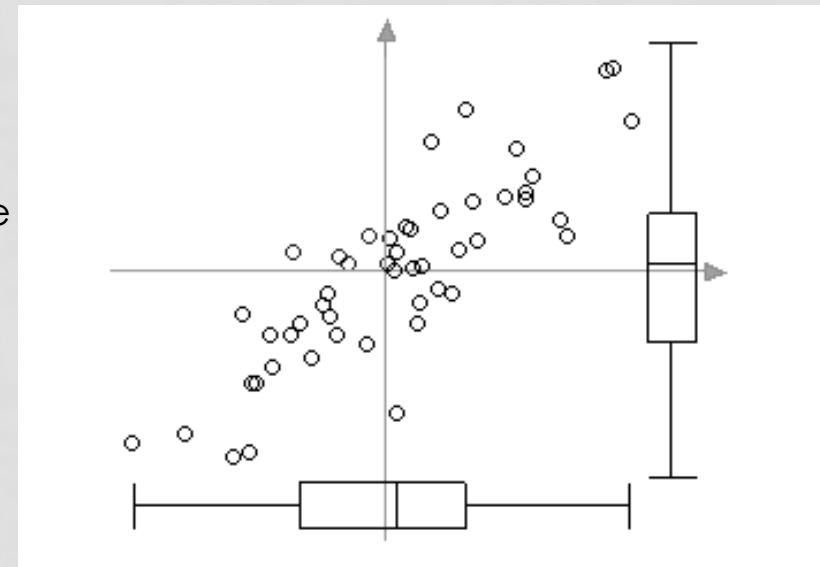
Streudiagramm (oder Streuungsdiagramm)

Ein Streudiagramm (engl. scatter plot) ist die graphische Darstellung von beobachteten Wertepaaren zweier Merkmale. Diese Wertepaare werden in ein kartesisches Koordinatensystem eingetragen, wodurch sich eine Punktwolke ergibt.

Bei beiden Boxplots stimmt der eingetragene Median fast mit der Koordinatenachse überein, es gibt also jeweils etwa gleich viele positive und negative Werte, gemeinsam nehmen die Variablen aber fast nur Werte im I. und im III. Quadranten ein.

Aus Lage und Form der dargestellten Punktwolke lassen sich die **Stärke** und die **Richtung** des Zusammenhangs der Merkmale ablesen.

Das Streudiagramm liefert erste Hinweise über eine mögliche Abhängigkeit zwischen Merkmalen.



KORRELATION

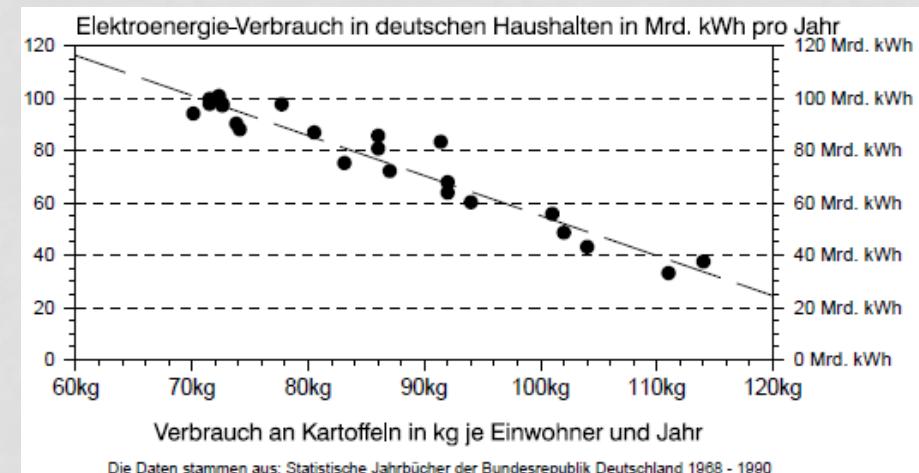
- **Korrelation** → zahlenmäßiger statistischer Zusammenhang zwischen zwei Merkmalen X und Y.
Eine **positive Korrelation** liegt vor, wenn die beiden Merkmale sich gleichförmig entwickeln → bei höheren Werten von X auch Y hohe Werte hat.
Eine **negative Korrelation** liegt vor, wenn X und Y sich gegenläufig entwickeln → bei höheren Werten von X liegen niedrigere Werte von Y vor.
- Ein **kausaler Zusammenhang** zwischen X und Y liegt vor, wenn es zwischen X und Y eine Ursache-Wirkungs-Beziehung gibt, d.h., wenn eine Veränderung des abhängigen Merkmals Y eindeutig auf eine Veränderung von X zurückzuführen ist.
- Eine Korrelation sagt nichts über einen kausalen Zusammenhang aus und auch nichts über eine Kausalitätsrichtung.

PROBLEME BEI DER ABHÄNGIGKEITSANALYSE

Problem: **Abhängigkeitsanalyse muss sinnvoll sein!! (Korrelation \neq Kausalität)**

Kausalität \rightarrow eine Ursache-Wirkungs-Beziehung zw. X und Y, d.h. wenn eine Veränderung des einen Merkmals eine Veränderung bei dem anderen Merkmal hervorruft.

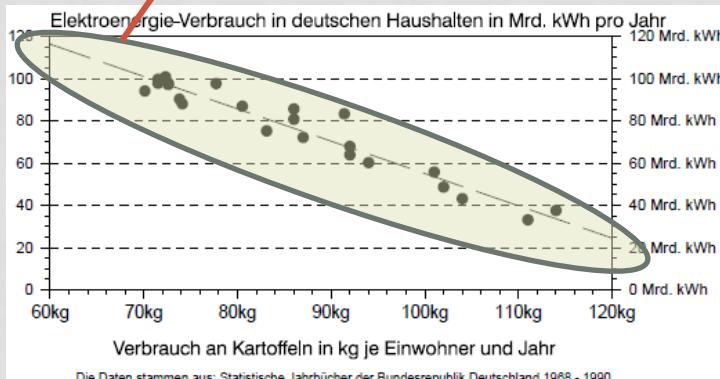
Korrelation \rightarrow ist eine notwendige aber keine hinreichende Voraussetzung für einen kausalen Zusammenhang. Quantifizierung erfolgt über den **Korrelationskoeffizienten**.



SCHEINKORRELATION

Korrelation
statistischer Zusammenhang

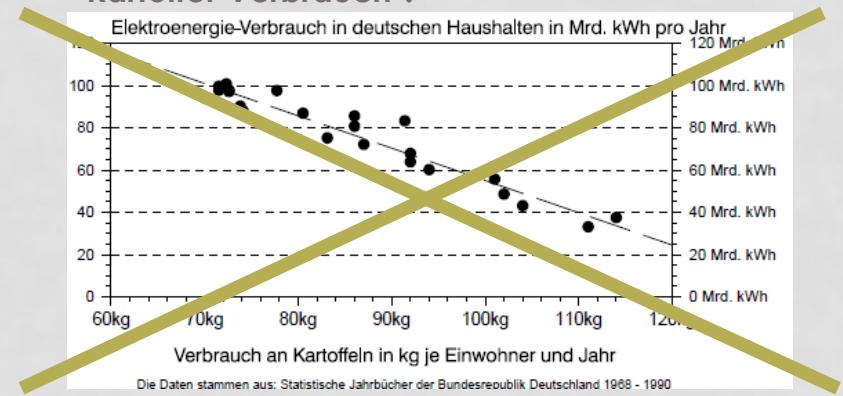
Energie-Sparen durch höheren
Kartoffel-Verbrauch ?



Kausalität
kausaler Zusammenhang =
Ursache-Wirkungs-Beziehung
... denn sonst könnte man den
Energie-Verbrauch durch Kartoffel-
Verbrauch beeinflussen ...

Scheinkorrelation
kein kausaler Zusammenhang

Energie-Sparen durch höheren
Kartoffel-Verbrauch ?



SCHEINKORRELATION

Das wohl berühmteste Beispiel für eine Scheinkorrelation: Der Storch bringt die Babys!

Der Wissenschaftler Robert Matthews fand 2001 eine **Korrelation** nicht unerheblicher Höhe von 0,62 zwischen der Geburtenrate eines Landes und der Anzahl dort lebenden Störche.

Wie kommt diese Korrelation zustande?

Bei Matthews basiert diese hohe Korrelation zu einem großen Teil auf der Größe des Landes: in größeren Ländern leben mehr Störche. Und dort werden mehr Kinder geboren als in kleineren Ländern. Auch eine andere Erklärung wäre denkbar → „**Urbanität vs. Ländlichkeit**“. In der Stadt leben weniger Störche als auf dem Land. Gleichzeitig ist auf dem Land aufgrund soziokultureller Unterschiede die Geburtenrate höher als in der Stadt. Daraus ergibt sich, dass in Gegenden, in denen viele Störche leben, auch die Geburtenrate höher ist.

Auf jeden Fall:

Ein kausaler Zusammenhang liegt nicht vor, der Storch bringt keine Kinder!



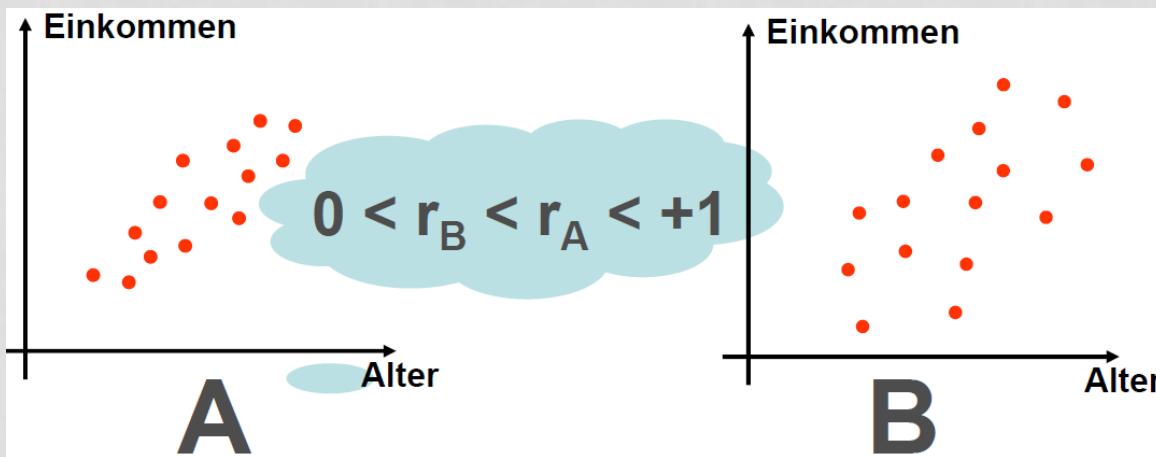
KORRELATION



KORRELATIONSKOEFFIZIENT

Korrelationskoeffizient r_{xy}

$$-1 \leq r_{xy} \leq +1$$



Korrelationskoeffizient → statistische Kennzahl, die informiert über

- die Stärke des linearen Zusammenhangs
- die Richtung des linearen Zusammenhangs

→ Korrelationskoeffizient ist ein dimensionsloses **Maß für den Grad des linearen Zusammenhangs**

KORRELATIONSKOEFFIZIENT

Korrelationskoeffizient r_{xy}

$$-1 \leq r_{xy} \leq +1$$

Positiver Zusammenhang $r > 0$: hohe Werte in der einen Variablen treten tendenziell gemeinsam mit hohen Werten in der anderen Variablen auf.

Negativer Zusammenhang $r < 0$: hohe Werte in der einen Variablen treten tendenziell gemeinsam mit niedrigen Werten in der anderen Variablen auf.

Korrelationskoeffizient $r = -1$: es liegt ein extrem starker negativer linearer Zusammenhang vor → die Punktwolke liegt auf einer Geraden mit negativer Steigung.

Korrelationskoeffizient $r = +1$: es liegt ein extrem starker positiver linearer Zusammenhang vor → die Punktwolke liegt auf einer Geraden mit positiver Steigung.

Korrelationskoeffizient $r = 0$: es liegt kein linearer Zusammenhang vor.

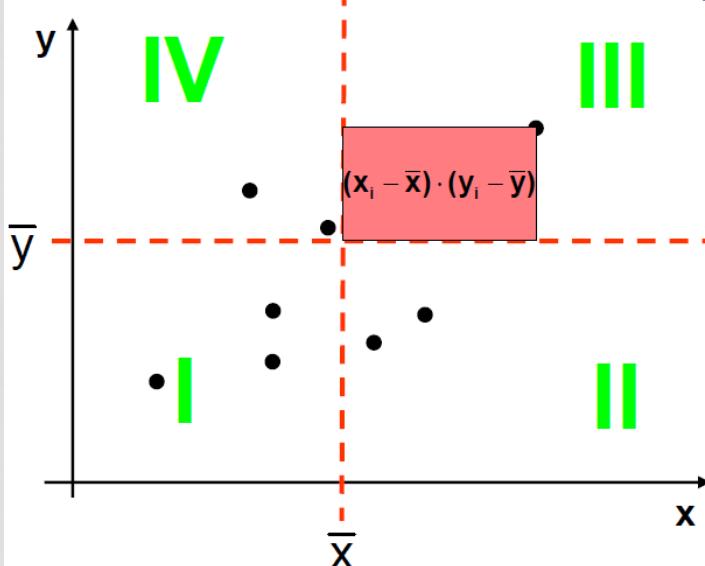
KORRELATIONSRECHNUNG (NACH PEARSON)

Definition:

Für zwei mindestens intervallskalierten Merkmale X und Y mit jeweils positiver Standardabweichung und Kovarianz ist der Korrelationskoeffizient (Pearson'scher Maßkorrelationskoeffizient) definiert durch:

$$r_{XY} = \frac{\text{COV}(X, Y)}{s_X \cdot s_Y}$$

Kovarianz $\text{COV}(X, Y) = \frac{1}{n} \cdot \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})$



(x_i, y_i) in Quadrant	Vorzeichen von $(x_i - \bar{x}) \cdot (y_i - \bar{y})$
I	+
II	-
III	+
IV	-

Die Kovarianz $\text{COV}(X, Y)$ informiert über die gemeinsame Variabilität der beiden Merkmale X und Y. Ist der Zusammenhang positiv, dann ist die Kovarianz positiv, ist der Zusammenhang negativ, dann ist die Kovarianz negativ. Gibt es keinen (linearen) Zusammenhang zwischen X und Y, dann liegt die Kovarianz in der Nähe von 0.

KORRELATIONSRECHNUNG (NACH PEARSON)

„Standardisierung“ der Kovarianz:

$$r_{XY} = \frac{\text{COV}(X, Y)}{s_X \cdot s_Y} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2} \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}} =$$
$$= \frac{\frac{1}{n} \sum_{i=1}^n x_i \cdot y_i - \bar{x} \cdot \bar{y}}{\sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2} \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n y_i^2 - \bar{y}^2}}$$

BEISPIEL

Beispiel:

Verkaufsfläche → Filialumsatz

Filiale Nr. i	x_i Verkaufsfläche (1000qm)	y_i Filialumsatz (Mio €)
1	3	30
2	2	10
3	6	40
4	5	20
Summe	16	100

KORRELATIONSRECHNUNG

Berechnungstabelle mit Hilfsgrößen

Filiale Nr. i	x_i Verkaufsfläche (1000qm)	y_i Filialumsatz (Mio €)	$x_i \cdot y_i$	x_i^2	y_i^2
1	3	30	90	9	900
2	2	10	20	4	100
3	6	40	240	36	1.600
4	5	20	100	25	400
Summe	16	100	450	74	3.000

$$r_{XY} = r = \frac{COV(X, Y)}{s_X \cdot s_Y} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{s_X \cdot s_Y} = \frac{\left(\frac{1}{n} \sum_{i=1}^n x_i \cdot y_i \right) - \bar{x} \cdot \bar{y}}{\sqrt{\left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2} \cdot \sqrt{\left(\frac{1}{n} \sum_{i=1}^n y_i^2 \right) - \bar{y}^2}}$$

KORRELATIONSRECHNUNG

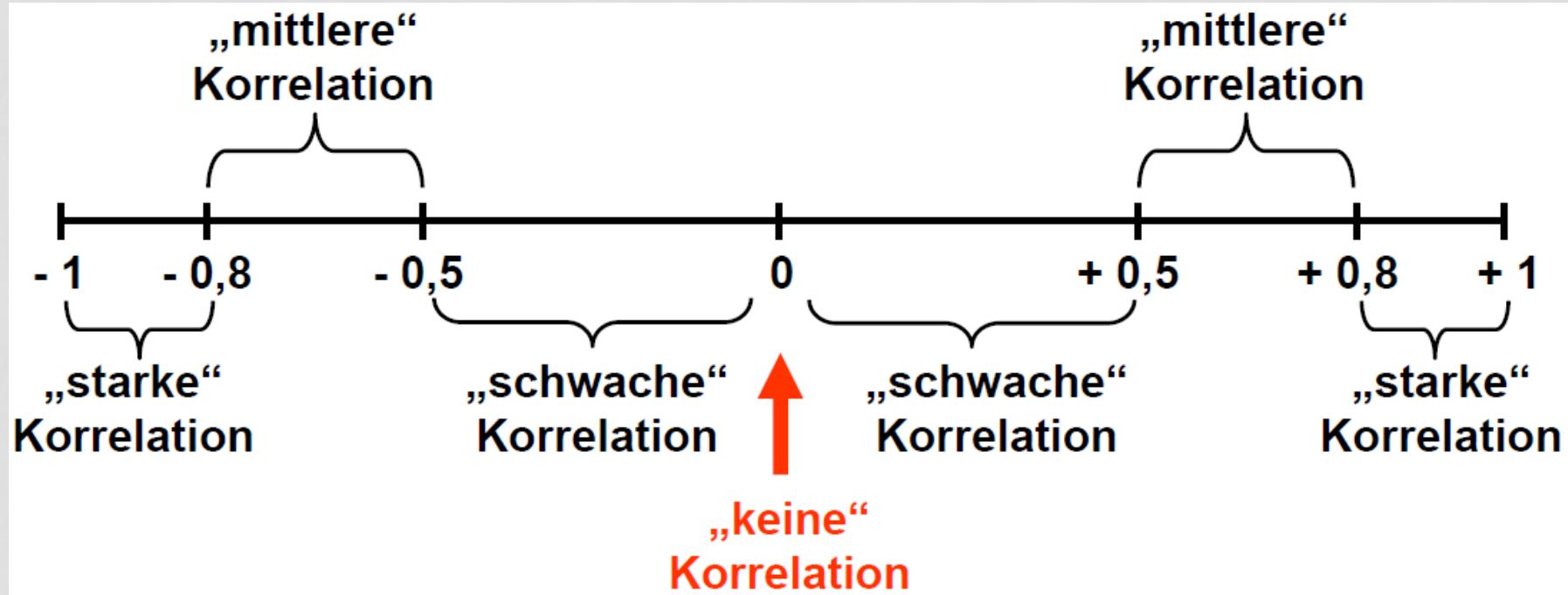
Berechnungstabelle mit Hilfsgrößen

Filiale Nr. i	x_i Verkaufsfläche (1000qm)	y_i Filialumsatz (Mio €)	$x_i \cdot y_i$	x_i^2	y_i^2
1	3	30	90	9	900
2	2	10	20	4	100
3	6	40	240	36	1.600
4	5	20	100	25	400
Summe	16	100	450	74	3.000

$$r_{XY} = \frac{\left(\frac{1}{n} \sum_{i=1}^n x_i \cdot y_i \right) - \bar{x} \cdot \bar{y}}{\sqrt{\left(\frac{1}{n} \sum_{i=1}^n x_i^2 \right) - \bar{x}^2} \cdot \sqrt{\left(\frac{1}{n} \sum_{i=1}^n y_i^2 \right) - \bar{y}^2}}$$

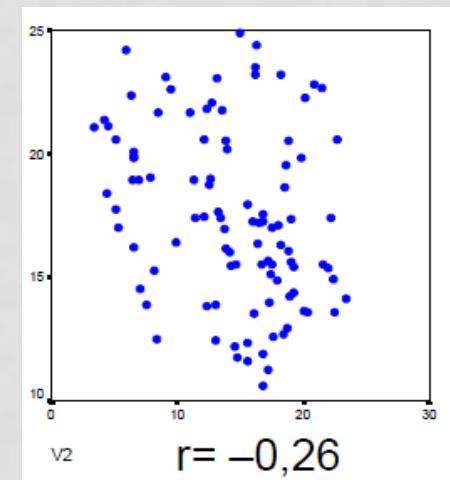
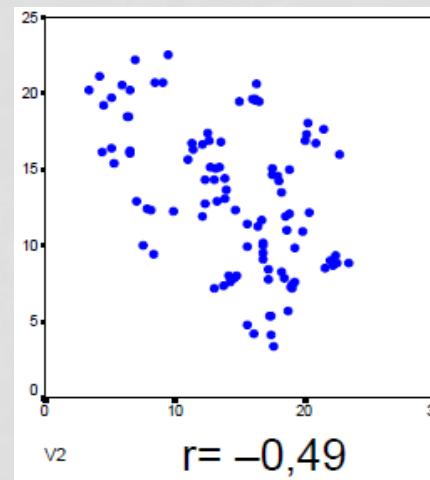
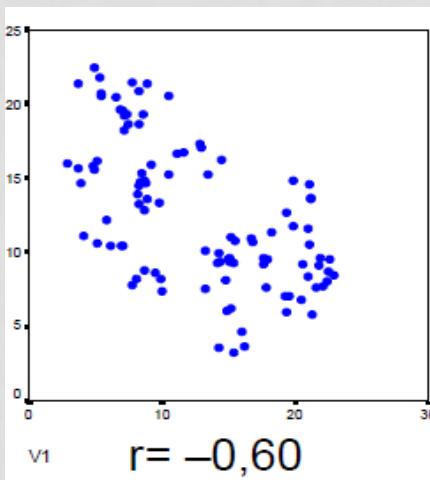
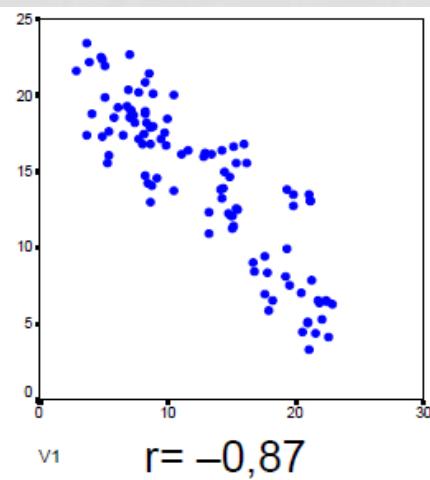
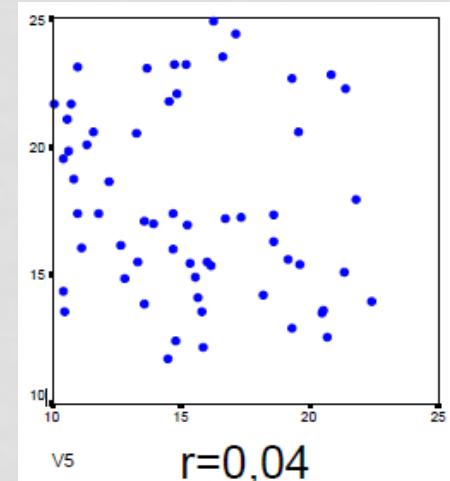
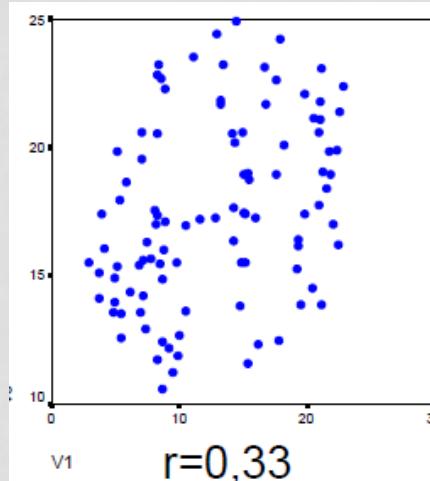
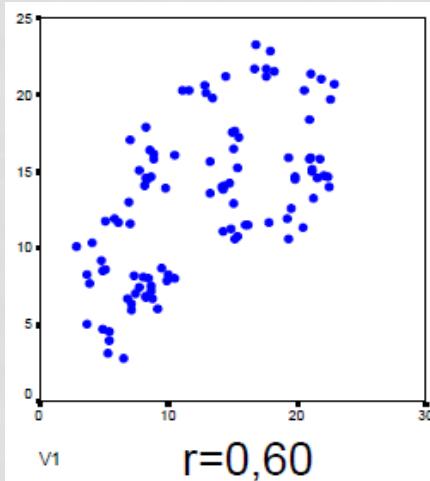
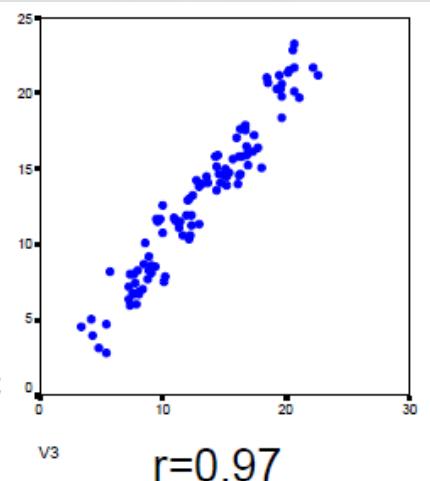
$$\frac{\frac{1}{4} \cdot 450 - 4 \cdot 25}{\sqrt{\frac{1}{4} \cdot 74 - 4^2} \cdot \sqrt{\frac{1}{4} \cdot 3000 - 25^2}} = \frac{112,5 - 100}{\sqrt{25} \cdot \sqrt{125}} = \frac{12,5}{1,581 \cdot 11,180} = \frac{12,5}{17,676} = +0,707$$

ZUR INTERPRETATION DES KORRELATIONSKoeffizientEN



Nur Orientierungswerte → Entscheidung ist immer problembezogen!!!

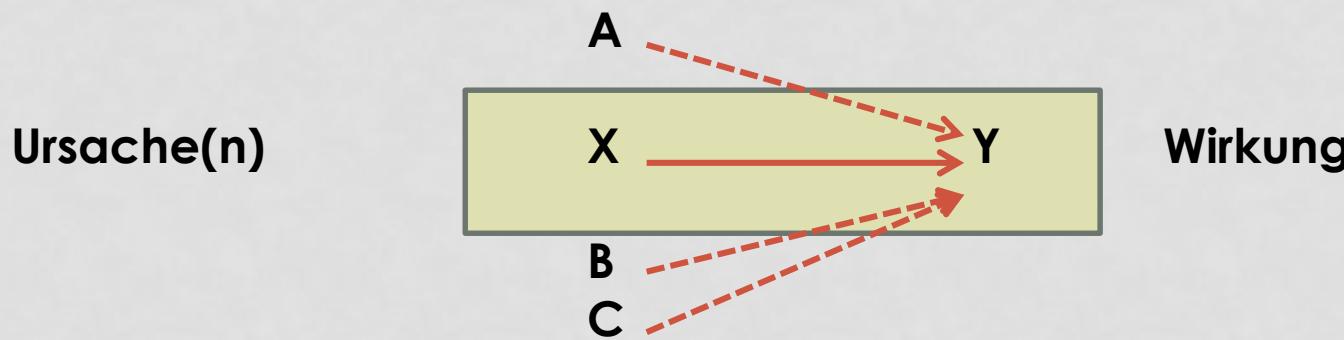
ZUR INTERPRETATION DES KORRELATIONSKOEFFIZIENTEN



PROBLEME BEI DER REGRESSIONSANALYSE

Problem:

**Es gibt meist nicht nur einen Einflussfaktor
(Probleme sind selten monokausal ...)**



- **einfache** Regressionsanalyse

→ zwei metrischen Größen: Einflussgröße X und Zielgröße Y. Es wird mithilfe von zwei Parametern eine Gerade durch eine Punktwolke gelegt, sodass der lineare Zusammenhang zwischen X und Y möglichst gut beschrieben wird.

- **multiple** Regressionsanalyse

→ eine Verallgemeinerung der einfachen linearen Regression mit k Regressoren, welche die abhängige Variable erklären sollen.

Mehrere metrischen Größen: mehrere Einflussgrößen X_1, \dots, X_k und eine Zielgröße Y.

REGRESSIONSFUNKTION

Voraussetzungen:

- X und Y quantitative (metrische) Merkmale
- $X \rightarrow Y$ (es existiert ein Zusammenhang)

Vorbereitende Arbeiten:

- Überprüfung, ob Abhängigkeitsanalyse sinnvoll ist
- Erhebung von Daten für X und Y $\rightarrow (x_1, y_1), \dots, (x_n, y_n)$

1. Schritt: Visualisierung im Streudiagramm

(qualitative Abhängigkeitsanalyse)

2. Schritt: Auswahl eines Funktionstyps

(hier: Beschränkung auf lineare Funktionen)

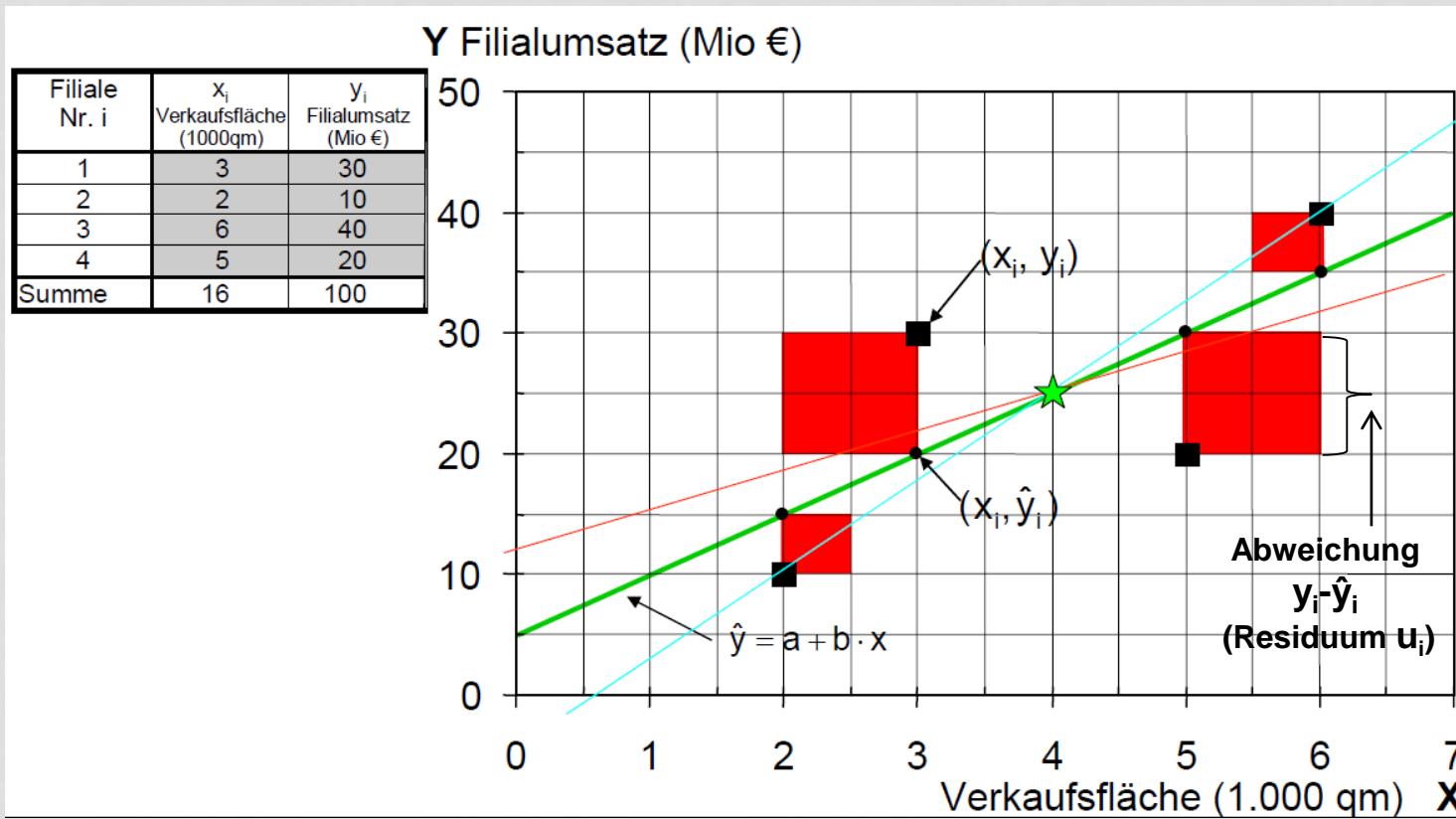
3. Schritt: Berechnung der Regressionsfunktion

(nach Methode der kleinsten Quadrate)

REGRESSIONSFUNKTION

Beispiel:

Verkaufsfläche → Filialumsatz



REGRESSIONSFUNKTION

Bestimmung der **Regressionsfunktion** $\hat{y} = a + b * x$ nach der **Methode der kleinsten Quadrate**:

Die **Regressionskoeffizienten a und b (Kurvenparameter)** werden so bestimmt, dass die Summe der quadratischen Abweichungen der Kurve von den beobachteten Punkten minimal ist:

$$OLS(a, b) = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - (a + b * x_i))^2 \rightarrow \min$$

Residuen u_i

Methode der kleinsten Quadrate (OLS = ordinary least squares)

Es wird die (partielle) Differentialrechnung genutzt, um die Extremwertaufgabe „Minimierung der Summe der Abweichungsquadrate“ zu lösen.

REGRESSIONSRECHNUNG

Berechnungstabelle mit Hilfsgrößen

Filiale Nr. i	x_i Verkaufsfläche (1000qm)	y_i Filialumsatz (Mio €)	$x_i \cdot y_i$	x_i^2	y_i^2
1	3	30	90	9	900
2	2	10	20	4	100
3	6	40	240	36	1.600
4	5	20	100	25	400
Summe	16	100	450	74	3.000

$$a = \frac{\sum_{i=1}^n x_i^2 \sum_{i=1}^n y_i - \sum_{i=1}^n x_i \sum_{i=1}^n x_i y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{74 \cdot 100 - 16 \cdot 450}{4 \cdot 74 - 16^2} = \frac{7400 - 7200}{296 - 256} = \frac{200}{40} = 5$$

$$b = \frac{n \sum_{i=1}^n x_i y_i - \sum_{i=1}^n x_i \sum_{i=1}^n y_i}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} = \frac{4 \cdot 450 - 16 \cdot 100}{4 \cdot 74 - 16^2} = \frac{1800 - 1600}{296 - 256} = \frac{200}{40} = 5$$

ANWENDUNG DER REGRESSIONSANALYSE

Regressionsverfahren haben viele praktische Anwendungen. Die meisten Anwendungen fallen in eine der folgenden beiden Kategorien:

- zum Erstellen eines **Vorhersagemodells**
- um die **Stärke des Zusammenhangs** zu quantifizieren: so können diejenigen x_j ermittelt werden, die gar keinen Zusammenhang mit y haben oder diejenigen Teilmengen x_i, \dots, x_j , die redundante Information über y enthalten.

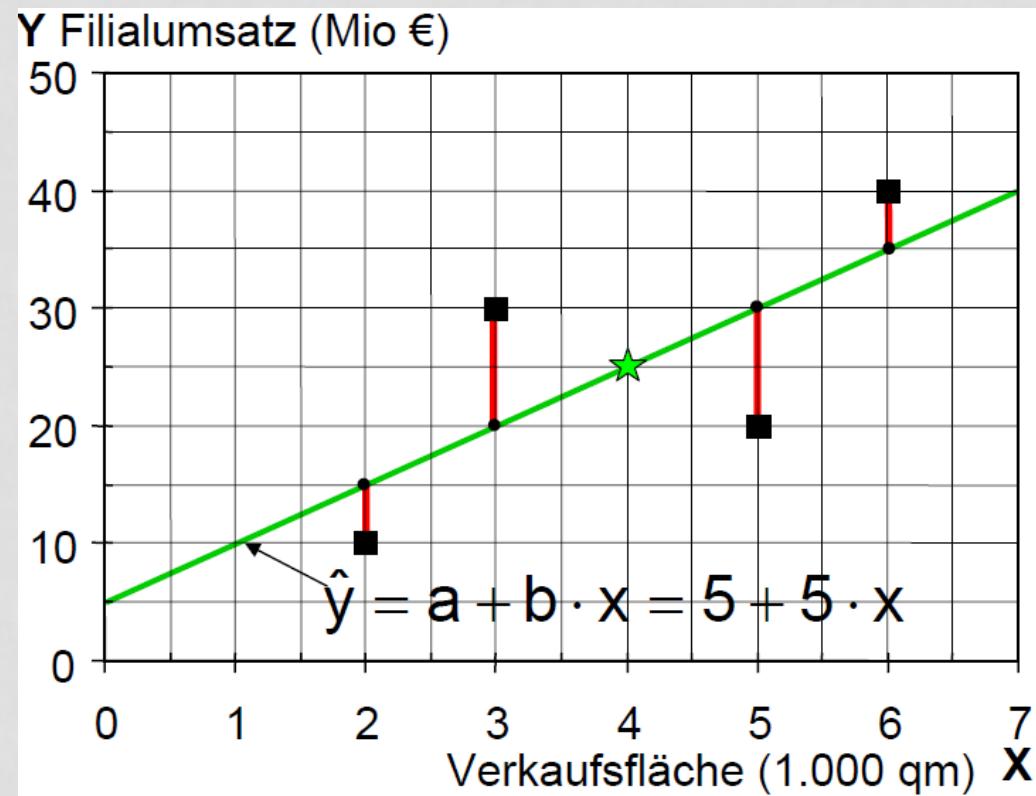
INTERPRETATION DER ERGEBNISSE

Regressionskoeffizienten a und b

Beispiel:

Fragen für die Schätzung:

- Umsatzprognose für neue Filiale mit 4.500 qm
- Lohnt sich eine Erweiterung um 1.000 qm?



INTERPRETATION DER ERGEBNISSE

Regressionskoeffizienten a und b

Beispiel:

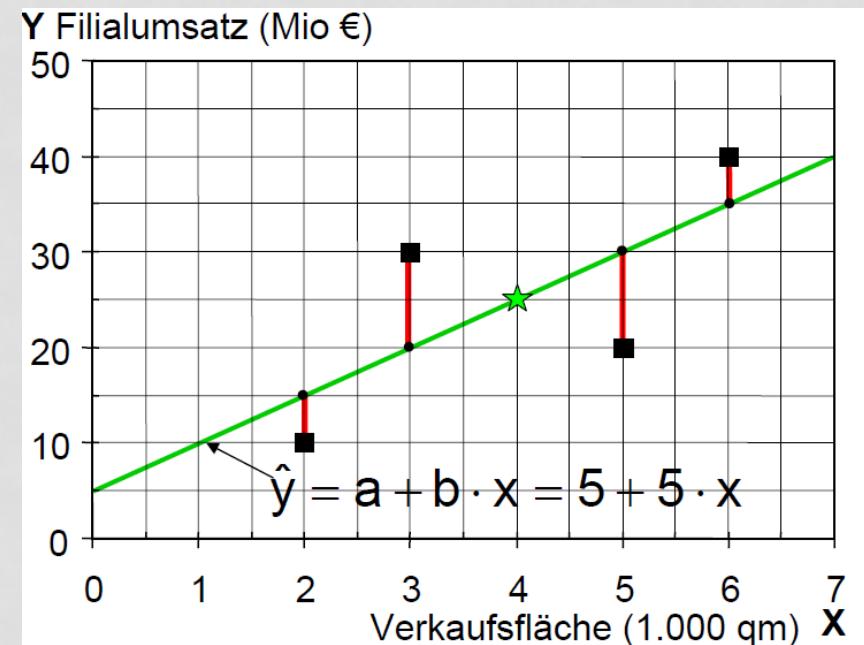
Fragen für die Schätzung:

- **Umsatzprognose für neue Filiale mit 4.500 qm Verkaufsfläche:**

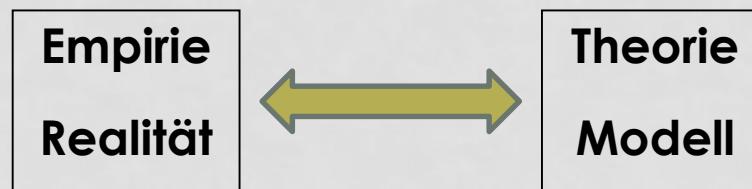
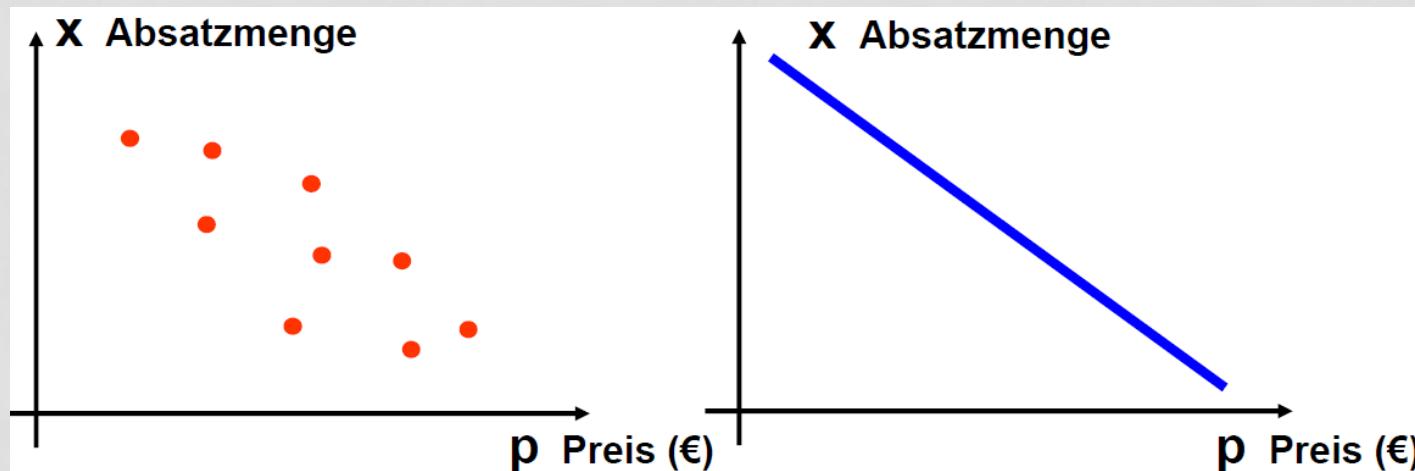
$$5 + 5 * 4,5 = 27,5 \text{ Mio €}$$

- **Lohnt sich eine Erweiterung um 1.000 qm Verkaufsfläche?**

$$\text{mit Erweiterung: } 5 + 5 * (4,5+1,0) = 5 + 5 * 5,5 = 32,5 \text{ Mio €}$$



MODELL VS. REALITÄT



Modell = vereinfachtes Abbild der Realität

- Wie gut beschreibt das Modell die Realität?
- Wie gut wird die Realität durch das Modell wiedergegeben?

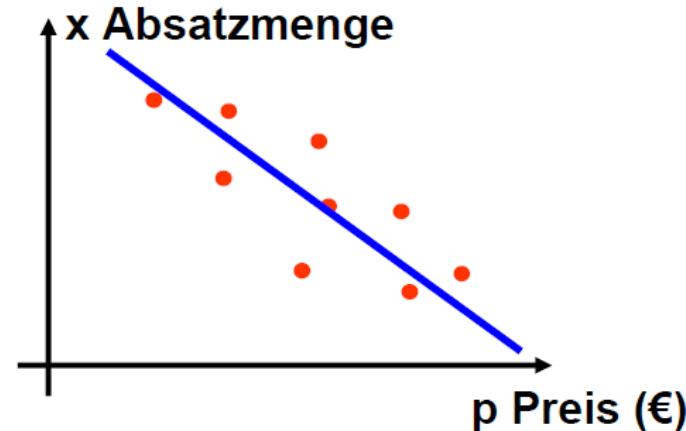
MODELL VS. REALITÄT

Regressionsrechnung:

$$\hat{x}(p) = a + b \cdot p = 100 - 5 \cdot p$$

Spezifikation
des Modells

Schätzung der
Parameter a, b



→ Wir brauchen **Gütemaße** für die Schätzung der Parameter:

- Wie gut ist die „goodness of fit“ (Anpassungsgüte)
- Wie gut beschreibt die Regressionsfunktion die Abhängigkeit?

MODELL VS. REALITÄT

Beispiel:

Verkaufsflächen →

unterschiedlich groß

Filialumsatz

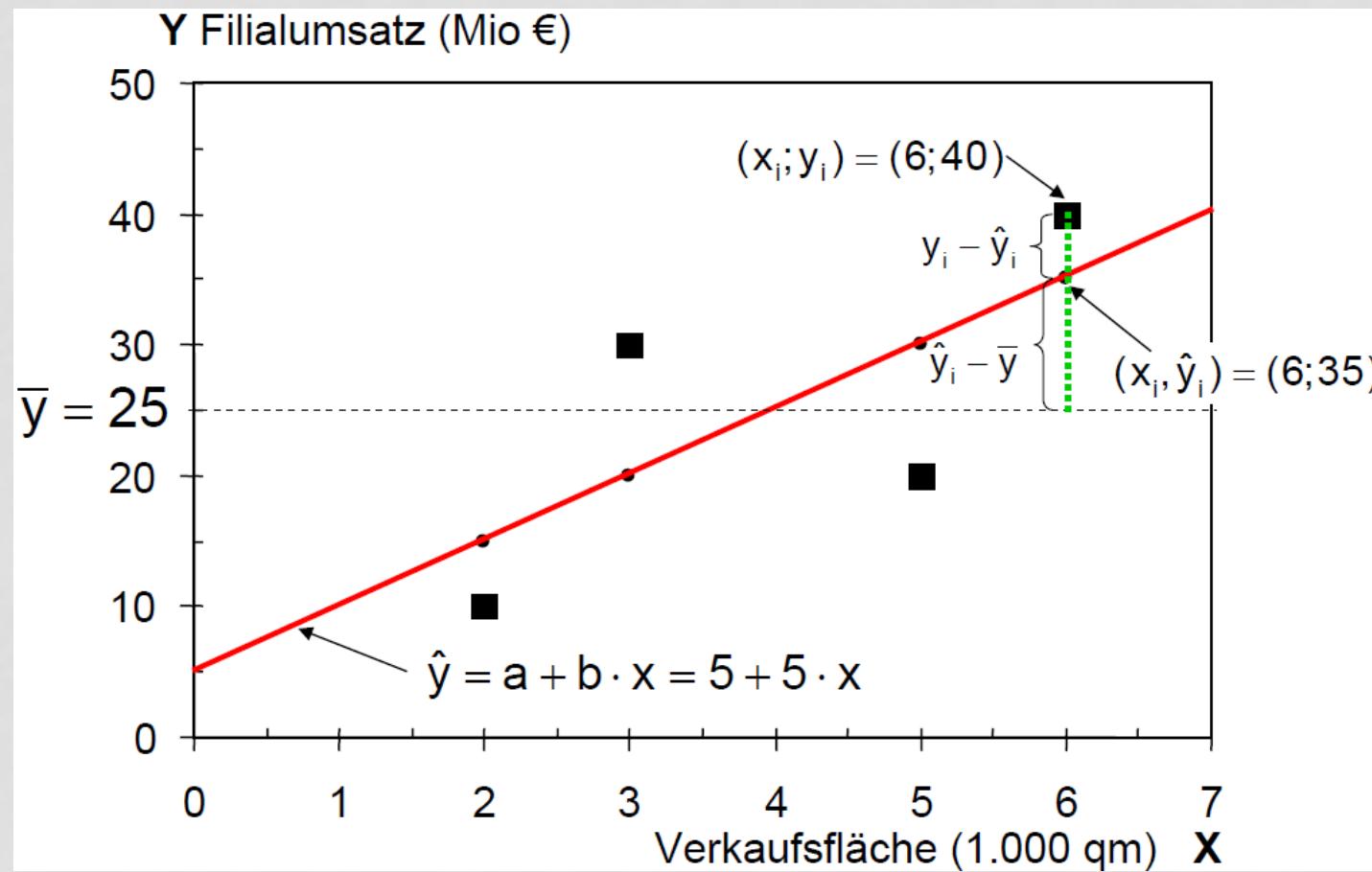
unterschiedlich hoch

WARUM?

- Wie gut erklären die Unterschiede bei den Verkaufsflächen die Unterschiede bei den Filialumsätzen?
→ **Wie viel Varianz wird durch das Modell nicht erklärt?**
- Wie gut erklärt die Regressionsfunktion die Abhängigkeit zwischen Verkaufsfläche und Filialumsatz?
→ **Wie hoch ist die Erklärungskraft des Modells?**

PROGNOSEWERTE UND RESIDUEN

Abweichungszerlegung: $(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$



(Vgl. mit Folie 17)

PROGNOSEWERTE UND RESIDUEN

Abweichungszerlegung: $(y_i - \bar{y}) = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y})$

Gesamtabweichung der Beobachtung y_i zum Mittelwert \bar{y}
 → diesen Fehler würden wir machen, wenn wir mit dem Mittelwert die entsprechende Beobachtung vorhersagen würden

Unerklärte Abweichung /Residuum der Beobachtung y_i zur Regressionsgeraden (Residuum u_i)
 → diesen Teil der Abweichung können wir auch durch Hinzunahme der unabhängigen Variablen x nicht vermeiden

Erklärte Abweichung der Regressionsgeraden zum Mittelwert \bar{y}
 → diesen Fehler können wir durch Hinzunahme der unabhängigen Variablen x vermeiden

Filiale Nr. i	x_i Verkaufsfläche (1000qm)	y_i Filialumsatz (Mio €)
1	3	30
2	2	10
3	6	40
4	5	20
Summe	16	100

$$(40 - 25) = (40 - 35) + (35 - 25) \quad (\text{Mio €})$$

$$15 = 5 + 10 \quad (\text{Mio €})$$

Filiale Nr. 3:

PROGNOSEWERTE UND RESIDUEN

Varianzzerlegung:

$$\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{n} \sum_{i=1}^n (u_i - \bar{u})^2 + \frac{1}{n} \sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2$$

$$s_y^2 = s_u^2 + s_{\hat{y}}^2$$

Die Varianz der Regressionswerte wird auch bestimmt durch die Varianz des unabhängigen Merkmals:

$$s_{\hat{y}}^2 = b^2 * s_x^2$$

PROGNOSEWERTE UND RESIDUEN

Berechnungstabelle mit Hilfsgrößen

Filiale Nr. i	x_i Verkaufsfläche (1000qm)	y_i Filialumsatz (Mio €)	$x_i \cdot y_i$	x_i^2	y_i^2	\hat{y}_i
1	3	30	90	9	900	20
2	2	10	20	4	100	15
3	6	40	240	36	1.600	35
4	5	20	100	25	400	30
Summe	16	100	450	74	3.000	100

$$\hat{y} = 5 + 5 \cdot x$$

$$s_y^2 = \left(\frac{1}{n} \cdot \sum_{i=1}^n y_i^2 \right) - \bar{y}^2 = \frac{1}{4} \cdot 3.000 - 25^2 = 750 - 625 = 125$$

$$s_x^2 = \left(\frac{1}{n} \cdot \sum_{i=1}^n x_i^2 \right) - \bar{x}^2 = \frac{1}{4} \cdot 74 - 4^2 = 18,5 - 16 = 2,5$$

$$s_{\hat{y}}^2 = \left(\frac{1}{n} \cdot \sum_{i=1}^n \hat{y}_i^2 \right) - \bar{\hat{y}}^2 = \frac{1}{4} \cdot (20^2 + 15^2 + 35^2 + 30^2) - 25^2 =$$

$$687,5 - 625 = 62,5$$

PROGNOSEWERTE UND RESIDUEN

Berechnungstabelle mit Hilfsgrößen

Filiale Nr. i	x_i Verkaufsfläche (1000qm)	y_i Filialumsatz (Mio €)	$x_i \cdot y_i$	x_i^2	y_i^2	\hat{y}_i	$u_i = (y_i - \hat{y}_i)$
1	3	30	90	9	900	20	+10
2	2	10	20	4	100	15	-5
3	6	40	240	36	1.600	35	+5
4	5	20	100	25	400	30	-10
Summe	16	100	450	74	3.000	100	0

Varianz der Residuen:

$$s_u^2 = \left(\frac{1}{n} \cdot \sum_{i=1}^n u_i^2 \right) - \bar{u}^2 = \frac{1}{4} \cdot (10^2 + (-5)^2 + 5^2 + (-10)^2) - 0^2 = \\ \frac{1}{4} \cdot 250 - 0 = 62,5$$

BESTIMMTHEITSMAß

Das **Bestimmtheitsmaß R²** (Erklärungskraft des Modells) ist ein **Gütemaß der linearen Regression**.

Das R² gibt an, wie gut die unabhängige Variable Y geeignet ist, die Varianz der abhängigen Variable X zu erklären.

(unbrauchbares Modell) **0% ≤ R² ≤ 100%** (perfekte Modellanpassung)

Das R² nutzt das Konzept der Varianzzerlegung und besagt, dass sich die Varianz des abhängigen Merkmals in erklärte Varianz und nicht erklärte Varianz (Residualvarianz) zerlegen lässt.

Bestimmtheitsmaß R² → Anteil der Varianz der abhängigen Variable, der sich durch die Varianz der unabhängigen Variable erklären lässt.

$$R^2 = \frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})_2}{\sum_{i=1}^n (y_i - \bar{y})^2} = \frac{\text{erklärte Variation}}{\text{Gesamtvariation}} = 1 - \frac{\text{unerklärte Variation}}{\text{Gesamtvariation}} = \frac{\sum_{i=1}^n u_i^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

BESTIMMTHEITSMAß

Es folgt: R^2 ist das Verhältnis aus der Streuung der Prognosewerte und der Gesamtstreuung der y-Werte (**s. Folie 38**):

$$R^2 = \frac{s_{\hat{y}}^2}{s_y^2} = \frac{62,5}{125} = 0,5 \iff 1 - \frac{s_u^2}{s_y^2} = 1 - \frac{62,5}{125} = 1 - 0,5 = 0,5$$

Achtung!

- Bei einer einfachen linearen Regression (nur eine unabhängige Variable) entspricht das Bestimmtheitsmaß dem Quadrat des Korrelationskoeffizienten nach Pearson r_{XY}

$$R^2 = (r_{XY})^2$$

Beispiel:

X = Verkaufsfläche, Y = Filialumsatz

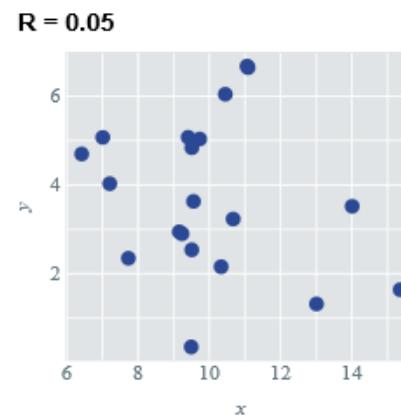
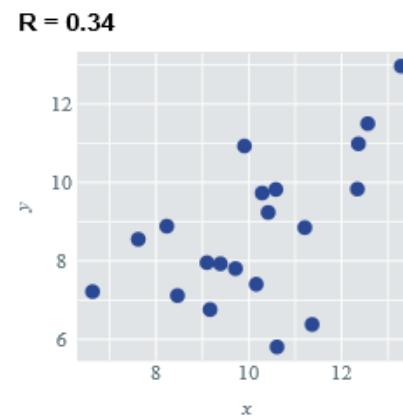
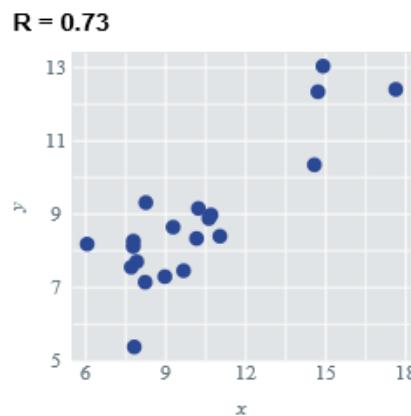
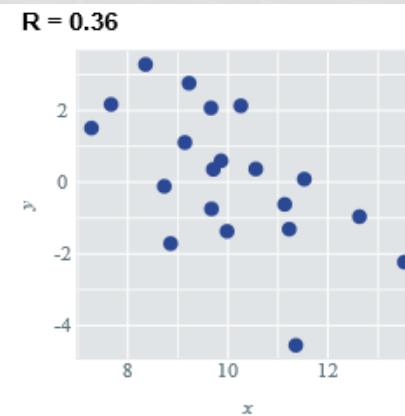
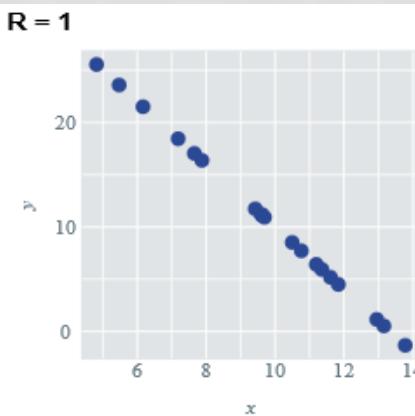
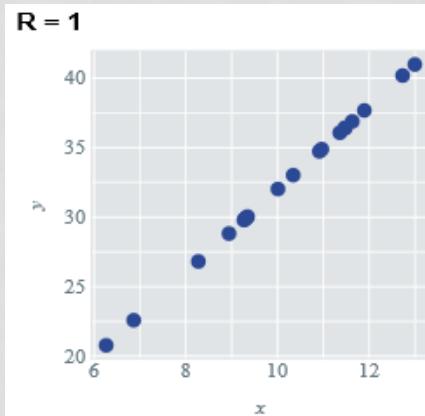
$$r_{XY} = 0,707 \rightarrow R^2 = 0,707^2 = 0,50 = 50\%$$

Bedeutung / Interpretation:

50 % der Varianz der Filialumsätze lassen sich durch die Varianz der Verkaufsflächen erklären. Die anderen 50 % lassen sich nur durch andere Einflussfaktoren erklären.

BESTIMMTHEITSMAß

Die folgende Grafiksammlung zeigt verschiedene Streudiagramme in Abhängigkeit des Wertes des R^2 . Je eher die Datenpunkte auf einer Linie liegen, desto höher ist das R^2 . Streuen die Datenpunkte ohne Zusammenhang im Raum, liegt das R^2 nahe 0.



BESTIMMTHEITSMAß

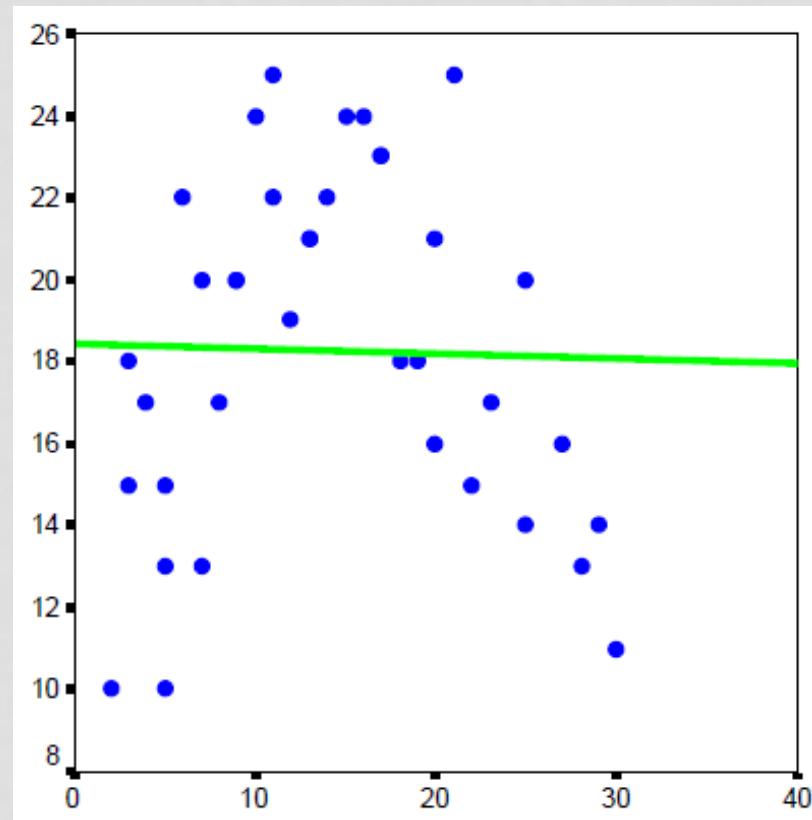
"Wie hoch muss mein R² sein?"

- Die übliche Größenordnung des R² variiert, je nach dem um welches Anwendungsgebiet es sich handelt. In Bereichen wie dem klassischen Marketing, in denen es hauptsächlich darum geht, menschliches Verhalten zu erklären bzw. vorherzusagen, sind meist geringe R² (deutlich kleiner 50%) zu erwarten. In anderen Bereichen wie bspw. der Physik sind höhere R² die Regel. Dies ist wenig überraschend, da auf das menschliche Verhalten zahlreiche und häufig nicht direkt messbare Einflüsse wirken. In der Physik hingegen werden oft Zusammenhänge zwischen wenigen exakt messbaren Größen untersucht. Dies geschieht zusätzlich meist unter experimentellen Bedingungen, unter denen sich Störeinflüsse minimieren lassen.
- Während auf der Mikro-Ebene in vielen Fällen bereits ein R² von 10% als gut gelten kann, erwarten viele bei stärker aggregierten Daten ein R² von 40% bis 80% oder sogar mehr. Ein Modell mit geringem R² - selbst bei stärker aggregierten Daten – ist nicht nutzlos, da die Alternative dazu oft gar kein Modell darstellt, was einem R² von 0 entspricht. Im übertragenen Sinne bedeutet das, dass eine systematische Prognose auf Basis eines Modells mit beschränktem R² oft schon besser ist als eine unsystematische Planung, die ausschließlich auf Bauchgefühl setzt. Generell ist die Aussagekraft von Modellen mit geringem R² nicht zwangsläufig schlecht.

PROBLEME BEI LINEAREN REGRESSION UND KORRELATION

Nur lineare Zusammenhänge werden erfasst!

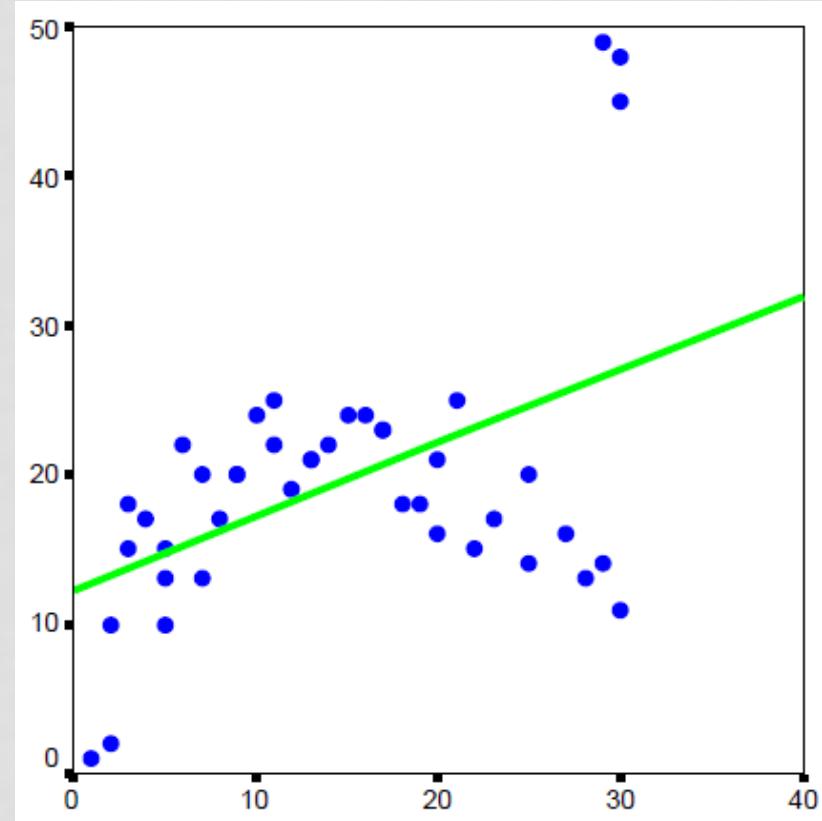
Die Gerade ist quasi horizontal – was nicht dem „eigentlichen“ Zusammenhang entspricht. Hier geht es um einen nicht linearen Zusammenhang, der nicht durch die lineare Regression beschrieben werden kann.



PROBLEME BEI LINEAREN REGRESSION UND KORRELATION

**Einzelne Fälle können starken Einfluss ausüben
(nicht zuletzt wegen dem Quadrieren)!**

Die gleichen Daten wie vorhin plus einige Extremwerte (links unten, rechts oben) erzeugen eine deutlich steigende Gerade



PROBLEME BEI LINEAREN REGRESSION UND KORRELATION

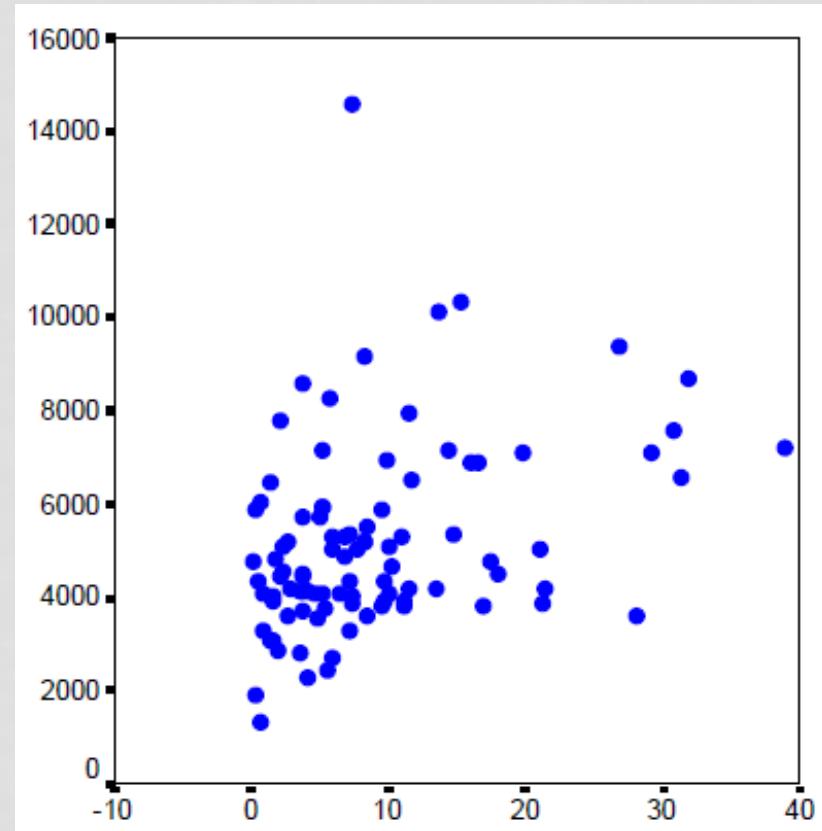
**Einzelne Fälle können starken Einfluss ausüben
(nicht zuletzt wegen dem Multiplizieren)!**

Korrelation über alle Fälle:

$$r=0,35$$

Korrelation ohne Extremfall
(y über 14.000):

$$r=0,39$$



WIRTSCHAFTSSTATISTIK

MODUL 7: WAHRSCHEINLICHKEITSRECHNUNG

WS 2021/22

DR. E. MERINS

MONTY-HALL-DILEMMA

Monty-Hall-Problem oder Ziegenproblem

USA-Spielshow „Let's Make a Deal“ → Deutsche Variante „Geh aufs Ganze!“

Angenommen Sie hätten die Wahl zwischen drei Toren. Hinter einem der Tore ist ein Auto, hinter den anderen sind Ziegen. Sie wählen ein Tor, z.B. Tor Nr. 1, und der Moderator, der weiß, was hinter jedem Tor ist, öffnet ein anderes Tor, z.B. Nr. 3, hinter dem eine Ziege steht. Er fragt Sie nun: „Möchten Sie auf das Tor Nr. 2 wechseln?“

Ist es von Vorteil, die Wahl des Tores zu ändern?

Antwort (ohne Berücksichtigung einer bestimmten Motivation des Moderators):

Ja, Sie sollten wechseln!

Das zuerst gewählte Tor hat die Gewinnchance von 1/3, aber das zweite Tor hat eine Gewinnchance von 2/3.

Hier ist ein Weg, sich das Geschehen vorzustellen: angenommen, es gäbe 1 Million Tore und Sie wählen Tor Nr. 1. Dann öffnet der Moderator, der das eine Tor mit dem Preis immer vermeidet, alle Tore bis auf das Tor Nummer 777.777. Sie würden doch sofort zu diesem Tor wechseln, oder?

BESCHREIBENDE STATISTIK UND WAHRSCHEINLICHKEITSTHEORIE

Die beschreibende Statistik kommt ohne den Begriff **Wahrscheinlichkeit** aus.

Beschreibende Statistik

Relative Häufigkeit

Häufigkeitsverteilung

Stichprobe

Mittelwert

Standardabweichung

Varianz

Median

Quantile

Wahrscheinlichkeitstheorie

→ Wahrscheinlichkeit

→ Wahrscheinlichkeitsverteilung

→ Zufallsvariablen

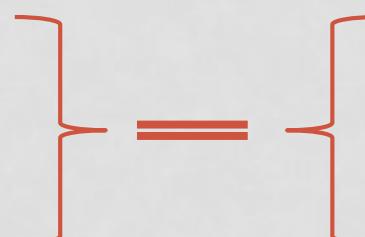
→ Erwartungswert

→ Streuung

Varianz

Median

Quantile



WAHRSCHEINLICHKEITSTHEORIE

Die Verbindung zur **Wahrscheinlichkeitstheorie** wird auch über den **Zufallsaspekt einer Stichprobe** hergestellt. Historisch ist die **Wahrscheinlichkeitsrechnung** eng mit dem **Glücksspiel** verbunden.

Ein (**Zufalls**-) Experiment ist ein beliebig oft (unter identischen Bedingungen) wiederholbarer Vorgang, dessen Ergebnis „vom Zufall abhängt“, d.h. nicht exakt vorhergesagt werden kann. Die verschiedenen möglichen Ergebnisse oder Realisationen des Experiments heißen **Elementarereignisse ω (,Klein-Omega‘)**. Sie bilden zusammen den **Ereignisraum Ω (,Groß-Omega‘)**.

Experiment → die Erhebung eines Merkmals an einem Merkmalsträger
Elementarereignisse → die Merkmalsausprägungen
Stichprobe vom Umfang n → die n-malige Wiederholung des Experiments

Annahme: die Ausgangssituation bei der n-fachen Wiederholung des Experiments ist immer dieselbe. In der Praxis ist dies jedoch unrealistisch. So sind z.B. bei einem Test zur Wirkung eines Medikaments an 20 Versuchspersonen die Bedingungen (Alter, frühere Krankheiten etc.) bei jeder der 20 Versuchswiederholungen (hier also Versuchspersonen) andere.

ZUFALLSEXPERIMENTE

Beispiele für Zufallsexperimente:

1. Bernoulli-Experiment: Werfen einer Münze: $\Omega = \{\text{Kopf, Wappen}\}$ oder $\Omega = \{0, 1\}$
2. Würfeln: $\Omega = \{1, 2, 3, 4, 5, 6\}$
3. Lotto 6 aus 49: $\Omega = \{ \omega \mid \omega = \{j_1, \dots, j_6\}, j_1, \dots, j_6 \in \{1, 2, 3, \dots, 48, 49\} \}$
Da Mengen nur verschiedene Elemente enthalten, gilt $|\{j_1, \dots, j_6\}| = 6$.
4. Anzahl der Anrufe in einer Telefonvermittlung pro Tag: $\Omega = N_0 := N \cup \{0\}$.
5. $\Omega = \{ \omega \mid \omega = \text{Matrikelnummer eines Studenten im WS 2019/20} \}$.
6. Verlauf der Körpertemperatur eines Lebewesens: $\{ \omega = (\text{id}, f) \mid \text{id} \in N, f \in {}^{\circ}\text{C}(R_+) \}$.

Ergebnis des Experiments ist die Identifikationsnummer id des Lebewesens und eine (beschränkte) stetige Funktion auf der nichtnegativen reellen Achse. $f(0)$ ist die Körpertemperatur bei der Geburt. Nach dem Tod ($T > 0$) des Lebewesens könnte man die Umgebungstemperatur zur Fortsetzung der Funktion f heranziehen.

ZUFALLSEXPERIMENTE

Fortsetzung...

Das letzte Beispiel zeigt, dass auch **Funktionen** als Ergebnisse eines Zufallsexperiments auftreten können.

Man interessiert sich also dafür, ob bei Durchführung des **Zufallsexperiments** bestimmte **Ereignisse** eintreten.

Zum Beispiel, ob:

1. beim Wurf einer Münze $A = \{\text{Kopf}\}$ gefallen ist
2. beim Würfeln eine 5 oder 6, d. h. $B = \{5, 6\}$ herauskam
3. im Lotto 6 aus 49 "sechs Richtige" angekreuzt wurden
4. mehr als 1000 Anrufe pro Tag in der Telefonvermittlung, $D = \{n \mid n > 1000\}$, auftraten
5. $K = \{\omega \mid \text{Matrikelnummer } \omega \text{ beginnt mit einer 7}\}$
6. die Körpertemperatur eines Lebewesens nie den Wert 40°C überschritten.

In jedem Beispiel handelt es sich bei **Ereignissen** um **Teilmengen** von Ω .

EREIGNISRAUM

Ereignisraum $\Omega \neq \emptyset$ → auch **Ergebnismenge** oder **Merkmalraum** genannt

$\Omega \neq \emptyset$ → eine nichtleere Menge **Ω** ist die Menge aller möglichen Ergebnisse eines mathematischen Zufallsexperiments, die sog. **Ergebnismenge** oder **Merkmalraum** oder **Ereignisraum**. Man spricht auch vom Stichprobenraum (sample space).

Die **Anzahl der Ergebnisse** der Menge **Ω** nennt man **Mächtigkeit** $|\Omega| = n$.

Ω kann **endlich, abzählbar** oder sogar **überabzählbar unendlich** sein.

Ω heißt **diskret**, falls **es höchstens abzählbar unendlich viele Elemente** hat.

EREIGNIS

Ein **Ereignis** (event) ist eine **Teilmenge A** des **Ereignisraums Ω** .

$$A \subset \Omega$$

A tritt ein, falls sich bei Versuchsdurchführung ein $\omega \in A$ ergibt.

Die einelementigen Teilmengen des Ereignisraums (oder die Elemente von Ω) heißen **Elementarereignisse** (singleton) $\{\omega\}$.

Ein **Gegenereignis** ist die Menge aller Ergebnisse, die nicht zum Ereignis gehören.

Ω heißt **sicheres Ereignis** → tritt also immer ein

\emptyset heißt **unmögliches Ereignis** → kann nie eintreten

A^c heißt **Komplementärereignis** → Gegenereignis, ohne A

Teilmengen A und B heißen **unvereinbar** oder **disjunkt**, falls $A \setminus B = \emptyset$

WAHRSCHEINLICHKEIT

Um exakte Voraussagen über die Begrenzung unserer Möglichkeiten zu treffen, brauchen wir einen **Maß für die Sicherheit** (oder **Unsicherheit**). Ein solches Maß ist die **Wahrscheinlichkeit p** (engl.: probability).

Die **Wahrscheinlichkeitsrechnung** ordnet jedem **Ereignis A** eines **Zufallsexperiments** eine **Wahrscheinlichkeit p(A)** (oder **p_A** oder **Prob(A)**) für sein Eintreten zu.

Beispiel:

Beim Münzwerfen gibt es nur zwei Elementarereignisse, die gleichmöglich sind.

$$p(\text{,Kopf'}) = \frac{1}{2}$$

$$p(\text{,Zahl'}) = \frac{1}{2}$$

$$p(\text{,Kopf oder Zahl'}) = 2/2 = 1 \rightarrow \text{entweder ,Kopf' oder ,Zahl' tritt beim Wurf ein}$$

$$p(\text{,Kopf und Zahl'}) = 0/2 = 0 \rightarrow \text{,Kopf' und ,Zahl' können nicht gleichzeitig eintreten}$$

WAHRSCHEINLICHKEIT UND RELATIVE HÄUFIGKEIT

Was bedeutet Wahrscheinlichkeit?

Im normalen Sprachgebrauch wird die Wahrscheinlichkeit eines Ereignisses auch häufig als Prozentzahl angegeben, indem sie mit 100 multipliziert und dafür mit dem Zusatz 'Prozent' versehen wird. Spätestens an dieser Stelle fällt die Parallelität zu den relativen Häufigkeiten eines Merkmals auf. In der späteren Anwendung werden unter anderem die Wahrscheinlichkeiten eines Ereignisses oder einer Merkmalsausprägung durch die entsprechenden relativen Häufigkeiten geschätzt, die über die n-fache Wiederholung des Experiments gewonnen werden. Allgemein lassen sich natürlich auf diese Weise die Wahrscheinlichkeiten beliebiger Ereignisse näherungsweise bestimmen, wenn sie sich zum Beispiel nicht elementar logisch oder physikalisch herleiten lassen.

WAHRSCHEINLICHKEIT UND RELATIVE HÄUFIGKEIT

Was bedeutet Wahrscheinlichkeit?

Würfel:

Das **Maß für die Sicherheit**, die höchste Augenzahl 6 zu würfeln, könnte so formuliert werden: "Ungefähr bei jedem sechsten Würfel-Versuch wird die Augenzahl 6 auftreten". Das bedeutet: "Unter 6 Würfel-Versuchen wird ungefähr 1 mal die Augenzahl 6 auftreten".

Ganz sicher können wir natürlich nicht sein, dass bei nur 6 Versuchen die gewünschte Augenzahl genau 1 mal eintritt, also würfeln wir öfter: "Unter 6000 Würfel-Versuchen wird ungefähr 1000 mal die Augenzahl 6 auftreten". Das klingt schon plausibler. Gehen wir noch einen Schritt weiter: "Unter einer sehr großen Zahl n von Würfel-Versuchen wird ungefähr $n/6$ mal die Augenzahl 6 auftreten".

WAHRSCHEINLICHKEIT UND RELATIVE HÄUFIGKEIT

Was bedeutet Wahrscheinlichkeit?

Genauer: Wenn wir ein Zufallsexperiment in identischer Weise n mal durchführen und dabei genau m mal das Ereignis A eintritt, so nennen wir den Quotienten m/n die **relative Häufigkeit $h(A)$** , mit der das Ereignis A eingetreten ist. Die relative Häufigkeit wird nicht bei jeder Reihe von n Versuchsdurchführungen gleich sein. Wenn aber n sehr groß ist, so ergibt sich jedes Mal ungefähr die gleiche relative Häufigkeit und wenn wir gedanklich n gegen unendlich wachsen lassen, so sollte die relative Häufigkeit einen fixen, nur vom Zufallsexperiment und dem betrachteten Ereignis A abhängigen Wert annehmen. Diesen Wert nennen wir die **Wahrscheinlichkeit des Ereignisses**.

$$P(A) := \lim_{n \rightarrow \infty} h_n(A)$$

→ **empirisches Gesetz der großen Zahlen**

WAHRSCHEINLICHKEIT UND RELATIVE HÄUFIGKEIT

Was bedeutet Wahrscheinlichkeit?

Zufallsexperiment:

Würfeln mit einem Zufallsgenerator: für drei Werte von n wird die Anzahl des Auftretens von Augenzahl 6 in diesen n Versuchsdurchführungen und die dazugehörige relative Häufigkeit ermittelt. Dabei wurde jeder dieser n Versuche 5 mal durchgeführt.

n	Augenzahl 6 tritt so oft auf	relative Häufigkeit
6	1 1 0 2 0	0,1667 0,1667 0 0,3333 0
60	7 9 8 10 9	0,1167 0,15 0,1333 0,1667 0,15
6000	1046 1026 993 963 986	0,174 0,171 0,166 0,161 0,164

Die relativen Häufigkeiten werden mit wachsendem n einander immer ähnlicher.

Die **Wahrscheinlichkeit** eines Ereignisses ist die für eine gegen unendlich strebende Anzahl n von Durchführungen des betreffenden Zufallsexperiments vorausgesagte **relative Häufigkeit** seines Eintretens.

→ **mathematische Idealisierung**, da n in der Wirklichkeit nicht "gegen unendlich strebt"

WAHRSCHEINLICHKEITSEIGENSCHAFTEN

Eigenschaften der Wahrscheinlichkeit:

1. Die relative Häufigkeit jedes Ereignisses A liegt im Bereich $0 \leq h(A) \leq 1$, und daher gilt dies auch für jede Wahrscheinlichkeit.

(Beweis: tritt das Ereignis bei n -maliger Durchführung des Zufallsexperiments m mal ein, so gilt $0 \leq m \leq n$, woraus die Behauptung folgt).

2. Tritt ein Ereignis A mit Sicherheit ein, so tritt es bei n -maliger Durchführung des Zufallsexperiments immer, also n mal, ein. Seine relative Häufigkeit ist dann
$$h(A) = n/n = 1 \rightarrow p(A) = 1$$
3. Tritt ein Ereignis A mit Sicherheit nicht ein, so tritt es bei n -maliger Durchführung des Zufallsexperiments nie, also 0 mal, ein. Seine relative Häufigkeit ist dann
$$h(A) = 0/n = 0 \rightarrow p(A) = 0$$

AXIOME VON KOLMOGOROW

Die axiomatische Begründung der Wahrscheinlichkeitstheorie wurde in den 1930er Jahren von **Andrei Kolmogorow** entwickelt.

Ein **Wahrscheinlichkeitsmaß** (kurz **W-Maß**) muss demnach folgende **drei Axiome** erfüllen:

1. Die **Wahrscheinlichkeit** für das Eintreten eines Ereignisses A ist immer eine **reelle Zahl** zwischen 0 und 1: $0 \leq p(A) \leq 1$
2. Das sichere Ereignis **Ω** hat die Wahrscheinlichkeit **1**:
 $p(A) = 1$ → A tritt mit Sicherheit ein
 $p(A) = 0$ → A tritt mit Sicherheit nicht ein
 $0 < p(A) < 1$ → die Werte dazwischen drücken **Grade an Sicherheit** aus. Je größer die Wahrscheinlichkeit $p(A)$, umso „eher“ ist anzunehmen, dass das Ereignis A eintritt.
3. Die Wahrscheinlichkeit einer Vereinigung abzählbar vieler disjunkter Ereignisse ist gleich der Summe der Wahrscheinlichkeiten der einzelnen Ereignisse
→ **σ -Additivität („Sigma‘-Additivität“)**:

$$P(A_1 \cup A_2 \cup \dots \cup A_n) = \sum_{i=1}^n P(A_i)$$

MENGENTHEORETISCHE KONZEPTE

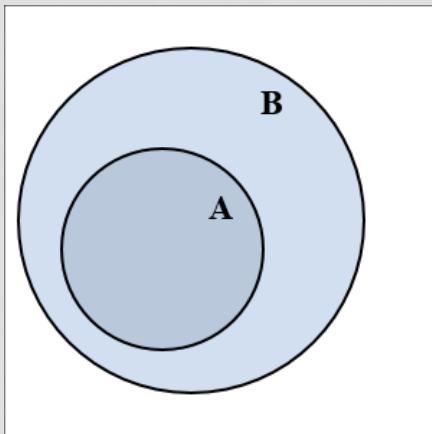
Ereignisse sind Teilmengen des Ereignisraums.

- Ereignisse können ihre Beziehungen in Begriffen der **Mengenlehre** ausdrücken
- Ereignisse können wie **Mengen** miteinander verknüpft werden

Mengenoperationen:

- \cap **Schnittmenge**
- U **Vereinigung**
- \setminus **Mengendifferenz**
- c **Komplementbildung**

MENGENTHEORETISCHE KONZEPTE



A ist eine (echte) **Teilmenge** von **B**

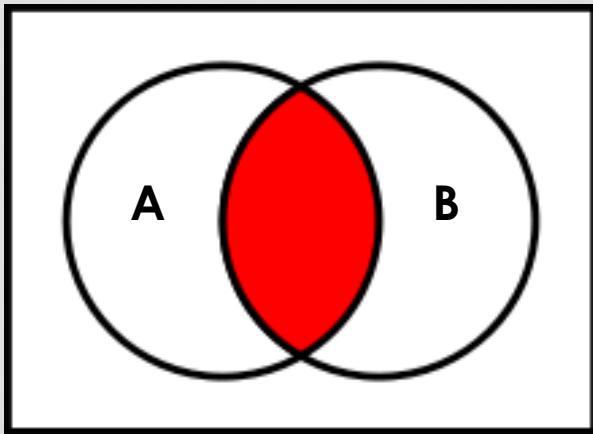
Eine Menge A heißt **Teilmenge** einer Menge B, wenn jedes Element von A auch Element von B ist.

Formal: $A \subseteq B : \Leftrightarrow \forall x (x \in A \rightarrow x \in B)$

Zwei Mengen heißen **gleich**, wenn sie dieselben Elemente enthalten.

Formal: $A = B : \Leftrightarrow \forall x (x \in A \leftrightarrow x \in B)$

MENGENTHEORETISCHE KONZEPTE

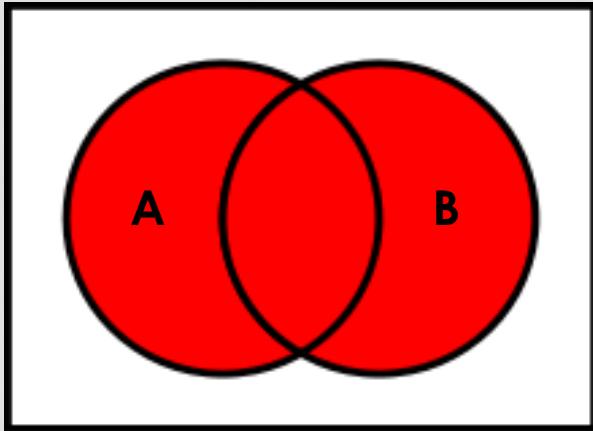


Schnittmenge von **A und B**

Die **Schnittmenge** von **A und B** ist die Menge der Objekte (eine nichtleere Menge), die sowohl in A als auch in B enthalten sind.

Formal: $A \cap B := \{ x \mid (x \in A \wedge x \in B) \}$

MENGENTHEORETISCHE KONZEPTE

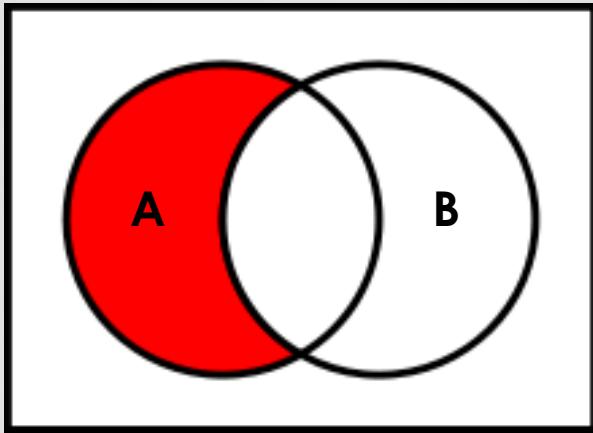


Vereinigungsmenge von **A** und **B**

Die **Vereinigungsmenge** von A und B ist die Menge (nicht notwendigerweise nichtleere) der Objekte, die in mindestens einem Element von A und B enthalten sind.

Formal: $A \cup B := \{x \mid (x \in A) \vee (x \in B)\}$

MENGENTHEORETISCHE KONZEPTE



A ohne B

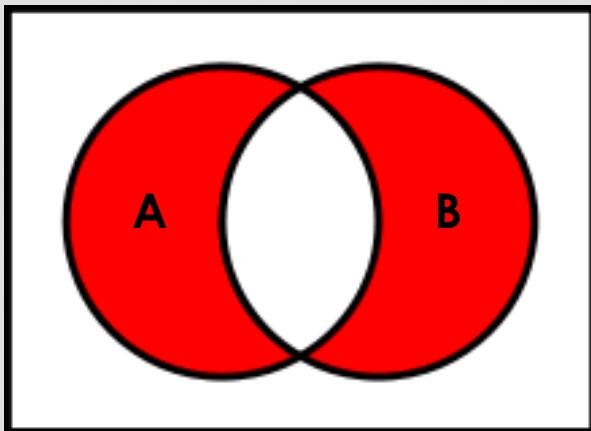
Die **Differenzmenge** (wird nur für 2 Mengen definiert) von A und B ist die Menge der Elemente, die in A aber nicht in B enthalten sind.

Formal: $A \setminus B := \{ x \mid (x \in A) \wedge (x \notin B) \}$

Komplement von B in Bezug auf A (A ohne B): ist B eine Teilmenge von A, spricht man einfach vom Komplement der Menge B.

Formal: $B^c := \{ x \mid x \notin B \}$

MENGENTHEORETISCHE KONZEPTE

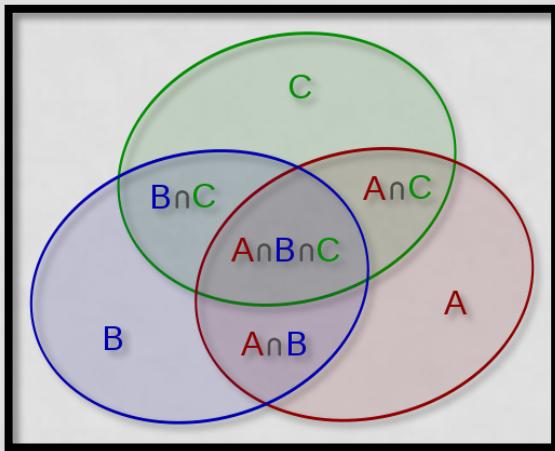


Symmetrische Differenz von A und B

Formal: $A \Delta B := \{x \mid (x \in A \wedge x \notin B) \vee (x \in B \wedge x \notin A)\}$

$$A \Delta B = (A \cup B) - (A \cap B)$$

MENGENTHEORETISCHE KONZEPTE



Einschluss-Ausschluss-Verfahren (auch Prinzip von Inklusion und Exklusion oder Prinzip der Einschließung und Ausschließung)

Für zwei endliche disjunkte Mengen A und B gilt (**Summenregel**):

$$|A \cup B| = |A| + |B| - |A \cap B|$$

Noch allgemeiner gilt die **Summenregel** für drei Mengen:

$$|A \cup B \cup C| = |A| + |B| + |C| - |A \cap B| - |A \cap C| - |B \cap C| + |A \cap B \cap C|$$

GESETZMÄßIGKEITEN

Für alle $A, B, C \subseteq X$ gilt:

Antisymmetrie: $A \subseteq B \text{ und } B \subseteq A \rightarrow A = B$

Transitivität: $A \subseteq B \text{ und } B \subseteq C \rightarrow A \subseteq C$

Die Mengen-Operationen Schnitt \cap und Vereinigung \cup sind kommutativ, assoziativ und zueinander distributiv:

Assoziativgesetz: $(A \cup B) \cup C = A \cup (B \cup C)$

$$(A \cap B) \cap C = A \cap (B \cap C)$$

Kommutativgesetz: $A \cup B = B \cup A$

$$A \cap B = B \cap A$$

Distributivgesetz: $A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$

$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$$

De Morgansche Gesetze: $(A \cup B)^c = A^c \cap B^c$

$$(A \cap B)^c = A^c \cup B^c$$

Absorptionsgesetz: $A \cup (A \cap B) = A$

$$A \cap (A \cup B) = A$$

GESETZMÄßIGKEITEN

Fortsetzung...

Für die Differenzmenge gilt:

Assoziativgesetze:

$$(A \setminus B) \setminus C = A \setminus (B \cup C)$$

$$A \setminus (B \setminus C) = (A \setminus B) \cup (A \cap C)$$

Distributivgesetze:

$$(A \cap B) \setminus C = (A \setminus C) \cap (B \setminus C)$$

$$(A \cup B) \setminus C = (A \setminus C) \cup (B \setminus C)$$

$$A \setminus (B \cap C) = (A \setminus B) \cup (A \setminus C)$$

$$A \setminus (B \cup C) = (A \setminus B) \cap (A \setminus C)$$

$$A \setminus B = A \cap B^c$$

GESETZMÄßIGKEITEN

Fortsetzung...

Für die **symmetrische Differenz** gilt:

Assoziativgesetz:

$$(A \Delta B) \Delta C = A \Delta (B \Delta C)$$

Kommutativgesetz:

$$A \Delta B = B \Delta A$$

Distributivgesetz:

$$(A \Delta B) \cap C = (A \cap C) \Delta (B \cap C)$$

$$A \Delta \emptyset = A$$

$$A \Delta A = \emptyset$$

$$A \Delta B = (A \cup B) \setminus (A \cap B)$$

$$A \Delta B = A^c \Delta B^c$$

$$(A \Delta B)^c = (A \cap B) \cup (A^c \cap B^c)$$

GESETZMÄÙIGKEITEN

Fortsetzung...

Für die Komplementmenge gilt:

$$A \cap A^C = \emptyset \quad A \cup A^C = \Omega$$

$$\emptyset^C = \Omega \quad \Omega^C = \emptyset$$

Doppeltes Komplement:

$$(A^C)^C = A$$

Idempotenzgesetze:

$$A \cap A = A \quad A \cup A = A$$

Neutralen Elemente:

$$A \cap \Omega = A \quad A \cup \emptyset = A$$

Dominanzgesetze:

$$A \cap \emptyset = \emptyset \quad A \cup \Omega = \Omega$$

REGELN DER WAHRSCHEINLICHKEITSRECHNUNG

$$P(\Omega) = 1$$

$$P(\neg A) = P(\Omega \setminus A) = 1 - P(A)$$

Ersetzt man bei endlichen Mengen die **Wahrscheinlichkeit P(A)** durch die Anzahl der Elemente von A, so sind alle Axiome die Aussagen der elementaren Mengenlehre:

1) $P(A \cup B) = P(A \setminus B) + P(A \cap B) + P(B \setminus A)$

$$P(A) = P(A \setminus B) + P(A \cap B)$$

$$P(B) = P(A \cap B) + P(B \setminus A)$$

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

- 2) sind A und B voneinander unabhängige Ereignisse:

$$P(A \cup B) = P(A \text{ oder } B) = P(A) + P(B)$$

$$P(A \cap B) = P(A \text{ und } B) = P(A) * P(B)$$

REGELN DER WAHRSCHEINLICHKEITSRECHNUNG

Einschluss–Ausschluss–Formel (Inklusion–Exklusion–Formel): (s. auch Folie 22)

$$P(A) = 1 - P(\neg A) = 1 - P(B_1 \cup B_2 \cup B_3) =$$

$$1 -$$

$$[P(B_1) + P(B_2) + P(B_3) - P(B_1 \cap B_2) - P(B_1 \cap B_3) - P(B_2 \cap B_3) + \\ P(B_1 \cap B_2 \cap B_3)]$$

Beispiel:

100 Studenten haben u.a. Kurse A, B und C belegt. 65 Studenten haben Kurs A belegt, 32 Kurs B, 18 Kurs C, 15 A und B, 9 B und C, 7 A und C, 3 haben alle drei Kurse belegt.

Wie viele Studenten haben keinen der Kurse A, B oder C belegt.

(Anzahl Studenten ohne Kurse A, B, C) =

$$100 - (65 + 32 + 18 - 15 - 9 - 7 + 3) = 100 - 87 = 13$$

BERECHNUNG VON WAHRSCHEINLICHKEITEN

Will man diese Formeln auf die **Berechnung von Wahrscheinlichkeiten** anwenden, so muss man annehmen, dass jedes Ziehungsergebnis gleich wahrscheinlich ist.

Bezeichnen wir mit Ω die von k und n abhängige Menge aller möglichen Ziehungsergebnisse und betrachten eine Teilmenge $E \subset \Omega$, so ist p die **Wahrscheinlichkeit**, dass ein Ziehungsergebnis zur „**Ergebnismenge**“ E gehört. Diese Wahrscheinlichkeit ist unter obiger Gleichwahrscheinlichkeits-Annahme **das Verhältnis der günstigen Möglichkeiten zu allen Möglichkeiten**:

$$p = \frac{|E|}{|\Omega|}$$

LAPLACE - EXPERIMENT

Ein **Laplace-Experiment** ist ein **Zufallsexperiment**, bei dem **jedes Ereignis die gleiche Wahrscheinlichkeit p** besitzt, d.h. alle Ergebnisse sind gleich wahrscheinlich.

Beispiel:

Ein nicht gezinkter (idealer) Würfel, also ein Würfel bei dem jede Augenzahl gleich wahrscheinlich ist, bezeichnet man als **Laplace-Würfel**. Bei diesem Würfel ist die Wahrscheinlichkeit jedes Elementarereignisses gleich.

Die entsprechende **Wahrscheinlichkeit** ist dann $(n/6)/n = 1/6$.

Das bedeutet, wenn der Laplace-Würfel 600 mal geworfen wird, dann erwartet man, dass jede Zahl 100-mal erscheint.

LAPLACE - EXPERIMENT

Zur Erinnerung: Ereignisse sind komplex, sie sind Zusammenfassungen von Versuchsausgängen.

Beispiel:

Für den (idealen) Würfel ist auch "Die Augenzahl ist gerade" ein Ereignis.
Wie groß ist die Wahrscheinlichkeit für sein Eintreten?

Unter den 6 möglichen Augenzahlen (\rightarrow **mögliche Fälle**) sind 3 Zahlen gerade (2, 4 und 6 \rightarrow **günstige Fälle**). Jeder einzelne günstige Fall (und auch jeder einzelne ungünstige Fall) tritt bei n-maligem Würfeln für sehr großes n gleich oft ein, nämlich $n/6$ mal, d.h. sein relativer Anteil ist $1/6$.

Der relative Anteil der günstigen Fälle zusammen ist dreimal so groß wie der relative Anteil jeder einzelnen geraden Augenzahl, also $3/6 = 1/2$. Also ist die Wahrscheinlichkeit, eine gerade Augenzahl zu würfeln, gleich $1/2$.

- Die Anzahl **aller möglichen Versuchsausgänge** eines Laplace-Experiments (d.h. die **Zahl der Elemente seines Ereignisraums**) wird die **Zahl der möglichen Fälle** genannt. Alle diese Fälle sind gleich wahrscheinlich.
- Sei A ein Ereignis. Dann ist die **Wahrscheinlichkeit** für das Eintreten des Ereignisses A gegeben durch den Quotienten:

$$p(A) = \text{Zahl der günstigen Fälle} / \text{Zahl der möglichen Fälle}$$

LAPLACE- EXPERIMENT

Beispiel 1:

Es werden zwei unterscheidbare (ideale) Würfeln geworfen. Wir stellen uns der Einfachheit halber vor, es handelt sich um einen **roten** und einen **blauen** Würfel. Dabei sollen die beiden Würfeln unabhängig voneinander fallen. Die möglichen Versuchsausgänge sind alle 36 möglichen geordneten Paare von Augenzahlen: $(1, 1)$, $(1, 2)$, $(1, 3)$, $(1, 4)$, $(1, 5)$, $(1, 6)$, $(2, 1)$, $(2, 2)$, $(2, 3)$, ..., $(6, 4)$, $(6, 5)$ und $(6, 6)$.

Wie ist die Wahrscheinlichkeit des Ereignisses „Die Summe der Augenzahlen ist gerade“?

Zahl der möglichen Fälle (Zahl aller möglichen Versuchsausgänge) = **36**

Zahl der günstigen Fälle (Zahl der möglichen Versuchsausgänge, bei denen die Summe der Augenzahlen gerade ist):

die Summe der Augenzahlen ist gerade, wenn beide Augenzahlen gerade oder wenn beide Augenzahlen ungerade sind. Da jeder Würfel 3 gerade und 3 ungerade Augenzahlen besitzt, gibt es 9 Versuchsausgänge der Form (gerade, gerade) und 9 Versuchsausgänge der Form (ungerade, ungerade). Insgesamt gibt es also **18** günstige Fälle.

Die Berechnung:

$$p(\text{Summe der Augenzahlen ist gerade}) = 18/36 = 1/2$$

→ diese Formel gilt nur für Laplace-Experimente!

LAPLACE- EXPERIMENT

→ Nicht jedes Zufallsexperiment ist ein Laplace-Experiment!

Beispiel 2:

In einer Urne befinden sich 10 rote, 15 blaue und 5 grüne Kugeln. Es wird eine Kugel zufällig ("blind") herausgegriffen. Kugeln gleicher Farbe werden nicht unterschieden.

Kein Laplace-Experiment! → die Versuchsausgänge rot, blau und grün (für die herausgegriffene Kugel) haben nicht die gleiche Chance einzutreten. Es lässt sich aber leicht mit einem kleinen Trick auf ein Laplace-Experiment zurückführen: wir nummerieren die Kugeln durch, so dass jede ihre eigene Identität besitzt. Nun wird jede Nummer mit der gleichen Wahrscheinlichkeit gezogen – wir haben aus dem Urnenbeispiel vorübergehend ein Laplace-Experiment gemacht:

Zahl der möglichen Fälle = 30 (die Anzahl der Kugeln in der Urne)

Versuchsausgänge rot: Zahl der günstigen Fälle = 10

Versuchsausgänge blau: Zahl der günstigen Fälle = 15

Versuchsausgänge grün: Zahl der günstigen Fälle = 5

Die Wahrscheinlichkeiten für die drei Versuchsausgänge:

$p(\text{rote Kugel wird gezogen}) = 10/30 = 1/3$

$p(\text{blaue Kugel wird gezogen}) = 15/30 = 1/2$

$p(\text{grüne Kugel wird gezogen}) = 5/30 = 1/6$

Die heimliche Nummerierung der Kugeln wird nun nicht mehr benötigt. Durch diese drei Zahlen (die genau den relativen Häufigkeiten der drei Kugelsorten in der Urne entsprechen) lassen sich die Wahrscheinlichkeiten aller Ereignisse des Zufallsexperiments ausdrücken (z.B. für das Ereignis, eine nicht-rote Kugel zu ziehen).

GEGENWAHRSCHEINLICHKEIT

Ist **A ein Ereignis** (eine **Teilmenge des Ereignisraums Ω**), so können wir seine **Komplementärmenge $\Omega \setminus A$** bilden, d.h. die Menge aller Versuchsausgänge, die nicht in A enthalten sind. Da sie wieder eine Teilmenge von Ω ist, ist sie ebenfalls ein Ereignis. Wir können es als „**A tritt nicht ein**“ oder kurz „**nicht-A**“ bezeichnen. Eine andere Schreibweise dafür ist $\neg A$ oder \bar{A} . Es heißt auch das **Gegenereignis** von A (oder die Negation von A).

Bemerkung: das Gegenereignis des Gegenereignisses ist wieder das ursprüngliche Ereignis: $\neg \neg A = A$

Die Wahrscheinlichkeit eines **Gegenereignisses** (die so genannte **Gegenwahrscheinlichkeit**) ist durch Komplementärmenge gegeben:
 $p(\neg A) = 1 - p(A)$

→ Die Summe aus der **Wahrscheinlichkeit** und der **Gegenwahrscheinlichkeit** eines Ereignisses ist gleich **1**
 $p(A) + p(\neg A) = 1$

GEGENWAHRSCHEINLICHKEIT

Beispiel 2 (s. Folie 33):

Urne mit Kugeln:

A = „Es wird eine **nicht-rote** Kugel gezogen“

B = „Es wird eine rote Kugel gezogen“.

Zahl der möglichen Fälle = 30 (die Anzahl der Kugeln in der Urne)

Versuchsausgänge rot: Zahl der günstigen Fälle = 10

p(B) = p(rote Kugel wird gezogen) = 10/30 = 1/3

Eine andere Methode **p(A)** zu berechnen: **A ist das Gegenereignis zu B**, dessen Wahrscheinlichkeit 1/3 ist. Dann ist

$$p(A) = 1 - p(B) = 1 - 1/3 = 2/3$$

die Wahrscheinlichkeit, dass eine nicht-rote Kugel (sondern **blaue** oder **grüne** Kugel) gezogen wird.

ELEMENTARE KOMBINATORIK

Erste systematische Untersuchungen zu Fragen der Wahrscheinlichkeitstheorie wurden im 17. Jahrhundert vor allem im Zusammenhang mit Glücksspielen durchgeführt (Bernoulli, Fermat, Laplace, Pascal, ...). Unter anderem spielten damals Abzählungsaufgaben eine wichtige Rolle.

Abzählende Kombinatorik ist ein Teilbereich der **Kombinatorik** und beschäftigt sich mit der Bestimmung der **Anzahl möglicher Anordnungen oder Auswahlen**

- unterscheidbarer oder nicht unterscheidbarer Objekte (d. h. „ohne“ bzw. „mit“ Wiederholung derselben Objekte)
- mit oder ohne Beachtung ihrer Reihenfolge (d. h. „geordnet“ bzw. „ungeordnet“)

	<u>Ohne Wiederholung bzw. Zurücklegen</u>	<u>Mit Wiederholung bzw. Zurücklegen</u>
<u>Mit Berücksichtigung der Reihenfolge und $k \leq n$</u>	Variation ohne Wiederholung (engl. k-permutation)	Variation mit Wiederholung
<u>Ohne Berücksichtigung der Reihenfolge und $k < n$</u>	Kombination ohne Wiederholung (engl. k-combination)	Kombination mit Wiederholung

BAUSTEINE DER KOMBINATORIK

Multiplikationsregel der Kombinatorik:

Es sei eine **mehrfache Auswahl** zu treffen, wobei es m_1 Möglichkeiten für die erste Wahl, m_2 Möglichkeiten für die zweite Wahl, m_3 für die dritte Wahl usw. gibt. Können alle Möglichkeiten nach Belieben kombiniert werden, so lautet die **Gesamtzahl aller möglichen Fälle**:

$$m_1 \cdot m_2 \cdot m_3 \cdot \dots$$

Wichtigste Bausteine von Kombinatorik-Formeln:

1. Fakultät

Für $n \in \mathbb{N}_0$ gibt $n!$ (n -Fakultät) die Anzahl der möglichen Permutationen (=Vertauschungen) von n verschiedenen Objekten an

$$n! = n \cdot (n - 1) \cdot \dots \cdot 1$$

$$0! = 1$$

$$1! = 1$$

2. Binomialkoeffizient

$\binom{n}{k}$ → Tupel → Schreibweise: „ n über k “, oder „ k aus n “

Man kann auf $\binom{n}{k}$ Weisen k Elemente aus einer Menge mit n Elementen auswählen. Dies entspricht genau der Anzahl der Ziehungsergebnisse ohne Zurücklegen und ohne Anordnung (K_{ow}).

GRUNDMUSTER DER KOMBINATORIK

Sei \mathbf{A} einer n -elementige Menge $\mathbf{A} = \{1, \dots, n\}$, aus der man nacheinander k Elemente auswählt (k Ziehungen). Dabei unterscheidet man mit und ohne Zurücklegen und mit und ohne Berücksichtigung der Reihenfolge (Anordnung).

Es ergeben sich **vier Grundmuster der Kombinatorik**:

1. mit Reihenfolge, mit Zurücklegen

$$|\Omega| = V_{mW} = n^k$$

2. mit Reihenfolge, ohne Zurücklegen

$$\begin{aligned} k < n: \quad |\Omega| = V_{oW} &= \binom{n}{k} k! = \frac{n!}{(n-k)!} \\ k = n: \quad |\Omega| = V_{oW} &= n! \end{aligned}$$

3. ohne Reihenfolge, ohne Zurücklegen (Binomialkoeffizient)

$$|\Omega| = K_{oW} = \binom{n}{k} = \frac{n!}{k! (n-k)!}$$

4. ohne Reihenfolge, mit Zurücklegen

$$|\Omega| = K_{mW} = \binom{n+k-1}{k} = \frac{(n+k-1)!}{k! (n-1)!}$$

1. MIT REIHENFOLGE, MIT ZURÜCKLEGEN

Beispiel 1.1:

Wie viele "Wörter" (auch unsinnige) können aus 5 Buchstaben zustande kommen aus einem Alphabet vom Umfang 26?

n = 26, k = 5 ($k < n \rightarrow$ Variationen mit Wiederholung)

$$|\Omega| = V_{mW} = n^k = 26^5 = 11.881.376$$

Beispiel 1.2:

Fünfmaliges Werfen eines Würfels (= oder Werfen von fünf Würfeln).

n = 6, k = 5 ($k < n \rightarrow$ Variationen mit Wiederholung)

$$|\Omega| = V_{mW} = n^k = 6^5 = 7.776$$

2. MIT REIHENFOLGE, OHNE ZURÜCKLEGEN

Beispiel 2.1:

Auf wie viele Arten können sich 5 Personen auf 5 freie (unterscheidbare) Plätze verteilen?

$n = 5, k = 5$ ($n = k \rightarrow$ Variation ohne Wiederholung)

$$|\Omega| = V_{ow} = n! = 5! = 120 \text{ (Arten)}$$

Beispiel 2.2:

Es gibt eine genaue Sitzordnung für die Personen aus Beispiel 2.1.

Wie groß ist die Wahrscheinlichkeit, dass sich jede Person auf den ihr zugedachten Platz zufällig gesetzt hat?

Achtung! Laplace-Experiment:

Die "Zahl der möglichen Fälle" ist 120, die "Zahl der günstigen Fälle" ist 1

$$p = \frac{|E|}{|\Omega|} = \frac{1}{120} = 0,008$$

2. MIT REIHENFOLGE, OHNE ZURÜCKLEGEN

Beispiel 2.3:

Wie ist die Wahrscheinlichkeit, beim viermaligen Werfen eines Würfels lauter verschiedene Augenzahlen zu erzielen?

$$n = 6, k = 4$$

Schritt 1: (Ergebnisraum)

$$|\Omega| = V_{mW} = n^k = 6^4 = 1.296$$

Schritt 2: (Menge der günstigen Ereignissen)

$$|E| = V_{oW} = \binom{n}{k} k! = \frac{n!}{(n-k)!} = \frac{6!}{(6-4)!} = 6 * 5 * 4 * 3 = 360$$

Schritt 3: (Wahrscheinlichkeit)

$$p = \frac{|E|}{|\Omega|} = \frac{360}{1.296} = \frac{5}{18} = 0,278$$

3. OHNE REIHENFOLGE, OHNE ZURÜCKLEGEN

Beispiel 3.1:

Experiment: gleichzeitiges Ziehen von 2 Kugeln aus der Urne U_6 (Urne mit 6 Kugeln) ohne Zurücklegen.

Wie viele Paare sind möglich wenn man die Kugeln durchnummeriert?

Ergebnisraum: $\Omega = \{\{1,2\}, \{1,3\}, \dots, \{5,6\}\}$

Mächtigkeit des Ergebnisraumes:

$$|\Omega| = K_{ow} = \binom{n}{k} = \frac{n!}{k!(n-k)!} = \binom{6}{2} = \frac{6!}{2!(6-2)!} = \frac{6!}{2! * 4!} = \frac{6 * 5 * 4 * 3 * 2}{2 * 4 * 3 * 2} = 15$$

3. OHNE REIHENFOLGE, OHNE ZURÜCKLEGEN

Beispiel 3.2:

Einer Urne mit 6 roten und 4 grünen Kugeln werden gleichzeitig 5 Kugeln entnommen.

Wie ist die Wahrscheinlichkeit, dass genau 2 der Kugeln rot sind?

$$n=10, n_1 = 6, n_2 = 4, k = 5$$

Schritt 1: (Ergebnisraum)

$$|\Omega| = K_{oW} = \binom{n}{k} = \binom{10}{5} = \frac{10!}{5!(10-5)!} = 252$$

Schritt 2: (Menge der günstigen Ereignissen: 2 Kugeln müssen aus der Menge der roten Kugeln und der Rest aus der Menge der grünen Kugeln stammen)

$$|E| = K_{oW} = \binom{6}{2} \binom{4}{3} = \frac{6!}{2! 4!} * \frac{4!}{3! 1!} = 60$$

Schritt 3: (Wahrscheinlichkeit)

$$p = \frac{|E|}{|\Omega|} = \frac{60}{252} = \frac{15}{64} = 0,24$$

4. OHNE REIHENFOLGE, MIT ZURÜCKLEGEN

Beispiel 4.1:

10 Sportlerinnen nehmen an 3 Wettbewerben teil, bei denen es jeweils genau eine Siegerin gibt. Auf wie viele Arten können die Preise verteilt werden?

$$n = 10, k = 3$$

$$|\Omega| = K_{mW} = \binom{n+k-1}{k} = \frac{(n+k-1)!}{k! (n-1)!} = \frac{12!}{3! 9!} = 220$$

ANORDNUNGEN DER ELEMENTE EINER MENGE MIT WIEDERHOLUNG

Werden bei einer Anordnung mit den k Elementen einer Menge das 1. Element n_1 -mal, das 2. Element n_2 -mal usw. verwendet, dann nennt man eine derartige Anordnung eine **Permutation mit Wiederholung**.

Ist $n = n_1 + n_2 + \dots + n_k$, dann ist die Anzahl der Permutationen

$$P_{mW} = \frac{n!}{n_1! n_2! \dots n_k!}$$

Beispiel:

Der Name **RAFAELLA** ist eine Anordnung der Buchstabenmenge {R, A, F, E, L} mit der Besonderheit, dass der Buchstabe L zweimal und der Buchstabe A dreimal auftritt. Es gibt also

$$P_{mW} = \frac{8!}{1! 3! 1! 1! 2!} = \frac{8 * 7 * 6 * 5 * 4 * 3 * 2}{3 * 2 * 2} = 3.360$$

Permutationen, d.h. man kann 3.360 unterschiedliche „Wörter“ aus dieser Buchstabenmenge zusammenstellen.

BEDINGTE WAHRSCHEINLICHKEIT

In zahlreichen Anwendungsfällen tritt das Problem auf, dass nur solche Versuchsausgänge eines Zufallsexperiments von Interesse sind, bei denen ein bestimmtes Ereignis eintritt.

Damit tritt eine neue Fragestellung auf:

Wie groß ist die Wahrscheinlichkeit für das Eintreten eines Ereignisses **A unter der Voraussetzung**, dass **B** eingetreten ist?

Die **bedingte Wahrscheinlichkeit** $p(A | B)$ des Ereignisses A "unter der Voraussetzung B" ist (nach empirischen Gesetzen der großen Zahlen, Folie 12) definiert als der Quotient aus der absoluten Häufigkeit H_{AB} für AB (das gleichzeitige Eintreten von A und B) und H_B , der absoluten Häufigkeit von B. **Voraussetzung: $P(B) > 0$**

$$P(A|B) := \frac{P(A \cap B)}{P(B)}$$

$$P(A \cap B) = P(B \cap A) = P(B)P(A|B) = P(A)P(B|A)$$

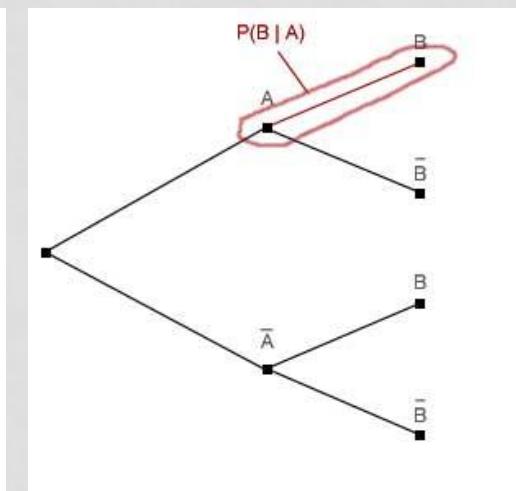
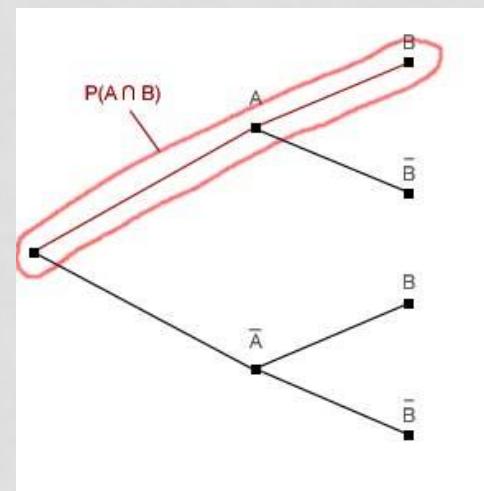
BEDINGTE WAHRSCHEINLICHKEIT

$P(A \cap B)$ und $P(B|A)$ sind unterschiedliche Wahrscheinlichkeiten.

$P(A \cap B)$ bezeichnet die Wahrscheinlichkeit, dass man mit allen möglichen Ausgängen des Zufallsexperiments startet, dann zuerst das Ereignis A und dann noch das Ereignis B erhält. Bei $P(B|A)$ ist bereits A eingetreten, die möglichen Ausgänge des Zufallsexperiments sind daher schon stark eingeschränkt. Jetzt ist nur noch wichtig, mit welcher Wahrscheinlichkeit B eintritt.

Beispiel zum Verständnis:

Sei A das Ereignis „weiblich“ und B das Ereignis „Hochschulabschluss“ bei einer zufällig herausgegriffenen Person im Hörsaal. Die „bedingte“ relative Häufigkeit bezieht sich auf den Anteil der Frauen unter den Akademikern.



TOTALE WAHRSCHEINLICHKEIT

Der **Satz von der totalen Wahrscheinlichkeit** ist ein Hilfsmittel, um mit Hilfe von bekannten Wahrscheinlichkeiten weitere zu ermitteln.

Für **zwei Ereignisse** A und B:

$$P(A) = P(A \cap B) + P(A \cap \bar{B}) = P(A|B)P(B) + P(A|\bar{B})P(\bar{B})$$

Für **endlich viele Ereignisse B_i** : sei $\{B_1, \dots, B_n\}$ eine Menge von **paarweise disjunkten Ereignissen**, dann gilt:

$$P(A) = \sum_{i=1}^n P(A|B_i) * P(B_i)$$

SATZ VON BAYES

Satz von Bayes (\rightarrow direkte Konsequenz aus dem **Satz von der totalen Wahrscheinlichkeit**)

Für **zwei Ereignisse A und B** mit $P(B) > 0$ gilt:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

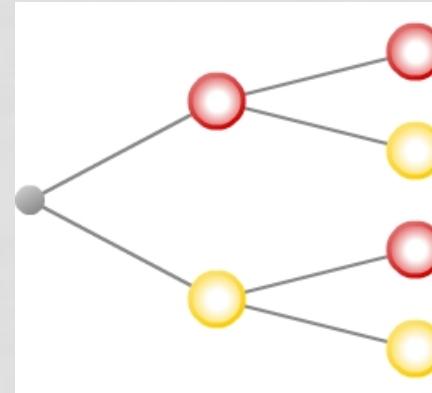
Für **endlich viele Ereignisse B_i** : seien B_i **paarweise disjunkt** und $A \subset \bigcup_{i=1}^n B_i$ und $P(A) \neq 0$, dann gilt:

$$P(B_i|A) = \frac{P(A|B_i) * P(B_i)}{P(A)} = \frac{P(A|B_i) * P(B_i)}{\sum_{i=1}^n P(A|B_i) * P(B_i)}$$

WAHRSCHEINLICHKEITSGRAPHEN

Wahrscheinlichkeitsgraph (W-Graph) → grafische Darstellungsform der Ermittlung von Wahrscheinlichkeiten; dient zur Beschreibung des Ablaufs eines mehrstufigen Zufallsexperimentes (eines Zufallsprozesses).

Ein **W-Graph** ist ein bewerteter gerichteter Graph mit Baumstruktur →**Baumdiagramm**



In einem **Baumdiagramm** werden die Ausgänge eines Zufallsexperiments als **Linien** dargestellt und die entsprechenden **Wahrscheinlichkeiten** dazugeschrieben. Die Linien entsprechen disjunkten (einander ausschließenden) Ereignissen. Die **Kugelsymbole** oder **Knoten** am Ende jeder Linie und die **Farben** kennzeichnen die einzelnen Versuchsausgänge (die Kugelsymbole können auch durch entsprechende **Beschriftungen** ersetzt werden).

BAUMDIAGRAMME

Pfadregeln für Baumdiagramme:

1. Multiplikationssatz:

Die **Wahrscheinlichkeit** für das Eintreten eines Pfades ist das **Produkt** aller der längs diesen Pfades verzeichneten Wahrscheinlichkeiten.

2. Satz von der totalen Wahrscheinlichkeit:

Die **Wahrscheinlichkeit** eines Ereignisses A ist gleich der **Summe** der Wahrscheinlichkeiten aller Pfade, die zu einem Zustand führen, bei dem das Ereignis A eintritt.

BAUMDIAGRAMME

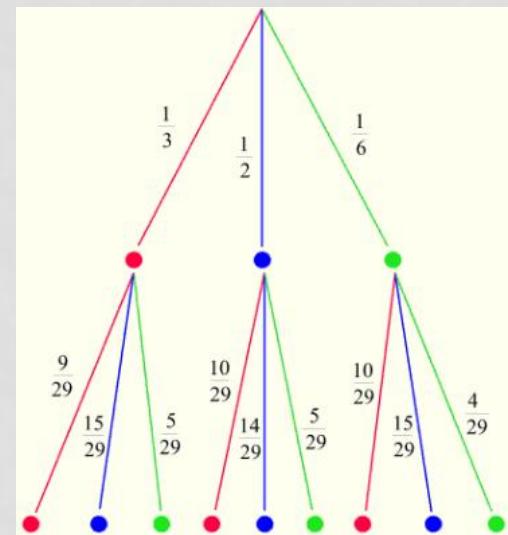
Beispiel aus dem Beispiel 2 (Folie 33):

Aus der Urne werden **hintereinander zwei Kugeln, ohne die erste zurückzulegen, gezogen.**

Mit welcher Wahrscheinlichkeit werden eine **rote** und eine **blaue** Kugel (egal in welcher Reihenfolge) gezogen?

Die Wahrscheinlichkeiten für die erste Ziehung sind dem Baumdiagramm zu entnehmen. Nach der ersten Ziehung sind nun nur 29 Kugeln in der Urne und die Wahrscheinlichkeiten für die zweite Ziehung hängen davon ab, welche Farbe die zuerst gezogene Kugel hat.

Das Prinzip des Baumdiagramms besteht nun darin, an das Ende jeder Linie, die einem Ausgang der ersten Ziehung entspricht, eine weitere Verzweigung anzuhängen, die die zweite Ziehung (unter den entsprechenden neuen Umständen) darstellt.



BAUMDIAGRAMME

Beispiel aus dem Beispiel 2 (Folie 33):

$$p(\text{rote Kugel}) = 1/3$$

$$p(\text{blaue Kugel}) = 1/2$$

$$p(\text{grüne Kugel}) = 1/6$$

Die Summe dieser Wahrscheinlichkeiten ist 1.

Multiplikationsregel für Baumdiagramme:

Die **Wahrscheinlichkeit** für das Eintreten eines Pfades ist das **Produkt der entlang ihm verzeichneten Wahrscheinlichkeiten**:

$$p(\text{erst rot, dann blau}) = (1/3) \times (15/29) = 5/29$$

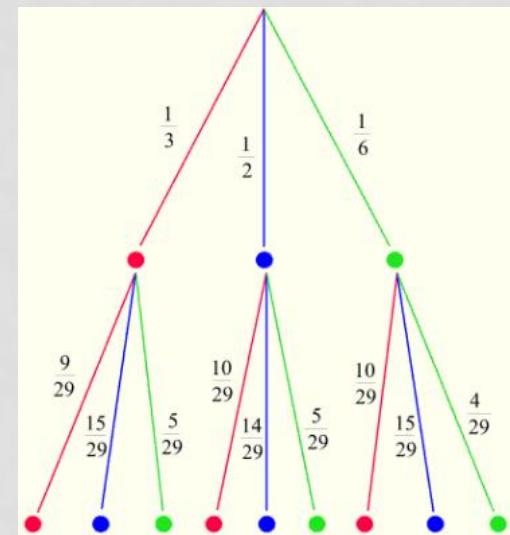
$$p(\text{erst blau, dann rot}) = (1/2) \times (10/29) = 5/29$$

Additionsregel für Baumdiagramme:

Da Pfade disjunkte Ereignisse darstellen, werden

Wahrscheinlichkeiten addiert:

$$p(A) = 5/29 + 5/29 = 10/29$$



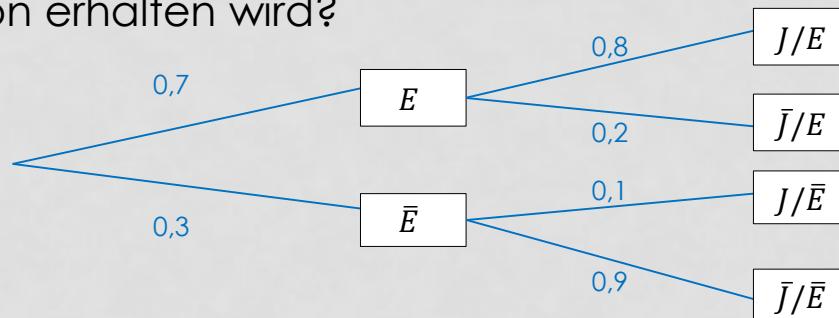
TOTALE WAHRSCHEINLICHKEIT

Beispiel:

Der Informatikstudent glaubt am Anfang seines Studiums, dass er dieses mit einer Wahrscheinlichkeit von 0,7 erfolgreich beenden wird. Mit erfolgreich abgeschlossenem Studium beträgt die Wahrscheinlichkeit, die gewünschte Position zu erhalten, 0,8, ohne Studienabschluss nur 0,1. Wie groß ist die Wahrscheinlichkeit, dass der Student die Position erhalten wird?

E: Ende des Studiums

J: Job



Tipp:
s. auch Folie 48

Gefragt ist hier nach $P(J)$.

Dies ist die **totale Wahrscheinlichkeit** dafür, die gewünschte Position zu erhalten. Dazu muss man die **bedingten Wahrscheinlichkeiten von J** unter allen möglichen Hypothesen mit den Wahrscheinlichkeiten der Hypothesen multiplizieren und die Ergebnisse aufzufaddieren (Regeln des Baumdiagramms):

$$P(J) = \sum_{i=1}^n P(J|E_i) * P(E_i) = 0,8 * 0,7 + 0,1 * 0,3 = 0,59$$

MONTY-HALL-DILEMMA ODER ZIEGENPROBLEM

Tor 1 gewählt	Tor 2	Tor 3	Moderator öffnet ...	Ergebnis beim Wechseln	Ergebnis beim Behalten
Auto	Ziege	Ziege	Tor 2 oder Tor 3	Ziege	Auto
Ziege	Auto	Ziege	Tor 3	Auto	Ziege
Ziege	Ziege	Auto	Tor 2	Auto	Ziege

6 Fälle für das Öffnen der Tore 2 und 3 → das entspricht einem Zufallsexperiment, bei dem die beiden Ziegen voneinander unterscheiden werden können und jede Verteilung von Auto und Ziegen hinter den drei Toren gleich wahrscheinlich ist (**Laplace-Experiment**).

Man sieht, dass in zwei der drei Fälle der Kandidat durch Wechseln das Auto gewinnt.

Der Kandidat hat also eine **durchschnittliche Gewinnchance** von $p = 2/3$. Demnach wäre es für einen Kandidaten, der mehrmals an dieser Spielshow teilnehmen dürfte, von Vorteil, die Wahl des Tors immer zu ändern.

Das gilt nur unter Voraussetzung, dass der Moderator immer eine Ziegentür öffnet und einen Wechsel anbietet → feste Spielregeln!

MONTY-HALL-DILEMMA ODER ZIEGENPROBLEM

Formelle mathematische Lösung mit dem Satz von Bayes

Folgende Situation: der Kandidat hat das **Tor 1 gewählt** und der Moderator hat daraufhin das **Tor 2 geöffnet**. Wie hoch ist die Gewinnwahrscheinlichkeit, wenn der Kandidat das Tor wechselt?

Moderator wählt zufällig mit der gleichen Wahrscheinlichkeit eines der beiden anderen Tore, hinter dem sich immer eine Ziege befindet.

Folgende Ereignisse sind definiert:

M_j : der Moderator hat das Tor j geöffnet ($j=1,2,3$)

G_i : der Gewinn ist hinter Tor i ($i=1,2,3$)

$$P(G_1) = P(G_2) = P(G_3) = \frac{1}{3}$$

$$P(M_2|G_1) = \frac{1}{2}$$

$$P(M_2|G_2) = 0$$

$$P(M_2|G_3) = 1$$

$$P(G_3|M_2) = \frac{P(M_2 \cap G_3)}{P(M_2)} = \frac{P(M_2|G_3)P(G_3)}{P(M_2|G_1)P(G_1) + P(M_2|G_2)P(G_2) + P(M_2|G_3)P(G_3)} = \frac{\frac{1}{2} * \frac{1}{3}}{\frac{1}{2} * \frac{1}{3} + 0 * \frac{1}{3} + 1 * \frac{1}{3}} = \frac{2}{3}$$

Die Gewinnchancen erhöhen sich mit dem Wechsel des Tors von anfangs 1/3 auf nun 2/3.

MONTY-HALL-DILEMMA ODER ZIEGENPROBLEM

Formelle mathematische Lösung mit dem Satz von Bayes

Folgende Situation: der Kandidat hat das **Tor 1 gewählt** und der Moderator hat daraufhin das **Tor 2 geöffnet**. Wie hoch ist die Gewinnwahrscheinlichkeit, wenn der Kandidat das Tor wechselt?

Moderator öffnet immer das Tor 2, vorausgesetzt hinter dem befindet sich eine Ziege und der Kandidat das Tor 2 nicht gewählt hat. Sonst wird das Tor mit der größeren Zahl geöffnet.

Folgende Ereignisse sind definiert:

M_j : der Moderator hat das Tor j geöffnet ($j=1,2,3$)

G_i : der Gewinn ist hinter Tor i ($i=1,2,3$)

$$P(G_1) = P(G_2) = P(G_3) = \frac{1}{3}$$

$$P(M_2|G_1) = 1$$

$$P(M_2|G_2) = 0$$

$$P(M_2|G_3) = 1$$

$$P(G_3|M_2) = \frac{P(M_2 \cap G_3)}{P(M_2)} = \frac{P(M_2|G_3)P(G_3)}{P(M_2|G_1)P(G_1) + P(M_2|G_2)P(G_2) + P(M_2|G_3)P(G_3)} = \frac{1 * \frac{1}{3}}{1 * \frac{1}{3} + 0 * \frac{1}{3} + 1 * \frac{1}{3}} = \frac{1}{2}$$

Die Gewinnchance ist 50% → also unabhängig von der Entscheidung das Tor zu wechseln.