

Inhaltsverzeichnis

1.	Skript – Einführung in die Statistik.....	4
1.1.	Allgemeines zur „Statistik“	4
1.2.	Gegenstand der Statistik	4
1.3.	Bereiche der Statistik.....	5
1.4.	Tabellen vs. Grafiken	6
1.5.	Fachterminologie.....	6
1.6.	Stichprobe	6
1.7.	Zufallsstichprobe	7
1.8.	Klumpenstichprobe	7
1.9.	Willkürliche und Bewusste Auswahlen	7
1.10.	Ablauf einer empirischen Untersuchung (5D's)	8
1.11.	"Operationalisierung" eines Begriffs	9
1.12.	Datenanalyse mit Statistik-Software und Vergleich.....	9
1.13.	Statistikbegriffe in Englisch.....	9
2.	Skript – Skalen und Klassierung	10
2.1.	Messbarkeit	10
2.2.	Grundbegriffe	10
2.3.	Merkmalstypen	11
2.4.	Skalenniveau.....	11
2.5.	Skalierung	12
2.6.	Klassierung	13
3.	Skript – Häufigkeiten und Häufigkeitsverteilung	14
3.1.	Datendokumentation	14
3.2.	Häufigkeitsverteilung.....	14
3.3.	Absolute und relative Häufigkeiten	14
3.4.	Eindimensionale Häufigkeitsverteilung	15
3.5.	Summenhäufigkeiten	15
3.6.	Zweidimensionale Häufigkeitsverteilung.....	16
3.7.	Empirische Verteilungsfunktion.....	17
3.8.	Eigenschaften der Häufigkeitsverteilung	17
3.9.	Grafische Darstellung	18
4.	Skript – Lageparameter	20
4.1.	Allgemein.....	20
4.2.	Modus oder Modalwert	20
4.3.	Median oder Zentralwert	21
4.4.	Arithmetisches Mittel	23
4.5.	Median vs. Mittelwert	24

4.6.	Neigung / Schiefe	25
4.7.	Quantile.....	25
5.	Skript – Streuungsparameter	26
5.1.	Allgemein.....	26
5.2.	Streuungsparameter.....	26
5.3.	Spannweite.....	26
5.4.	Quartil	27
5.5.	Quartilsabstand vs. Spannweite	27
5.6.	Boxplot	28
5.7.	Varianz.....	29
5.8.	Standardabweichung.....	31
6.	Skript – Korrelation und Regression	32
6.1.	Gemeinsame Analyse mehrerer Merkmale	32
6.2.	Zusammenhangsanalyse (Interdependenzanalyse).....	32
6.3.	Abhängigkeitsanalyse (Dependenzanalyse).....	33
6.4.	Multivariate Analysemethode	33
6.5.	Streudiagramm (oder Streuungsdiagramm)	34
6.6.	Korrelation.....	34
6.7.	Probleme bei der Abhängigkeitsanalyse (Scheinkorrelation)	34
6.8.	Korrelation.....	35
6.9.	Probleme bei der Regressionsanalyse	38
6.10.	Regression	38
6.11.	Modell vs. Realität.....	40
6.12.	Prognosewerte und Residuen.....	41
6.13.	Bestimmtheitsmaß	42
6.14.	Probleme bei linearer Regression und Korrelation	43
7.	Skript – Wahrscheinlichkeitsrechnung	44
7.1.	Begriffe	44
7.2.	Wahrscheinlichkeitstheorie	44
7.3.	Zufallsexperimente.....	44
7.4.	Ereignisraum.....	45
7.5.	Ereignis	45
7.6.	Wahrscheinlichkeit	45
7.7.	Mengentheoretische Konzepte	47
7.8.	Gesetze	49
7.9.	Laplace.....	50
7.10.	Gegenwahrscheinlichkeit.....	50
7.11.	Elementare Kombinatorik.....	51
7.12.	Permutation mit Wiederholung.....	53

7.13.	Bedingte Wahrscheinlichkeit	54
7.14.	Totale Wahrscheinlichkeit	54
7.15.	Satz von Bayes	55
7.16.	Wahrscheinlichkeitsgraphen	55
8.	Formelsammlung.....	56
9.	Klausuraufgaben WS19/20	59

1. Skript – Einführung in die Statistik

1.1. Allgemeines zur „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).**

1.2. Gegenstand der Statistik

1.2.1. 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** → $n < N$

1.2.2. Datenanalysen

- Anwendung statistischer Verfahren zum Zweck der Erkenntnisgewinn

1.2.3. 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

1.2.4. Datenbeurteilung

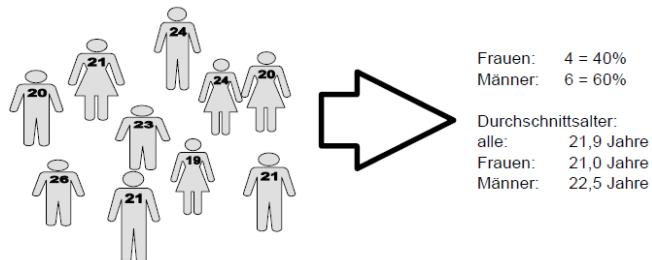
- Die Beurteilung von Daten erfolgt durch:
 - Schlüsse auf der Basis unvollständiger Daten, z. B. Schlüsse von der Stichprobe auf ihre Grundgesamtheit
 - Allgemeiner: auf der Basis unsicherer Daten, unter Anwendung der Wahrscheinlichkeitsrechnung. Dies ist Gegenstand der induktiven (schließenden) Statistik.

1.2.5. Datenaufbereitung

- Ordnung, Zusammenfassung und Darstellung des erhobenen statistischen Datenmaterials in Datendateien, Tabellen und/oder geeigneten Grafiken.
- 1.2.6. 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

1.3. Bereiche der Statistik

1.3.1. deskriptive oder beschreibende Statistik



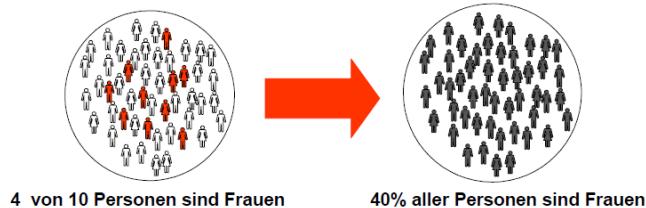
Die deskriptive Statistik (lat.: *descriptio* Beschreibung) dient der Betrachtung der Daten an sich. Die gewonnenen Daten werden verdichtet bzw. so dargestellt, dass das Wesentliche deutlich hervortritt. Für eine übersichtliche Darstellung muss das, oft sehr umfangreiche, Material auf geeignete Art und Weise zusammengefasst werden.

Dazu werden insbesondere die drei Darstellungsformen benutzt:

- Tabellen
- grafische Darstellungen
- charakteristische Maßzahlen

1.3.2. 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“).

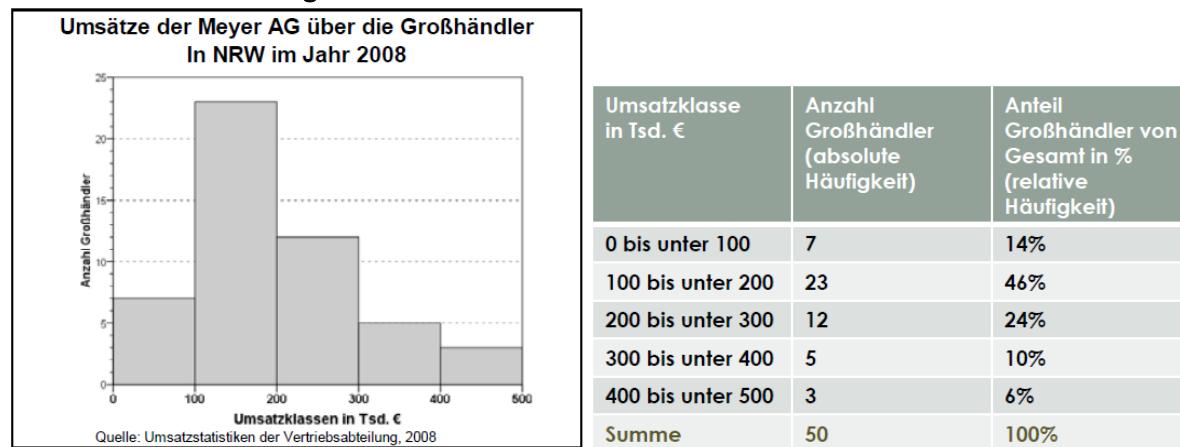
Weitere Synonyme für induktive Statistik: analytische oder inferentielle Statistik.

1.4. 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.

1.5. 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

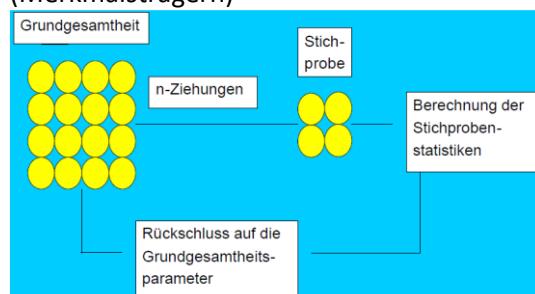
Information der Tabelle

(klassierte) Häufigkeitsverteilung mit absoluten und relativen (Klassen-)Häufigkeiten

1.6. 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.

1.7. Zufallsstichprobe

1.7.1. 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.

1.7.2. 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.

1.8. 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.

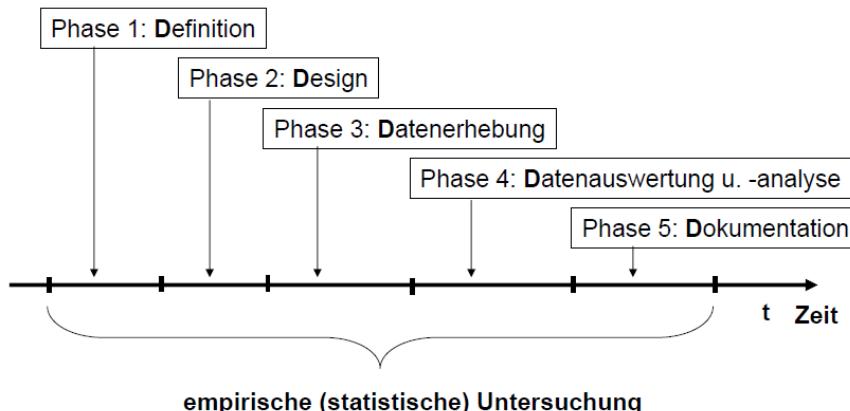
Beispiel:

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.

1.9. 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)

1.10. Ablauf einer empirischen Untersuchung (5D's)



1.10.1. 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!

1.10.2. 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
- „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

1.10.3. Datenerhebung (Phase 3)

- entsprechend der getroffenen Entscheidungen

1.10.4. 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

1.10.5. 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

1.11. "Operationalisierung" eines Begriffs

Seite 25 und 26 im Dokument [Modul 1 Einführung in die Statistik](#)

1.12. Datenanalyse mit Statistik-Software und Vergleich

Seite 27 und 33 im Dokument [Modul 1 Einführung in die Statistik](#)

1.13. Statistikbegriffe in Englisch

Deutsch	englisch
Grundgesamtheit	population
Stichprobe	sample
arithmetisches Mittel	mean
Modus	mode
Spannweite	range
Varianz	variance
Standardabweichung	Standard deviation (stddev)

2. Skript – Skalen und Klassierung

2.1. 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.



2.2. Grundbegriffe

Grundbegriffe

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

2.3. Merkmalstypen

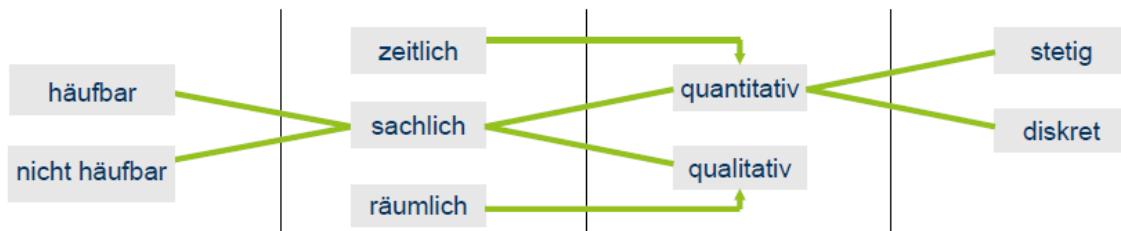
2.3.1. Qualitative Merkmale

Merkmale lassen sich nicht mit Zahlen messen Codierung und Rangordnung möglich) z.B. Geschlecht, Güteklaasse

- Diskrete Merkmale
 - Qualitative Merkmale sind immer diskret, da sie von Natur aus nur eine abzählbare Menge möglicher Merkmalswerte haben

2.3.2. Quantitative = metrische Merkmale

- Stetige Merkmale
 - Menge der Merkmalsausprägungen überabzählbar, Intervall der reellen Zahlen (es gibt zwischen zwei Ausprägungen immer noch weitere Zwischenwerte) z.B. Gewicht, Alter, Fahrzeit
- Diskrete Merkmale
 - Menge der Merkmalsausprägungen endlich bzw. abzählbar (i.d.R. ganze Zahlen) z.B. Kinderzahl, Sitzplätze, das monatliche Gehalt



2.4. 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**:

- Nicht metrische Skalen → Anwendung bei qualitativen Merkmalen. Keine Rechenoperationen mit den Merkmalsausprägungen zulässig:
 - **Nominalskala**
 - **Ordinalskala**
- **Metrische Skalen** (Kardinalskalen) → Anwendung bei quantitativen Merkmalen. Skala hat Nullpunkt und Maßeinheit. Rechenoperationen sind zulässig:
 - **Intervallskala**
 - **Verhältnisskala** (Ratioskala)
 - **Absolutskala**

2.5. 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 keinen natürlichen Nullpunkt, keine Verhältnisse können gebildet werden. Daten können alle (<i>unendlich viele</i>) 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 skalierten 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

Beispiele:

Merkmal	Menge der Merkmalsausprägungen	Messinstrument	Skala	Merkmalstyp
Familienstand	{ledig, verheiratet, verwitwet, geschieden}	Frage	Nominalskala	qualitatives Merkmal
Hotelpünktelklasse	{*****; ****; ***; **; *; };	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 (°C)	I	Thermometer	Metrische Skala = Intervallskala	Quantitative Merkmale
Körpergröße	{x x ∈ I und x > 0}	cm-Maß	Metrische Skala = Verhältnisskala	= metrische Merkmale
Kinderzahl	I N ∪ {0}	Frage	Metrische Skala = Absolutskala	

2.6. Klassierung

2.6.1. Qualitative Merkmale

Beispiel:

Merkmal: Beruf

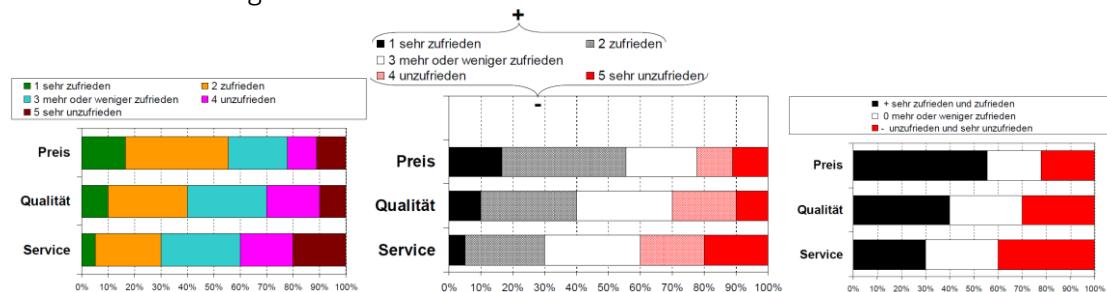
Merkmalsausprägung:

→ Berufsgruppe: Handwerker = Klasse von z.B.

- Maurer
- Dachdecker
- Schreiner
- Fliesenleger

Zielkonflikt: Übersichtlichkeit versus Informationsverlust

2.6.2. Rangmerkmale



2.6.3. Metrische Merkmale

diskret

z.B. Einwohnerzahl

stetig

z.B. Körpergröße

Klassierung	Klassierung	Klassierung
1. 0 – 19.999	0 – 99 cm	0 – 100 cm
2. 20.000 – 49.999	100 – 39 cm	100 – 140 cm
3. 50.000 – 99.999	140 – 159 cm	140 – 150 cm
4. 100.000 – 249.999 usw.	160 – 169 cm usw.	150 – 170 cm usw.

wg. Lücken wg. Über-
schneidungen

richtig !!!

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

2.6.4. Entscheidung bei der 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“

3. Skript – Häufigkeiten und Häufigkeitsverteilung

3.1. Datendokumentation

Formen als Dokumentation der Daten:

Einzelwerte (Einzelbeobachtungen)

→ ungeordnete Reihe (Urliste, Rohdaten, Primärdaten)

→ 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 Datensätzen 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

3.2. Häufigkeitsverteilung

Die Daten einer Urlistemüssen 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 klassierten Daten → **Häufigkeitsverteilung** für klassierte Daten (klassierte Verteilung)
- **Schritt 3:** Darstellen der klassierten Daten → **Häufigkeitstabelle** für klassierte Daten

3.3. Absolute und relative Häufigkeiten

Merkmalsausprägung und zugehörige Häufigkeit

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

absolute Häufigkeit

$$x_j \longrightarrow f(x_j) = \frac{h(x_j)}{n} \quad j = 1, \dots, m$$
$$f(x_j) = \frac{h(x_j)}{n} \cdot 100 (\%)$$

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

3.4. Eindimensionale Häufigkeitsverteilung

Beispiel 1:

- n = 20 Personen wurden gefragt nach dem Merkmal X: Familienstand mit den j=4 Merkmalsausprägungen:
x1= ledig, x2= verheiratet, x3= geschieden, x4= 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

- 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

- Beispiel 2:
 - Frage: Wo wohnen Sie?
 - Antworten: B C A B C B BBA AD 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%

3.5. Summenhäufigkeiten

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)

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

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

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

...

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

.....

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

3.5.1. Eindimensionale Häufigkeitsverteilung mit Summenhäufigkeiten

- sinnvoll nur für Rangmerkmale und metrische Merkmale

Index	Merkmalsausprägungen	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%	-	-	-

3.5.2. Eindimensionale klassierte Häufigkeitsverteilung mit Summenhäufigkeiten

Klasse Nr.	Klasse	absolute Häufigkeit	relative Häufigkeit in %	absolute Summenhäufigkeit	relative Summenhäufigkeit	Klassenbreite	Klassenmitte
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%	$170 + 30 = 200$	$85 + 15 = 100\%$	$200 - 100 = 100$	$(200 + 100) / 2 = 150$
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 der Klasse i
 - $m_i = 1/2(x_{k-1} + x_k)$

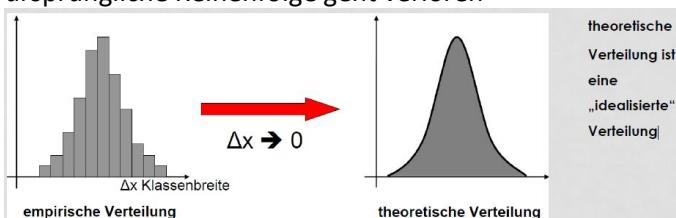
3.6. Zweidimensionale Häufigkeitsverteilung

Zweidimensionale Häufigkeitsverteilung: G → M Kreuztabelle

		Randverteilung: eindim. Häufigkeitsverteilung von G		
		w	m	Σ
		400	800	1.200
		40,0%	80,0%	60%
		33,3%	66,7%	40,0%
		20,0%	40,0%	
		600	200	800
		60,0%	20,0%	40%
		75,0%	25,0%	
		30,0%	10,0%	
		1.000	1.000	2.000
		50%	50%	100%
Absolute Häufigkeiten der Merkmalsausprägungskombinationen				
Relative Spaltenhäufigkeiten (bedingte relative Häufigkeiten)				
Relative Zeilenhäufigkeiten (bedingte relative Häufigkeiten)				
Relative Häufigkeiten der Merkmalsausprägungskombinationen				

3.7. Empirische Verteilungsfunktion

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

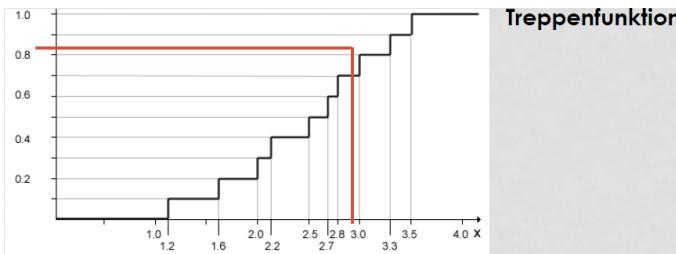


In der mathematischen Statistik klingt das dann in etwa wie folgt:

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

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 nicht fallend (steigt oder ist konstant)
- $F(x)$ ist eine Treppenfunktion mit Sprungstellen bei x_1, x_2, \dots, x_i
- Die Größe der Sprünge beträgt $f_i = F(x_i) - F(x_{i-1})$



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

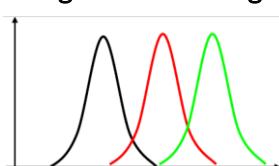
Greift man auf der x-Achse den Wert 3 heraus, so lässt sich der dazugehörige y-Wert 0.8 wie folgt interpretieren: 80% der Abiturienten haben im schlechtesten Fall den Notendurchschnitt 3 bekommen.

Anders formuliert:

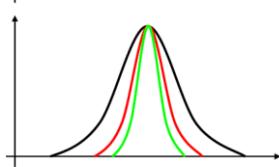
Der Notendurchschnitt ist bei 80 % der Schüler kleiner oder gleich 3.

3.8. Eigenschaften der Häufigkeitsverteilung

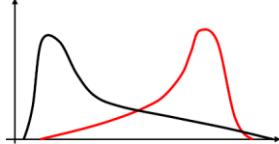
Lage



Streuung = Wölbung, Form



Schiefe



3.9. Grafische Darstellung

- 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)

3.9.1. Säulendiagramm

- höhenproportionale Darstellungsform einer Häufigkeitsverteilung durch auf der **x-Achse senkrecht stehende, nicht aneinandergrenzende 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.

3.9.2. Stabdiagramm

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

3.9.3. 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)

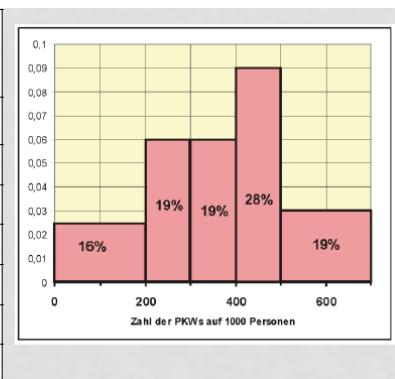
3.9.4. Kreisdiagramm

- **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

3.9.5. 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 = $c \cdot f(x_i)$, wobei $f(x_i)$ 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).

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	-	-

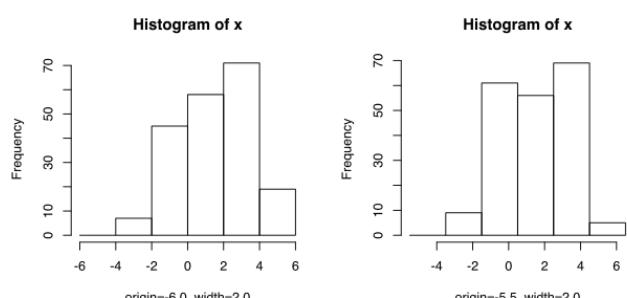


Beispiel:

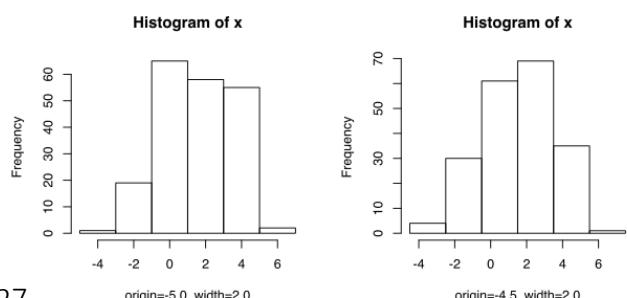
Vier Histogramme für den gleichen Datensatz:

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



3.9.6. Weitere siehe [Modul 3](#) ab Seite 27

- Segmente
- Wärmekarte
- Punkte-Wärmekarte
- Clusteranalyse mit Zentroiden
- Netzwerkanalyse in einer geographischen Karte

4. Skript – Lageparameter

4.1. Allgemein

- Lageparameter beschreiben die "Lage" der Elemente der Grundgesamtheit bzw. der Stichprobe in Bezug auf die Messskala
- noch Lokationsmaße genannt
- Allgemeine Mittelwerte:
 - Modus \bar{X}_D
 - Median \bar{X}_z
 - arithmetisches Mittel \bar{X}
 - Quantil \bar{X}_p
- Spezielle Mittelwerte:
 - geometrisches Mittel \bar{X}_G
 - harmonisches Mittel \bar{X}_H

4.2. Modus oder Modalwert

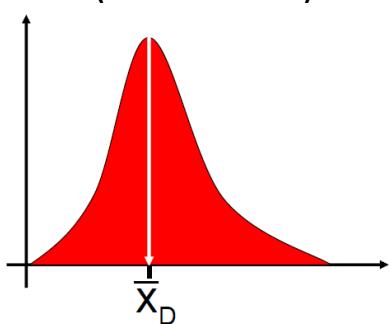
Der Modus oder Modalwert ist die am häufigsten auftretende Merkmalsausprägung (maximale Häufigkeit). Er wird hauptsächlich für nominale Merkmale verwendet, ist aber auch für alle anderen (diskreten) Merkmalstypen sinnvoll.

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

Bemerkung:

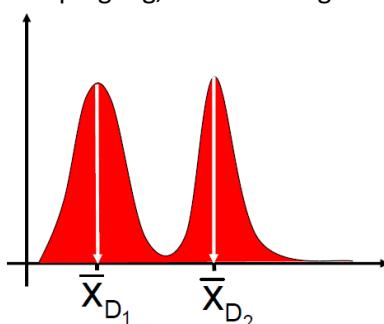
Gibt es mehrere Merkmalsausprägungen mit der gleichen maximalen Häufigkeit, so existieren mehrere Modalwerte → Multimodale Verteilungen (bimodale Verteilung: zwei Modalwerte; trimodale Verteilung: drei Modalwerte; usw.)

Modus (oder Modalwert): Merkmalsausprägung, die am häufigsten vorkommt



unimodale Verteilung

Dichtekurve hat nur
ein lokales Maximum Maxima



multimodale Verteilung

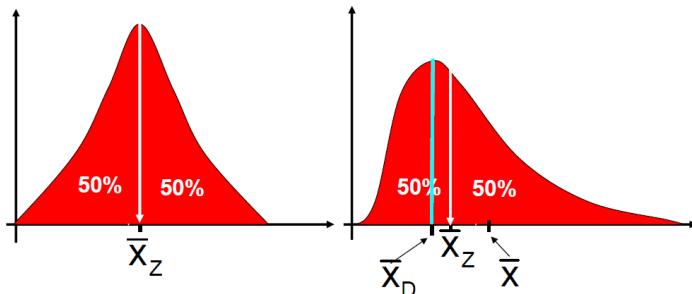
Dichtekurve hat mehrere lokale
(bimodale Verteilung, Trimodale Verteilung usw.)

4.3. Median oder Zentralwert

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.

Wert in der Mitte der geordneten Reihe. 50% der Beobachtungswerte liegen unter dem Median, 50% darüber.



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

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

Beispiel 1: n ungerade

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.

Beispiel 2: n gerade

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.

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

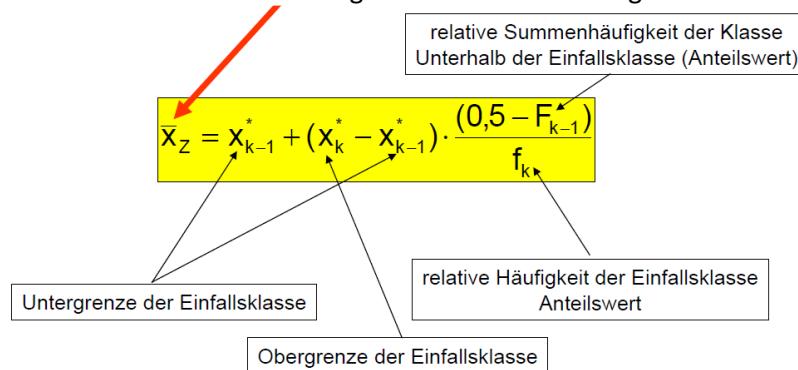
4.3.1. Median bei klassierten Daten

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

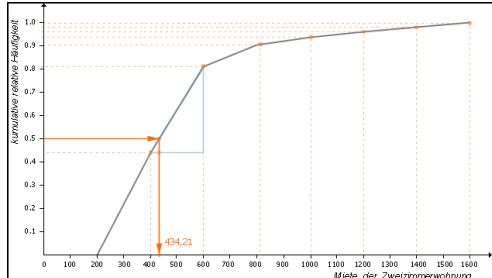
$$\bar{x}_z := x_{k-1} + (x_k - x_{k-1}) * \frac{0,5 - F_{k-1}}{f_k}$$

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

- 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.



Beispiel 1:



Ergebnis:

Näherungswert für Median aus klassierten Daten
Median = 434,21 €.

Der tatsächliche Median der Daten ist 423 €.
Achtung!! Es existiert Fehler der Näherung.

Beispiel 2:

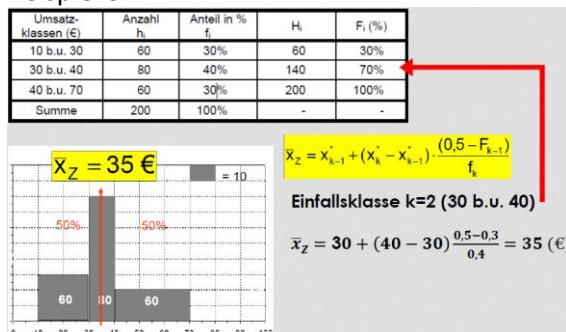
Klassen-Nr. i	Größenklassen (cm)	h_i	$f_i (\%)$	H_i	$F_i (\%)$
1	100 b.u. 150	40	40%	40	40%
2	150 b.u. 170	40	40%	80	80%
3	170 b.u. 200	20	20%	100	100%
Summe		100	100%	-	-

Einfallsklasse: $k = 2$

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

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

Beispiel 3:



4.4. Arithmetisches Mittel

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

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

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 schiefen Verteilungen!)

4.4.1. Arithmetisches Mittel aus Häufigkeitstabellen

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

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

Fall 1: Absolute Häufigkeit $h(x_i)$

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

$h(x_i)$ absolute Häufigkeit der Merkmalsausprägung x_i
 n Summe der absoluten Häufigkeiten
 j Anzahl der Merkmalsausprägung x_i

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:

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

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

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:

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

Durchschnittsnote:

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

$f(x_i)$ relative Häufigkeit der Merkmalsausprägung x_i
 n Summe der absoluten Häufigkeiten
 j Anzahl der Merkmalsausprägung x_i

4.4.2. 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.

Fall 1: Absolute Häufigkeit h_i

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:

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

26

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

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:

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

Durchschnittsgewicht:

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

4.5. 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
-----	-----	-------	-------	-------	-------	-------	-------	-------

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

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

4.6. Neigung / Schiefe

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

Rechtsschiefe (linkssteile) Häufigkeitsverteilung:

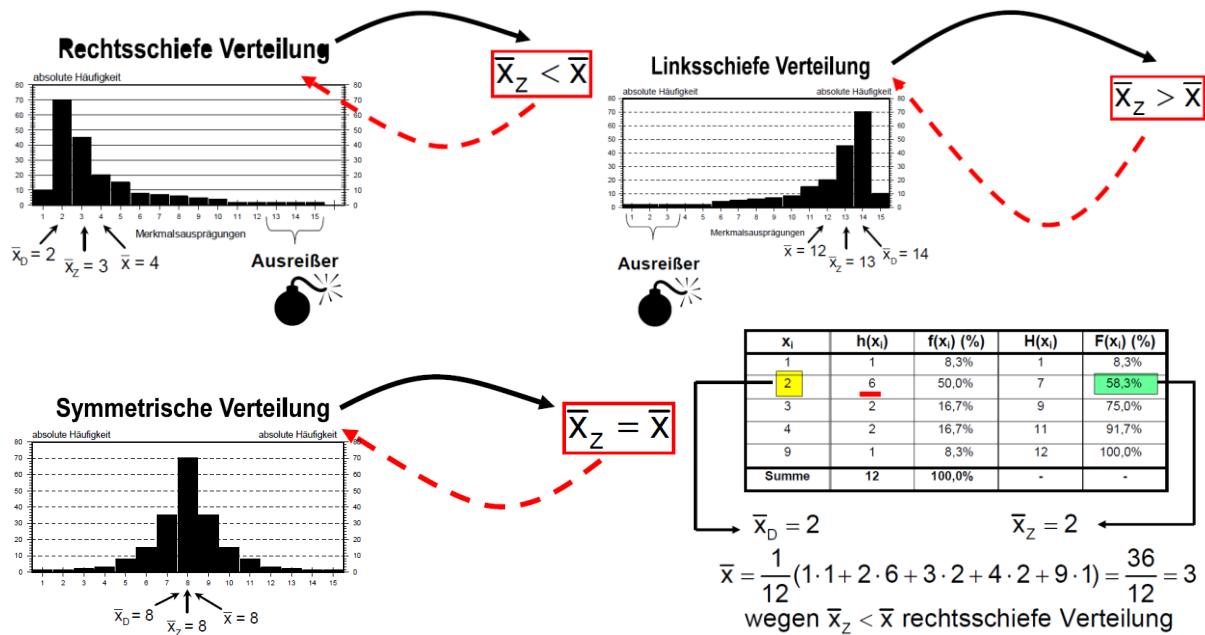
Modus < Median < arithmetisches Mittel $\bar{x}_D < \bar{x}_Z < \bar{x}$

linksschiefe(rechtssteile) Häufigkeitsverteilung:

Modus > Median > arithmetisches Mittel $\bar{x}_D > \bar{x}_Z > \bar{x}$

Unimodale symmetrische Häufigkeitsverteilung:

Modus \approx Median \approx arithmetisches Mittel $\bar{x}_D \approx \bar{x}_Z \approx \bar{x}$

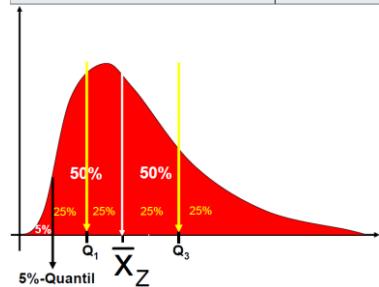


4.7. Quantile

Ein Quantil ist ein Lagemaß in der Statistik.

Quantile teilen eine Verteilung in Abschnitte gleicher Häufigkeit.

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



5. Skript – Streuungsparameter

5.1. Allgemein

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.

- die Berechnung des Durchschnitts ist nicht immer sinnvoll
- der Durchschnitt kann offensichtlich nicht immer alles beschreiben

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

5.2. 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

Spannweite w:

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

(Inter)Quartilsabstand:

$$Q_A = IQR = Q_3 - Q_1$$

Varianz:

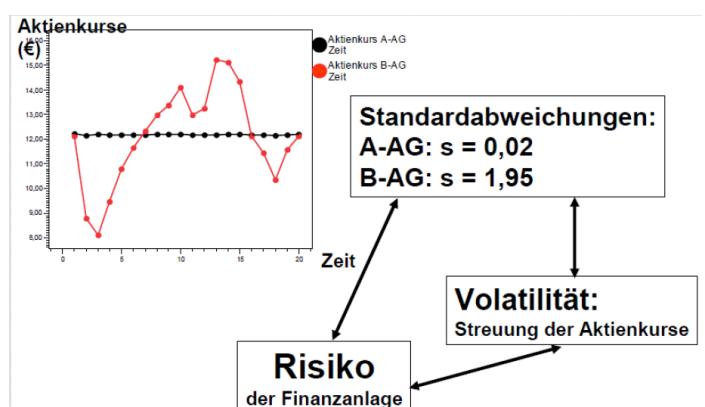
$$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

Standardabweichung:

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

Variationskoeffizient:

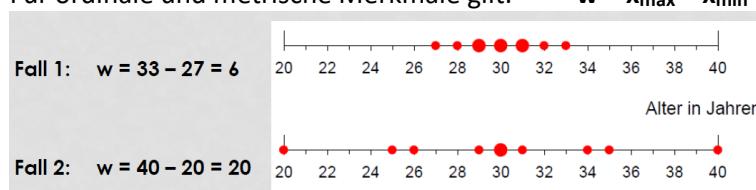
$$v = \frac{s}{\bar{x}}$$



5.3. Spannweite

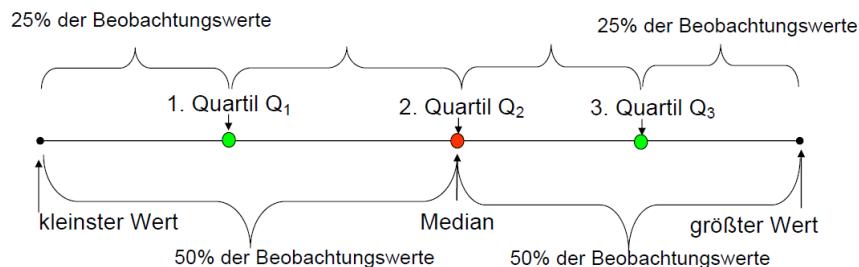
Spannweite (oder Variationsbreite) w: Ausdehnung der Werte (Maß für die Breite des Streubereichs einer Häufigkeitsverteilung)

Für ordinale und metrische Merkmale gilt: $w = x_{\max} - x_{\min}$



5.4. Quartil

5-Punkte Zusammenfassung der geordneten statistischen Reihe:



Der (Inter-) **Quartilsabstand** (engl.: *interquartilerange*, IQR) bezeichnet die Differenz zwischen dem oberen und dem unteren Quartil $Q_3 - Q_1$ und umfasst daher 50% der Verteilung.

Der Quartilsabstand wird als **Streuungsmaß** verwendet.

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.

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
	25%	25%	25%	25%									
		1. Quartil Q_1		2. Quartil Q_2		3. Quartil Q_3							
				50%									
				Quartilsabstand									
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
	25%	25%	25%	25%									
		1. Quartil Q_1		2. Quartil Q_2		3. Quartil Q_3							
				50%									
				Quartilsabstand									

$\bar{x}_2 = Q_2 = x_{\frac{n+1}{2}} = x_{\frac{13+1}{2}} = x_7 = 1,70$

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

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

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

5.5. Quartilsabstand vs. Spannweite

Quartilsabstand

Von Ausreißern unabhängig

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

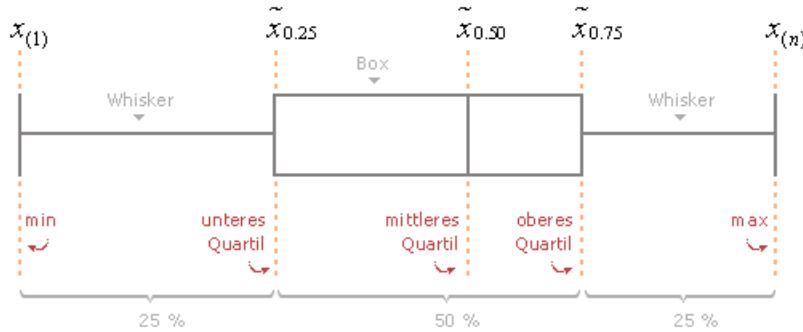
Spannweite

Vom kleinsten und größten Wert abhängig

Gibt die Gesamtbreite an, in dem alle Werte liegen

5.6. Boxplot

- Die grafische Darstellung der 5-Punkte-Zusammenfassung heißt
 - Box-and-Whisker-Plot
- Die 5-Punkte-Zusammenfassung besteht aus:
 - Minimum, Q1, Median, Q3, Maximum



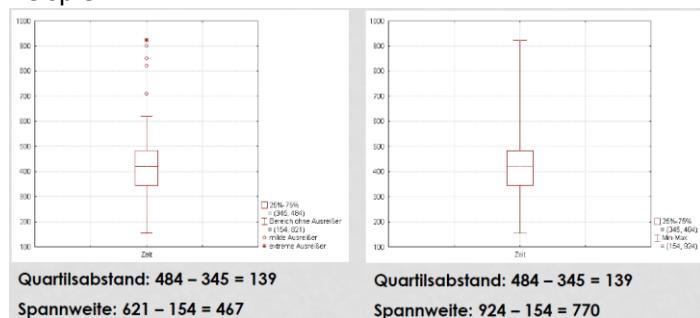
Aus einem **Boxplot** lassen sich Informationen über die:

- 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.

Beispiel:



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

5.7. 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.

Konstruktion der Varianz:

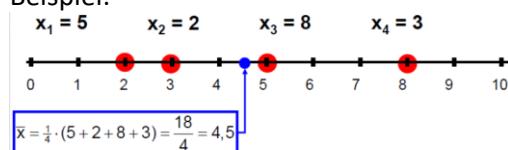
Bezugspunkt:	\bar{x}
Einzelstreuung/Einzelabweichung:	$(x_i - \bar{x})$
Summe der Einzelabweichungen:	$\sum_{i=1}^n (x_i - \bar{x})$
Summe der quadratischen Abweichungen:	$\sum_{i=1}^n (x_i - \bar{x})^2$
Varianz:	$s^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$
$s^2 = \frac{1}{n} ((x_1 - \bar{x})^2 + \dots + (x_n - \bar{x})^2) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = (\underbrace{\frac{1}{n} \sum_{i=1}^n x_i^2}_{\text{Formel (1)}}) - \underbrace{\bar{x}^2}_{\text{Formel (2)}}$	

Bemerkung:

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

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

Beispiel:



Berechnung der Varianz

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

5.7.1. 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

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

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

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:

$$n = \sum_{i=1}^6 h_i = 5 + 8 + 14 + 16 + 5 + 2 = 50$$

Berechnung der Varianz über die absolute Häufigkeit:

i	x_i	h_i	$x_i h_i$	\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

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

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	$\bar{x} = 3,28$			$s^2 = 1,562$

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

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

Fall 1: Absolute Häufigkeit h_i

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

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

h_i	absolute Häufigkeit der i -ten Klasse
n	Summe der absoluten Häufigkeiten
k	Anzahl der Klassen
m_i	Klassenmitte der i -ten Klasse

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

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:

$$n = \sum_{i=1}^4 h_i = 9 + 12 + 7 + 2 = 30$$

Berechnung der Varianz über die absolute Häufigkeit:

i	Klasse x_i	m_i	h_i	$m_i h_i$	\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

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

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	$\bar{x}=165,67$			80,51

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

5.8. 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 $[\bar{x} - 2s, \bar{x} + 2s]$

6. Skript – Korrelation und Regression

6.1. Gemeinsame 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)
→ **univariateAnalyse**
- Mehrere Merkmale werden gemeinsam ausgewertet →**multivariate Analyse**

$$X \leftarrow \rightarrow Y$$

Zusammenhang (z. B. : Lohnniveau $\leftarrow \rightarrow$ Preisniveau)

$$X \rightarrow Y$$

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

- 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?

6.2. 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**
- **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

6.2.1. 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

6.3. 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

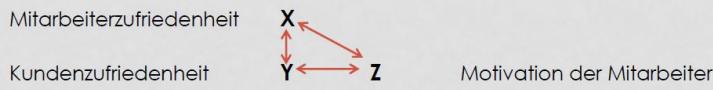
- Abhängigkeitsanalyse → Regressionsanalyse
- Abhängigkeitsmaß → Regressionsfunktion $\hat{y} = a + b*x$
- Grafische Darstellung → Streudiagramm + Regressionsgerade

6.4. Multivariate Analysemethode

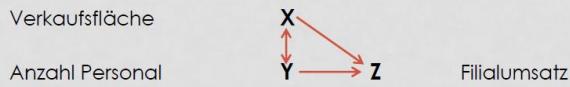
X, Y, Z Merkmale

Beispiel:

Zusammenhangsanalyse (Interdependenzanalyse)



Abhängigkeitsanalyse (Dependenzanalyse)



6.5. Streudiagramm (oder Streuungsdiagramm)

Ein Streudiagramm (*engl. scatterplot*) ist die graphische Darstellung von beobachteten Wertepaaren zweier Merkmale. Diese Wertepaare werden in ein kartesisches Koordinatensystem eingetragen, wodurch sich eine Punktfolge 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 Punktfolge lassen sich die **Stärke** und die **Richtung** des Zusammenhangs der Merkmale ablesen. Das Streudiagramm liefert erste Hinweise über eine mögliche Abhängigkeit zwischen Merkmalen.

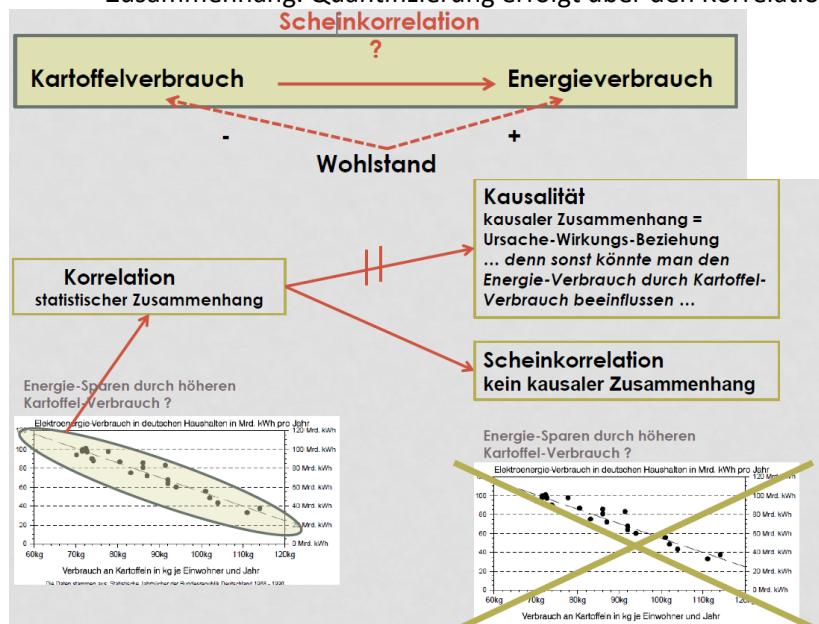
6.6. Korrelation

- 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.

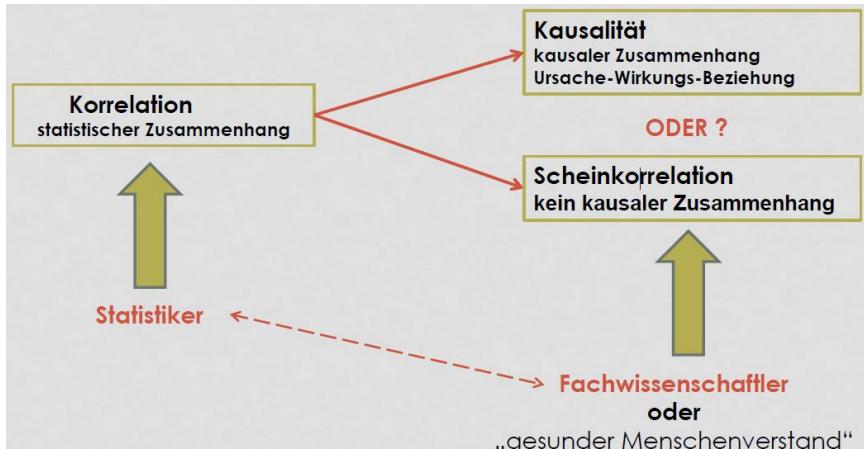
6.7. Probleme bei der Abhängigkeitsanalyse (Scheinkorrelation)

Problem: Abhängigkeitsanalyse muss sinnvoll sein!!
(Korrelation ≠ Kausalität)

- Kausalität → eine Ursache-Wirkungs-Beziehung zw. X und Y, d.h. wenn eine Veränderung des einen Merkmals eine Veränderung bei dem anderen Merkmal hervorruft.
- Korrelation → ist eine notwendige aber keine hinreichende Voraussetzung für einen kausalen Zusammenhang. Quantifizierung erfolgt über den Korrelationskoeffizienten.

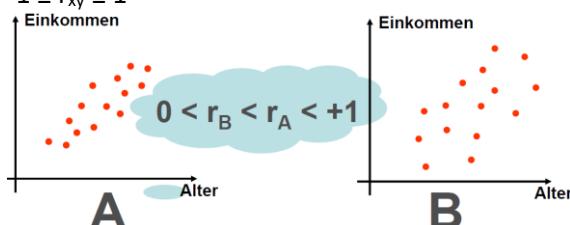


6.8. Korrelation



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
- **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.

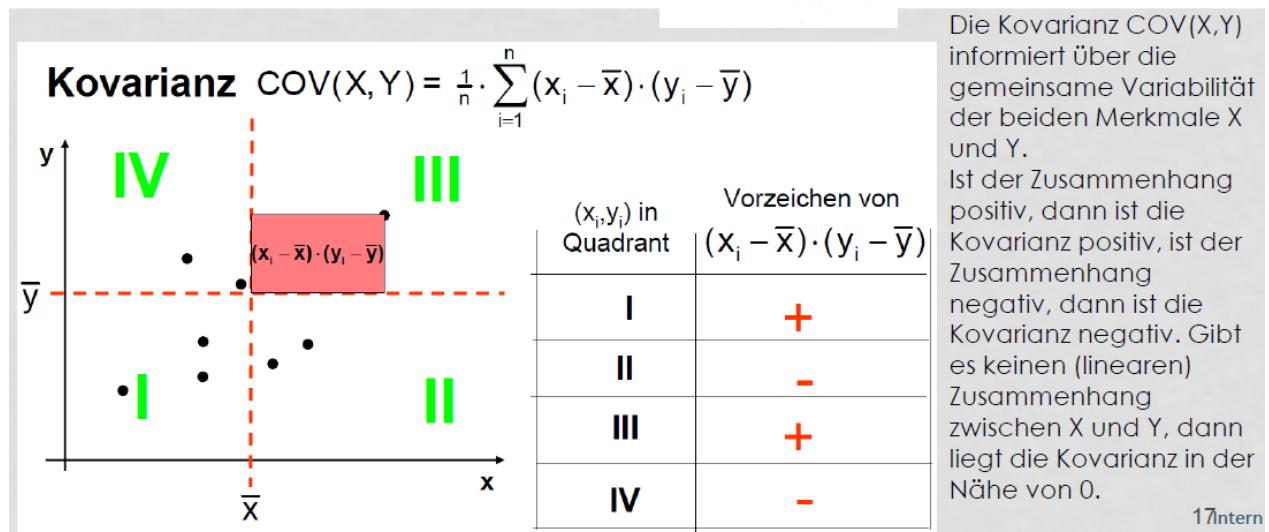
6.8.1. der Korrelationskoeffizient (nach Pearson)

Definition:

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

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

durch:



„Standardisierung“ der Kovarianz:

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

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

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

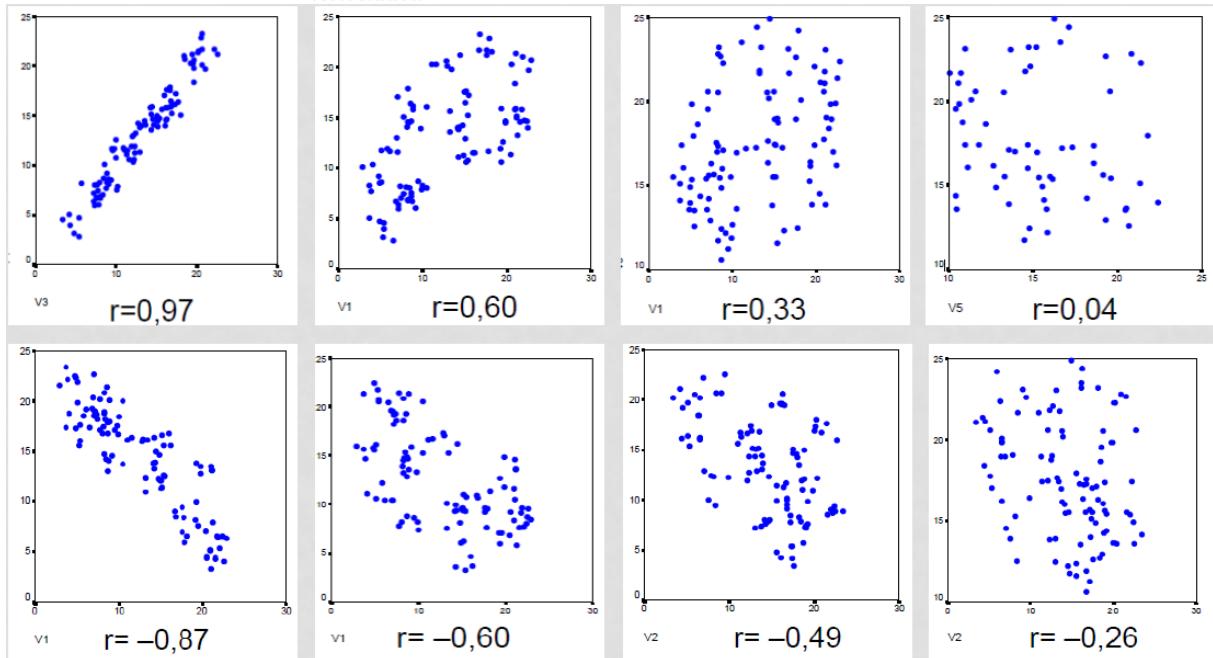
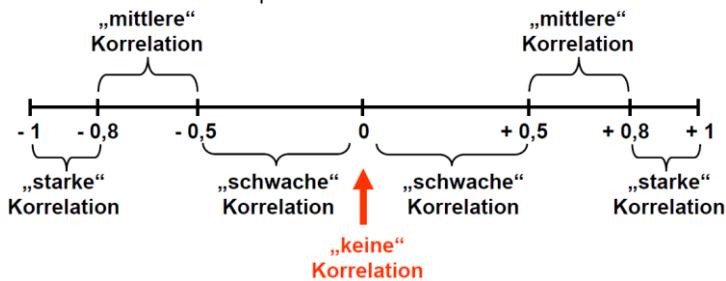
$$r_{XY} = r = \frac{\text{COV}(X, Y)}{s_X \cdot s_Y} = \frac{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{s_X \cdot s_Y} = \frac{\frac{1}{n} \sum_{i=1}^n x_i \cdot y_i - \bar{x} \cdot \bar{y}}{\sqrt{\frac{1}{n} \sum_{i=1}^n x_i^2 - \bar{x}^2} \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n y_i^2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

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

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

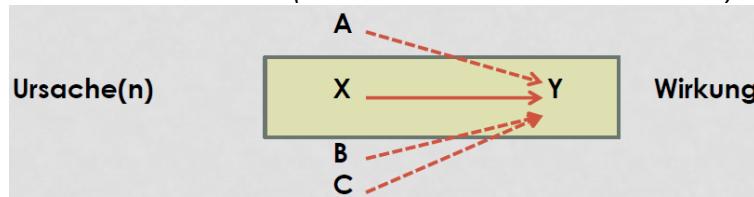
$$\frac{\frac{1}{4} \cdot 450 - 4 \cdot 25}{\sqrt{\frac{1}{4} \cdot 74 - 4^2} \cdot \sqrt{\frac{1}{4} \cdot 3000 - 25^2}} = \frac{112,5 - 100}{\sqrt{2,5} \cdot \sqrt{125}} = \frac{12,5}{1,581 \cdot 11,180} = \frac{12,5}{17,676} = +0,707$$

6.8.2. Interpretation



6.9. 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.

6.10. Regression

Voraussetzungen:

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

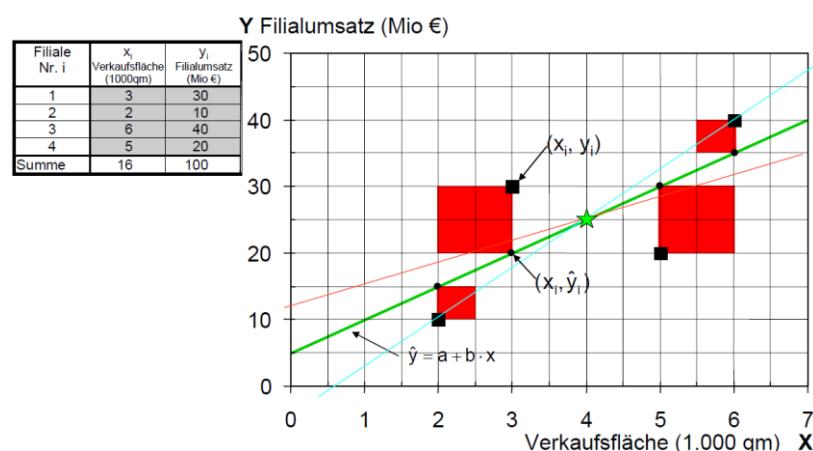
Vorbereitende Arbeiten:

- 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)

Beispiel:

Verkaufsfläche \rightarrow Filialumsatz



Bestimmung der Regressionsfunktion $\hat{y} = a + b \cdot x$ 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:

Residuen u_i

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

↑

Methode der kleinsten Quadrate (OLS = ordinary least squares)

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

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

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

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

6.10.1. 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.

6.10.2. Interpretation

Regressionskoeffizienten a und b

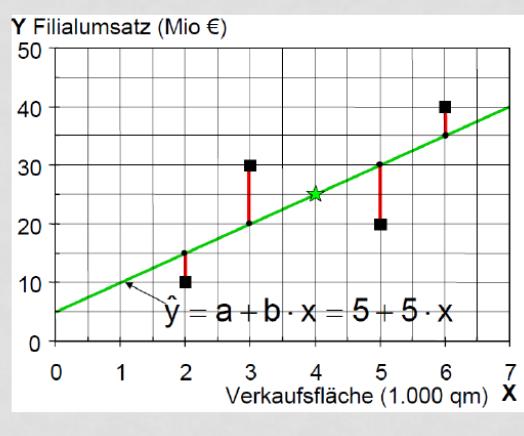
Beispiel:

Fragen für die Schätzung:

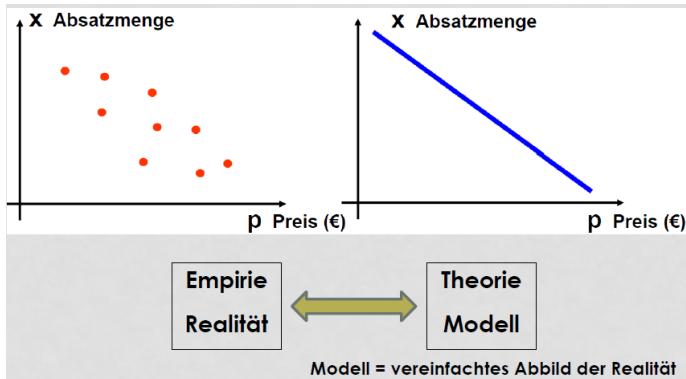
- **Umsatzprognose für neue Filiale mit 4.500 qm Verkaufsfläche:**
 $5 + 5 \cdot 4,5 = 27,5$ Mio €

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

mit Erweiterung: $5 + 5 \cdot (4,5+1,0) = 5 + 5 \cdot 5,5 = 32,5$ Mio €



6.11. Modell vs. Realität



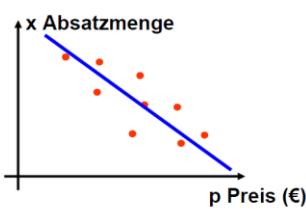
- Wie gut beschreibt das Modell die Realität?
- Wie gut wird die Realität durch das Modell wiedergegeben?

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?

Beispiel:

Verkaufsflächen
unterschiedlich groß



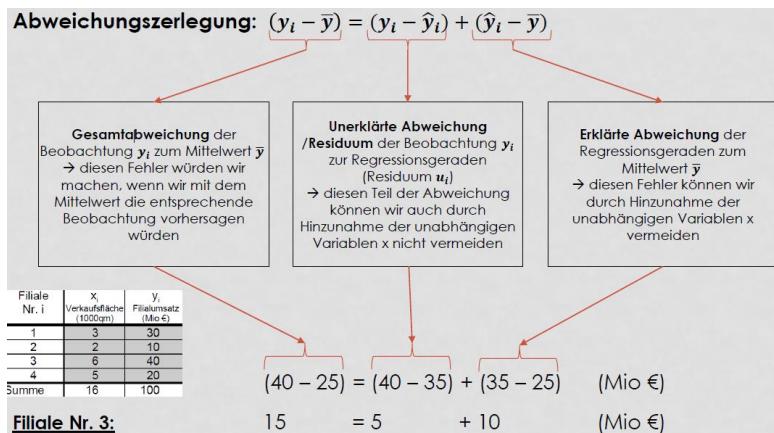
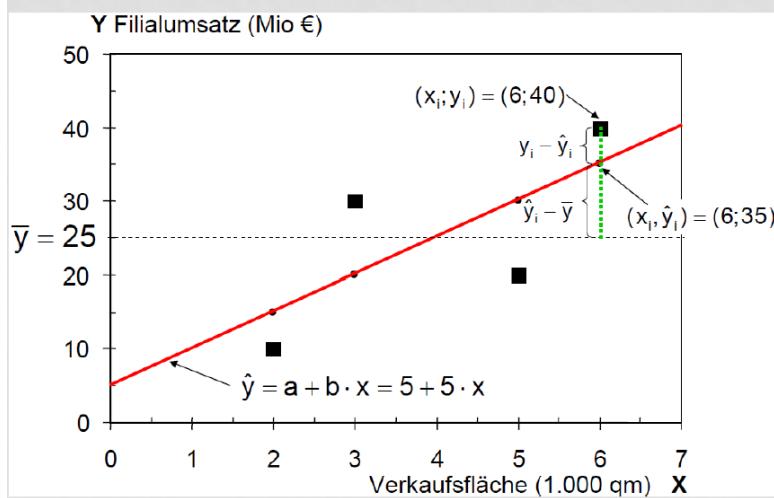
Filialumsatz
unterschiedlich hoch

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?

6.12. Prognosewerte und Residuen

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



Varianzzerlegung:

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

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

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

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

Filiale Nr. i	X Verkaufsfläche (1000qm)	Y 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

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

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

$s_{\hat{y}}^2 = (\frac{1}{n} \cdot \sum_{i=1}^n \hat{y}_i^2) - \bar{\hat{y}}^2 = \frac{1}{4} \cdot (20^2 + 15^2 + 35^2 + 30^2) - 25^2 = 687,5 - 625 = 62,5$

Filiale Nr. i	X Verkaufsfläche (1000qm)	Y 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:

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

6.13. 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.

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

Es folgt: R² ist das Verhältnis aus der Streuung der Prognosewerte und der Gesamtstreuung der y-Werte

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

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:

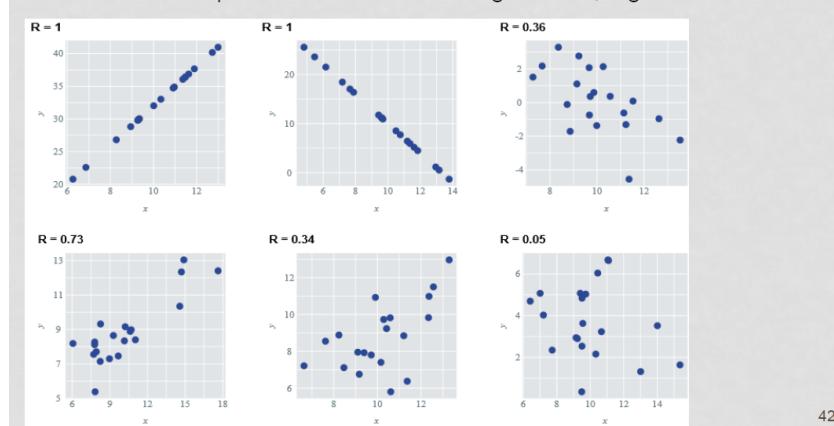
X = Verkaufsfläche, Y = Filialumsatz

r_{XY} = 0,707 → R² = 0,707² = 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.

Die folgende Grafiksammlung zeigt verschiedene Streudiagramme in Abhängigkeit des Wertes des R². Je eher die Datenpunkte auf einer Linie liegen, desto höher ist das R². Streuen die Datenpunkte ohne Zusammenhang im Raum, liegt das R² nahe 0.



"Wie hoch muss mein R² sein?"

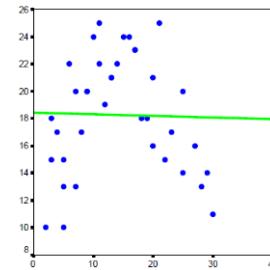
- Die übliche Größenordnung des R² variiert, je nach dem um welches Anwendungsgebiet es sich handelt. In Bereichen wie dem klassischen Marketing, in denen es hauptsächlich darum geht, menschliches Verhalten zu erklären bzw. vorherzusagen, sind meist geringe R² (deutlich kleiner 50%) zu erwarten. In anderen 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.

6.14. Probleme bei linearer Regression und Korrelation

Nur lineare Zusammenhänge werden erfasst!

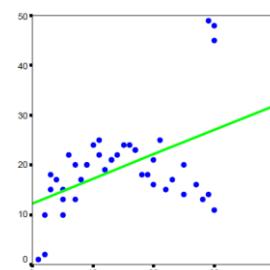
Die Gerade ist quasi horizontal –was nicht dem „eigentlichen“ Zusammenhang entspricht.

Hier geht es um einen nicht linearen Zusammenhang, der nicht durch die lineare Regression beschrieben werden kann.



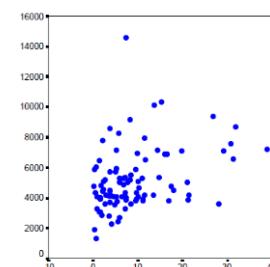
Einzelne Fälle können starken Einfluss ausüben (nicht zuletzt wegen dem Quadrieren)!

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



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



7. Skript – Wahrscheinlichkeitsrechnung

7.1. Begriffe

Die beschreibende Statistik kommt ohne den Begriff Wahrscheinlichkeit aus.

Beschreibende Statistik	Wahrscheinlichkeitstheorie
Relative Häufigkeit	Wahrscheinlichkeit
Häufigkeitsverteilung	Wahrscheinlichkeitsverteilung
Stichprobe	Zufallsvariablen
Mittelwert	Erwartungswert
Standardabweichung	Streuung
Varianz	Varianz
Median	Median
Quantile	Quantile

7.2. 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

7.3. Zufallsexperimente

Beispiele für Zufallsexperimente:

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

7.4. Ereignisraum

- Ereignisraum $\Omega \rightarrow$ auch Ergebnismenge oder Merkmalraum genannt
- $\Omega \neq \emptyset \rightarrow$ 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.

7.5. Ereignis

- Ein Ereignis (event) ist eine Teilmenge A des Ereignisraums Ω .
- 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 \rightarrow tritt also immer ein
- \emptyset heißt unmögliches Ereignis \rightarrow kann nie eintreten
- A^c heißt Komplementäreignis \rightarrow Gegenereignis, ohne A
- Teilmengen A und B heißen unvereinbar oder disjunkt, falls $A \setminus B = \emptyset$

7.6. Wahrscheinlichkeit

Um exakte Voraussagen über die Begrenzung unserer Möglichkeiten zu treffen, brauchen wir ein 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 $\text{Prob}(A)$) für sein Eintreten zu.

7.6.1. Wahrscheinlichkeit und relative Häufigkeit

Modul 7 Seite 10-13

7.6.2. Eigenschaften

Eigenschaften der Wahrscheinlichkeit:

- 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).
- 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 $h(A) = n/n = 1 \rightarrow p(A) = 1$
- 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 $h(A) = 0/n = 0 \rightarrow p(A) = 0$

7.6.3. axiomatische Begründung

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:

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

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

7.6.4. Regeln

$$P(\Omega) = 1$$

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

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:

$$\begin{aligned} 1) \quad P(A \cup B) &= P(A \setminus B) + P(A \cap B) + P(B \setminus A) \\ P(A) &= P(A \setminus B) + P(A \cap B) \\ P(B) &= P(A \cap B) + P(B \setminus A) \\ P(A \cup B) &= P(A) + P(B) - P(A \cap B) \end{aligned}$$

2) sind A und B voneinander unabhängige Ereignisse:

$$\begin{aligned} P(A \cup B) &= P(A \text{ oder } B) = P(A) + P(B) \\ P(A \cap B) &= P(A \text{ und } B) = P(A) * P(B) \end{aligned}$$

Einschluss-Ausschluss-Formel (Inklusion-Exklusion-Formel):

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

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) =

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

7.6.5. Berechnung

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

Gleichwahrscheinlichkeitsannahme das Verhältnis der günstigen Möglichkeiten zu allen Möglichkeiten:

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

7.7. Mengentheoretische Konzepte

Ereignisse sind Teilmengen des Ereignisraums.

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

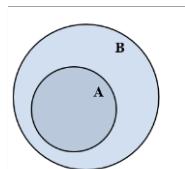
Mengenoperationen:

\cap Schnittmenge

\cup Vereinigung

\setminus Mengendifferenz

C Komplementbildung



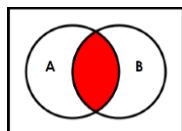
A ist eine (echte) Teilmenge von B

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.

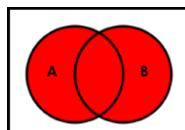
Formal: $A = B: \Leftrightarrow \forall x (x \in A \Leftrightarrow x \in B)$



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.

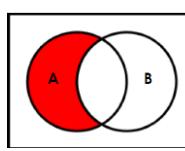
Formal: $A \cap B := \{x | (x \in A \wedge x \in B)\}$



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



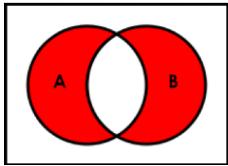
A ohne B

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

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

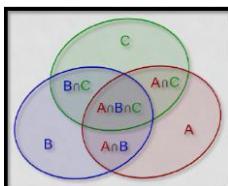
Formal: $B^C := \{x | x \notin B\}$



Symmetrische Differenz von A und B

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

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



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:

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

7.8. Gesetze

Für alle $A, B, C \subseteq X$ gilt:

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

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

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: (Regeln von de Morgan)	$(A \cup B)^c = A^c \cap B^c$
	$(A \cap B)^c = A^c \cup B^c$

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

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

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

Für die symmetrische Differenz gilt:

Assoziativgesetz:	$(A \Delta B) \Delta C = A \Delta (B \Delta C)$
-------------------	---

Kommutativgesetz:	$A \Delta B = B \Delta A$
-------------------	---------------------------

Distributivgesetz:	$(A \Delta B) \cap C = (A \cap C) \Delta (B \cap C)$
--------------------	--

$$A \Delta \emptyset = A$$

$$A \Delta A = \emptyset$$

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

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

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

Für die Komplementmenge gilt:

$A \cap A^c = \emptyset$	$A \cup A^c = \Omega$
$\emptyset^c = \Omega$	$\Omega^c = \emptyset$

Doppeltes Komplement:

$$(A^c)^c = A$$

Idempotenzgesetze:

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

Neutrale Elemente:

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

Dominanzgesetze:

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

7.9. Laplace

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

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

- 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:

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

→ 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:

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

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

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

7.10. Gegenwahrscheinlichkeit

Ist A ein Ereignis (eine Teilmenge des Ereignisraums Ω), so können wir seine Komplementärmenge $\Omega - A$ 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

$$p(A) + p(\neg A) = 1$$

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

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

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.

7.11. 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“)

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

7.11.1. Bausteine

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:

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

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

$$0! = 1$$

$$1! = 1$$
- 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}).

7.11.2. Grundmuster

Sei A eine n elementige Menge $A = \{1, \dots, n\}$, 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:

- mit Reihenfolge, mit Zurücklegen

$$|\Omega| = V_{mw} = n^k$$
- mit Reihenfolge, ohne Zurücklegen

$$k < n: |\Omega| = V_{ow} = \binom{n}{k} k! = \frac{n!}{(n - k)!}$$

$$k = n: |\Omega| = V_{ow} = n!$$
- ohne Reihenfolge, ohne Zurücklegen (Binomialkoeffizient)

$$|\Omega| = K_{ow} = \binom{n}{k} = \frac{n!}{k! (n - k)!}$$
- ohne Reihenfolge, mit Zurücklegen

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

7.11.3. 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)

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

7.11.4. 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

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

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)

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

7.11.5. Ohne Reihenfolge, ohne Zurücklegen

Beispiel 3.1:

Experiment: gleichzeitiges Ziehen von 2 Kugeln aus der Urne U₆ (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:

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

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)

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

43

7.11.6. 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?

$n = 10, k = 3$

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

7.12. Permutation mit Wiederholung

Werden bei einer Anordnung mit den k Elementen einer Menge das 1. Element n₁ mal das 2. Element n₂-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.

7.13. 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) 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: $P(B) > 0$

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

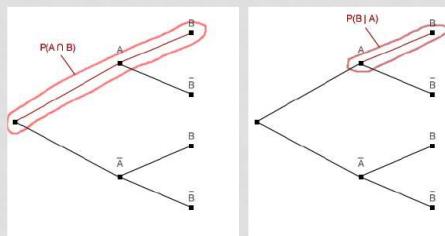
$$P(A \cap B) = P(B \cap A) = P(B)P(A|B) = P(A)P(B|A)$$

$P(A \cap B)$ 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.



7.14. 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:

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

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):

$$P(J) = \sum_{i=1}^n P(J|E_i) * P(E_i) = 0.8 * 0.7 + 0.1 * 0.3 = 0.59$$

7.15. 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:

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

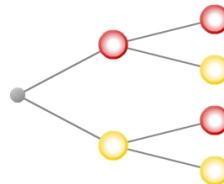
Für endlich viele Ereignisse B; seien B_i paarweise disjunkt und $A \subset \bigcup_{i=1}^n B_i$ und $P(A) \neq 0$, dann gilt:

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

7.16. Wahrscheinlichkeitsgraphen

Wahrscheinlichkeitsgraph (W Graph) \rightarrow grafische Darstellungsform der Ermittlung von Wahrscheinlichkeiten; dient zur Beschreibung des Ablaufs eines mehrstufigen Zufallsexperiments (eines Zufallsprozesses).

Ein W Graph ist ein bewerteter gerichteter Graph mit Baumstruktur \rightarrow 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).

7.16.1. Baumdiagramme

Pfadregeln für Baumdiagramme

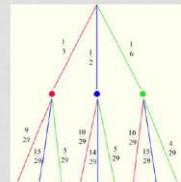
- Multiplikationssatz:
 - Die Wahrscheinlichkeit für das Eintreten eines Pfades ist das Produkt aller der längs diesen Pfades verzeichneten Wahrscheinlichkeiten
- Satz von der totalen Wahrscheinlichkeit:
 - Die Wahrscheinlichkeit eines Ereignisses A ist gleich der Summe der Wahrscheinlichkeiten aller Pfade, die zu einem Zustand führen, bei dem das Ereignis A eintritt.

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 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 den Ende jeder Linie, die einem Ausgang der ersten Ziehung entspricht, eine weitere Verzweigung anzuhängen, die die zweite Ziehung (unter den entsprechenden neuen Umständen) darstellt.



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.

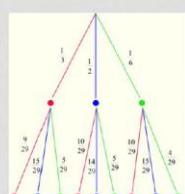
Multiplikationsregel für Baumdiagramme:
 Die **Wahrscheinlichkeit** für das Eintreten eines Pfaes ist das **Produkt der entlang ihm verzeichneten Wahrscheinlichkeiten**:

$$p(\text{erst rot, dann blau}) = (1/3) \times (10/29) = 5/29$$

$$p(\text{erst blau, dann rot}) = (1/2) \times (15/29) = 5/29$$

Additionsregel für Baumdiagramme:
 Da Pfade disjunkte Ereignisse darstellen, werden **Wahrscheinlichkeiten addiert**:

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



8. Formelsammlung

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 : $b_i = x_{k-1} - x_k$

Klassenmitte m_i : $m_i = 1/2(x_{k-1} + x_k)$

Beispielberechnung

i	1	2	3	4	5
X_i	-4	-1	0	1	4

Varianzberechnung: $((x_{i1}-\text{Mittelwert})^2 + (x_{i2}-\text{Mittelwert})^2 + \dots) / \text{Anzahl}$

$$((-4-0)^2 + (-1-0)^2 + (0-0)^2 + (1-0)^2 + (4-0)^2) / \text{Anzahl} = 16+1+0+1+16 = 34/5 = 34/5 = 6,8$$

Arithmetisches Mittel (Mittelwert): $(x_{i1} + x_{i2} + \dots) / \text{Anzahl}$

$$((-4-1+0+1+4) / \text{Anzahl}) = 0/5 = 0$$

Modus

Häufigste Wert $\rightarrow 1x50$ $1x40$ $2x20$ $1x10$

$\rightarrow 20$

Median

Mittlere Wert $\rightarrow 50,40,20,20,10$ (Mitte)

$\rightarrow 20$

Spannweite

Maximum-Minimum $\rightarrow 50-10=40$

Standardabweichung

Wurzel aus Varianz

Variationskoeffizient

Varianz/arithmetische Mittel

Kombinatorik:

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

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

- mit Reihenfolge, mit Zurücklegen
 $|\Omega| = V_{mW} = n^k$
- mit Reihenfolge, ohne Zurücklegen
 $k < n: |\Omega| = V_{ow} = \binom{n}{k} k! = \frac{n!}{(n-k)!}$
 $k = n: |\Omega| = V_{ow} = n!$
- ohne Reihenfolge, ohne Zurücklegen
 $|\Omega| = K_{ow} = \binom{n}{k} = \frac{n!}{k!(n-k)!}$
- ohne Reihenfolge, mit Zurücklegen
 $|\Omega| = K_{mW} = \binom{n+k-1}{k} = \frac{(n+k-1)!}{k!(n-1)!}$

Permutation mit Wiederholungen:

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

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

Aufgabe 1.10

$$10!/(1!*2!*3!*4!) = 12600$$

Alle Elemente der Grundmenge für die Aufgabe relevant?

JA \Rightarrow Permutation

Elemente unterscheidbar? Ohne Wiederholung? Ohne Zurücklegen?

JA \Rightarrow Permutation ohne Wiederholung

NEIN \Rightarrow Permutation mit Wiederholung

NEIN \Rightarrow Variation oder Kombination

Reihenfolge ist zu berücksichtigen?

JA \Rightarrow Variation

Elemente unterscheidbar? Ohne Wiederholung? Ohne Zurücklegen?

JA \Rightarrow Variation ohne Wiederholung

NEIN \Rightarrow Variation mit Wiederholung

NEIN \Rightarrow Kombination

Elemente unterscheidbar? Ohne Wiederholung? Ohne Zurücklegen?

JA \Rightarrow Kombination ohne Wiederholung

NEIN \Rightarrow Kombination mit Wiederholung

Kombination ohne Wiederholung	Kombination mit Wiederholung	Variation ohne Wiederholung	Variation mit Wiederholung	Permutation ohne Wiederholung	Permutation mit Wiederholung
$\binom{n}{k}$	$\binom{n+k-1}{k}$	$\frac{n!}{(n-k)!}$	n^k	$n!$	$\frac{n!}{k!}$

Regressionsfunktion:

i	X_i	Y_i	$X_i * Y_i$	X_i^2	Y_i^2
1	1	0	0	1	0
2	2	2	4	4	4
3	3	3	9	9	9
4	4	6	24	16	36
5	5	9	45	25	81
Summe	15	20	82	55	130

$$a = (X_i^2 * Y_i - X_i * (X_i * Y_i)) / (i * X_i^2 - (X_i)^2)$$

$$a = (55 * 20 - 15 * 82) / (5 * 55 - 15^2)$$

$$a = -130 / 50 = -2,60$$

$$b = (i * (X_i * Y_i) - X_i * (Y_i)) / (i * X_i^2 - (X_i)^2)$$

$$b = (5 * 82 - 15 * 20) / (5 * 55 - 15^2)$$

$$b = 110 / 50 = 2,20$$

Regressionsfunktion $\hat{y} = a + b * x$

$$\hat{y} = -2,60 + 2,20x$$

Korrelationskoeffizient:

$$r_{xy} = \text{Zähler} / \text{Nenner}$$

$$\text{Zähler} = (1/i) * (X_i * Y_i) - (X_i/i) * (Y_i/i)$$

$$\text{Zähler} = (1/5) * 82 - (15/5) * (20/5)$$

$$\text{Zähler} = 0,2 * 82 - 3 * 4 = 4,4$$

$$\text{Nenner} = \text{Wurzel aus}((1/i) * X_i^2 - (X_i/i)^2) * \text{Wurzel aus}((1/i) * Y_i^2 - (Y_i/i)^2)$$

$$\text{Nenner} = \text{Wurzel aus}((1/5) * 55 - (15/5)^2) * \text{Wurzel aus}((1/5) * 130 - (20/5)^2)$$

$$\text{Nenner} = \text{Wurzel aus}(0,2 * 55 - 9) * \text{Wurzel aus}(0,2 * 130 - 16)$$

$$\text{Nenner} = \text{Wurzel aus}(2) * \text{Wurzel aus}(10) = 4,47$$

$$r_{xy} = 4,4 / 4,47 = 0,98$$

Bestimmtheitsmaß:

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

$$R^2 = (0,98)^2 = 0,96$$

9. Klausuraufgaben WS19/20

Aufgabe 1.1

b

Aufgabe 1.2

c

Aufgabe 1.3

d

Aufgabe 1.4

a

Aufgabe 1.5

c

Aufgabe 1.6 Histogrammaufgabe → entscheidend ist die Anzahl der Kästchen

b → Bereich 1 < Bereich 2 = Bereich 3

Aufgabe 1.7

c

Aufgabe 1.8 Gegeben sind drei Ereignisse und es gilt $X \cap Y = Z \neq \emptyset$. Welche der folgenden Aussagen ist falsch?

a → $P(X \cup Y) = P(X) + P(Y)$ → Schnittmenge von X und Y ist beim addieren doppelt

Aufgabe 1.9 Auf wie viele Arten können 7 Fahrräder an 7 Personen verliehen werden? (1x/Person)

a → mit Reihenfolge, ohne Zurücklegen: $7! = 5.040$

Aufgabe 1.10 Kerstin fädelt 1 schwarze, 2 rote, 3 blaue und 4 weiße Perlen auf eine Schnur.

Auf wie viele versch. Arten kann Kerstin ihre Kette gestalten?

d → Permutation mit Wiederholungen: $10!/(1!*2!*3!*4!) = 12.600$

Aufgabe 2 Eine Umfrage unter 200 Studierenden auf dem Campus der Ruhr-Universität ergab:
130 Studierende besitzen ein Auto, 160 einen Führerschein und 128 sowohl Auto als auch Führerschein.

Aufgabe 2.a Stellen Sie anhand dieser Angaben eine zweidimensionale Kontingenztabelle auf.

Auto/Führerschein	Hat Führerschein	Kein Führerschein	Summe
Hat Auto	128 abs. H. 64% rel. H. 80% rel. H. Sp. 98,46% rel. H. Ze.	2 1% 5% 1,54%	130 65%
Kein Auto	32 16% 20% 45,71%	38 19% 95% 54,29%	70 35%
Summe	160 80%	40 20%	200 100%

Aufgabe 2.b Wie viel Prozent der Studierenden besitzen kein Auto?

35%

Aufgabe 2.c Wie viel Prozent der Führerscheinhaber besitzen ein Auto?

80%

Aufgabe 2.d Wie viel Prozent der Autobesitzer haben keinen Führerschein?

1,54%

Aufgabe 2.e Wie viel Prozent der Studierenden besitzen weder Auto noch Führerschein?

19%

- Aufgabe 3** Im Rahmen einer Passantenbefragung wurden 140 Personen befragt, wie oft sie im letzten Halbjahr ein bestimmtes Kino besucht haben.
 50 geben an nur einmal in dem Kino gewesen zu sein, 40 Personen waren zweimal in dem Kino, 20 Personen dreimal, 10 regelmäßige Kinobesucher sogar neunmal und noch 20 Personen hatten das Kino im letzten Halbjahr nicht besucht.

Aufgabe 3.a

i	x _i	h _i	f _i (%)	H _i	F _i (%)
1	0x	20	14,29%	20	14,29%
2	1x	50	35,71%	70	50,00%
3	2x	40	28,57%	110	78,57%
4	3x	20	14,29%	130	92,86%
5	9x	10	7,14%	140	100,00%
Summe		140	100,00%		

Aufgabe 3.b

- Modus $1 \times 50 \quad 1 \times 40 \quad 2 \times 20 \quad 1 \times 10 \rightarrow 20$
 Median $50, 40, 20, 20, 10$ (Mitte) $\rightarrow 20$
 Arithmetische Mittel $140/5 = 28$
 Die Form der obigen Verteilung Linksschiefe Verteilung
 Spannweite Maximum-Minimum $\rightarrow 50-10=40$
 Varianz $((50-28)^2 + (40-28)^2 + (20-28)^2 + (20-28)^2 + (10-28)^2)/5 =$
 $((22)^2 + (12)^2 + (-8)^2 + (-8)^2 + (-18)^2)/5 =$
 $(484+144+64+64+324)/5 = 1080/5 = 216$
 Standardabweichung Wurzel aus 216 = 14,6969
 Variationskoeffizient $14,6969/28 = 0,5249$

Aufgabe 5

Aufgabe 5.a

$$6/10 * 4/10 = 24/100 = 24\%$$

Aufgabe 5.b

$$6/10 * 6/10 = 36/100 = 60\%$$

Aufgabe 5.c

$$4/10 * 3/9 = 12/90 = 13,33\%$$

Aufgabe 6

Aufgabe 6.a

A: Mail ist Spam

B: Mail enthält das Wort „Viagra“ und ist Spam

C: Mail enthält das Wort „Viagra“ und ist kein Spam

Aufgabe 6.b

$$(P(A)*P(B))/(P(A)*P(B)+P(C)*P(A)) = 99,98\%$$

Aufgabe 4

Aufgabe 4.a

i	X _i	Y _i	X _i * Y _i	X _i ²	Y _i ²
1	1	0	0	1	0
2	2	2	4	4	4
3	3	3	9	9	9
4	4	6	24	16	36
5	5	9	45	25	81
Summe	15	20	82	55	130

$$a = (X_i^2 * Y_i - X_i * (X_i * Y_i)) / (i * X_i^2 - (X_i)^2) = (55*20 - 15*82) / (5*55 - 15^2) = -130 / 50 = -2,60$$

$$b = (i * (X_i * Y_i) - X_i * Y_i) / (i * X_i^2 - (X_i)^2) = (5*82 - 15*20) / (5*55 - 15^2) = 110 / 50 = 2,20$$

Regressionsfunktion $\hat{y} = a + b*x$

$$\hat{y} = -2,60 + 2,20x$$

Aufgabe 4.b

X _i	\hat{y}
1	-0,40
2	1,80
3	4,00
4	6,20
5	8,40

Aufgabe 4.c

Aufgabe 4.d

$$\hat{y} = -2,60 + 2,20 * 10$$

$$\hat{y} = 19,4$$

Aufgabe 4.e

$$r_{xy} = \text{Zähler} / \text{Nenner}$$

$$\text{Zähler} = (1/i) * (X_i * Y_i) - \text{Mittelwert}(X_i) * \text{Mittelwert}(Y_i)$$

$$\text{Zähler} = (1/5) * 82 - (15/5) * (20/5) = 0,2 * 82 - 3 * 4 = 4,4$$

$$\text{Nenner} = \text{Wurzel aus}((1/i) * X_i^2 - \text{Mittelwert}(X_i)^2) * \text{Wurzel aus}((1/i) * Y_i^2 - \text{Mittelwert}(Y_i)^2)$$

$$\text{Nenner} = \text{Wurzel aus}((1/5) * 55 - (15/5)^2) * \text{Wurzel aus}((1/5) * 130 - (20/5)^2)$$

$$\text{Nenner} = \text{Wurzel aus}(0,2 * 55 - 9) * \text{Wurzel aus}(0,2 * 130 - 16)$$

$$\text{Nenner} = \text{Wurzel aus}(2) * \text{Wurzel aus}(10) = 4,47$$

$$r_{xy} = 4,4 / 4,47 = 0,98$$

Aufgabe 4.f

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

$$R^2 = (0,98)^2 = 0,96$$