THM TECHNISCHE HOCHSCHULE MITTELHESSEN

BWL1: Grundlagen und Unternehmens-software

Beschreibende Statistik

Lernziele

- Sie können die Grundbegriffe der Statistik anwenden und von einander abgrenzen
- Sie können Lage- und Streumaße bestimmen und interpretieren
- Sie können Häufigkeitsverteilungen erstellen und analysieren
- Sie können Korrelationen bestimmen
- Sie sind in der Lage Regressionsanalysen durchzuführen

Statistik ist das Handwerkszeug in vielen Bereichen der BWL wie Marktforschung, Logistik, Controlling …

Abbildungsquelle: https://www.handicapverbessern.com/wp-content/uploads/2017/08/Ziel-im-golf-visualisieren.jpg

Beschreibende Statistik

Grundlagen	Häufigkeits-verteilungen	Lage- & Streumaße	Zweidim-ensionale Häufigkeits-verteilungen
Korrelation & Regression			

Ursprünge der Statistik

- **Babylon (ca. 3.800 v. Chr.)**
  - Schätzung zur Erhebung von Steuern
  - Schätzung, wie viele Soldaten ein Land bereit stellen kann.
- **Griechische Antike**
  - Aufzeichnungen über Getreideeinfuhr und zollpflichtige Ware
- **Römisches Reich**
  - regelmäßige Volkszählungen

Moderne Statistik

- **John Graunt (17. JH)**
  - Analyse von Sterbe- und Geburtenraten
- **Cardano, Pascal, Fermat (17. JH)**
  - Untersuchungen zu mathematischen Grundlagen im Glücksspiel
  - Wahrscheinlichkeitsrechnung
- **Kolmogorow (19. JH)**
  - Begründer der modernen Wahrscheinlichkeitsrechnung

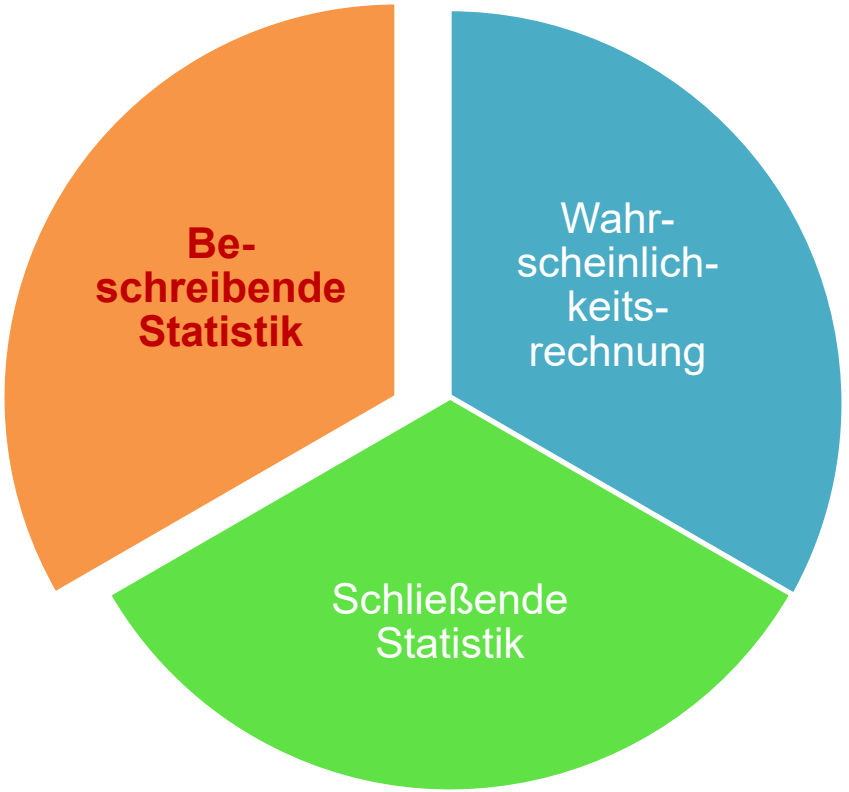
Definition & Gebiete der Statistik

1) Zusammenstellung von Daten

→ informatorische Fundierung von Entscheidungen

2) Methodenlehre

→ Herausfiltern von entscheidungsrelevanten Informationen aus den verfügbaren Daten



Missbrauch der Statistik

Traue keiner Statistik, die Du nicht selbst gefälscht hast.

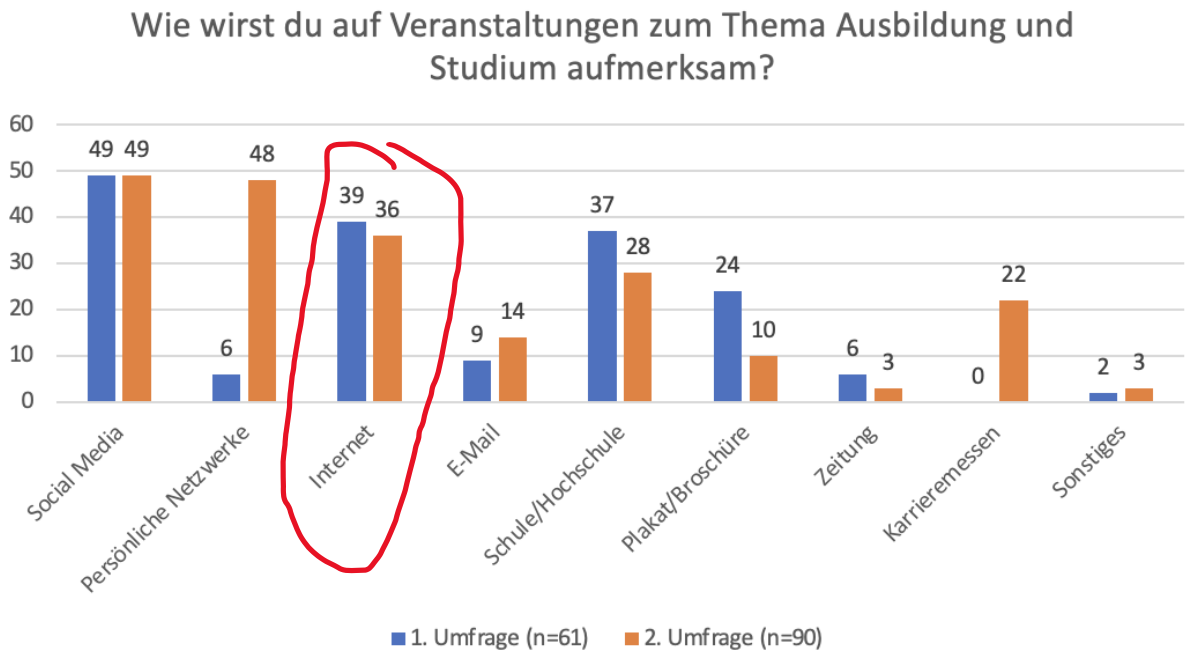
(vermeintlich Winston Churchill, wahrscheinlich Joseph Goebbels)

Quelle: Schwarze, Grundlagen d. Statistik

Beispiel für Fehlinterpretation

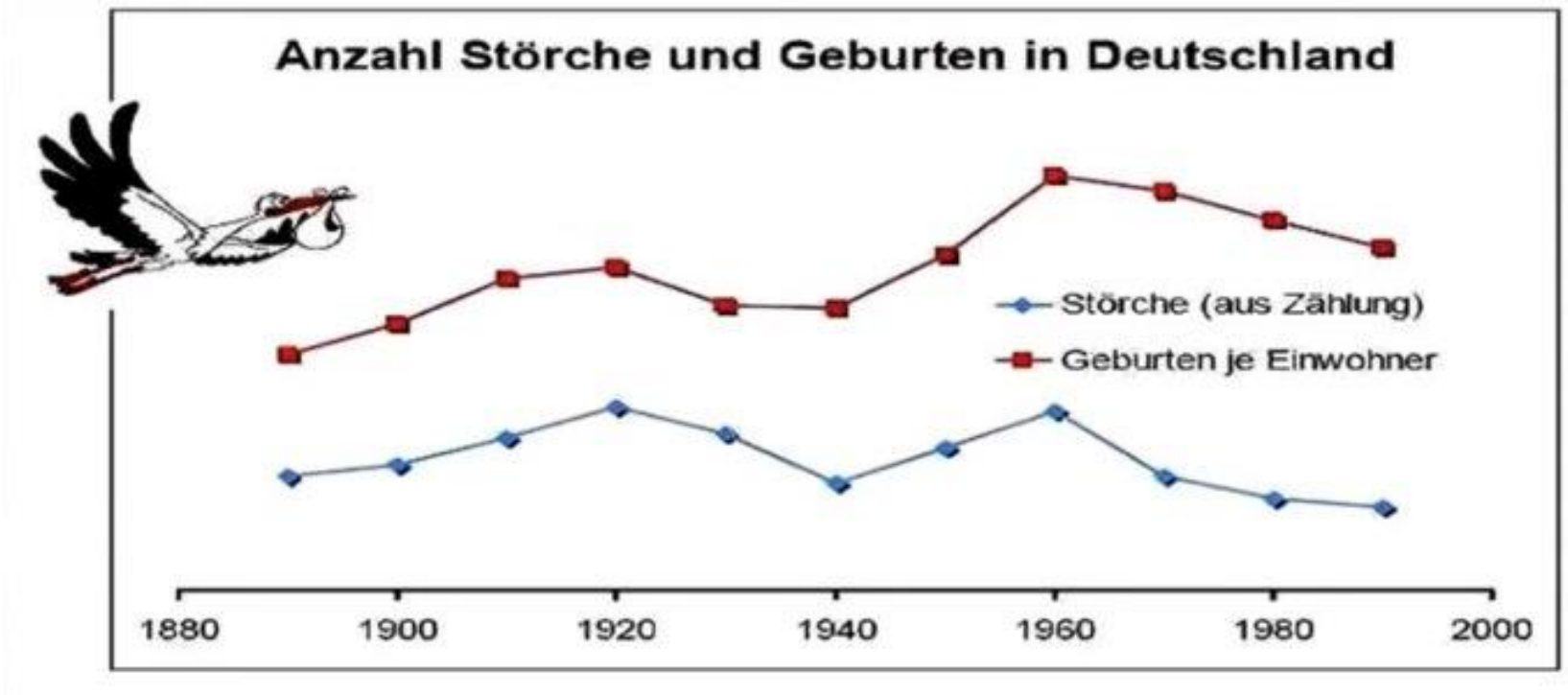
Umfrage unter Schülern und Schülerinnen

„Das Internet bleibt als Informationsquelle Nr. 3 für Ausbildung und Studium stabil.“



Achtung: die Anzahl hat zwar nur um 3 (von 39 auf 36) abgenommen. Allerdings haben an der 2. Umfrage ca 30% mehr teilgenommen. Die Internetnutzung ist also im Vergleich zur ersten Umfrage erheblich gesunken.

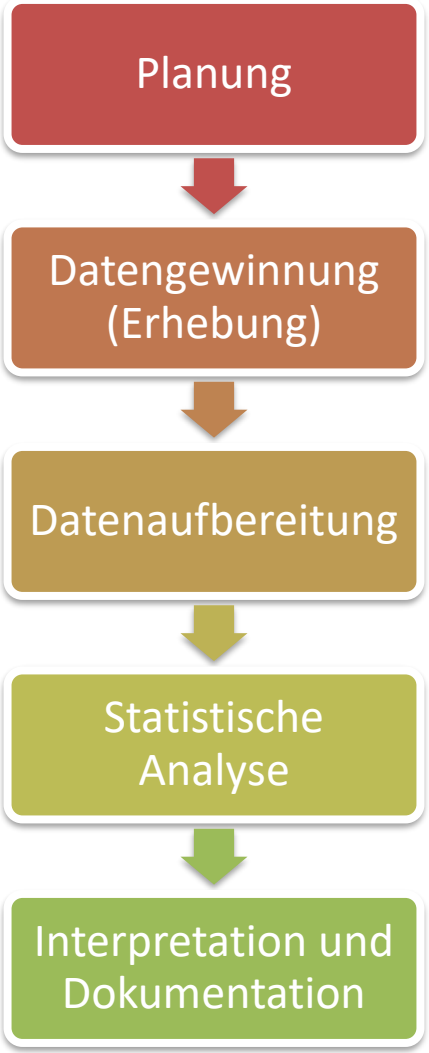
Beispiel für Scheinkorrelation



Fehlinterpretation:
Zusammenhang zwischen der Anzahl an Horstpaaren und den Geburten in Deutschland

Quelle: Statistisches Bundesamt

Ablauf statistischer Untersuchungen



1. Planung

- Formulierung der Ziele der Untersuchung
- Detaillierte Definition des Untersuchungsgegenstands
- Abgrenzung des Untersuchungsgegenstands
  - Sachlich
  - Räumlich
  - Zeitlich
- Berücksichtigung von Vorgaben und eventuellen Randbedingungen

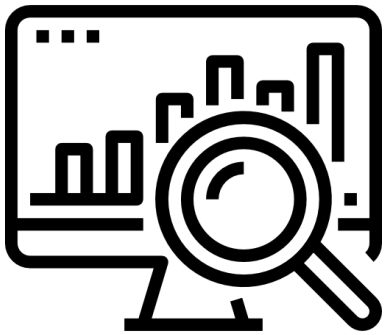
2. Datengewinnung (Erhebung)

- Primärerhebung (schriftlich oder mündlich)
  - Unmittelbar durch
  - Experiment
  - Beobachtung
  - Befragung

- Sekundärerhebung
  - greift auf bestehende Daten zurück

Problematik: z.B. Fragestellung, ohne Antworten zu beeinflussen (Suggestivfragen)

Datengewinnung (Erhebung)



3. Datenaufbereitung

- Erfassung der Daten
- Kontrolle und eventuelle Fehlerkorrektur
- Ordnen der Daten für weitere Verarbeitung
- Eventuell erste Darstellung in Tabellen oder Grafiken

Datenaufbereitung

4. Statistische Analyse

- Auswertung und Analyse der Daten
- Anwendung mathematischer und statistischer Verfahren und Methoden
- z.B. Ermittlung von Mittelwerten und Häufigkeiten
- Untersuchung von Zusammenhängen

Statistische Analyse

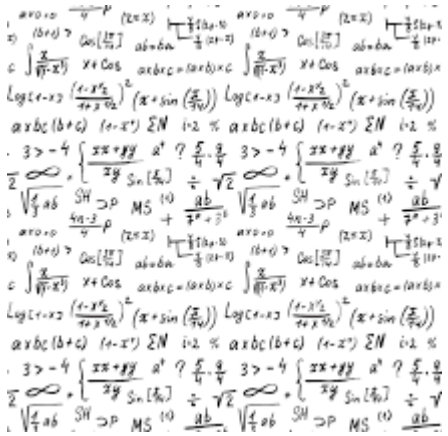
5. Interpretation & Dokumentation

- Präsentation der Ergebnisse
- Darstellung grafisch, tabellarisch oder in Einzelwerten
- Kommentierung der Daten
- Nutzbarmachung z.B. für Entscheidungsprozesse

Interpretation und Dokumentation

Grundbegriffe der Statistik

- **Merkmalsträger t (t=1,…,n)**
  = Einzelobjekte einer statistischen Untersuchung
- **Statistische Masse**
  = Gesamtheit der Merkmalsträger
- **Statistisches Merkmal**
  = bei der statistischen Masse untersuchte Eigenschaft
- **Beobachtungswerte (Merkmalswerte) x_t (t=1,…, n)**
  =bei den Merkmalsträgern festgestellte Ergebnisse
- **Merkmalsausprägungen x_i (i=1,...,m)**
  = theoretisch mögliche Ergebnisse



Grundbegriff: Merkmalsträger (statistische Einheit)

Die **statistische Einheit** (Merkmalsträger) ist das **Einzelobjekt** einer statistischen Untersuchung. Sie ist ***Träger*** der Information(en), für die man sich bei der Untersuchung interessiert.

Identifikationskriterien für statistische Einheiten:

- Sachlich
- Räumlich
- Zeitlich

Beispiele:

- Untersuchung der Abfüllmengen eines Automaten zum Füllen von Bierflaschen

→ statistische Einheit: die abgefüllten Bierflaschen

- Untersuchung der Lebensdauer von Autoreifen
  → statistische Einheit: die einzelnen Autoreifen

Beispiel Merkmalsträger (statistische Einheit)

Prognose des Wahlergebnisses einer Kommunalwahl in einer Stadt

→ statistische Einheit: der Bürger

Identifikationskriterien:

- Sachlich: wahlberechtigter Bürger
- Räumlich: Gebiet der Großstadt
- Zeitlich: Tag der Umfrage

Grundbegriff: Statistische Masse

Auch Grundgesamtheit, Population, Kollektiv, …

Eine **statistische Masse** ist eine **Gesamtheit** (Menge) von *statistischen Einheiten* mit übereinstimmenden *Identifikationskriterien*.

Die sachlichen, räumlichen und zeitlichen Identifikationskriterien ergeben sich aus der Zielsetzung bzw. Aufgabenstellung der statistischen Untersuchung.

Beispiel:

Alle wahlberechtigten Bürger einer Kommune

Rechenbeispiel: Statistische Einheit (Merkmalsträger) und statistische Masse (Grundgesamtheit)

Ermittlung der durchschnittlichen Studiendauer von Studierenden an öffentlichen deutschen Hochschulen bis zum Abschluss der ersten berufsqualifizierenden Prüfung (Diplom, Bachelor, Magister, Staatsexamen) im Jahr 2025.

Statistische Einheit (Merkmalsträger): _____

Statistische Masse (Grundgesamtheit): _____

Identifikationskriterien

- Sachlich: _____
- Räumlich: _____
- Zeitlich: _____

Lösung: Statistische Einheit (Merkmalsträger) und statistische Masse (Grundgesamtheit)

Ermittlung der durchschnittlichen Studiendauer von Studierenden an öffentlichen deutschen Hochschulen bis zum Abschluss der ersten berufsqualifizierenden Prüfung (Diplom, Bachelor, Magister, Staatsexamen) im Jahr 2025.

Statistische Einheit (Merkmalsträger):	Studierende
Statistische Masse (Grundgesamtheit):	Alle Studierenden

Identifikationskriterien

• Sachlich:	Arten der Abschlüsse
• Räumlich:	öffentliche deutsche Hochschulen
• Zeitlich:	Jahr 2025

Grundbegriff: Stichprobe

Eine **Stichprobe** ist eine nach bestimmten Methoden ausgewählte Teilmenge der Grundgesamtheit (Statistische Masse), die möglichst repräsentativ sein sollte. Wird bei einer statistischen Untersuchung nur ein Teil der interessierenden statistischen Masse erfasst, dann heißt dieser Teil **Stichprobe**.

Bemerkungen:

- Häufig ist die Untersuchung der Grundgesamtheit z.B. aus zeitlichen oder Kostengründen nicht möglich. Dann erfolgt die Untersuchung auf der Stichprobe.
- Mit welchen Fehlern können die Ergebnisse und Aussagen von der Stichprobe auf die Grundgesamtheit übertragen werden?
- Eine einfache Übertragung oder Verallgemeinerung von der Stichprobe auf die Grundgesamtheit ist unzulässig.
- Weiterführend in BWL1 nicht behandelt

Beispiel:

- Befragung einer „***repräsentativen***“ Anzahl von Wahlberechtigten für eine Wahlprognose.

Grundbegriff: Statistisches Merkmal (Beobachtungsmerkmal)

Eine bei einer statistischen Untersuchung interessierende **Eigenschaft** einer *statistischen Einheit* heißt **Merkmal**. Die statistischen Einheiten heißen auch **Merkmalsträger**.

Beispiele:

a) Statistische Einheit: Student

→ Merkmale: *Alter, Schulabschluss, Studienfach, Statistiknote, …*

b) Statistische Einheit: Landwirtschaftlicher Betrieb

→ Merkmale: *Ackernutzfläche, Rinderbestand, Umsatz, Anzahl Mitarbeiter, …*

Grundbegriff: Merkmalswert (Beobachtungswert)

Eine bei einer statistischen Untersuchung an einer bestimmten *statistischen Einheit* festgestellte *Merkmalsausprägung* heißt **Merkmalswert** oder **Beobachtungswert**.

Bemerkung:

Bei einer statistischen Analyse sind die (konkreten tatsächlichen) Merkmalswerte die Daten, die ausgewertet werden.

Beispiel:

Merkmalswerte für Studierende eines bestimmten Kurses

<u>Merkmale</u>

Schwerpunkt:	Management, Medien, IT
Alter:	20, 21, 24, 26
Semester:	3.

Grundbegriff: Merkmalsausprägungen

Die *möglichen* Ausprägungen (Werte oder Kategorien), die ein Merkmal annehmen kann, heißen **Merkmalsausprägungen**.

Bemerkung:
Die Merkmalswerte sind eine Teilmenge der Merkmalsausprägungen.

Beispiele:

Merkmal Geschlecht:	Ausprägungen: männlich, weiblich.
Merkmal Studienfach:	Ausprägungen: BWL, Jura, Wirtschaftsinformatik, …
Merkmal Alter:	Ausprägungen: 1, 2, 3, 4, 5, … , 20, 21, 22, …
Merkmal Semester:	Ausprägungen: 1., 2., 3., 4., 5., 6., 7., ….
Merkmal Note:	Ausprägungen: 1, 2, 3, 4, 5, (6)
Merkmal Bundesland:	Ausprägungen: NRW, Bayern, RLP, Hessen, …

Grundbegriff: Häufbarkeit

Merkmale sind häufbar

Bei manchen Merkmalen kann es vorkommen, dass einzelne Merkmalsträger **mehrere Merkmalsausprägungen pro Merkmal** aufweisen. Solche Merkmale nennen man **häufbar**.

Beispiele:

Doppelbeschäftigung
Mehrere Wohnorte
Mehrere Staatsbürgerschaften
Mehrere Fahrzeuge

Eigenschaften von häufbaren Merkmalen

- Können in einer Häufigkeitsverteilung dargestellt werden
- Können in Kategorien oder Klassen eingeteilt werden können, sodass man die Häufigkeit jeder Kategorie oder Klasse zählen kann.
- Kategoriale als auch diskrete Merkmale sind typischerweise häufbar

Sie sind gefragt!

THM
TECHNISCHE HOCHSCHULE MITTELHESSEN

Frage

Welche der folgenden Merkmale eines Unternehmens sind häufbar?

- Standort
- Gründungsjahr
- Rechtsform
- Branche
- Zahl der Beschäftigten
- Umsatz

→ **Particify https://ars.particify.de Raum-Nr.: 7207 7982**

Sie sind gefragt!

THM
TECHNISCHE HOCHSCHULE MITTELHESSEN

Frage

Welche der folgenden Merkmale eines Unternehmens sind häufbar?

- Standort
- Gründungsjahr
- Rechtsform
- Branche
- Zahl der Beschäftigten
- Umsatz

→ **Particify https://ars.particify.de Raum-Nr.: 7207 7982**

Grundbegriff: Klassierung von Merkmalsausprägungen

Eine **Klassierung** (Zusammenfassung, Gruppierung) von Merkmalsausprägungen wird in der Statistik häufig vorgenommen, wenn die **Anzahl** der verschiedenen **Merkmalsausprägungen** zu einer **unübersichtlichen** Auswertung führt.

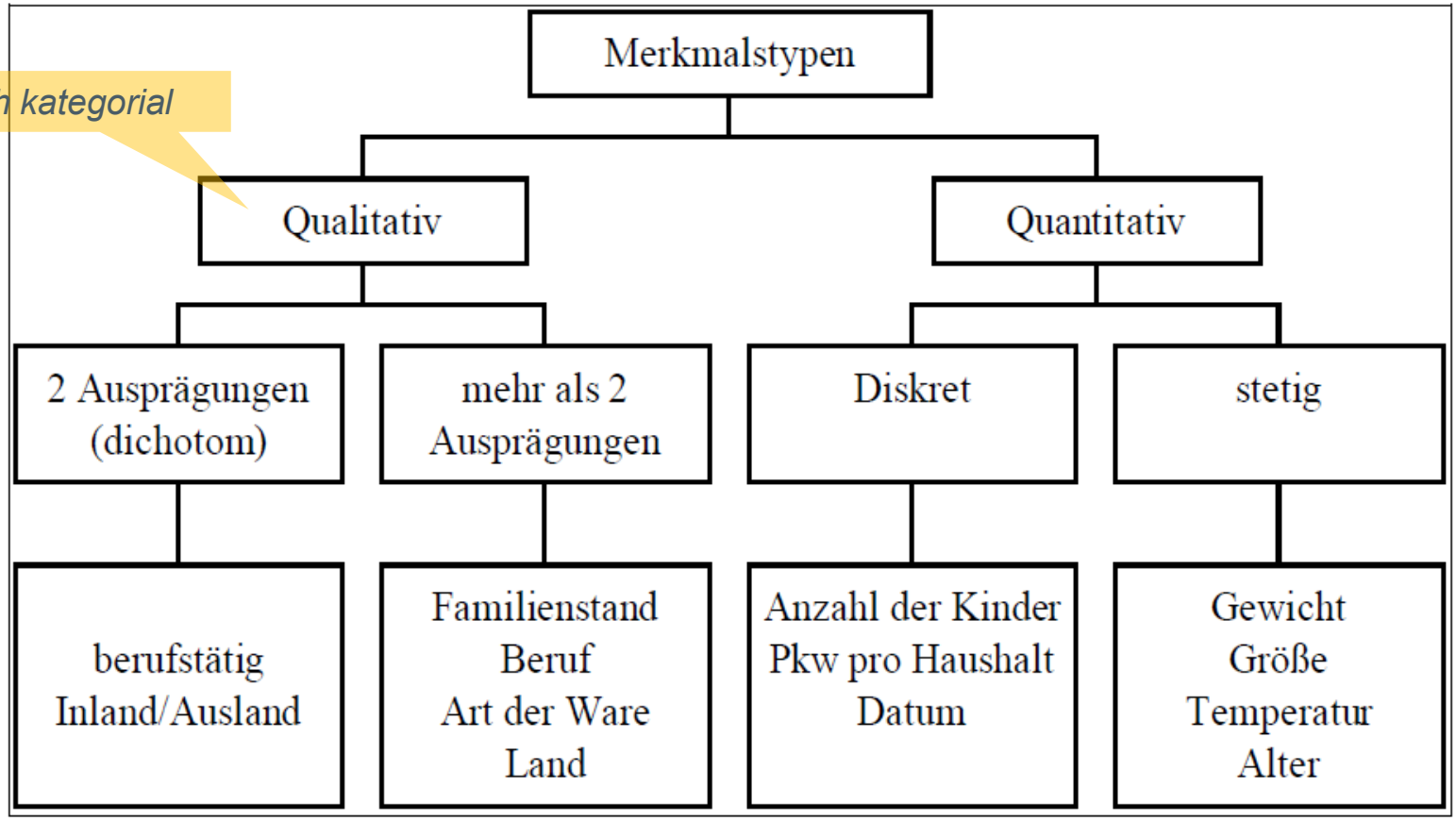
Beispiele:

- Umsatz
- Alter
- Leistung (Kfz)

Beschreibende Statistik

Grundlagen	Häufigkeits-verteilungen	Lage- & Streumaße	Zweidim-ensionale Häufigkeits-verteilungen
Korrelation & Regression			

Merkmalstypen der Statistik



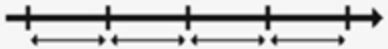
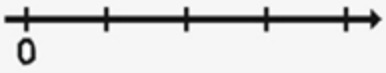
Quelle: Prof. Dr. M. Schmidt, THM FB Wirtschaft: Skript Statistik für BWL

Skalierungen I

Skala	Beschreibung	Beispiel
Nominalskala *qualitativ*	einfache Aufzählung,**keine Rangfolge**	Geburtsdatum, Farben, Berufe
Ordinalskala *qualitativ*	es besteht eine natürliche Rangfolge, aber die **Abstände** sind **nicht aussagefähig**	Schulnoten
Intervallskala *quantitativ*	Differenz zwischen zwei Angaben ist aussagefähig, aber **kein eindeutiger Null-punkt**	Temperatur
Verhältnisskala *quantitativ*	Verhältnis zwischen zwei Werten ist aussagefähig,ebenso der Nullpunkt	Umsatz, Gewinn

Quelle: Prof. Dr. M. Schmidt, THM FB Wirtschaft: Skript Statistik für BWL

Skalierungen II

Merkmalsart	Relation zwischen den Merkmals-ausprägungen	Skalierung	Beispiele
qualitatives Merkmal	Verschiedenheit $x_i \neq x_j$	Nominalskala 	Familienstand, Geschlecht, Beruf, Postleitzahl
komparatives Merkmal	Rangfolge $x_i < x_j$	Ordinalskala 	Handelsklasse, Schulnoten, Rating-Urteile
quantitatives Merkmal	Abstände $(x_i - x_j)$ sinnvoll	Intervallskala 	Temperatur in [°C], Geburtsjahrgang
	Verhältnisse $(x_i : x_j)$ sinnvoll	Verhältnisskala 	Preis, Umsatz, Einkommen, Alter

Quelle: Wewel, Statistik für BWL

Beschreibende Statistik

Grundlagen	**Häufigkeits-verteilungen**	Lage- & Streumaße	Zweidim-ensionale Häufigkeits-verteilungen
Korrelation & Regression			

Häufigkeitsverteilung

Die grundlegende Aufbereitungsform statistischer Daten ist die Darstellung in Form einer Häufigkeitsverteilung. Beziehen sich die Daten auf nur **ein** Merkmal der statistischen Erhebung, so spricht man von einer **eindimensionalen Häufigkeitsverteilung**. Die Darstellungsform erfolgt typischerweise in **tabellarischer** oder **grafischer** Form.

Liegen mehrere Merkmale in der ***Urliste*** vor, so ergibt sich eine **Datenmatrix**, wobei sich jede Zeile auf einen Merkmalsträger und jede Spalte auf ein Merkmal bezieht. Der Inhalt einer Matrix-Zelle ist der jeweilige Beobachtungswert.

Erste typische Fragenstellungen sind für die Merkmale

- Anzahl der Merkmalsausprägungen
- Durchschnitt (Mittelwert)
- Summen
- Abweichungen

Beispiel einer Häufigkeitsverteilung

Beispiel: Personalerhebung

Als Beispiel dient die Personaldatei eines kleinen Unternehmens. Für die Personalerhebung werden für die insgesamt **n = 25 Beschäftigten (= Merkmalsträger)** u.a. die folgenden Merkmale erfasst.

Merkmale für Personalerhebung:

- **Personalnummer**
- derzeitige **Abteilung**szugehörigkeit
- **Ausbildung** (höchster Schul- bzw. Hochschulabschluss)
- **Jahr** des Eintritts in die Firma
- derzeitiges monatliches (Brutto-)**Gehalt**

Quelle: Wewel, Statistik für BWL

Urliste Personalerhebung

Personal-nummer	Abteilung	Ausbildung	Eintrittsjahr	Monatliches Bruttogehalt [€]
560426	Finanzen	Abitur	2006	13200
590303	Vertrieb	Mittlere Reife	2008	13500
611117	Entwicklung	Promotion	2008	17600
620212	Geschäftsführung	Master	2006	18000
620624	Entwicklung	Master	2007	17400
640530	Geschäftsführung	Mittlere Reife	2009	7200
681212	Entwicklung	Bachelor	2009	10800
700525	Test/Anwendungen	Bachelor	2010	7200
711204	Geschäftsführung	Bachelor	2007	13600
730119	Finanzen	Bachelor	2009	9600
730523	Entwicklung	Master	2010	7400
730526	Schulung	Promotion	2009	18200
741222	Vertrieb	Abitur	2008	9600
750227 *)	Geschäftsführung	Mittlere Reife	2010	2800
750705	Schulung	Master	2009	9200
760104	Test/Anwendungen	Bachelor	2010	5000
760930	Entwicklung	Bachelor	2010	4800
780522	Finanzen	Abitur	2010	6400
780920	Entwicklung	Master	2010	5400
790820	Entwicklung	Bachelor	2010	6800
791112 *)	Schulung	Bachelor	2010	2400
811030 *)	Test/Anwendungen	Abitur	2010	2000
820512	Vertrieb	Bachelor	2010	5500
850624	Vertrieb	Abitur	2010	4000
890823	Test/Anwendungen	Abitur	2010	3400

*) Halbtagskräfte

Quelle: Wewel, Statistik für BWL

Bezeichnungen in einer Häufigkeitsverteilung

Am Beispiel: Personalerhebung

Symbol	Bezeichnung	Beispiel
X	Merkmal	Abteilung
x_i	Merkmalsausprägungen	Finanzen, Vertrieb
n_i	Absolute Ausprägungen	Anzahl Vertriebsmitarbeiter
n (oder m)	Gesamtzahl der Elemente	25 (alle Mitarbeiter)

Quelle: Wewel, Statistik für BWL

Drei Arten von Häufigkeitsverteilungen

Häufigkeiten werden unterschieden in

- **absolute** Häufigkeiten n_i

Beispiel: Ein Unternehmen verkauft in einem Monat 500 Einheiten eines bestimmten Produkts. Die absolute Häufigkeit der Verkäufe beträgt 500.

$$n_i (i = 1 \ldots m) \text{mit} \sum_{i=1}^{m} n_i = n$$

- **relative** Häufigkeiten h_i

Beispiel: In einer Umfrage geben 80 von 200 befragten Kunden an, mit dem Service zufrieden zu sein. Die relative Häufigkeit der zufriedenen Kunden beträgt 0.4 (80 / 200).

$$h_i = \frac{n_i}{n} \text{mit} \sum_{i=1}^{m} h_i = 1$$

- **prozentuale** Häufigkeiten

Beispiel: In einem Unternehmen sind von 100 Mitarbeitern im Durchschnitt 5 Mitarbeiter krankgeschrieben. Die prozentuale Häufigkeit der Abwesenheiten beträgt 5% (5 / 100 * 100).

$$\frac{n_i}{n} \cdot 100$$

Quelle: Wewel, Statistik für BWL

Beurteilung der Merkmale

Für jedes Merkmal wird die Anzahl der Ausprägungen, Skalierung und Merkmalsart bestimmt

Merkmal	Anzahl der Merkmals-ausprägungen	Skalierung	Merkmalsart
Abteilung	5 (z. B. Vertrieb, IT, HR, Produktion, Einkauf)	Nominalskala	Qualitativ
Ausbildung	4 (z. B. Bachelor, Master, Ausbildung, kein Abschluss)	Ordinalskala	Qualitativ
Eintrittsjahr	35 (z. B. 1995–2025)	Intervallskala	Quantitativ – diskret
Bruttogehalt	100+ (z. B. in 100 €-Schritten)	Verhältnisskala	Quantitativ – stetig

Quelle: Wewel, Statistik für BWL

Tabellarische Darstellung

Die tabellarische Darstellung eignet sich für qualitative Merkmale

Nr.	Merkmals-ausprägungen x_i	Strichliste	Häufigkeiten n_i	h_i
1	E: Entwicklung	||||| ||||	9	0.18
2	F: Finanzen	||||| ||	7	0.14
3	G: Geschäftsführung	||||| ||||	9	0.18
4	S: Schulung	||||| ||||	9	0.18
5	T: Test/Anwendungen	||||| ||||	9	0.18
6	V: Vertrieb	||||| ||	7	0.14
(Summe)			50	1.0

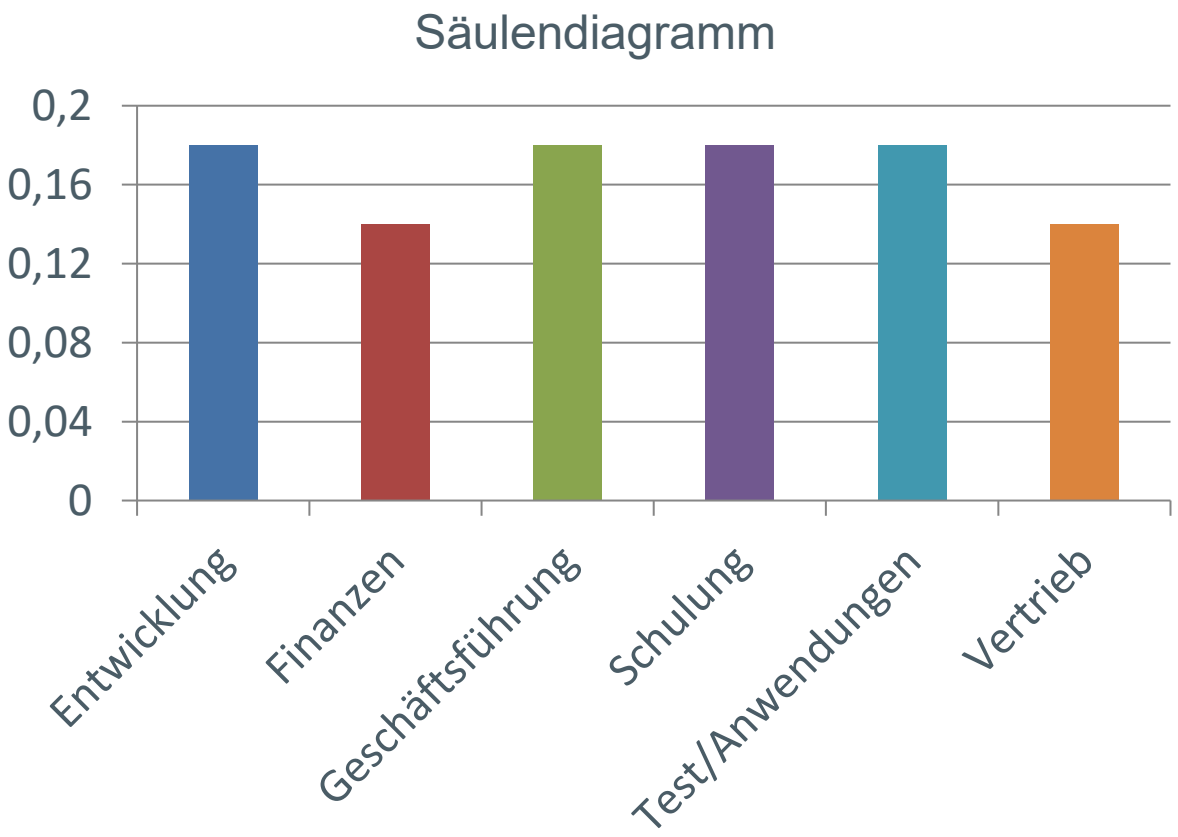
9/50

7/50

Quelle: Wewel, Statistik für BWL

Darstellung als Säulendiagramm

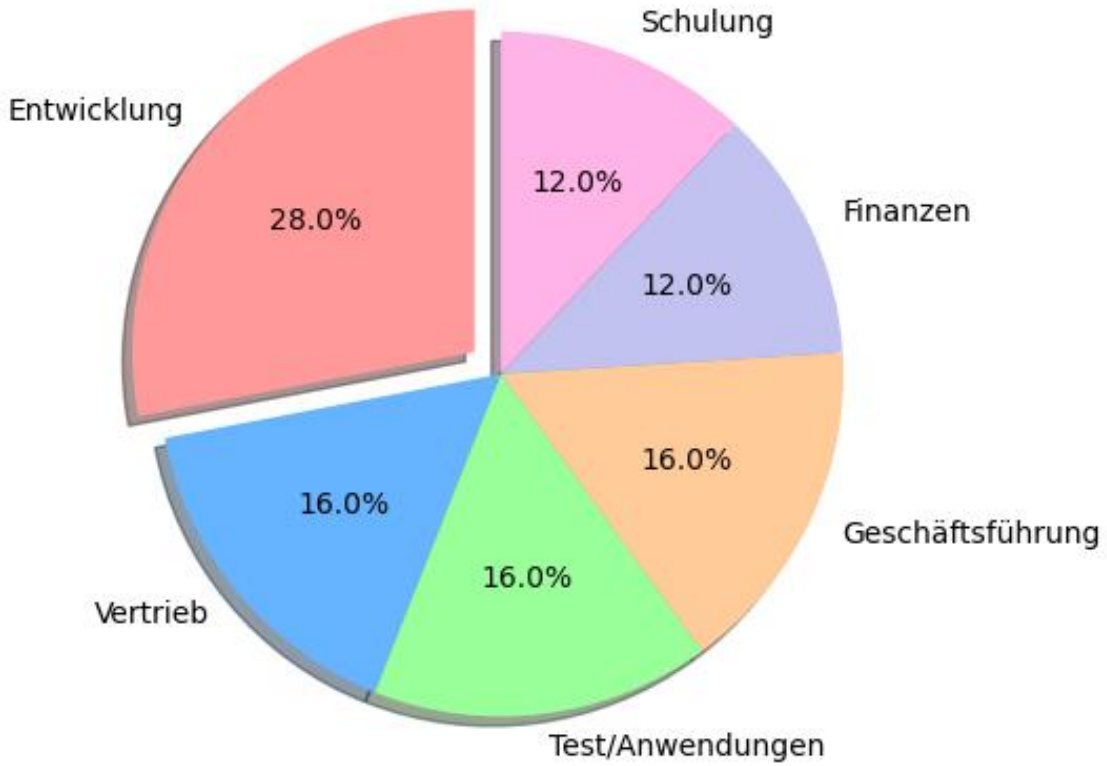
Die Darstellung als Säulendiagramm eigenet sich für qualitative Merkmale.
Hier mit relativen Häufigkeiten



Quelle: Wewel, Statistik für BWL

Darstellung als Kreisdiagramm

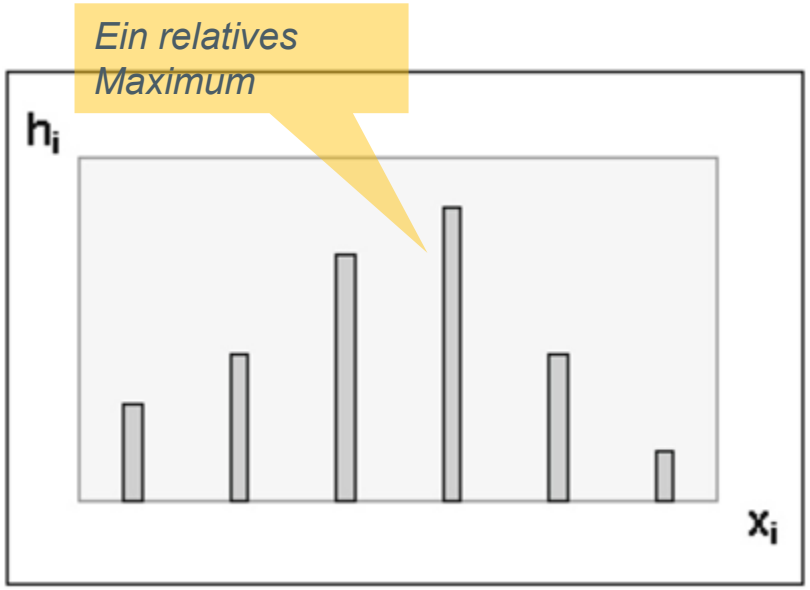
Die Darstellung als Kreisdiagramm zeigt die prozentualen Häufigkeiten



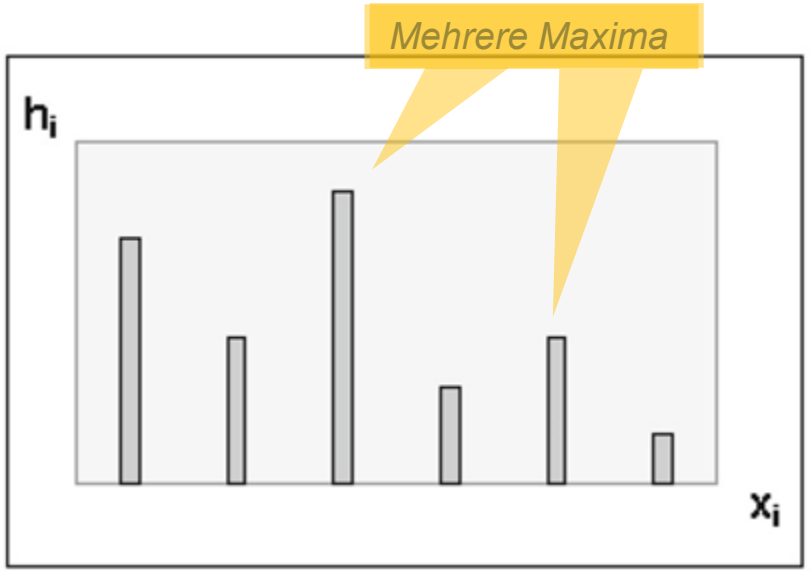
Quelle: Wewel, Statistik für BWL

Komparative Merkmale

- **Komparative Merkmale** weisen **eine eindeutige Rangfolge** auf und können daher der **Größe nach aufsteigend sortiert** werden.
- Beispiele: Schulnoten, Handelsklassen, Bonitätsrating
- Eine Häufigkeitsverteilung ist **unimodal**, wenn die Folge der Häufigkeiten nur ein relatives Maximum aufweist, sie ist **multimodal**, wenn die Folge mehrere relative Maxima aufweist.



Unimodale Verteilung



Multimodale Verteilung

Quelle: Wewel, Statistik für BWL

Kumulierte komparative Merkmale

- **Komparative Merkmale** können aufgrund der Rangfolge (Sortierung) auch kumuliert (d.h. aufsummiert) dargestellt werden.

- **Kumulierte absolute Häufigkeiten**:

$$N_i = \sum_{k=1}^{i} n_k = N_{i-1} + n_i \; (i = 1 \ldots m) \; (N_0 = 0)$$

- **Kumulierte relative Häufigkeiten**:

$$H_i = \sum_{k=1}^{i} h_k = H_{i-1} + h_i \; (i = 1, \ldots, m) \; (H_0 = 0)$$

Quelle: Wewel, Statistik für BWL

Tabellarische Darstellung komparativer Merkmale

i	x_i	n_i	h_i	N_i	H_i
1	Mittlere Reife	30	0.3	30	0.3
2	Abitur	25	0.25	55	0.55
3	Bachelor	20	0.2	75	0.75
4	Master	15	0.15	90	0.9
5	Promotion	10	0.1	100	1.0
Summe		100	1.0		

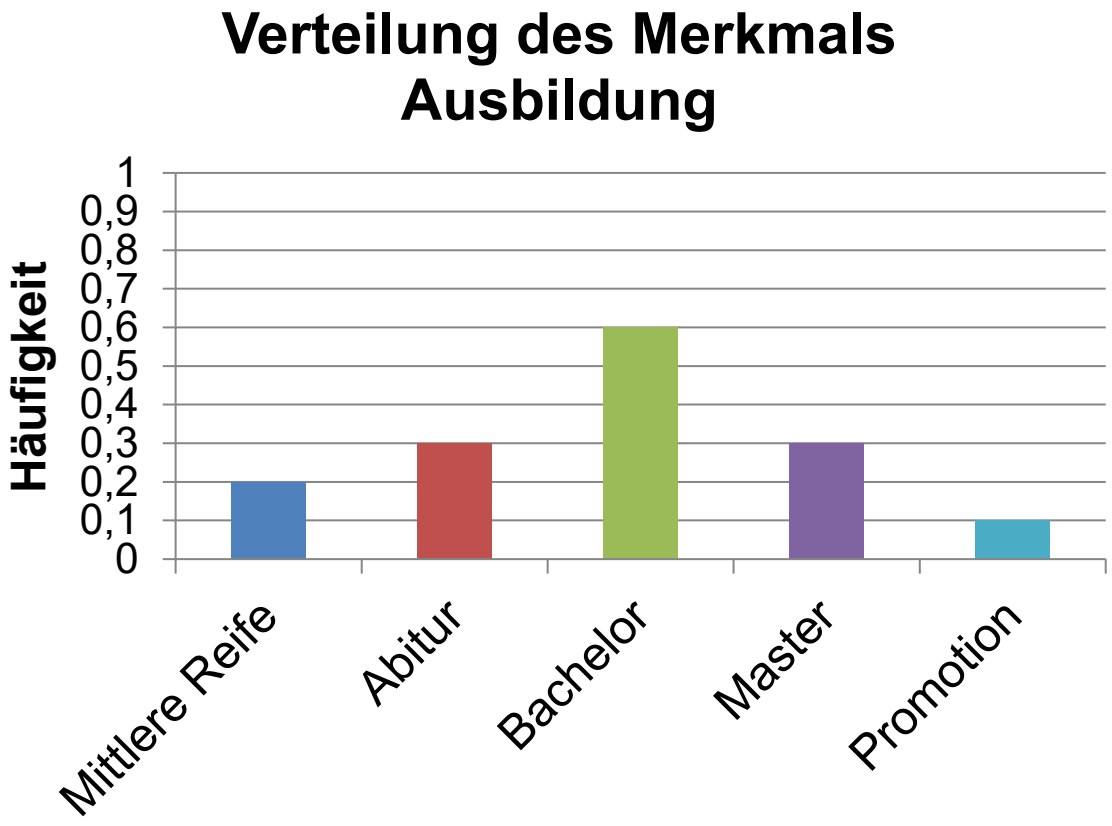
Symbol	Bezeichnung
x_i	Merkmalsausprägungen
n_i	Absolute Ausprägungen
h_i	Relative Ausprägung
n	Gesamtzahl der Elemente
N_i	Kummulierte absolute Häufigkeit
H_i	Kummulierte relative Häufigkeit

Quelle: Wewel, Statistik für BWL

Säulendiagramm nicht kummuliert

Verteilung des Merkmals „Ausbildung“ mit relativen Häufigkeiten.

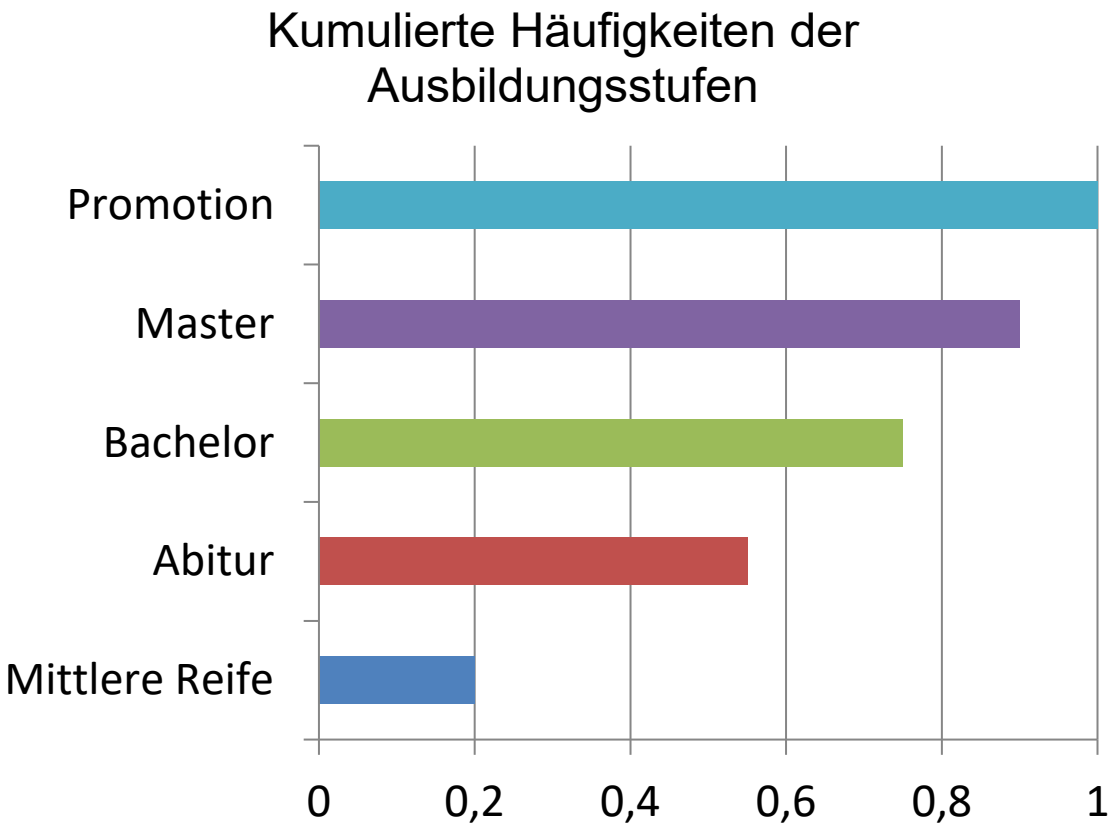
Noch nicht kummuliert !



Quelle: Wewel, Statistik für BWL

Balkendiagramm kummulierter Merkmale

Verteilung des komparativen Merkmals „Ausbildung“ als kummulierte relative Häufigkeiten



Quelle: Wewel, Statistik für BWL

Zweidimensionale Häufigkeiten

Werden zwei Merkmale in einer statistischen Auswertung betrachtet, so heißt diese Statistik **zweidimensionale Häufigkeitsverteilung**. Die Summe in der ***Randzeile*** bzw. ***Randspalte*** nennt man **Randverteilungen**.

Alternative Bezeichnungen:

- Kontingenztabelle
- Korrelationstabelle
- **Kreuztabelle** (Marktforschung)
- **Pivot-Tabelle (Excel)**

Merkmal Geschlecht

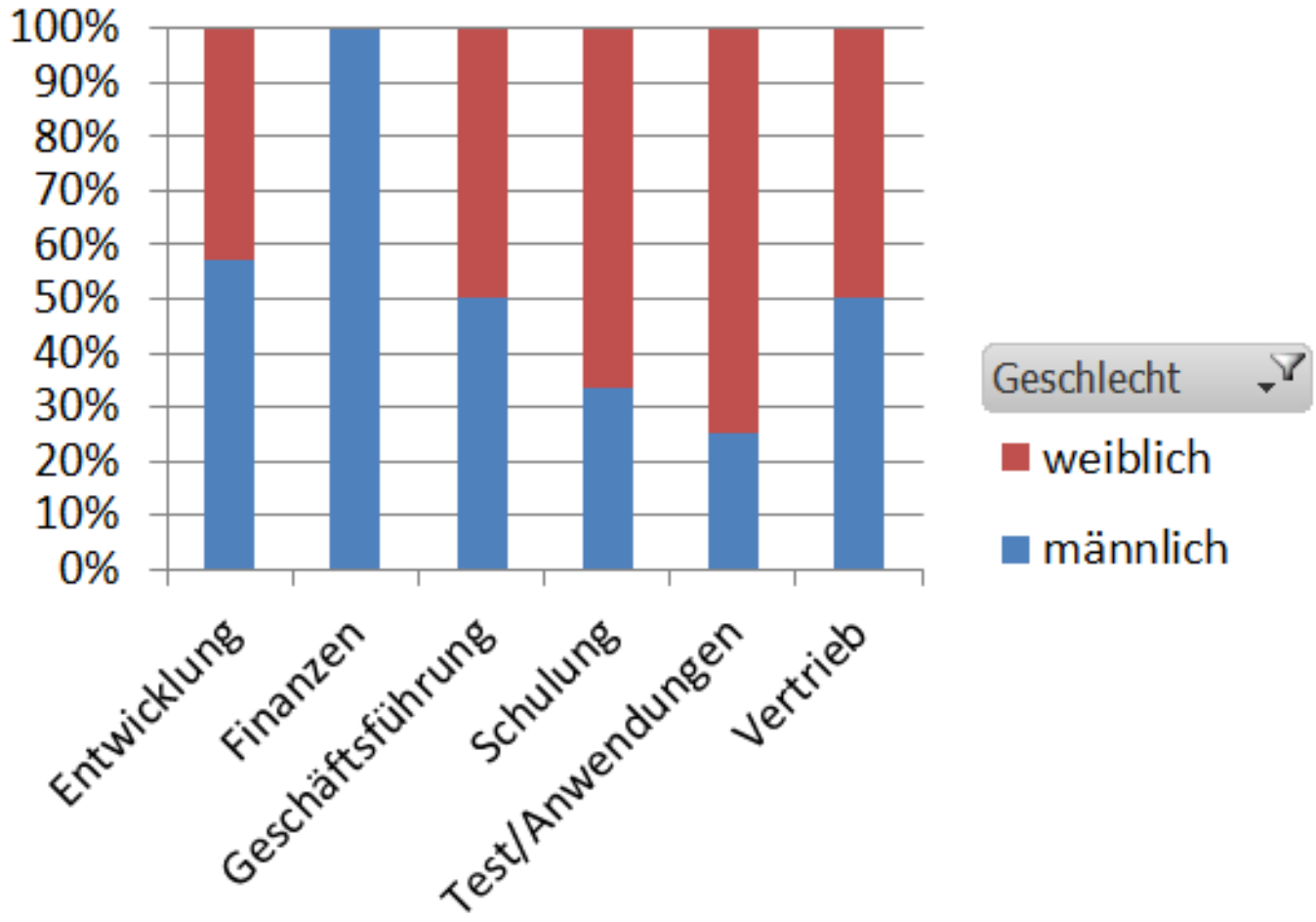
Randspalte

Merkmal Abteilung

Randzeile

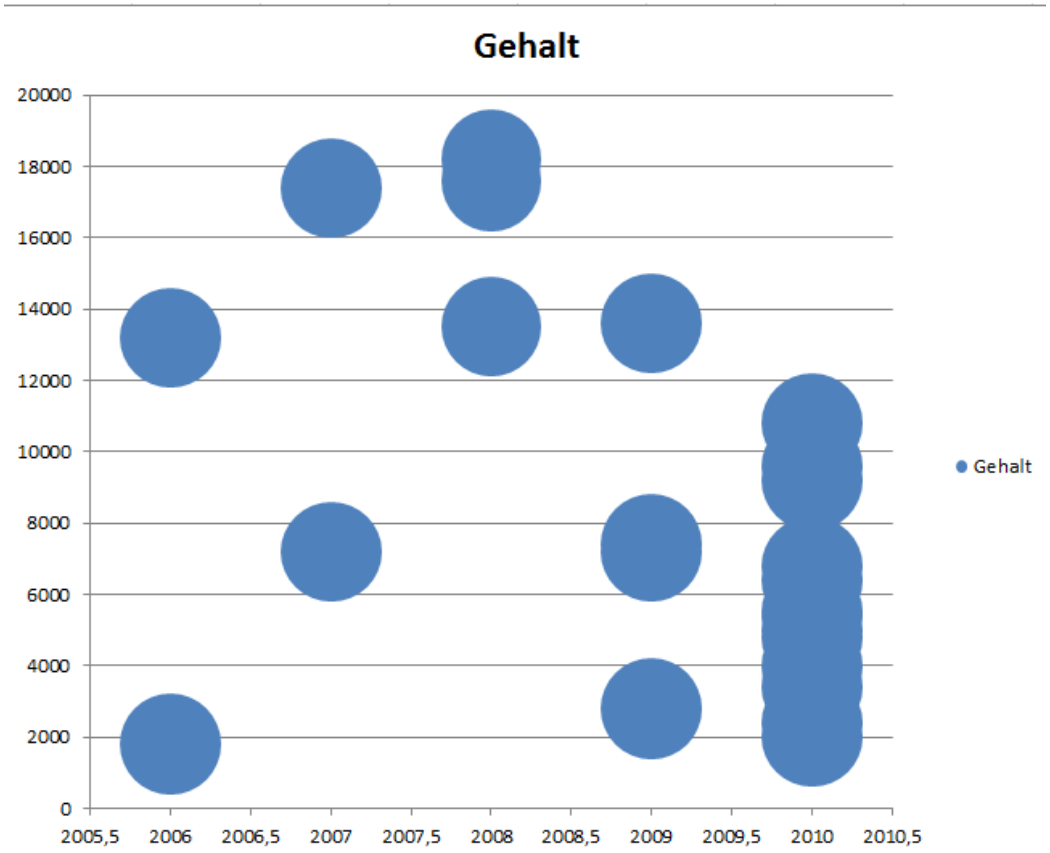
Zeilenbeschriftungen	männlich	weiblich	Gesamt
Entwicklung	4	3	7
Finanzen	3		3
Geschäftsführung	2	2	4
Schulung	1	2	3
Test/Anwendungen	1	3	4
Vertrieb	2	2	4
Gesamtergebnis	**13**	**12**	**25**

Kreuztabelle als „gestapeltes Balkendiagramm“



Streudiagramm

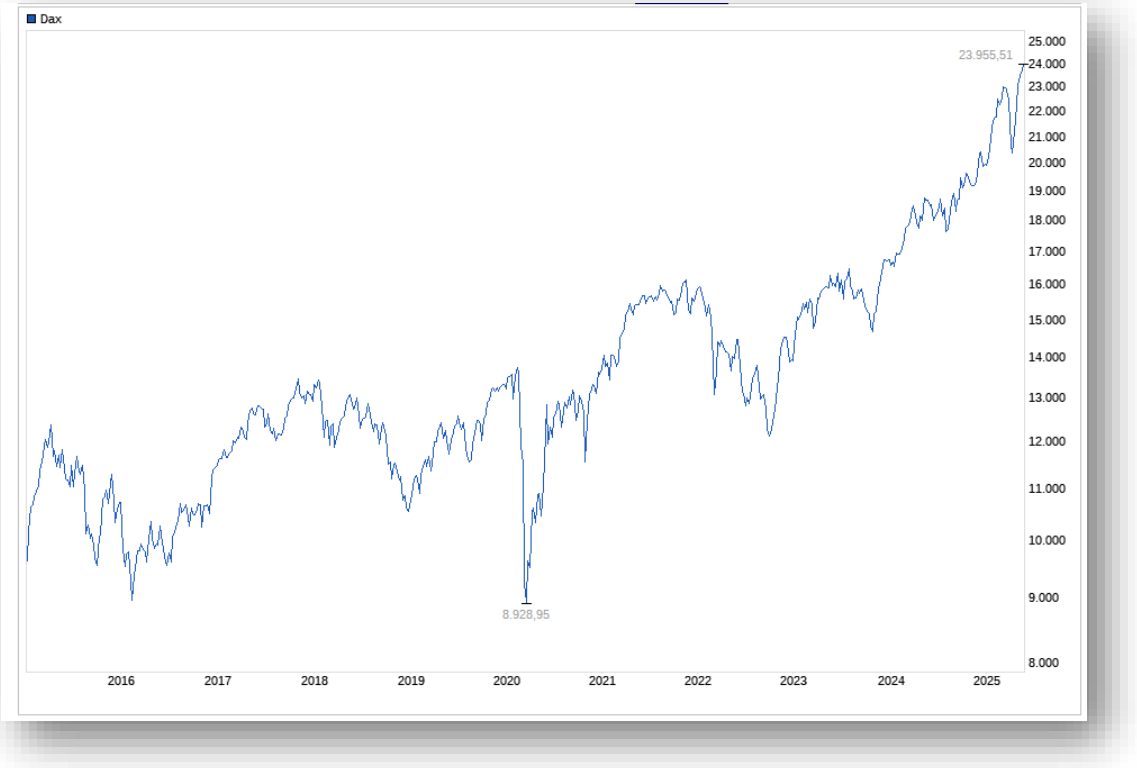
Möchte man eine eventuelle Beziehung zwischen zwei Merkmalen pro Merkmalsträger beobachten, so verwendet man **Streudiagramme** (auch **Blasendiagramme**).



Darstellung der Korrelation zwischen Eintrittsjahr und Gehalt als „**Blasendiagramm**“

Stetige Merkmale

Bei **stetigen** (oder quasi-stetigen) Merkmalen ist der Vergleich zwischen den einzelnen Werten sinnvoll. Zur Darstellung von solchen Zeitreihen verwendet man häufig **Liniendiagramme**.



Dax-Entwicklung der letzten 10 Jahre

Klassierung bei stetigen Merkmalen

- Bei **stetigen quantitativen** Merkmalen existieren i.d.R. sehr viele Merkmalsausprägungen (Umsatz pro Unternehmen, Bilanzsumme, Preis pro Fahrzeug, verkaufte Menge, …)
- Bei einer Auswertung der Häufigkeitsverteilung wären fast alle n_i gleich 1.
- Für eine übersichtliche Darstellung der Häufigkeiten bildet man daher **Klassen**

Klassen:

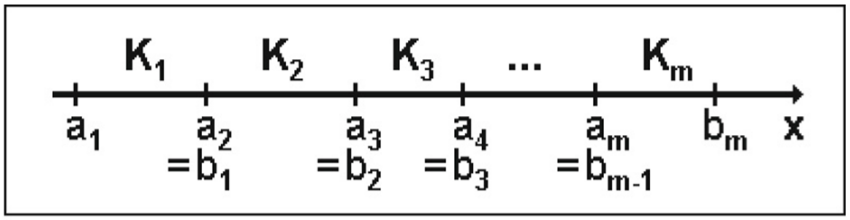
$$K_i = [a_i; b_i[\qquad (i = 1, ..., m)$$

Klassenmitten:

$$x_i^{\star} = \frac{a_i + b_i}{2} \qquad (i = 1, ..., m)$$

Klassenbreiten:

$$w_i = b_i - a_i \qquad (i = 1, ..., m)$$



Quelle: Wewel, Statistik für BWL

Tabellarische Darstellung quantiatives stetiges Beispiel

i	$K_i = [a_i; b_i[$	x_i	w_i	n_i	h_i	H_i
1	[0; 3000[	1500	3000	20	0.2	0.2
2	[3000; 5000[	4000	2000	25	0.25	0.45
3	[5000; 7000[	6000	2000	30	0.3	0.75
4	[7000; 10000[	8500	3000	15	0.15	0.9
5	[10000; 20000[	15000	10000	10	0.1	1.0
Summe				100	1.0	

Klasse von 10.000 inklusive bis 20.000 exklusive

Symbol	Bezeichnung
x_i	Merkmalsausprägungen
n_i	Absolute Ausprägungen
w_i	Klassenbreite
h_i	Relative Ausprägung
n	Gesamtzahl der Elemente
N_i	Kummulierte absolute Häufigkeit
H_i	Kummulierte relative Häufigkeit

Quelle: Wewel, Statistik für BWL

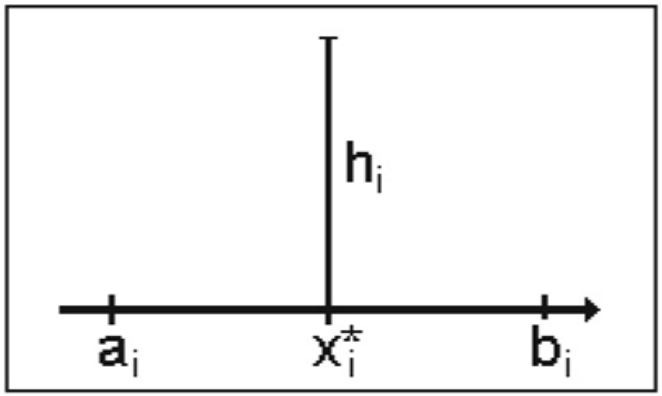
Verteilung stetiger Merkmale

Ein Problem bei Klassierungen liegt im *Informationsverlust*, wo die einzelnen Beobachtungswerte in der Klasse liegen (verteilt sind). Übliche und plausible Annahmen zur Verteilung der Werte innerhalb einer Klasse sind:

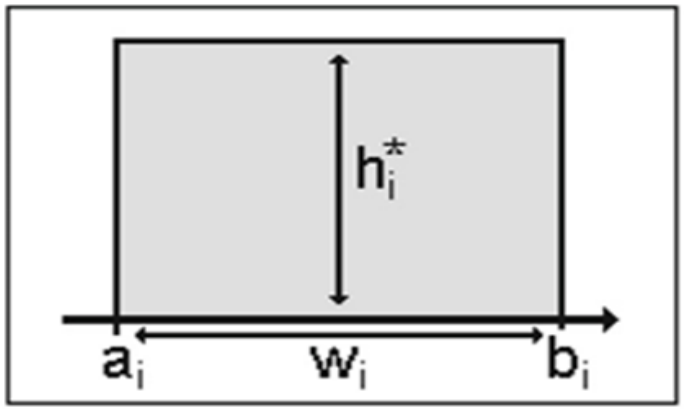
- **Einpunkt-Verteilung**: Werte liegen auf einem Punkt in der Klassenmitte
- **Rechteck-Verteilung**: stetige gleichmäßige Verteilung (Streuung) in der Klassenbreite (Dichte der Beobachtungswerte durch Flächeninhalt gegeben).

Normierte relative Häufigkeit:

$$h_i^\star = \frac{h_i}{w_i} \qquad (i = 1, \ldots, m)$$



Einpunkt-Verteilung



Rechteck-Verteilung

Quelle: Wewel, Statistik für BWL

Klassierungs-Funktionen

Die grafische Darstellung einer klassierten Häufigkeit erfolgt meist mittels eines **Histogramms** oder der **klassierten empirischen Verteilungsfunktion**.

Die zugehörige **Histogrammfunktion** lautet:

$$h^{\star}(x) = \begin{cases} h_i^{\star} & \text{für} \quad a_i \le x < b_i \quad (i = 1, \ldots, m) \\ 0 & \text{sonst} \end{cases}$$

Die (**stetige**, **monoton steigende**) **klassierte empirische Verteilungsfunktion** lautet:

$$H(x) = \begin{cases} 0 & \text{für } x < a_1 \\ H_i - h_i^{\star} \cdot (b_i - x) & \text{für } a_i \le x < b_i \quad (i = 1, \ldots, m) \\ 1 & \text{für } x \ge b_m \end{cases}$$

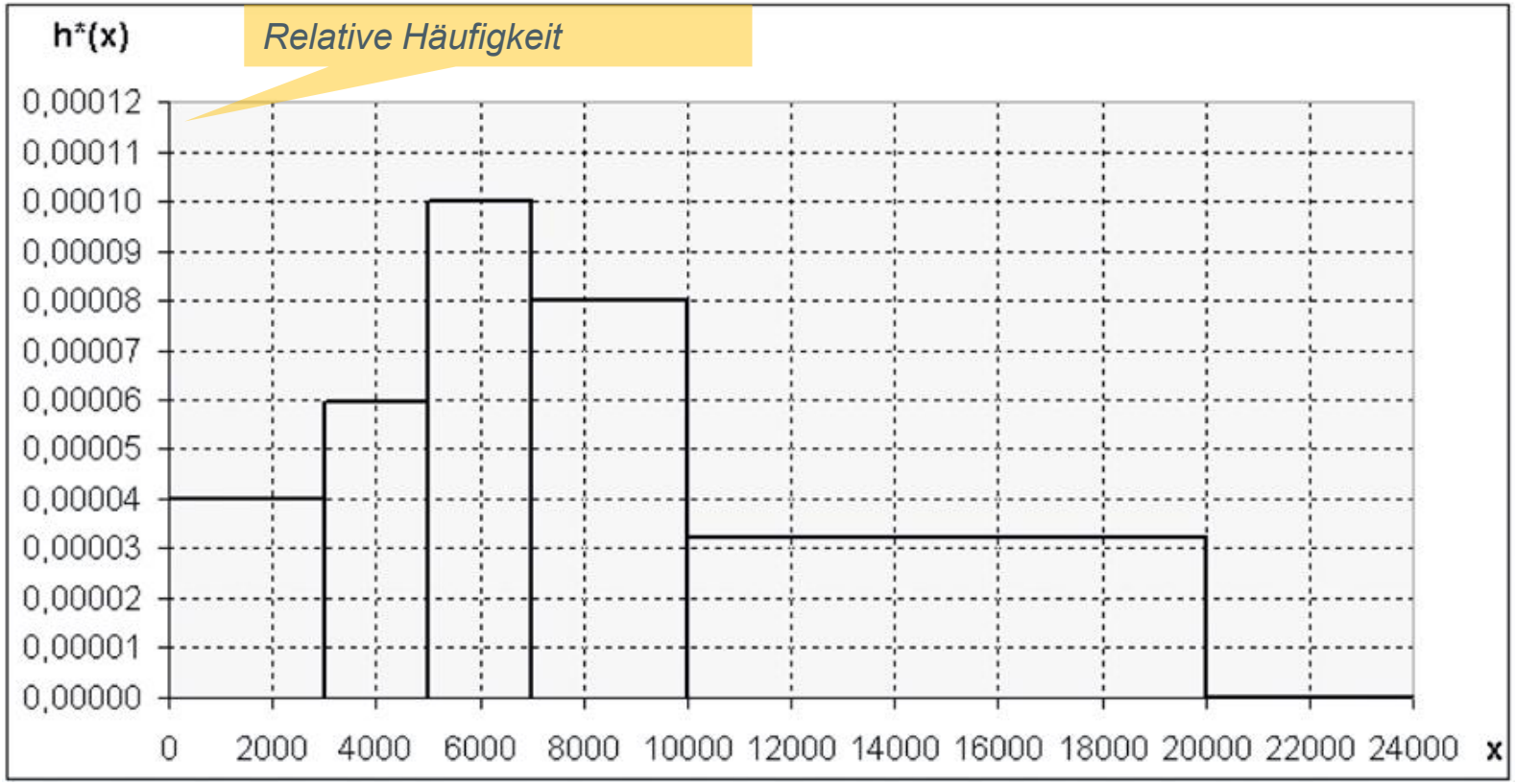
Es gilt folgender Zusammenhang:

- h*(x) ist die **Ableitung** von H(x)
- H(x) ist die **Stammfunktion** von h*(x)

Quelle: Wewel, Statistik für BWL

Histogrammfunktion für stetige Merkmale

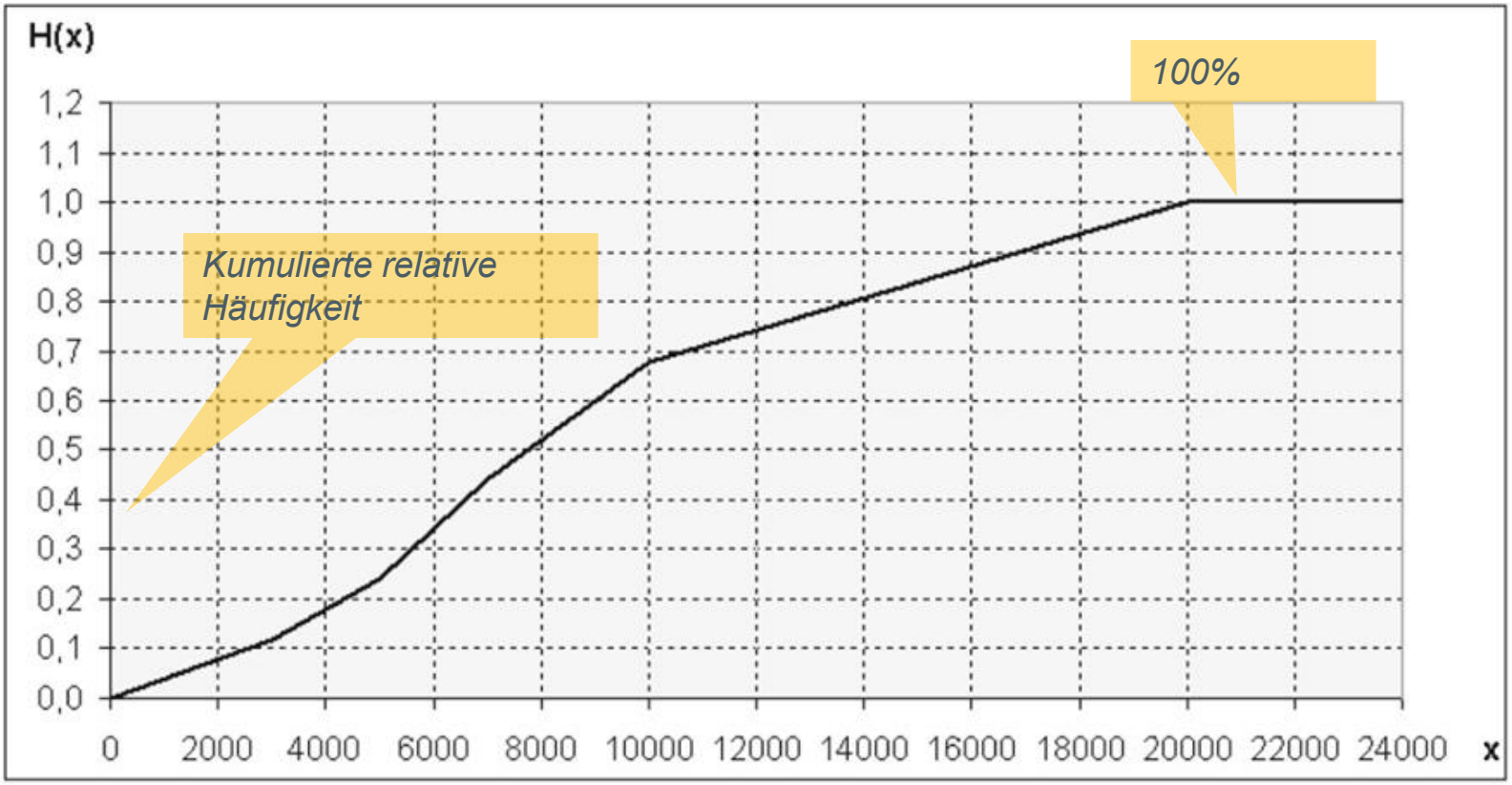
Histogramm der Verteilung des Merkmals „Bruttogehalt“



Quelle: Wewel, Statistik für BWL

Klassierte empirische Verteilungsfunktion

Empirische Verteilungsfunktion des Merkmals Bruttogehalt



Quelle: Wewel, Statistik für BWL

Beschreibende Statistik

Grundlagen	Häufigkeits-verteilungen	**Lage- & Streumaße**	Zweidim-ensionale Häufigkeits-verteilungen
Korrelation & Regression			

Lagemaße & Streuungsmaße

Aufgabe der statistischen Datenanalyse ist es, aus dem i.d.R. unübersichtlichen Datenmaterial (*Urliste*) die *relevanten spezifischen Informationen* herauszufiltern, die für ein zu lösendes Entscheidungsproblem benötigt werden. Die Daten werden entweder direkt aus den **Beobachtungswerten** oder aus den **gruppierten** bzw. **klassierten Häufigkeitsverteilungen** ermittelt.

Die wichtigsten Kenngrößen sind **Lagemaße** (Mittelwerte) und **Streuungsmaße**.

Lagemaße geben eine zentrale Tendenz der Beobachtungswerte an.

→ wichtigstes Maß: das **arithmetische Mittel**

Streuungsmaße geben an, wie stark sich die Beobachtungswerte unterscheiden (streuen).

→ wichtigste Streuungsmaße: **Varianz** / **Standardabweichung**

Lagemaß: Modus

Das einfachste Lagemaß ist der **dichteste Wert** (engl. *dense*) oder **Modus**.

Der **Modus** ist die **Merkmalsausprägung**, die am **häufigsten** vorkommt.

$$\overline{x}_D = x_k \quad \text{mit} \quad h_k = \max h_i$$

Der **dichteste Wert** (Modus) ist unmittelbar aus einem Balkendiagramm (Säulendiagramm) ersichtlich und **der einzig sinnvolle Mittelwert für qualitative Merkmale.**

Beispiel: diskrete Merkmale der Personaldaten

- Abteilung: _____
- Ausbildung: _____
- Eintrittsjahr: _____

Lagemaße: Median oder Zentralwert

- Der **Median** (oder auch **Zentralwert**) ist der **Wert**, der in einer *geordneten* Beobachtungsreihe in der **Mitte** steht.

$$\tilde{x} = \begin{cases} x_{\frac{n+1}{2}} & n \text{ ungerade} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) & n \text{ gerade.} \end{cases}$$

Beispiel:

Der **Median** für die Werte 4, 2, 1, 1, 37 lautet: **2**
(geordnete Reihe: 1, 1, 2, 4, 37)

Bemerkung:

In einer geordneten Reihe gibt der **Median** (der Wert, nicht der Rangplatz) an, dass die Hälfte der Beobachtungswerte kleiner (oder gleich) als der Median und die andere Hälfte größer als der Median ist.

Der **Median** steht also in der Mitte aller Beobachtungswerte.

Lagemaße: arithmetisches Mittel

Das **arithmetische Mittel** (auch *Mittelwert* oder *Durchschnitt*) ist das am häufigsten verwendete Maß. Es ist ein *summarisches* Maß und somit nur bei quantitativen Merkmalen anwendbar.

N: Anzahl der Werte
x_i: Wert i
i: Zählvariable

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Bemerkungen:

- Das arithmetische Mittel hat eine sogenannte Schwerpunkteigenschaft, d.h. die Summe der Abweichungen der einzelnen Ausprägungen vom Mittelwert ist gleich 0:

$$\sum_{i=1}^{n} (x_i - \overline{X}) = 0$$

- Die Summe der quadrierten Abweichungen vom Mittelwert ist minimal:

$$\sum_{i=1}^{n} (x_i - \overline{X})^2 = \text{Min}$$

Beispiele zum arithmetischen Mittel

Altersdurchschnitt der Studierenden:

- Annette (22), Bernd (23), Michael (19), Thomas (24):
- → (22 + 23 + 19 + 24) / 4 = 88 / 4 = **22 J**

Kaufpreis eines PKW:

- Audi (45 T€), BMW (40 T€), Opel (25 T€), Porsche 100T€):
  → (45 + 40 + 25 + 100) / 4 = 210 / 4 = **52,5 T€**

Das arithmetische Mittel: Gruppierung

Das arithmetische Mittel kann bei gruppierten Verteilungen nicht einfach aus den Gruppenausprägungen abgeleitet werden. Hier hilft das **gewichtete arithmetische Mittel**.

Beispiel: Personalbogen

Eintrittsjahr
2006
2007
2008
2009
2010

Ansatz 1: (2006 + 2007 + 2008 + 2009 + 2010) / 5 = 10.040 / 5 = 2008

Fehler: Die Anzahl der jeweiligen Ausprägungen (**Gewichtung**) ist hier nicht berücksichtigt.

Lösung: das **gewichtete arithmetische Mittel**

$$\overline{X} = \frac{\sum_{i=1}^{m} (n_i \cdot x_i)}{\sum_{i=1}^{m} n_i}$$

Eintrittsjahr	Anzahl ID
2006	2
2007	2
2008	3
2009	4
2010	14
Gesamtergebnis	25

Ansatz 2 (gewichtet):

$$\frac{(2 \cdot 2006 + 2 \cdot 2007 + 3 \cdot 2008 + 4 \cdot 2009 + 14 \cdot 2010)}{2 + 2 + 3 + 4 + 14}$$

$$= \frac{50226}{25} = \boxed{2009,04}$$

Das arithmetische Mittel: Klassierung

Die Berechnung des arithmetische Mittel bei **klassierten** Verteilungen kann als Spezialfall des gewichteten arithmetischen Mittels interpretiert werden. Da man die einzelnen Ausprägungen nicht kennt, nimmt man zur *näherungsweisen* Ermittlung die Häufigkeiten sowie die jeweiligen Klassenmitten.

$$\overline{X} = \sum_{i=1}^{m} h_i \cdot x_i^{\star}$$

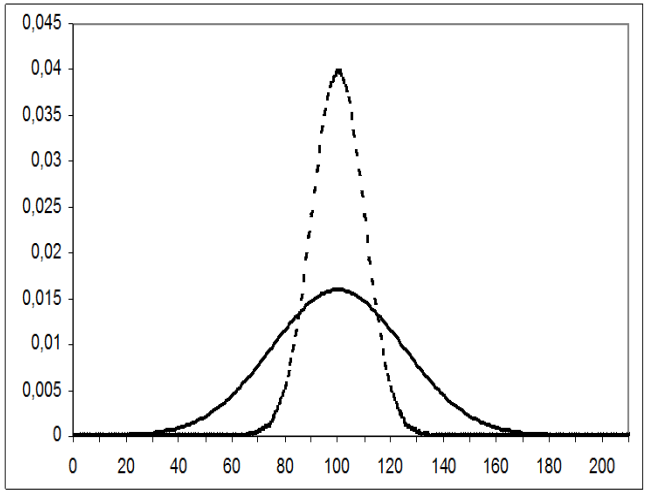
m: Anzahl der Werte
h_i: relative Häufigkeit
x_i: einzelner Wert
i: Zählvariable

Beispiel:

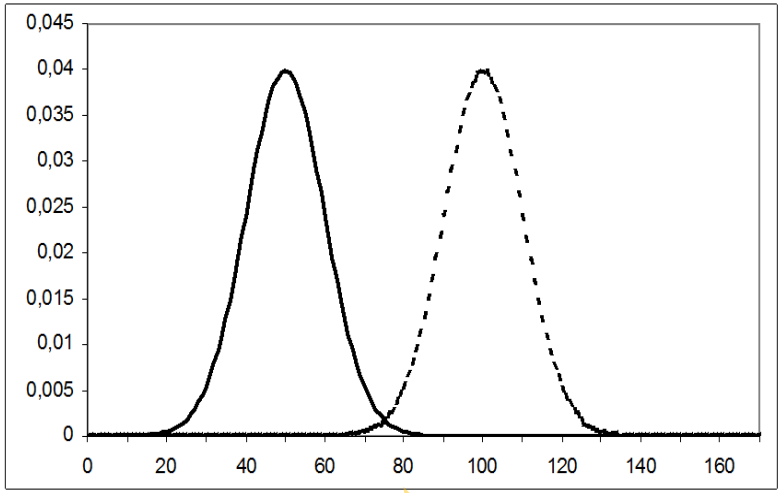
K_i	x_i^*	h_i	$h_i x_i^*$
[0; 3000[	1500	0,12	
[3000; 5000[	4000	0,12	
[5000; 7000[	6000	0,20	
[7000; 10000[	8500	0,24	
[10000; 20000[	15000	0,32	
---	---	1,00	

Streuungsmaße

Mit Hilfe der **Streuungsmaße** erhalten wir Informationen, wie weit die einzelnen Beobachtungswerte zusammen liegen bzw. wie weit sie **gestreut** (verteilt) sind.



2 Verteilungen mit unterschiedlichen Streuungen und gleichem Mittelwert



2 Verteilungen mit gleicher Streuung und unterschiedlichem Mittelwert

Quelle: Prof. Dr. M. Schmidt, THM FB Wirtschaft: Skript Statistik für BWL

Spannweite einer Verteilung

- Erstes einfaches Maß ist die **Spannweite**
- Grober Anhaltspunkt über die Streuung der Werte
- Messung: **Differenz** (*Abstand*) zwischen dem größten und kleinsten Beobachtungswert.

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

Eigenschaften:

- Einfach zu ermitteln
- Sehr abhängig von „Ausreißern“ der Beobachtungswerte nach unten bzw. nach oben
- Aussagekraft für die gesamte Verteilung sehr begrenzt
- Typischer Anwendungsbereich:
  Angabe von Höchst- und Tiefstwerten bei *Aktienkursen* pro Tag oder pro Jahr.

Quelle: Wewel, Statistik für BWL

Quantile, Quartile und Quartilsabstand

Quantile (häufig auch *Perzentile*) geben in einer vom kleinsten zum größten Wert geordneten Beobachtungsreihe mit n Werten denjenigen Beobachtungswert an, bis zu dem ein Anteil **p** (Prozent) der Beobachtung angefallen ist.

$$q_p = \begin{cases} = \frac{1}{2}(x_{n\cdot p} + x_{n\cdot p+1}) & \text{für n*p ganzzahlig} \\ = x_{[n\cdot p]+1} & \text{für n*p nicht ganzzahlig} \end{cases}$$

Beispiel:

x_3 x_4

$$x_1, \ldots, x_{10} = (1,1,1,3,4,7,9,11,13,13),\ p = 0{,}3$$

$$n \cdot p = 10 \cdot 0{,}3 = 3 \text{ ist ganzzahlig}$$

$$\rightarrow \tilde{x}_{0,3} = \frac{1}{2}(x_{n\cdot p} + x_{n\cdot p+1}) = \frac{1}{2}(x_3 + x_4) = \frac{1}{2}(1+3) = 2$$

Alternativ: p = 0,75

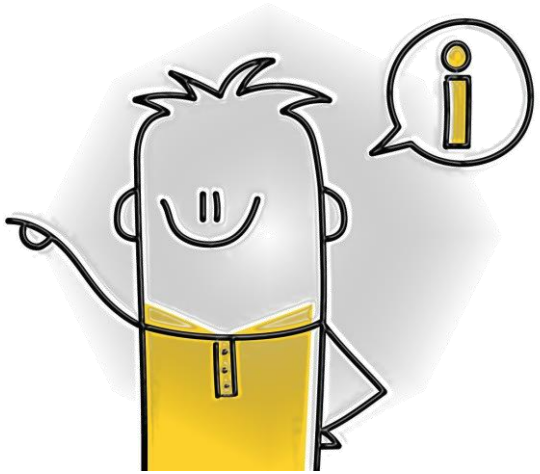
$$n \cdot p = 10 \cdot 0{,}75 = 7{,}5 \text{ ist nicht ganzzahlig} \rightarrow \tilde{x}_{0,75} = x_{\lceil n\cdot p \rceil} = x_{\lceil 7{,}5 \rceil} = x_8 = 11$$

Anwendungsbeispiele für Quantile

- Marktanteile: Die größten drei Unternehmen einer Branche haben einen Marktanteil von 80%

- Verteilung von Vermögen, Einkommen, Grundbesitz, …
  Die x-% Ärmsten eines Landes haben ein Einkommen bis zu y Euro.
  Die x-% Reichsten (Vermögendsten) Einwohner eines Landes halten y % des Gesamtvermögens.

- Ergonomie:
  In der Automobilbranche werden die Autos auf 90% der Bevölkerung ausgerichtet: Die 5% kleinsten Frauen und 5% größten Männer bleiben unberücksichtigt.



Quartile als spezielle Quantile

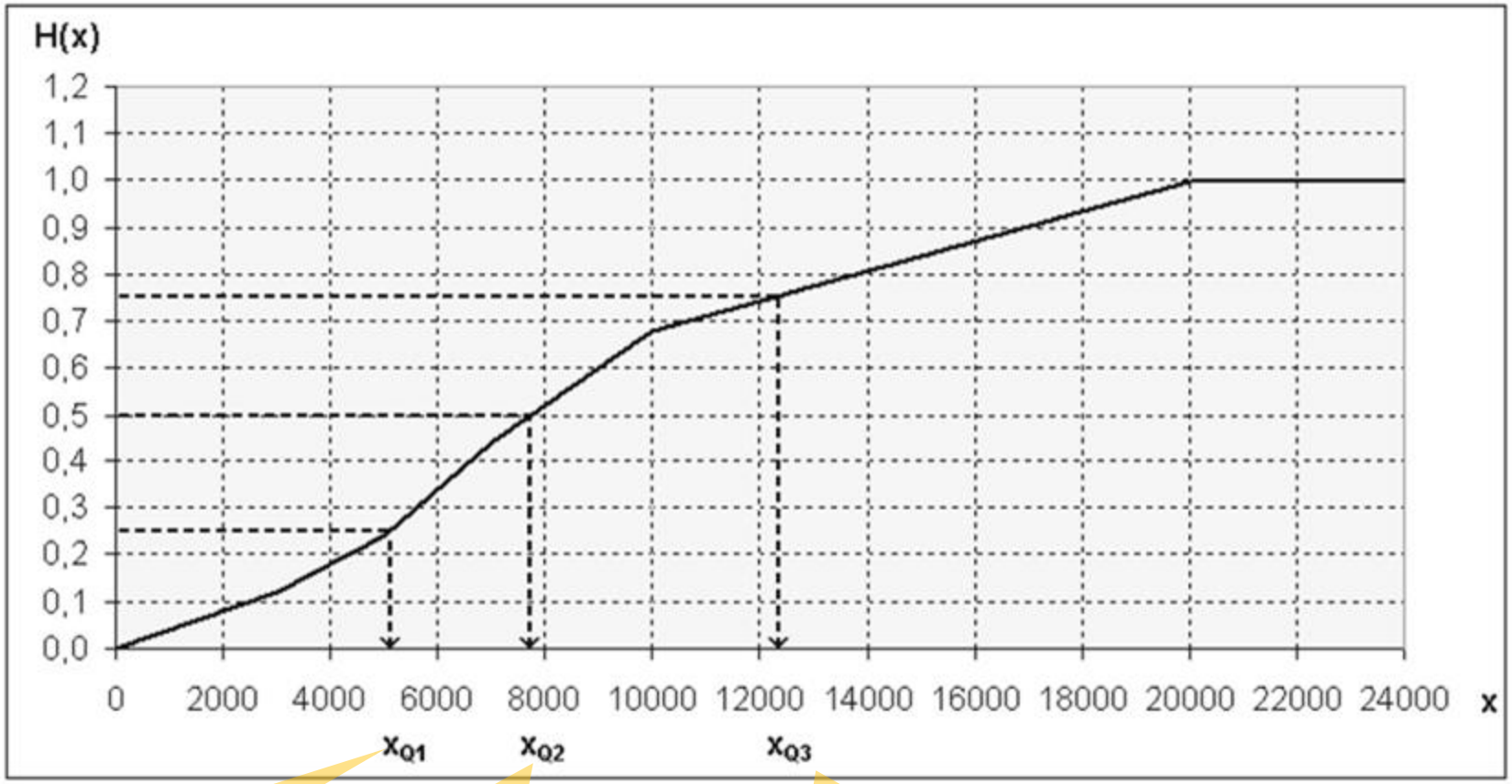
Es werden drei **Quartile** (*latein.: Viertel*) betrachtet, die jeweils als ein Vielfaches von 25% definiert sind:

- Das **erste** Quartil $q_{0,25}$ mit Werten bis **25%**
- Das **zweite** Quartil $q_{0,5}$ mit Werten bis **50%**
- Das **dritte** Quartil $q_{0,75}$ mit Werten bis **75%**

Bemerkung:
Das **zweite** Quartil entspricht dem **Median**.

Grafisches Beispiel für drei Quartile

Bestimmung der Quartile für das Bruttogehalt (x)



1. Quartil: 25%

2. Quartil: 50%

3. Quartil: 75%

Quelle: Wewel, Statistik für BWL

Quartilsabstand

Der **Quartilsabstand Q** (auch Inter-Quartilsabstand) ist definiert als der Abstand zwischen dem **dritten** und dem **ersten** Quartil:

$$Q \ (= IQA) = q_{0,75} - q_{0,25}$$

Der **Quartilsabstand** gibt die **Spannweite** der mittleren 50% der Beobachtungswerte an, d.h. die 25%-kleinsten sowie die 25%-größten Werte bleiben unberücksichtigt.

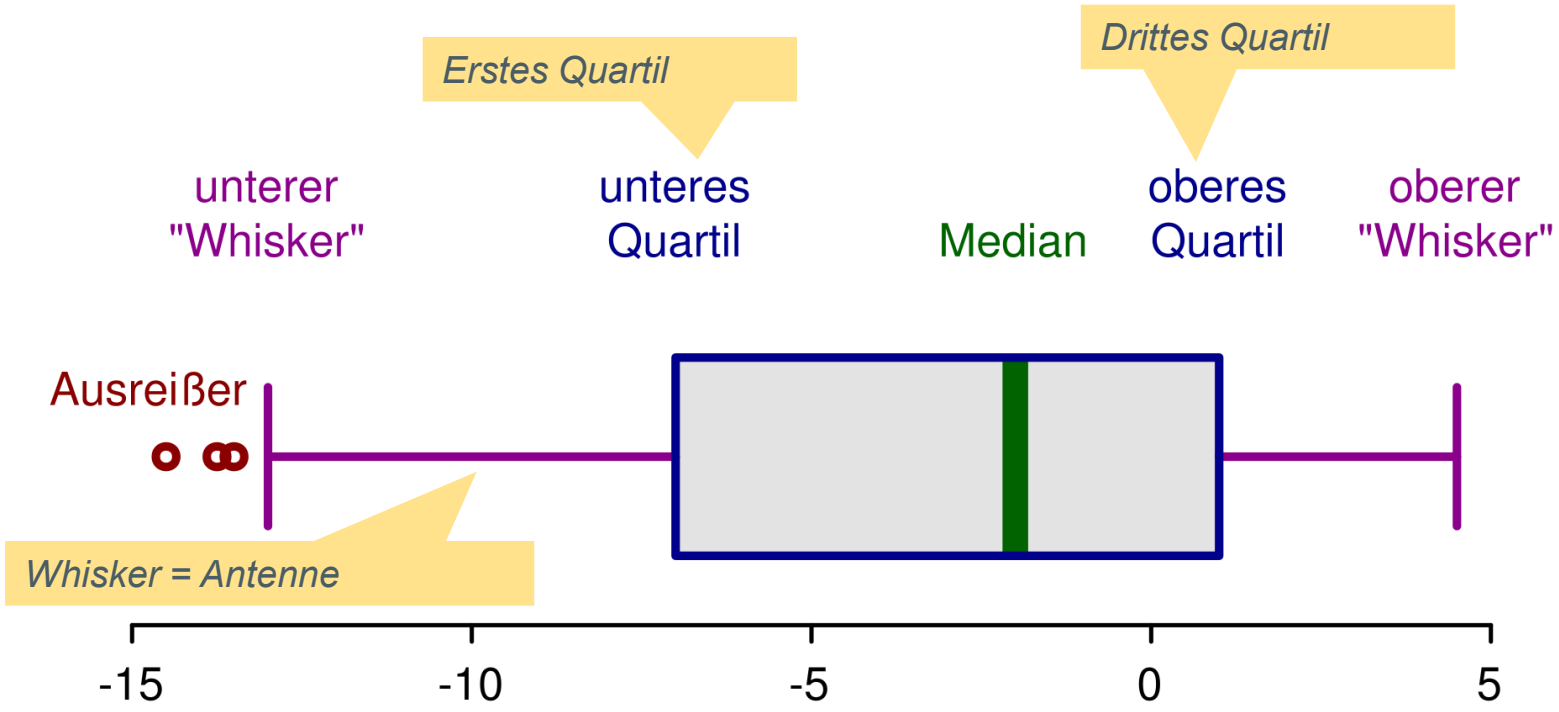
Vorteil:
Die Ausreißer nach unten bzw. nach oben verzerren nicht so stark den Wert der Spannweite, d.h. die Angabe der *häufigsten* Beobachtungswerte ist aussagekräftiger.

Beschreibende Statistik

Grundlagen	Häufigkeits-verteilungen	**Lage- & Streumaße**	Zweidim-ensionale Häufigkeits-verteilungen
Korrelation & Regression			

Box-Whisker-Plot (auch: Box-Plot)

Für eine erste schnelle Übersicht der Beobachtungswerte wird häufig ein sogenannter **Box-Whisker-Plot** eingesetzt

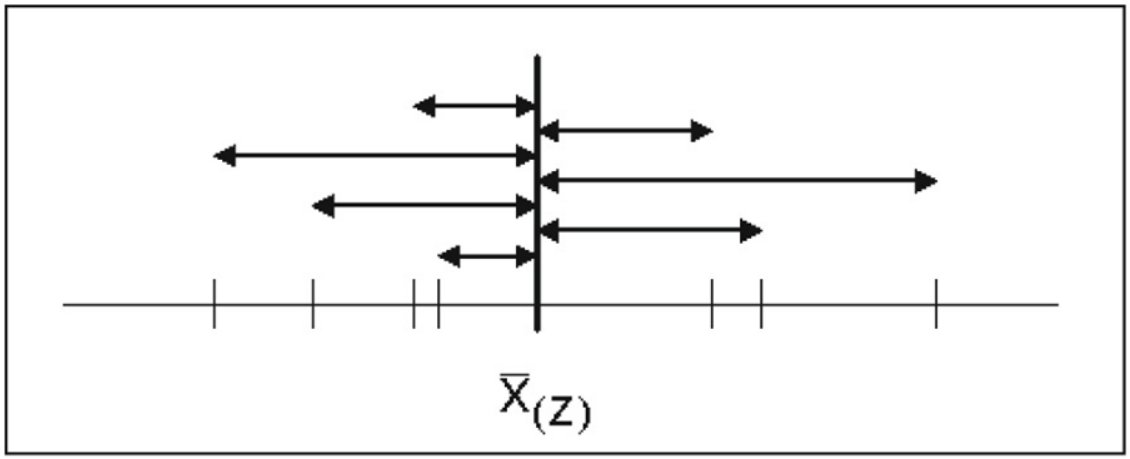


Durchschnittliche absolute Abweichung (1/2)

Fragestellung:
Wie kann man die Streuung der Werte messen, d.h. der Streuung eine Maßzahl zuordnen?
Idee: „Messen der Abstände“ (zu einer Referenzzahl)

Zur Erinnerung:
Die Summe der Abstände aller Beobachtungswerte vom Mittelwert ist 0.

$$\sum_{i=1}^{n} (x_i - \overline{X}) = 0$$



Streuung als Abweichung vom Mittelwert

Durchschnittliche absolute Abweichung (2/2)

Fragestellung:
Wie kann man die Streuung der Werte messen, d.h. der Streuung eine Maßzahl zuordnen?

Lösung (1): „Messen der **absoluten** Abstände“ (zu einer Referenzzahl)
→ Es werden die absoluten Abstände vom **arithmetischen Mittelwert** bzw. **Median** ermittelt und summiert.

Durchschnittliche absolute Abweichung

vom Mittelwert:

$$d_{\overline{X}} = \frac{1}{n} \cdot \sum_{i=1}^{n} |x_i - \overline{X}|$$

vom Median

$$d_{med} = \frac{1}{n} \cdot \sum_{i=1}^{n} |x_i - \overline{X_{med}}|$$

Betragszeichen = Ergebnis immer positiv

Varianz und Standardabweichung

Fragestellung:
Wie kann man die Streuung der Werte messen, d.h. der Streuung eine Maßzahl zuordnen?

Lösung (2): „Messen der **quadrierten absoluten** Abstände“ (zu einer Referenzzahl). Eine der am häufigsten eingesetzten Methoden.
Es werden die quadrierten absoluten Abstände vom **arithmetischen Mittelwert** ermittelt und summiert:

Varianz:

$$V \ (= S^2) \ = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{X})^2$$

Standardabweichung:

$$\sigma \ (= \sqrt{V}) \ = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \overline{X})^2}$$

Wurzel aus der Varianz

Variationskoeffizient

Sollen **Streuungsvergleiche** für *verhältnisskalierte* Merkmale angestellt werden, bei denen die Beobachtungswerte in den verglichenen Verteilungen verschiedene Größenordnung haben, so ist es meistens sinnvoll, die **absoluten Streuungsmaße** mit einem passenden **Mittelwert** zu normieren. Dadurch entsteht ein **relatives Streuungsmaß**.

→ Am gebräuchlichsten ist der **Variationskoeffizient** als **Quotient** aus **Standardabweichung** und **arithmetischem Mittel**:

Variationskoeffizient:

$$VK = \frac{\sigma}{\overline{X}}$$

Standardabweichung

Arithmetisches Mittel

Eine typische Anwendung ist der Vergleich der Streuung von Einkommen der Bevölkerung in Industrieländern im Vergleich zu Entwicklungsländern.

Beschreibende Statistik

Grundlagen	Häufigkeits-verteilungen	Lage- & Streumaße	**Zweidim-ensionale Häufigkeits-verteilungen**
Korrelation & Regression			

Zweidimensionale Häufigkeitsverteilungen

- Bisher **n Merkmalsträger** mit **m Merkmalen** betrachtet
- Konzentration auf jeweils ein Merkmal

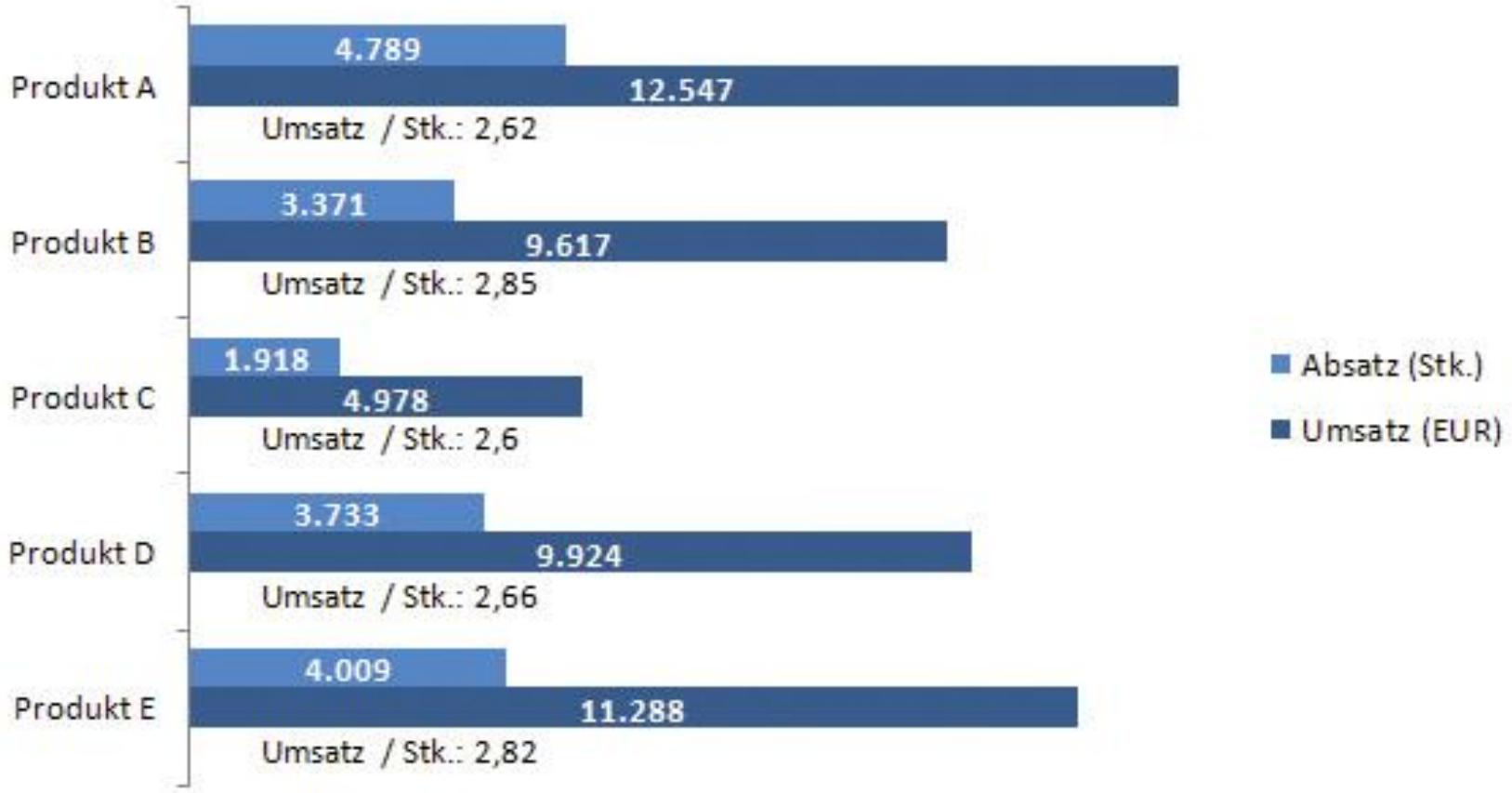
- Häufige Fragestellung: ob und welche Zusammenhänge es zwischen unterschiedlichen Merkmalen pro Merkmalsträger gibt.

- Zusammenhang geht bei isolierter Auswertung pro Merkmal verloren

- Informationen daher in Form einer **mehrdimensionalen Häufigkeitsverteilung**

Beispiel 2dim-Häufigkeitsverteilungen

Umsatz und Stückzahl pro Produkt



Quelle: PimpMyChart.com

Beobachtungswerte und Merkmalsausprägungen

Zweidimensionale Beobachtungswerte (bei den Merkmalsträgern jeweils **festgestellte** Kombinationen) der Merkmale **X** und **Y**.

$$(x_t, y_t) \qquad (t = 1, \ldots, n)$$

Symbol	Bezeichnung
x_t	Merkmalsausprägung von X
y_t	Merkmalsausprägung von Y
(x_i,y_j)	Wertepaar x und i
p	Menge aller Ausprägungen von X
q	Menge aller Ausprägungen von Y

Zweidimensionale Merkmalsausprägungen (bei den Merkmalsträgern jeweils **theoretisch mögliche** Kombinationen) der Merkmale **X** und **Y**.

$$(x_i, y_j) \qquad (i = 1, \ldots, p; j = 1, \ldots, q)$$

Da jedes x mit jedem y kombiniert werden kann, gibt es p x q zweidimensionale Ausprägungen v von (X, Y)

Beispiel zweidimensionale Häufigkeitsverteilung

Stellen wir uns vor, wir untersuchen den Zusammenhang zwischen **Bildung** (X) und **Einkommensklasse** (Y) bei einer Gruppe von Personen.

Merkmal X (Bildung):

- x_1: Kein Abschluss
- x_2: Schulabschluss
- x_3: Hochschulabschluss
  → **p=3**

Merkmal Y (Einkommen):

- y_1: < 1.000 €
- y_2: 1.000–2.000 €
- y_3: > 2.000 €
  → **q=3**

Mögliche Kombinationen: 3·3=**9** wie z. B.:

- (x_1,y_1): Kein Abschluss, < 1.000 €
- (x_3,y_3): Hochschulabschluss, > 2.000 €
- usw.

Absolute und relative Häufigkeiten

zweidimensionale **absolute Häufigkeiten**:

$$n_{ij} \qquad (i = 1, \ldots, p; j = 1, \ldots, q)$$

Symbol	Bezeichnung
n_{ij}	Absolute Häufigkeit der Merkmalskombination von x und y an der Stelle i, j
h_{ij}	relative Häufigkeit der Merkmalskombination von x und y an der Stelle i, j
n	Gesamtzahl der Elemente

zweidimensionale **relative Häufigkeiten:**

$$h_{ij} = \frac{n_{ij}}{n} \qquad (i = 1, \ldots, p; j = 1, \ldots, q)$$

Besipiel 2dim-Verteilung Personalerhebung (1/4)

Die **Urliste** der Personaldaten enthält für die 25 Beschäftigten (Merkmalsträger) des Unternehmens jeweils die **Merkmale** *Abteilung*, *Ausbildung*, *Eintrittsjahr* und *Gehalt*. Wir erhalten damit eine **vierdimensionale** Matrix mit Beobachtungswerten, um z.B. folgendes zu analysieren:

In welcher Abteilung ist der Anteil der Hochschulabsolventen am höchsten?

Gibt es deutliche Gehaltsunterschiede zwischen den Abteilungen?

Steigen die Gehälter mit zunehmender Qualifikation und/oder Betriebszugehörigkeit?

Besipiel 2dim-Verteilung Personalerhebung (2/4)

Zur besseren Übersicht fassen wir die Merkmale *Ausbildung* und *Abteilung* zu zwei übergeordneten **Klassen** zusammen:

Ausbildungen zu **Qualifikation**:

- **Nicht-Akademiker** (*Mittlere Reife, Abitur*)
- **Akademiker** (*Bachelor*, *Master*, *Diplom*, *Promotion*)

Abteilungen zu **Geschäftsbereich**:

- **Bereich I**: (*Geschäftsleitung*, *Finanzen*)
- **Bereich II**: (*Entwicklung* und *Test/Anwendungen*)
- **Bereich III**: (*Schulung* und *Vertrieb*)

Besipiel 2dim-Verteilung Personalerhebung (3/4)

absolute und relative Häufigkeitsverteilung

absolut

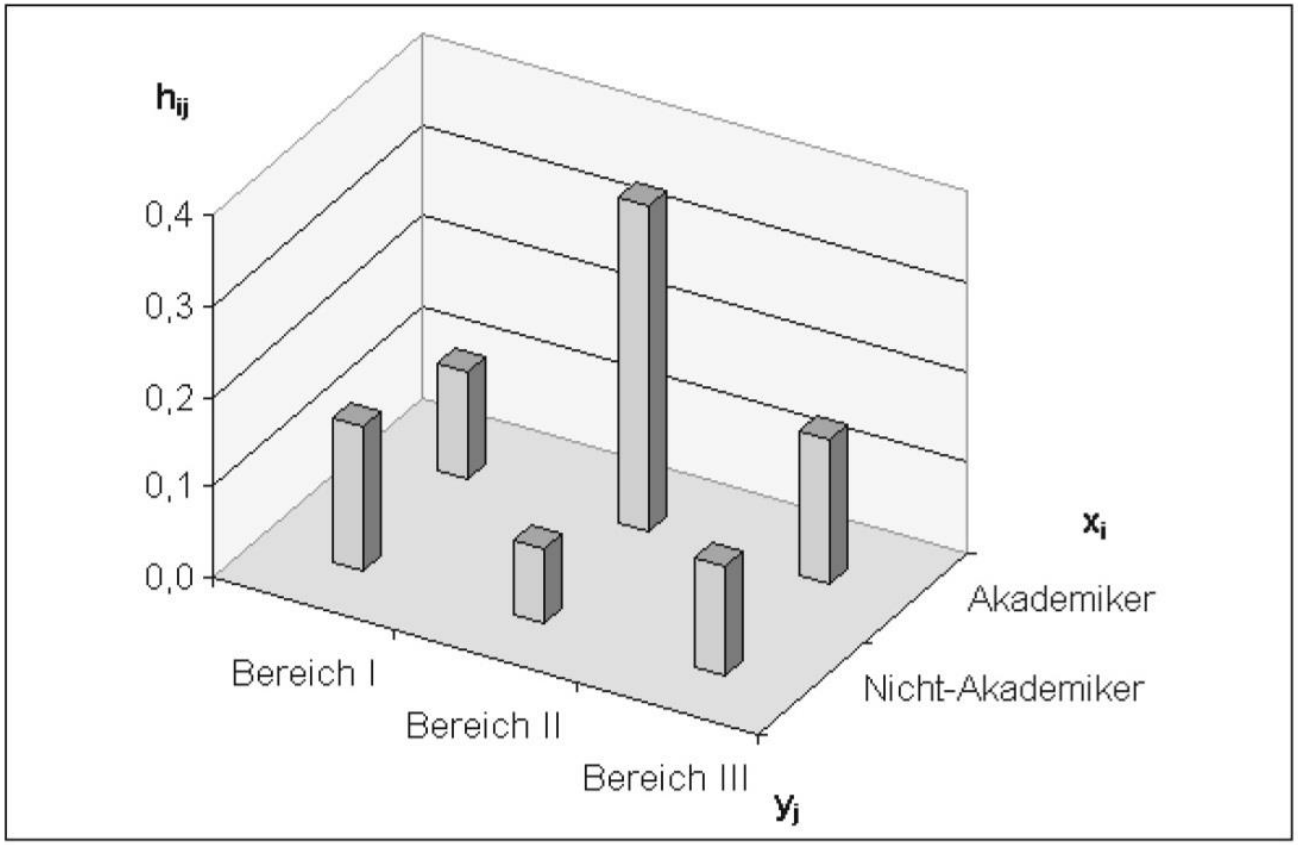
Bereich (y_j) / Qualifikation (x_i)	I	II	III	$n_{i.}$ ↓
Nicht-Akademiker				
Akademiker				
$n_{.j}$ →				

relativ

Bereich (y_j) / Qualifikation (x_i)	I	II	III	$h_{i.}$ ↓
Nicht-Akademiker				
Akademiker				
$h_{.j}$ →				

Beispiel 2dim-Verteilung Personalerhebung (4/4)

zweidimensionale Verteilung



Quelle: Wewel, Statistik für BWL

Beschreibende Statistik

Grundlagen	Häufigkeits-verteilungen	Lage- & Streumaße	Zweidim-ensionale Häufigkeits-verteilungen
Korrelation & Regression			

Korrelation

Bei **quantitativen** Merkmalen verwendet man zur Beurteilung des Zusammenhangs zwischen zwei Merkmalen die **Korrelationsmaße**. Dabei wird sowohl die **Intensität** (Stärke; Enge der Veribindung) als auch die **Richtung** des Zusammenhangs (positiv/negativ) gemessen:

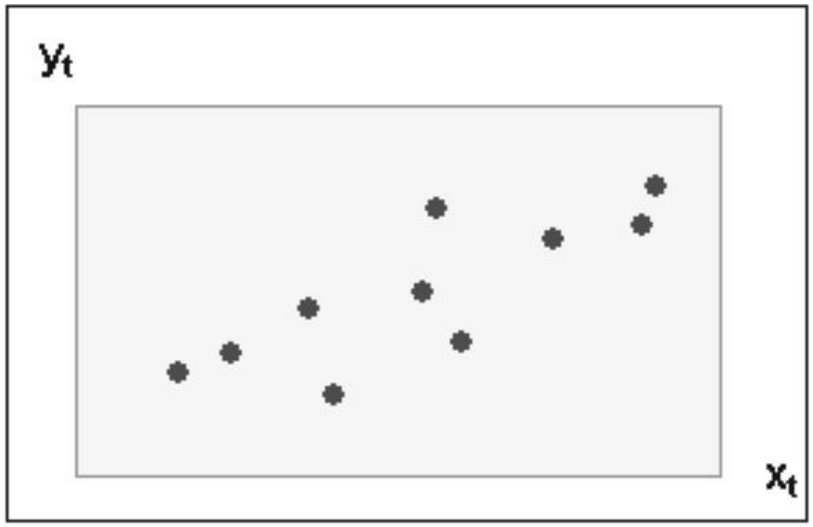
Positive Korrelation liegt vor, wenn die *zweidimensionalen* Beobachtungspaare (x_i, y_i) **gleichsinnig** sind, d.h. es gilt tendenziell

→ Je **größer** x_i, desto **größer** ist auch y_i
bzw. je kleiner x_i desto kleiner auch y_i

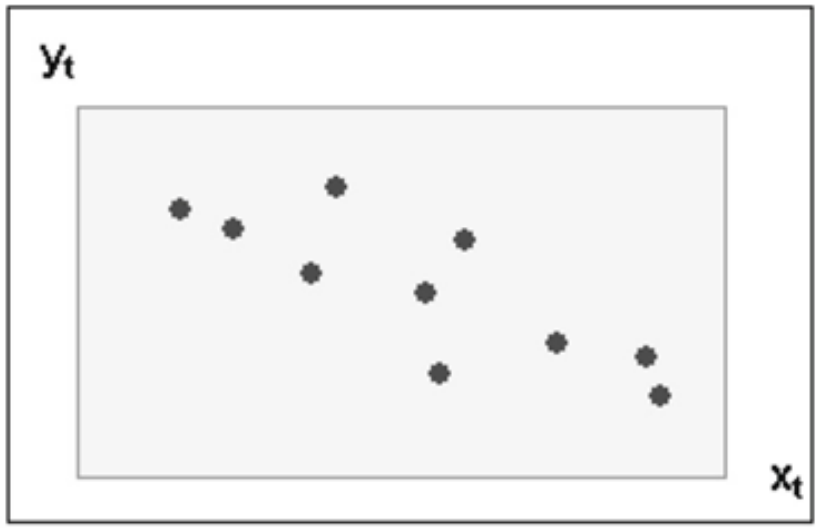
Negative Korrelation liegt vor, wenn die *zweidimensionalen* Beobachtungspaare (x_i, y_i) **gegensinnig** sind, d.h. es gilt tendenziell

→ Je **größer** x_i, desto **kleiner** ist y_i.

Positive vs. negative Korrelation



Streuungsdiagramm positiv korrelierter Beobachtungswerte

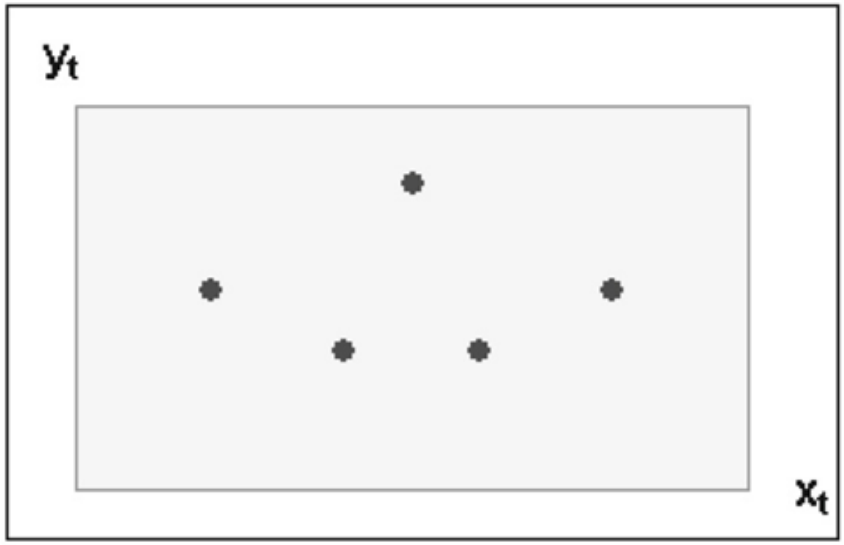


Streuungsdiagramm negativ korrelierter Beobachtungswerte

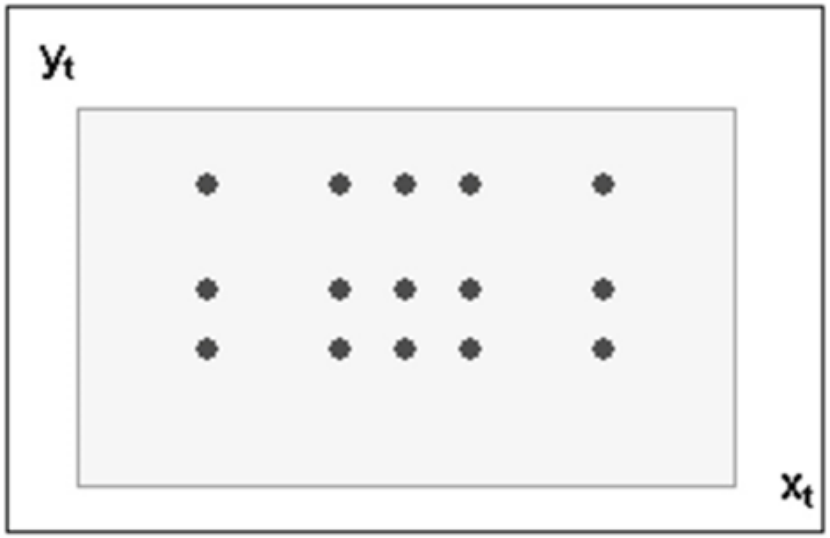
Quelle: Wewel, Statistik für BWL

Beispiele für unkorrelierte Verteilungen

Beispiele für **unkorrelierte** Verteilungen



Unkorreliertheit (keine lineare Beziehung)



Unabhängigkeit (somit auch Unkorreliertheit)

Stärkeres Konzept als Unkorrelierrtheit

Quelle: Wewel, Statistik für BWL

Kovarianz

Die grundlegende Kennzahl für die Korrelationsanalyse zweier quantitativer Merkmale X und Y ist die **Kovarianz**. Sie ist wie folgt definiert

$$\sigma_{xy} = \frac{1}{n} \cdot \sum_{t=1}^{n} (x_t - \overline{x}) \cdot (y_t - \overline{y})$$

h_{ij}: Die **Häufigkeit**, mit der die Kombination der Merkmalsausprägungen x_i und y_j beobachtet wurde.
x_i: Die i-te Ausprägung des Merkmals X
y_j: Die j-te Ausprägung des Merkmals Y
$\bar{x}$: Der **arithmetische Mittelwert** (Durchschnitt) aller x-Werte
$\bar{y}$: Der **arithmetische Mittelwert** aller y-Werte
p: Anzahl der Ausprägungen von X
q: Anzahl der Ausprägungen von Y

$$= \sum_{i=1}^{p} \sum_{j=1}^{q} h_{ij} \cdot (x_i - \overline{x}) \cdot (y_i - \overline{y})$$

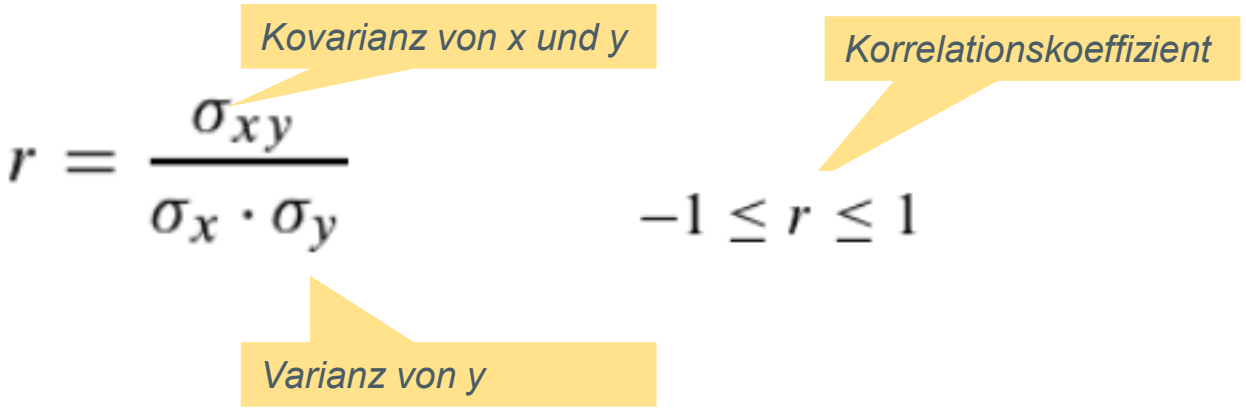
Im Gegensatz zur Varianz kann die Kovarianz auch negative Werte annehmen. Das Vorzeichen gibt die Richtung des Zusammenhangs der beiden Merkmale an:

$\sigma_{xy} > 0 \Rightarrow$ **positive** Korrelation $\quad$ $\sigma_{xy} < 0 \Rightarrow$ **negative** Korrelation

$\sigma_{xy} = 0 \Rightarrow$ **keine** Korrelation

Korrelationskoeffizient

Der **absolute** Wert der **Kovarianz** liefert keinen guten Anhaltspunkt für eine Interpretation der Intensität. Dazu wird die Kovarianz mittels der jeweiligen Standardabweichungen der Merkmale X und Y **normiert** und man erhält den **Korrelationskoeffizienten r**:

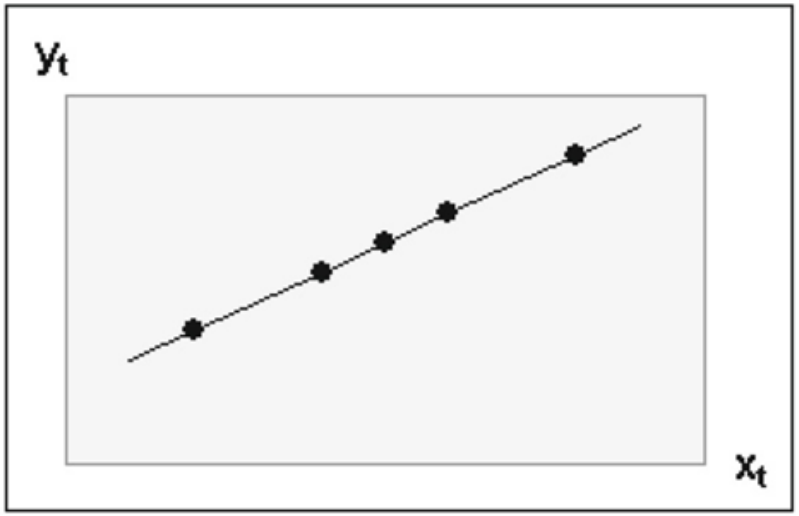


Bemerkung:

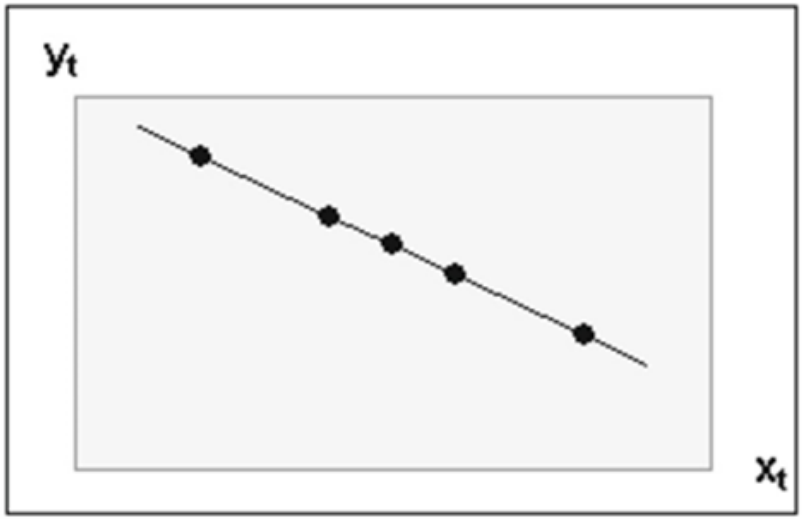
Liegen die Beobachtungspaare (x, y) auf einer **steigenden** bzw. **fallenden** Geraden, so nimmt der Korrelationskoeffizient r den Extremwert **1** (**steigend**) bzw. **-1** (**fallend**) an.

Korrelationsverläufe

Korrelationskoeffizient mit extremer Korrelation (r=1 bzw. r = -1)



Extreme positive Korrelation



Extreme negative Korrelation

Quelle: Wewel, Statistik für BWL

Richtung und Intensität

Faustregeln zur Interpretation des **Korrelationskoeffizienten** r bzgl. **Richtung (positiv, negativ)** und **Intensität (stark, schwach)**:

X und Y sind unabhängig verteilt → $\sigma_{xy} = r = 0$

$$
\begin{aligned}
-1 \leq r < -0,6 &\Rightarrow \textbf{starke negative Korrelation} \\
-0,6 \leq r < 0 &\Rightarrow \textbf{schwache negative Korrelation} \\
r = 0 &\Rightarrow \textbf{keine Korrelation} \\
0 < r \leq 0,6 &\Rightarrow \textbf{schwache positive Korrelation} \\
0,6 < r \leq 1 &\Rightarrow \textbf{starke positive Korrelation}
\end{aligned}
$$

Regressionsanalyse

- Weiterentwicklung der Korrelationsanalyse für **quantitative** Merkmale
- Annahme: Abhängigkeit eines Merkmals Y (Zielgröße: *Umsatz*, *Gewinn*, …) von einem Merkmal X (Instrumentgröße: *Absatzpreis*, *Werbeausgaben*, …).

- Diesen Zusammenhang beschreibt die **Regressionsfunktion *y = f(x)***

- Tatsächlichem Beobachtungswert y_t und berechneter Funktionswert $f(x_t)$ i.d.R. auch eine Abweichung (Fehler oder Störung: **Residuum**) **e**

- **lineare Regressionsmodell**: $\hat{y} = a + b \cdot x$

- mit dem Schätzfehler e

$$e_i = y_i - \hat{y}_i$$

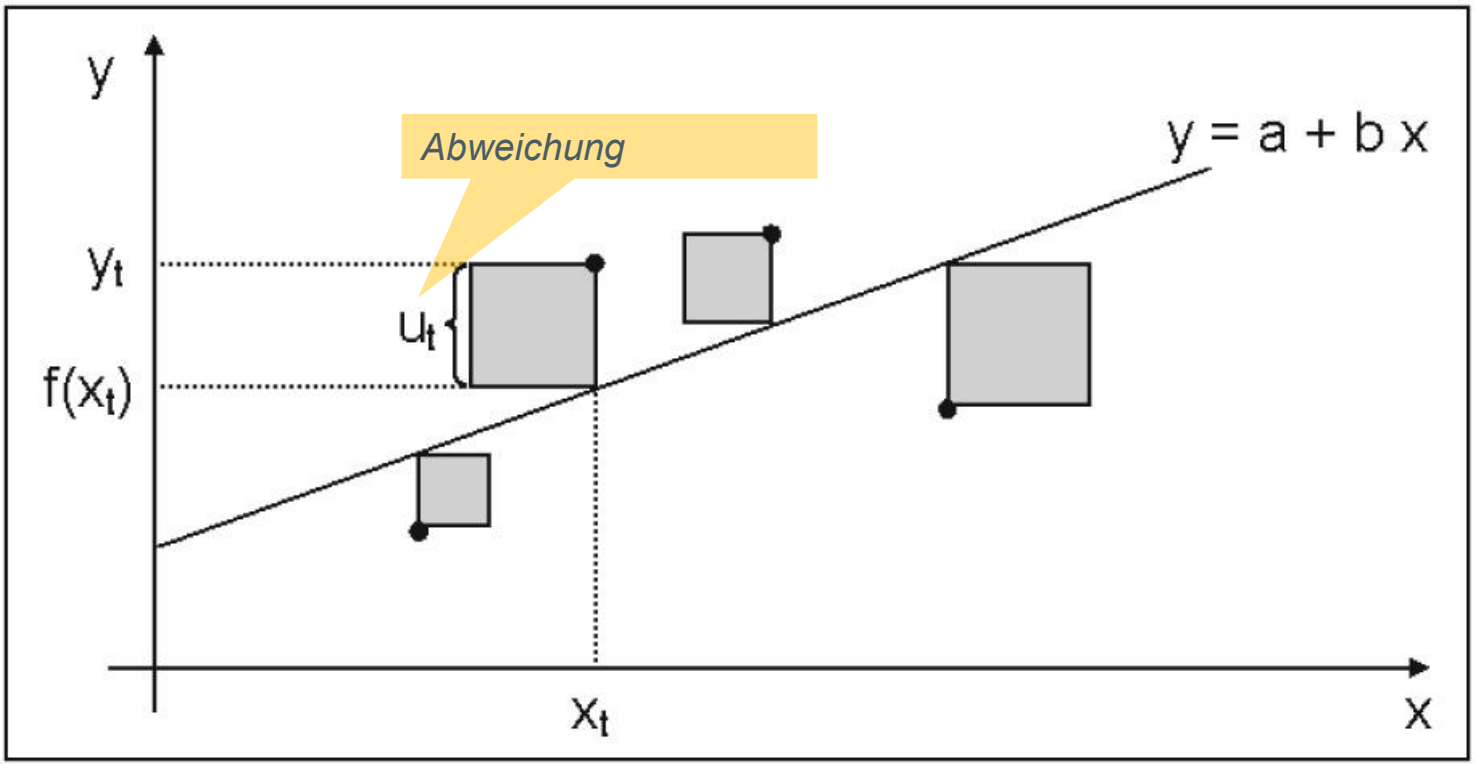
Schätzwert

Wert

- Anforderung:
  - Bestimmung der **Regressionskoeffizienten a** und **b**.
  - Beurteilung der **Aussagefähigkeit** (**Güte**) der Regressionsfunktion für Prognosen

Kleinst-Quadrate-Prinzip

- Als Optimalitätskriterium wird das **Kleinst-Quadrate-Prinzip** angewandt, d.h. die **Summe der quadrierten Fehler** ist **minimal**.
- Gerade so durch die Punktwolke legen, daß die Summe der horizontalen Abstandsquadrate minimiert wird



Quelle: Wewel, Statistik für BWL

Regressionskoeffizient & Regressionsgerade

Ermittlung der Regressionskoeffizienten. Es gilt zu minimieren:

$$q(a,b) := \sum_{t=1}^{n} e_t^2 = \sum_{t=1}^{n} (y_t - a - bx_t)^2 \quad \Rightarrow min!$$

Es ergibt sich als **Regressionsgerade**: $\hat{y} = a + b \cdot x$

mit den Regressions-koeffizienten a, b:

$$a = \overline{y} - b\,\overline{x}$$

$$b = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \overline{x}^2}$$

*Arithmet. Mittel aus x*y*

Quadrat des Mittelwerts

Arithmet. Mittel aus den quadrierten x-Werten

Der Steigungsparameter *b* für den Anstieg der Regressionsgeraden zeigt folgenden Zusammenhang zur Korrelation zwischen den Merkmalen X und Y:

- **steigende** Regressionsgerade ⇔ **positive** Korrelation
- **fallende** Regressionsgerade ⇔ **negative** Korrelation
- **horizontale** Regressionsgerade ⇔ **keine** Korrelation

Regressionsanalyse: Bestimmtheitsgard

Das Regressionsmodell ist umso besser, je größer der Anteil der erklärten Streuung an der Gesamtstreuung ist. Dieser Anteil **r^2** heißt **Bestimmungsgrad**:

$$= 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

mit

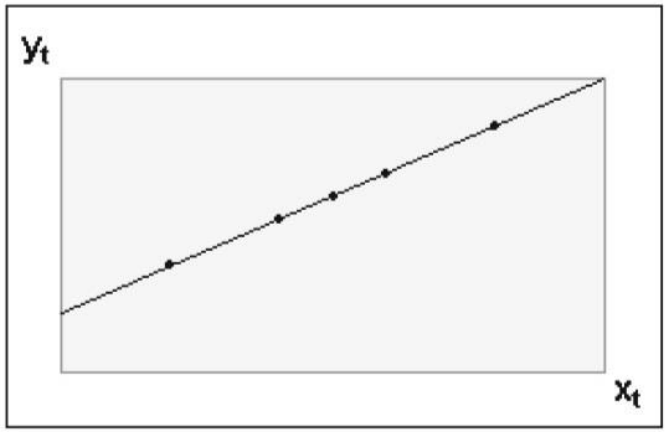
$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} y_i^2 - b_1 \sum_{i=1}^{n} y_i - b_2 \sum_{i=1}^{n} x_i y_i$$

und

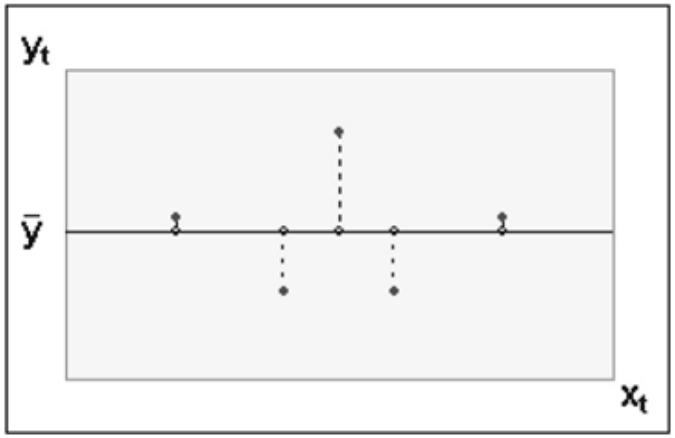
$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n} \left(\sum_{i=1}^{n} y_i \right)^2$$

Fallunterscheidung Zusammenhang

Fallunterscheidung: perfekter vs. nicht-linearer Zusammenhang



Perfekter linearer Zusammenhang



Kein linearer Zusammenhang

Grobe Orientierung für die Interpretation des **Bestimmungsgrades**:

$$0 \leq r^2 < \frac{1}{3} \Rightarrow \textbf{kein linearer Zusammenhang} \text{ zwischen X und Y}$$

$$\frac{1}{3} \leq r^2 \leq \frac{2}{3} \Rightarrow \textbf{schwach ausgeprägter Zusammenhang} \text{ zwischen X und Y}$$

$$\frac{2}{3} < r^2 \leq 1 \Rightarrow \textbf{stark ausgeprägter Zusammenhang} \text{ zwischen X und Y}$$

Quelle: Wewel, Statistik für BWL

Beispiel Regressionsanalyse (1/4)

Zusammenhang Jahresumsatz und Ladenfläche von 12 Filialen

Filiale i	Verkaufsfläche (in Tsd. qm) x_i	Jahresumsatz (in Mio. €) y_i
1	0,31	2,93
2	0,98	5,27
3	1,21	6,85
4	1,29	7,01
5	1,12	7,02
6	1,49	8,35
7	0,78	4,33
8	0,94	5,77
9	1,29	7,68
10	0,48	3,16
11	0,24	1,52
12	0,55	3,15

Errechnete Arbeitstabelle

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	0,31	2,93	0,0961	8,5849	0,9083
2	0,98	5,27	0,9604	27,7729	5,1646
3	1,21	6,85	1,4641	46,9225	8,2885
4	1,29	7,01	1,6641	49,1401	9,0429
5	1,12	7,02	1,2544	49,2804	7,8624
6	1,49	8,35	2,2201	69,7225	12,4415
7	0,78	4,33	0,6084	18,7489	3,3774
8	0,94	5,77	0,8836	33,2929	5,4238
9	1,29	7,68	1,6641	58,9824	9,9072
10	0,48	3,16	0,2304	9,9856	1,5168
11	0,24	1,52	0,0576	2,3104	0,3648
12	0,55	3,15	0,3025	9,9225	1,7325
$\sum$	10,68	63,04	11,4058	384,6660	66,0307

Quelle: Bleymüller. Statistik für Wirtschaftswissenschaftler

Beispiel Regressionsanalyse (2/4)

$$n = 12;\ \sum x_i = 10{,}68;\ \sum x_i^2 = 11{,}4058;$$

$$\sum y_i = 63{,}04;\ \sum y_i^2 = 384{,}6660;\ \sum x_i y_i = 66{,}0307.$$

Aus der Arbeitstabelle

$$a = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

Formel Regressionskoeffizient a

$$= \frac{11{,}4058 \cdot 63{,}04 - 10{,}68 \cdot 66{,}0307}{12 \cdot 11{,}4058 - 10{,}68 \cdot 10{,}68}$$

$$= \frac{13{,}813756}{22{,}8072} = 0{,}605675 \approx 0{,}6057$$

$$b = \frac{n \sum x_i y_i - \sum x_i \sum y_i}{n \sum x_i^2 - (\sum x_i)^2}$$

Formel Regressionskoeffizient b

$$= \frac{12 \cdot 66{,}0307 - 10{,}68 \cdot 63{,}04}{12 \cdot 11{,}4058 - 10{,}68 \cdot 10{,}68}$$

$$= \frac{119{,}1012}{22{,}8072} = 5{,}222088 \approx 5{,}2221.$$

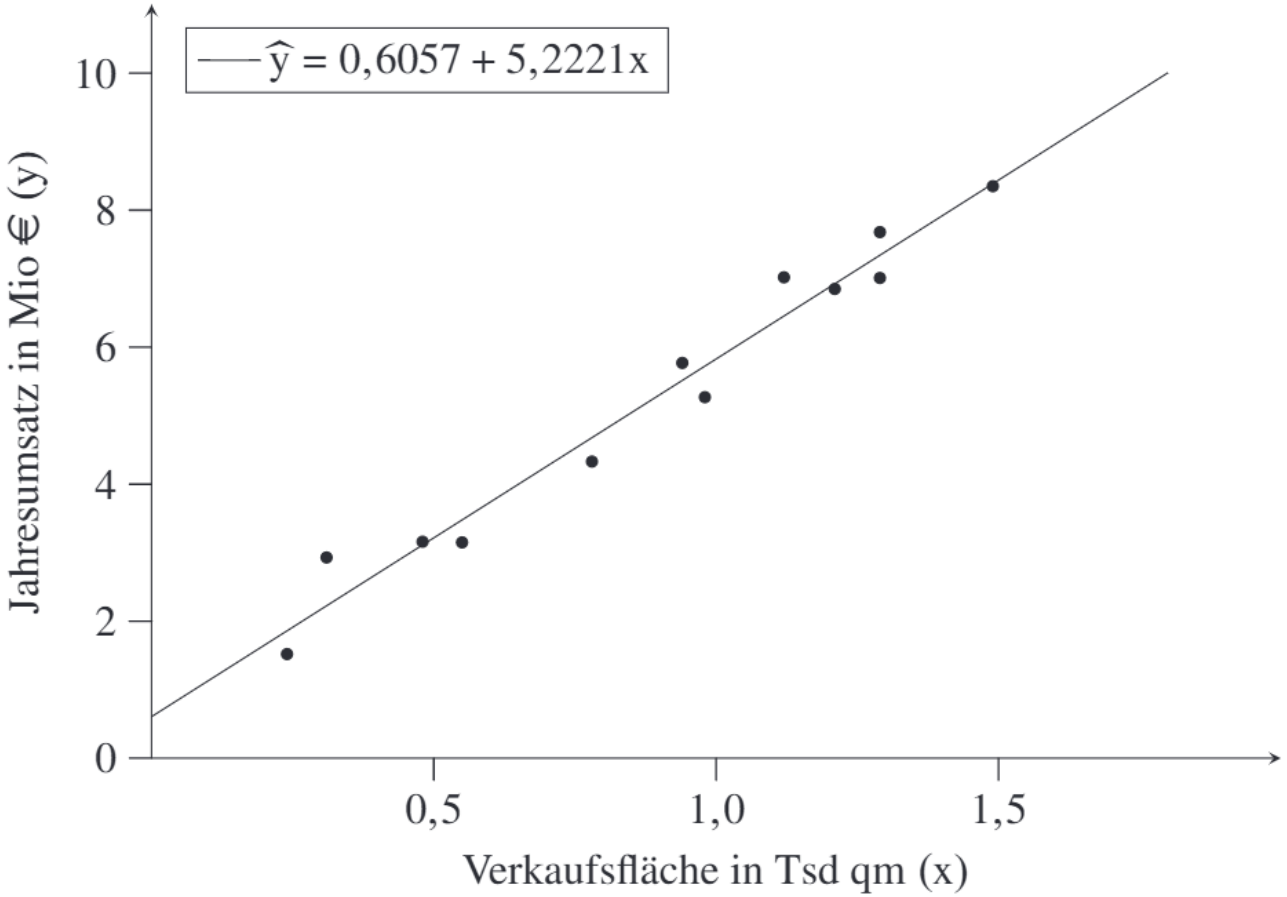
$$\hat{y} = 0{,}6057 + 5{,}2221x$$

Regressionsfunktion

Quelle: Bleymüller. Statistik für Wirtschaftswissenschaftler

Beispiel Regressionsanalyse (3/4)

Regressionsfunktion



Quelle: Bleymüller. Statistik für Wirtschaftswissenschaftler

Beispiel Regressionsanalyse (4/4)

Bestimmtheitsgrad r^2

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
$\sum$	10,68	63,04	11,4058	384,6660	66,0307

$$\hat{y} = 0{,}6057 + 5{,}2221x$$

$$\sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} y_i^2 - b_1 \sum_{i=1}^{n} y_i - b_2 \sum_{i=1}^{n} x_i y_i$$

$$\sum e_i^2 = 384{,}6660 - 0{,}605675 \cdot 63{,}04 - 5{,}222088 \cdot 66{,}0307$$
$$= 1{,}666$$

$$\sum_{i=1}^{n} (y_i - \bar{y})^2 = \sum_{i=1}^{n} y_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n} y_i\right)^2$$

$$\sum (y_i - \bar{y})^2 = 384{,}6660 - \frac{(63{,}04)^2}{12}$$
$$= 53{,}496$$

$$= 1 - \frac{\sum_{i=1}^{n} e_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

$$r^2 = 1 - \frac{1{,}666}{53{,}496} = 0{,}969$$

96% der Variation der Jahresumsätze werden durch die Regressionsfunktion erklärt.

Quelle: Bleymüller. Statistik für Wirtschaftswissenschaftler

Anwendungsbeispiele Regressionsanalyse

- **Umsatzprognose**: Beziehung zwischen Werbeausgaben und Umsatz untersuchen, um vorherzusagen, wie sich Änderungen im Marketingbudget auf den Umsatz auswirken
- **Preissetzung**: Einfluss des Preises eines Produkts auf die Verkaufszahlen bestimmen. Dies hilft Unternehmen, optimale Preisstrategien zu entwickeln.
- **Kundenzufriedenheit**: Zusammenhang zwischen verschiedenen Faktoren (z. B. Servicequalität, Wartezeit) und der Kundenzufriedenheit analysieren, um Verbesserungsmaßnahmen abzuleiten
- **Marktforschung**: Einfluss von demografischen Faktoren (z. B. Alter, Einkommen) auf das Kaufverhalten untersuchen