1 Einführung in die Statistik

Statistik ist die Kunst der Problemlösung spezieller Fragestellungen mittels vieler Daten. Mit dem Begriff „Statistik“ werden oft Tabellen und grafische Darstellungen assoziiert, die Sachverhalte aufklären sollen aber auch falsche Schlüsse suggerieren können. Dabei wird oft an folgende scherzhafte Sätze gedacht, welche zum Ausdruck bringen, mit Statistiken kritisch umzugehen.

„Trau keiner Statistik, die du nicht selber gefälscht hast.“

„Es gibt drei Arten von Lügen, die Notlüge, die gemeine Lüge und die Statistik.“

Statistik wird immer dann benötigt, wenn Daten erhoben, dargestellt und analysiert werden sollen. Dabei gibt es drei Bedeutungen des Wortes **Statistik**:

Statistik

- Wissenschaftliche Disziplin, d.h. Gesamtheit aller Methoden zur Untersuchung von Massenerscheinungen.
- Zusammenstellung von Zahlen oder Daten, d.h. Darstellung von Ergebnissen statistischer Untersuchungen in Form von Tabellen und/oder Grafiken.
- Stichprobenfunktionen, wie z.B. eine Teststatistik, die zur Überprüfung einer Hypothese verwendet wird.

Folgende Beispiele geben häufige Fragestellungen bzw. Bereiche an, in denen statistische Methoden angewandt werden.

- Zusammenhänge volkswirtschaftlicher Daten, etwa Bruttoinlandsprodukt und Arbeitslosigkeit; Prognose des Wirtschaftswachstums.
- Einfluß der Gestaltung von Bankfilialen auf die Kundenzufriedenheit; Kreditwürdigkeit von Bankkunden in Abhängigkeit vom Einkommen und persönlicher Eigenschaften (Kredit-Scoring).
- Finanzmarktanalyse (Untersuchung von Aktien-, Wechsel- und Zinskursen).
- Politische Umfragen („Welche Partei würden Sie wählen, wenn am nächsten Sonntag Bundestagswahl wäre?“).
- Klinische Studien (statistischer Nachweis der Wirksamkeit eines neuentwickelten Medikamentes).

- Vergleich zweier Düngemittel bezüglich ihres Ernteertrages.
- Qualitätskontrolle („Erfüllt das Produkt den Anforderungen des Produzenten bzw. des Konsumenten?“).

Wie die obigen Beispiele zeigen, ist die Statistik eine Wissenschaft, die mit vielen anderen Fachgebieten wie z.B. der Wirtschaftswissenschaft, der Biologie, der Medizin, der Psychologie und der Soziologie verknüpft wird. In neuentwickelten Wissenschaften, wie z.B. der Ökonometrie, der Biometrie, der Psychometrie und der Soziometrie sind spezielle Verfahren der Statistik konstruiert worden.

Die Statistik lässt sich in drei methodische Bereiche untergliedern:

1. **Beschreibende oder deskriptive Statistik**
   Die deskriptive Statistik dient der Erhebung, Aufbereitung und explorativen Datenanalyse von Datenmaterial.
2. **Wahrscheinlichkeitsrechnung**
   Die Wahrscheinlichkeitsrechnung ist die Grundlage der Inferenz-Statistik.
3. **Inferenz-Statistik (schließende oder induktive Statistik)**
   Die Inferenz-Statistik umfasst Verfahren, die den Rückschluss von einer Stichprobe auf die Grundgesamtheit betrachten.

Gegenstand dieser Kurseinheit ist die deskriptive Statistik. Die beiden anderen Bereiche werden in Kurseinheit 2 und Kurseinheit 3 behandelt.

1.1 Grundbegriffe der Datenerhebung

1.1.1 Statistische Einheit

Eine der bekanntesten statistischen Erhebungen ist der Mikrozensus, bei der jährlich ein Prozent aller in Deutschland befindlichen Haushalte über ihre wirtschaftliche und soziale Situation befragt werden. Die einzelnen bei der Befragung erfassten Haushalte sind die sogenannten statistischen Einheiten.

statistische Einheit

Die **statistische Einheit** (Element, Untersuchungseinheit, Merkmalsträger) ist der Träger der Informationen, die in einer statistischen Untersuchung von Interesse sind.

Beispiel 1.1.1:

- *Bei der Ermittlung von Aktienkursen sind die statistischen Einheiten die Aktien.*
- *In einer Untersuchung der Lebensdauer von Autoreifen ist die statistische Einheit der einzelne Autoreifen.*

Für jede statistische Untersuchung werden die statistischen Einheiten mittels Identifikationskriterien eindeutig festgelegt.

Die statistischen Einheiten müssen **sachlich**, **räumlich** und **zeitlich** voneinander abgegrenzt werden und somit eindeutig identifizierbar sein.

Identifikations-kriterien

Aus der betrachteten Fragestellung ergeben sich die speziellen Kriterien für die Abgrenzung statistischer Einheiten.

Beispiel 1.1.2:

Identifikationskriterien einer Wahlprognose der nächsten Bundestagswahl.

sachlich: *Wahlberechtigte Bürger*

räumlich: *Bundesrepublik Deutschland*

zeitlich: *Tag der Umfrage*

Die Definition statistischer Einheiten hängt immer vom Zweck der Untersuchung ab. Einer Betriebsstättenzählung können z.B. rechtliche (Verlagsgruppe), räumliche bzw. örtliche (Betriebsstätte in Hamburg, München etc.) oder technische (Druckerei, Redaktion) Einheiten zugrunde gelegt werden.

1.1.2 Merkmale

In einer statistischen Analyse sind i.d.R. nicht direkt die statistischen Einheiten, sondern spezielle Eigenschaften dieser Einheiten von Interesse.

Als **Merkmal** wird eine Eigenschaft einer statistischen Einheit bezeichnet, die in der statistischen Analyse betrachtet wird.

Merkmal

Variable

Merkmalsträger

Die Bezeichnung der Merkmale (mathematisch: **Variable**) erfolgt üblicherweise mit großen lateinischen Buchstaben: X, Y, Z, ... Die statistische Einheit wird dabei auch als **Merkmalsträger** bezeichnet.

***Beispiel 1.1.3**:*

a) *statistische Einheit:* *Angestellter*
Merkmale: *Alter, Familienstand, Wohnort*

b) *statistische Einheit:* *Autofabrik*
Merkmale: *Betriebsgröße, Anzahl der Angestellten, produzierte Stückzahlen*

Merkmals-ausprägung

Die Werte oder Kategorien, die ein Merkmal annehmen kann, werden als **Merkmalsausprägungen** bezeichnet.

Die m möglichen Ausprägungen eines Merkmals X werden in dieser Kurseinheit mit x_j $(j = 1, ..., m)$ gekennzeichnet.

***Beispiel 1.1.4**:*

statistische Einheit: Student

Merkmal	*Merkmalsausprägung*
Geschlecht	*männlich, weiblich*
Klausurnote	1 2 3 4 5 6
Körpergröße	*reelle Zahl aus dem Intervall* $[a, b]\ \forall\ a, b > 0$

Merkmalswert
Beobachtungswert

Eine an einer bestimmten statistischen Einheit festgestellte Merkmalsausprägung wird **Merkmalswert** oder **Beobachtungswert** genannt.

Die Merkmalswerte sind somit identisch mit dem Datenmaterial, welches für eine statistische Untersuchung verwendet wird.

***Beispiel 1.1.5**:*

Während einer Befragung gibt ein Bundesbürger an, er sei männlich und von Beruf Betriebswirt. Dann ist bzw. sind

- *statistische Einheit:* *befragte Person*
- *Merkmale:* *Geschlecht / Beruf*
- *Merkmalswerte:* *männlich / Betriebswirt*

1.2 Methoden der Datenerhebung

Erhebung

Für die **Erhebung** der Merkmalswerte **(Datenerhebung, statistische Erhebung)** der einzelnen statistischen Einheiten gibt es folgende Möglichkeiten:

- Befragung (persönlich, telefonisch, postalisch)
- Beobachtung
- Experiment

persönliche Befragung

Die **persönliche Befragung** durch Interviewer vor Ort hat den Vorteil, dass missverständliche und falsche Antworten durch direktes Nachfragen der Interviewer ausgeschlossen werden können. Als Nachteil erweist sich unter anderem der hohe finanzielle und organisatorische Aufwand.

telefonische Befragung

Die **telefonische Befragung** besitzt die gleichen Vorteile einer persönlichen Befragung. Zusätzliche Vorteile sind die schnelle Durchführung und die geringeren Kosten. Jedoch kann die Anzahl an Antwortverweigerungen recht hoch sein.

postalische Befragung

In einer **postalischen Befragung** mittels eines Fragebogens ist eine nachträgliche Korrektur der Antworten nicht möglich. Systematische Verzerrungen können entstehen, wenn die Rücklaufquote zu gering ist. Im Vergleich zur persönlichen Befragung sind hier z.B. die geringeren Kosten und die Vermeidung eines Interviewereinflusses bzw. Vermeidung von Interviewerfehlern als Vorteile zu nennen.

Beobachtung

Die einfachste Form der Datenerhebung ist die **Beobachtung** durch Zählung oder Messung, wobei die Problematik der Messungenauigkeit zu berücksichtigen ist. Messfehler können systematisch oder zufällig auftreten. Der zufällige oder statistische Fehler stellt meistens ein Konglomerat vieler Elementarfehler dar (z.B. Umgebungsveränderungen, Ablesefehler, usw.).

Experiment

Ein **Experiment** ist eine besondere Form der Beobachtung. Hierbei wird eine Hypothese aufgestellt, die mittels der durchgeführten Beobachtungen untermauert oder widerlegt wird. In den Wirtschaftswissenschaften ist Experimentieren meistens sehr aufwendig oder nur schwer möglich.

Beispiel 1.2.1:

Wenn ein Betrieb wissen will, wie produzierte Menge und Kosten zusammenhängen, wird er nicht experimentieren und die Produktionsmenge willkürlich verändern, da hier z.B. auf Marktgegebenheiten Rücksicht genommen werden muss.

1.2.1 Vollerhebung und Teilerhebung

Vollerhebung (Totalerhebung)

Werden in einer statistischen Erhebung **alle** Einheiten einer statistischen Masse berücksichtigt, dann handelt es sich um eine **Vollerhebung** (Totalerhebung).

Teilerhebung (Stichprobenerhebung)

Wird dagegen nur ein Teil der Einheiten erfasst, so entspricht dies einer **Teilerhebung** (Stichprobenerhebung).

Stichprobe

Grundgesamtheit

Der durch eine Teilerhebung bestimmte Teil der statistischen Masse wird **Stichprobe** genannt. Aus einer **Grundgesamtheit** (Gesamtmasse der statistischen Einheiten) mit N Elementen wird eine Stichprobe der Größe n gezogen. Die dabei vorliegenden Beobachtungen eines Merkmals X werden in dieser Kurseinheit mit x_i $(i = 1, ..., n)$ bezeichnet (nicht zu verwechseln mit den Merkmalsausprägungen x_j $(j = 1, ..., m)$).

Zu beachten ist, dass eine **Stichprobenuntersuchung** nur Aussagen über den tatsächlich betrachteten Teil der statistischen Masse zulassen kann. Rückschlüsse von einer Stichprobe auf die Grundgesamtheit können nur unter bestimmten Voraussetzungen gezogen werden. Diese Vorgehensweise ist Gegenstand der **Inferenz-Statistik**.

Vorteile einer Teilerhebung sind die geringen Kosten und die kurze Erhebungsdauer. Sinnvoll sind Teilerhebungen bei der Undurchführbarkeit einer Vollerhebung, z.B. zerstörende Prüfung. Nachteilig ist dagegen der mögliche Informationsverlust. Werden seltene Ereignisse untersucht, bietet sich eine Vollerhebung an. Liegen gesetzliche Vorschriften für die Durchführung einer Vollerhebung vor, kann die Durchführung einer Teilerhebung zu einer Nichtanerkennung der Resultate führen.

1.2.2 Primär- und Sekundärerhebung

Eine statistische Erhebung, in der das Datenmaterial mittels Befragung, Beobachtung oder durch ein Experiment gewonnen wird, heißt **Primärerhebung**.

Primärerhebung

Wird dagegen auf vorhandenes Datenmaterial zurückgegriffen, so handelt es sich um eine **Sekundärerhebung**. Von einer **registergestützten Erhebung** wird gesprochen, wenn auf bereits bestehende Datenbanken zurückgegriffen wird. Daten können vom **Statistischen Bundesamt** oder anderen Institutionen (Betriebe, Verwaltungen, Verbände,...) bezogen werden.

Sekundärerhebung

Die bekannteste Publikation des statistischen Bundesamtes ist das **Statistische Jahrbuch für die Bundesrepublik Deutschland**. Es erscheint jährlich und enthält umfangreiches Datenmaterial aus allen wirtschaftlichen und gesellschaftlichen Bereichen der Bundesrepublik.
Auf internationaler Ebene sind z. B. das **Europäische Amt für Statistik (Eurostat)**, der **Statistische Dienst der Vereinten Nationen** und die **OECD** (**O**rganisation for **E**conomic **C**o-operation and **D**evelopment) zu nennen.

Sekundärerhebungen haben den Vorteil, dass die Datenbeschaffung erheblich kostengünstiger ist als bei Primärerhebungen. Dafür bieten Primärerhebungen den Vorzug, den speziellen Bedürfnissen der jeweiligen statistischen Untersuchung genau angepasst zu sein.

1.3 Charakterisierung von Merkmalen

Voraussetzung für die Anwendung geeigneter statistischer Verfahren ist die Charakterisierung, d.h. die systematische Ordnung und Klassifizierung der zu untersuchenden Merkmale, denn die Auswahl der möglichen Analysemethoden richtet sich nach dem vorliegenden Merkmalstyp.

1.3.1 Skalierung der Merkmalsausprägungen

Skala

Merkmalswerte werden auf einer **Skala** gemessen. Die verschiedenen Skalen unterscheiden sich nach Ordnungskriterien, wobei das Skalenniveau eine wichtige Eigenschaft von Merkmalen ist. Die Skala mit dem niedrigsten Niveau ist die Nominalskala.

Nominalskala

Eine **Nominalskala** unterscheidet Merkmale nur nach Gleichheit oder Verschiedenheit. Es existiert keine Rangordnung.

Ein nominales Merkmal kann nur auf einer Nominalskala gemessen werden.

Beispiel 1.3.1:

Nominale Merkmale sind:

- *Geschlecht*
- *Nationalität*
- *Studienfach*
- *Aktienart*

qualitative Merkmale

Die Unterscheidung nominaler Merkmale ist qualitativer Art. Daher heißen nominalskalierte Merkmale auch **qualitative Merkmale**.

binäre (dichotome) Variable

Liegt ein Merkmal vor, das nur die Ausprägung „Eigenschaft vorhanden“ bzw. „Eigenschaft nicht vorhanden“ besitzt, so werden diese Ausprägungen oft mit den Zahlen 1 und 0 kodiert. In diesem Fall handelt es sich um eine **binäre (dichotome, 0-1) Variable**.

Beispiel 1.3.2:

Geschlecht:	$1 \mathrel{\widehat{=}}$ *weiblich*	$0 \mathrel{\widehat{=}}$ *männlich*
Körpergröße:	$1 \mathrel{\widehat{=}} < 170$ *cm*	$0 \mathrel{\widehat{=}} \geq 170$ *cm*
Aktienkurs:	$1 \mathrel{\widehat{=}}$ *gestiegen*	$0 \mathrel{\widehat{=}}$ *nicht gestiegen*

Eine **Ordinal-** oder **Rangskala** liegt vor, wenn die Merkmalswerte neben der qualitativen Unterschiedlichkeit eine natürliche Rangordnung besitzen.

Ordinalskala

Oft werden ordinale Merkmale auch als **intensitätsmäßige Merkmale** bezeichnet.

intensitätsmäßige Merkmale

Beispiel 1.3.3:

Ordinal messbare Merkmale sind:

- *Nach dem AAA Rating klassifizierte Merkmale (Triple A Rating: Verfahren zur Einschätzung bzw. Bewertung von Unternehmen).*
- *Klausurnoten*
- *Härtegrade von Bleistiften*
- *Tarifklassen*
- *Tabellenplätze der Bundesliga*

Ordinale Merkmale können bei Inkaufnahme eines Informationsverlustes auch auf einer Nominalskala gemessen werden. Die Umkehrung ist nicht möglich.

Ein wichtiger Aspekt ist, dass für eine Ordinalskala durch die Rangordnung keine konkrete Aussage über den absoluten Wert der Merkmalsausprägung gemacht werden kann und dass keine Informationen über die Abstände der Merkmalsausprägungen gegeben sind. Somit ist die Unterscheidung ordinaler Merkmale streng genommen qualitativer Art. Andererseits können die Ausprägungen ordinalskalierter Merkmale in eine monotone Reihenfolge gebracht werden, so dass eine schwache Quantität nicht abzustreiten ist. Für gewöhnlich werden ordinalskalierte Merkmale jedoch den qualitativen Merkmalen zugeordnet.

Beispiel 1.3.4:

- *Ein Bleistift mit Härtegrad* 2 *muss nicht doppelt so hart sein wie einer mit Härtegrad* 1*.*
- *Die Differenz zwischen den Noten* 2 *und* 3 *muss nicht genau so groß sein wie die zwischen* 3 *und* 4*.*

metrische Skala (Kardinalskala)

Besitzt ein Merkmal die Eigenschaften eines ordinalen Merkmals und ist zusätzlich noch die **Interpretation der Abstände** zweier verschiedener Merkmalsausprägungen möglich, so kann das Merkmal auf einer **metrischen Skala (Kardinalskala)** gemessen werden.

Die metrische Skala kann wie folgt unterteilt werden:

Intervallskala

Intervallskala:

Auf der Intervallskala können die Abstände zwischen den Ausprägungen verglichen werden. Es liegt kein natürlicher, sondern ein relativer Nullpunkt vor, z.B. die Temperatur, gemessen in Celsius oder Fahrenheit. Die gleiche Differenz auf dieser Skala bedeutet die gleichen Unterschiede in der Temperatur. Zu beachten ist jedoch, dass eine Temperatur von $50°C$ nicht doppelt so warm ist wie $25°C$.
Ein weiteres Beispiel ist der gregorianische Kalender. Hier wird der Nullpunkt auf das Jahr der Geburt Christi festgelegt.

Verhältnisskala

Verhältnisskala:

Zusätzlich zur Intervallskala können hier auch Verhältnisse verglichen werden. Es liegt ein natürlicher Nullpunkt vor, z.B. Temperatur in Kelvin, Größe, Gewicht, Lebensdauer.

Absolutskala

Absolutskala:

Es liegt eine Verhältnisskala vor, die nicht von den Einheiten abhängt, d.h. es existiert eine natürliche Einheit, z.B. Stückzahlen, Anzahl von Personen.

quantitative Merkmale

Metrische Merkmale unterscheiden sich durch ihre Größe und werden daher auch als **quantitative Merkmale** bezeichnet. Zu beachten ist, dass ordinal skalierte Variablen tatsächlich qualitativ sind, jedoch oft als quantitativ aufgefasst werden.

Bei Inkaufnahme eines Informationsverlustes können metrisch messbare Merkmale auch auf einer Ordinalskala oder Nominalskala gemessen werden. Auch hier ist die Umkehrung nicht möglich, d.h. ein Merkmal

kann immer auf einer Skala mit niedrigerem Niveau gemessen werden, jedoch nie auf einer mit höherem Niveau.

Des Weiteren können Merkmale zwischen stetige und diskrete Merkmale unterschieden werden.

diskrete Merkmale

Ein Merkmal heißt **diskret**, wenn es nur endlich viele oder höchstens abzählbar unendlich viele Ausprägungen besitzt (nominal skalierte Merkmale; Merkmale, deren Wert durch Zählen bestimmt wird). Nominal- und ordinalskalierte Merkmale sind stets diskret.

stetige Merkmale

Dagegen heißt ein metrisches Merkmal **stetig (kontinuierlich)**, wenn es überabzählbar viele Ausprägungen hat, d.h. wenigstens in einem bestimmten Bereich können unendlich viele Werte angenommen werden. Z.B. wird das Merkmal Einkommen als stetiges Merkmal behandelt, da es in Berechnungen mit vielen Nachkommastellen eingeht und somit überabzählbar viele Ausprägungen vorliegen können. Die tatsächliche Angabe erfolgt dagegen meistens nur mit zwei Nachkommastellen. Ebenso kann das Merkmal Wohnfläche auf viele Nachkommastellen gemessen werden, wobei oft nur ganze m^2 angegeben werden.

***Beispiel 1.3.5**:*

- *Diskrete Merkmale:* *Einwohnerzahl, Steuerklasse, Lottozahlen, Geschlecht*
- *Stetige Merkmale:* *Lebensdauer, Größe, Gewicht, Temperatur*

1.3.2 Skalentransformation

Oft ist es für eine statistische Analyse notwendig, z.B. in einem Vergleich verschiedener Längenmaße, die vorliegenden Daten geeignet zu transformieren.

Skalen-transformation

Eine **Skalentransformation** kann als Abbildung von einer Menge Merkmalsausprägungen in eine anderere Menge Merkmalsäusprägungen angesehen werden. Dabei ist zu beachten, dass die Ordnungseigenschaften der Skala erhalten bleiben.

***Beispiel 1.3.6**:*

a) Die Ausprägungen (Stammaktie, Vorzugsaktie) des nominalen Merkmals „Art der Aktie nach Stimmrecht“ können in die Werte S und V transformiert werden.

b) Kodierung der Ausprägungen des Merkmals „Beschäftigungsverhältnis“ durch Zahlen, z.B. Selbständig: 1*, Angestellter:* 2*.*

Je nach Skalenniveau sind verschiedene Transformationen zulässig, d.h. durch die Transformation darf keine Information verloren gehen.

eineindeutige

Bei einer **eineindeutigen Skalentransformation**, wird jedem Wert der alten Skala genau ein Wert der neuen Skala (und umgekehrt) zugeordnet.

monotone

Eine **monotone Skalentransformation** liegt vor, wenn die Rangordnung der Skalenwerte erhalten bleibt.

lineare Skalen-transformation

Lineare Skalentransformationen nutzen lineare Funktionen der Form $y = a+bx$, wobei das Verhältnis der Abstände zwischen den Skalenwerten erhalten bleibt.

***Beispiel 1.3.7**:*

a) Die Noten „sehr gut“, „gut“, „befriedigend“, „ausreichend“ und „mangelhaft“ werden mit 1*,* 2*,* 3*,* 4 *und* 5 *dargestellt (monotone Transformation).*

b) Die Umrechnung von Dollar nach Euro in der Form $y = bx$ (lineare Transformation mit b als Devisenkurs).

Folgende Tabelle gibt die Transformierbarkeit der verschiedenen Skalen an.

Skala		**sinnvolle Operationen**	**Transformier-barkeit**
Nominalskala		$= \mid \neq$	eineindeutig
Ordinalskala		$= \mid \neq; < \mid >$	streng monoton
metrische Skala	Intervall-Skala	$= \mid \neq; < \mid >; + \mid -$	linear $y_i = ax_i + b$
	Verhältnis-Skala	$= \mid \neq; < \mid >; + \mid -; \cdot \mid :$	linear $y_i = ax_i$
	Absolut-Skala	$= \mid \neq; < \mid >; + \mid -; \cdot \mid :$ (es liegt eine natürliche Einheit vor)	identisch $y_i = x_i$

Tabelle 1.3.1: *Übersicht über Skalentransformationen*

Tabelle 1.3.1 kann wie folgt interpretiert werden:

- Jede zulässige Skalentransformation muss eineindeutig sein.
- Jede zulässige Transformation einer Ordinalskala muss wenigstens streng monoton sein.
- Jede zulässige Transformation einer metrischen Skala muss linear sein.

Beispiel 1.3.8*:*

Untersucht wird die Intelligenz von zwei Probanden vor und nach Absolvierung einer bestimmten Trainingsmethode. Bei der Messung der Intelligenz liegt kein natürlicher Nullpunkt vor, d.h. Überlegungen, die sich auf Verhältnisse beziehen sind nicht zulässig. Ein Proband mit einem IQ von 100 *ist nicht doppelt so intelligent wie ein Proband mit einem IQ von* 50*. Sei*

$$
\begin{aligned}
y_{1v} &= ax_{1v} + b \quad \textit{Intelligenz von Proband 1 vor dem Training,} \\
y_{1n} &= ax_{1n} + b \quad \textit{Intelligenz von Proband 1 nach dem Training,} \\
y_{2v} &= ax_{2v} + b \quad \textit{Intelligenz von Proband 2 vor dem Training,} \\
y_{2n} &= ax_{2n} + b \quad \textit{Intelligenz von Proband 2 nach dem Training.}
\end{aligned}
$$

Hier sind Vergleiche mittels Quotientenbildung erst zulässig, wenn durch Differenzenbildung die Konstante b eliminiert wird.

$$y_{1n} - y_{1v} = a(x_{1n} - x_{1v}) \quad \textit{Intelligenzsteigerung von Proband 1}$$
$$y_{2n} - y_{2v} = a(x_{2n} - x_{2v}) \quad \textit{Intelligenzsteigerung von Proband 2}$$

Nun liegt ein natürlicher Nullpunkt vor und es kann die Relation, die Intelligenzsteigerung bei Proband 2 *ist doppelt so hoch wie bei Proband* 1*, betrachtet werden. Liegt eine Verhältnisskala vor, sind Vergleiche mittels Quotienten direkt, d.h. ohne Differenzenbildung, zulässig.*

1.3.3 Klassierung von Merkmalsausprägungen

Liegen zu viele verschiedene Merkmalsausprägungen vor, so dass das Datenmaterial unübersichtlich erscheint, z. B. bei stetigen Merkmalen, kann es sinnvoll sein, mehrere Ausprägungen zusammenzufassen.

Klasse

Die **Klassierung** von Merkmalsausprägungen stellt eine Zusammenfassung benachbarter Merkmalsausprägungen zu einer **Klasse** dar, wobei die vorgegebene Ordnung erhalten bleibt. Dabei sollten disjunkte Klassen mit möglichst gleicher Breite (Ausnahme: Randklassen) ausgewählt werden.

Ein gewisser Informationsverlust muss bei Klassenbildung in Kauf genommen werden, da später die exakten Ausprägungen oft nicht mehr festgestellt werden können. Es wird die Annahme getroffen, dass sich die Merkmalswerte gleichmäßig über die Klassen verteilen.

Für diskrete Merkmale gilt als Faustregel, dass bei Vorliegen von n Merkmalswerten die Anzahl der Klassen $\sqrt{n}$ betragen soll. Die Gesamtzahl der Klassen sollte aufgrund der Übersichtlichkeit die Zahl 20 nicht überschreiten. Eine Klasse wird mittels der **Klassengrenzen**, der **Klassenbreite** und der **Klassenmitte** eindeutig festgelegt.

untere und obere Klassengrenze

Die **untere Klassengrenze** der j-ten Klasse wird mit $\boldsymbol{x^*_{j-1}}$ und die **obere Klassengrenze** mit $\boldsymbol{x^*_j}$ bezeichnet. Die Gesamtzahl der Klassen wird mit m angegeben.

In der Praxis werden die Klassen oft so gebildet, dass die untere Grenze in die Klasse fällt. Für theoretische Überlegungen (siehe Summenhäufigkeitsverteilungen bei Klassenbildung) ist es sinnvoller, dass die obere Grenze in die Klasse fällt, d.h. betrachtet werden die Intervalle $(x^*_{j-1}, x^*_j]$. Deshalb wird hier die Regel „über... bis“ verwendet.

Klassenbreite

Die **Klassenbreite** $\boldsymbol{\Delta x_j = x^*_j - x^*_{j-1}}$ der j-ten Klasse ist als Differenz zweier aufeinanderfolgender Klassengrenzen definiert.

Wie bereits erwähnt, sollten die Klassenbreiten möglichst gleich sein. Dies ist jedoch nicht möglich, wenn viele Merkmale in einem engen Bereich streuen und der Rest in einem weiten Bereich. In diesem Fall wird im engstreuenden Bereich feiner klassifiziert.

Klassenmitte

Die **Klassenmitte x_j** wird mit $x_j = \frac{x^*_{j-1} + x^*_j}{2}$ angegeben.

Da i.d.R. die Klassenmitte den repräsentativen Wert der einzelnen Klassen darstellt, wird diese mit dem gleichen Symbol bezeichnet wie die Merkmalsausprägung bei nicht klassierten Merkmalen. Bei Altersklassen dagegen entspricht der repräsentative Wert x_j oft der oberen Grenze x^*_j. Z.B. wird die Altersklasse der 30-jährigen mit $(29; 30]$ angegeben. Schwierigkeiten bei der Analyse klassierter Merkmalsausprägungen treten auf, wenn im unteren und/oder im oberen Bereich nur sehr wenige Beobachtungen liegen.

offene Randklasse

Als **offene Randklasse** wird die erste oder letzte der geordneten Klassen bezeichnet, wenn keine untere bzw. obere Klassengrenze vorhanden ist.

Eine Klassenmitte kann für offene Randklassen nicht ohne Probleme angegeben werden. In diesem Fall kann wie folgt vorgegangen werden:

- Angabe des nächsten Wertes, der sich bei gleichen Abständen aller Klassenmitten ergibt.
- Angabe eines Schätzwertes oder mutmaßlichen Wertes.
- Explizite Berechnung eines Mittelwertes aus den ursprünglichen Merkmalswerten.

***Beispiel 1.3.9**:*

Zugrundegelegt werden die Umsatzzahlen einzelner Betriebe einer bestimmten Branche.

Klasse (in Mio. €)	(0;10]	(10;20]	(20;30]	> 30
Klassenmitte (€)	5	15	25	?

Die Klassenmitte der oberen offenen Randklasse kann mit 35 *angegeben werden oder es wird beispielsweise* 40 *„vermutet“. Bei Vorliegen der Ursprungsdaten, kann die Klassenmitte der offenen Randklasse exakt aus diesen Werten berechnet werden.*

1.4 Häufigkeitsverteilungen

statistische Reihe

Die zur Analyse erhobenen Daten x_1,...,x_n liegen in Form einer ungeordneten oder geordneten **statistischen Reihe** vor. Die Notation einer geordneten Reihe lautet $x_{(1)}, x_{(2)}, ..., x_{(n)}$.

Beispiel 1.4.1:

10 *Kreditnehmer wurden nach der aktuellen Laufzeit ihres Kredites befragt.*

ungeordnete Reihe:	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
	9	2	15	6	1	7	12	10	5	18
geordnete Reihe:	$x_{(1)}$	$x_{(2)}$	$x_{(3)}$	$x_{(4)}$	$x_{(5)}$	$x_{(6)}$	$x_{(7)}$	$x_{(8)}$	$x_{(9)}$	$x_{(10)}$
	1	2	5	6	7	9	10	12	15	18

Werden viele Beobachtungen erhoben, so liegen die einzelnen Merkmalsausprägungen mehrmals vor. Für nominalskalierte Merkmale ist die Reihenfolge aufgrund des Nichtvorliegens einer natürlichen Rangordnung willkürlich.

absolute Häufigkeit

Die **absolute Häufigkeit h_j** der Merkmalsausprägung x_j mit $h_j = h(x_j)$ ist die Anzahl der Beobachtungswerte, die diese Ausprägung aufweisen.

Wird die absolute Häufigkeit in Relation zu der Gesamtanzahl der Beobachtungswerte n gesetzt, so ergibt sich die sogenannte relative Häufigkeit.

relative Häufigkeit

Die **relative Häufigkeit f_j** mit $f_j = f(x_j) = \frac{h_j}{n}$ ist der relative (prozentuale) Anteil der Häufigkeit einer Merkmalsausprägung x_j an der Gesamtanzahl der Beobachtungswerte.

Die relativen Häufigkeiten können auch als **Prozentzahlen** angegeben werden.

Beispiel 1.4.2:

Von 60 *befragten Personen fahren* 15 *mit dem öffentlichen Nahverkehr zur Arbeit.*

Merkmalsausprägung x_j:	*Person nutzt öffentlichen Nahverkehr*
absolute Häufigkeit:	$h_j = 15$
relative Häufigkeit:	$f_j = \frac{15}{60} = 0.25$ *oder* 25%.

Die **Häufigkeitsverteilung** eines Merkmals ist eine Zuordnung, die zu jeder vorhandenen Merkmalsausprägung oder Merkmalsklasse angibt, wie häufig diese vorkommt (**absolute** oder **relative Häufigkeitsverteilung**).

Häufigkeitsverteilung

Beispiel 1.4.3:

20 *Personen wurden nach dem Schulabschluss befragt (H = Hauptschule, FO = Fachoberschule, A = Abitur, FH = Fachhochschule, U = Universität):*

A A FO H H FH U U FH H FH U FO A U FH H U FO H

Als Verteilung ergibt sich

Schulabschluss x_j	H	FO	A	FH	U	$\sum$
absolute Häufigkeit h_j	5	3	3	4	5	20
relative Häufigkeit f_j	25%	15%	15%	20%	25%	100%

Liegt eine statistische Reihe vor, deren Beobachtungen aus nur einem Merkmal bestehen, so wird eine **eindimensionale (univariate) Häufigkeitsverteilung** aufgestellt.

eindimensionale Häufigkeitsvtlg.

Eine **mehrdimensionale (multivariate) Häufigkeitsverteilung** ergibt sich, wenn mehrere Merkmale gleichzeitig betrachtet werden.

mehrdimensionale Häufigkeitsvtlg.

Beispiel 1.4.4:

Bei den 20 *Personen aus dem vorherigen Beispiel wurde zusätzlich auch das Merkmal „Geschlecht“ erhoben. Für die gemeinsam betrachteten Merkmale („Geschlecht“, „Schulabschluss“) lässt sich das Ergebnis in einer zweidimensionalen (bivariaten) Häufigkeitstabelle darstellen.*

(m,A) (w,A) (w,FO) (m,H) (m,H) (m,FH) (w,U) (w,U)
(m,FH) (w,H) (m,FH) (m,U) (w,FO) (w,A) (m,U) (w,H)
(m,FO) (w,U) (m,FO) (w,H)

	H	FO	A	FH	U	$\sum$
m	3	1	1	3	2	10
w	2	2	2	1	3	10
$\sum$	5	3	3	4	5	20

1.5 Grafische Darstellung von Daten

Neben der Zahlenzusammenstellung in Form von Tabellen, wobei hier auf Übersichtlichkeit, leichte Lesbarkeit und unmissverständliche Bezeichnungen zu achten ist, können statistische Daten mittels verschiedener Diagrammformen veranschaulicht werden.[1]
Grafische Darstellungen sollten jedoch keine Alternative zur Tabelle sein, sondern als Ergänzung bzw. als optisches Hilfsmittel dienen.

Stabdiagramm
Säulendiagramm
Balkendiagramm

Ein **Stab-** bzw. **Säulendiagramm** veranschaulicht bei Vorliegen einer horizontalen Achse eine höhenproportionale Darstellung der Häufigkeiten mittels Stäben bzw. Säulen.
Balkendiagramme besitzen eine vertikale Achse mit waagerecht aufgetragenen Balken (längenproportionale Darstellung).

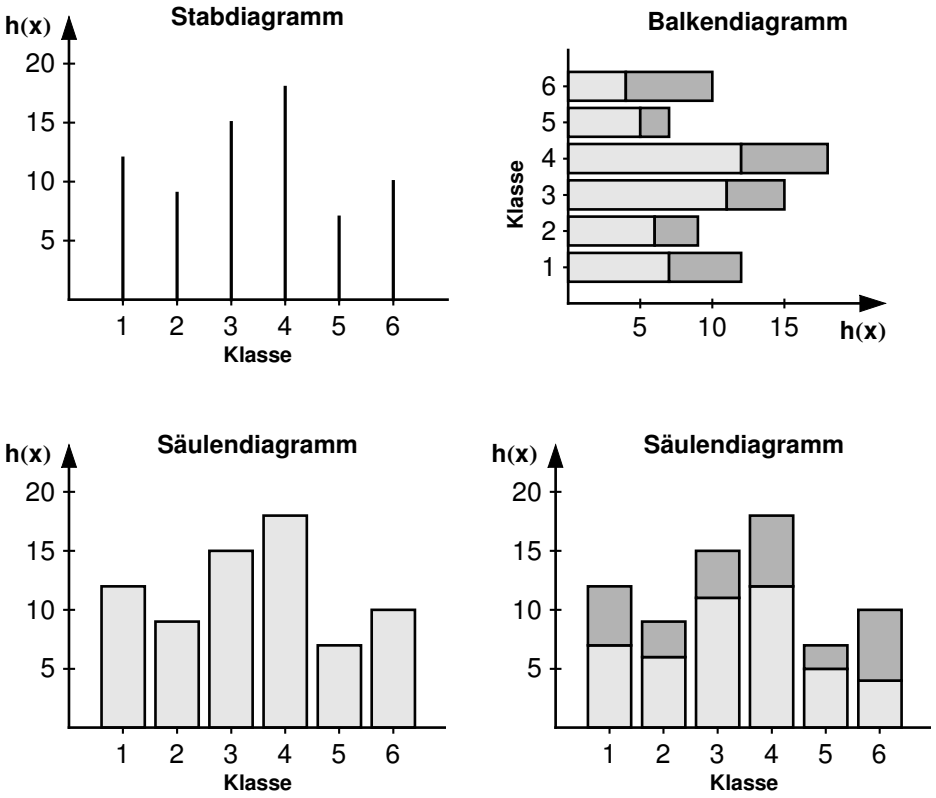


Abbildung 1.5.1: *Stab-, Säulen- und Balkendiagramm (rechts: gruppiertes Merkmal)*

Die Balken bzw. Säulen können gegebenenfalls nach einem weiteren Merkmal unterteilt werden. In diesem Fall liegen gruppierte Merkmale

[1] Zur Zusammenstellung von Daten mittels Tabellen siehe DIN 55301 „Gestaltung statistischer Tabellen“.

vor. Üblicherweise werden die Stäbe bzw. Säulen mit einem Zwischenraum aufgetragen. Stab- bzw. Säulendiagramme sind vor allem für nominal- und ordinalskalierbare Merkmale geeignet. In der Literatur wird oft nicht zwischen Säulen- und Balkendiagramm unterschieden, sondern beide werden einheitlich als Balkendiagramm bezeichnet.

Kreisdiagramm

Ein **Kreisdiagramm** ist eine grafische Darstellung von Häufigkeiten durch sektorale Aufteilung der Kreisfläche. Die Flächen der Sektoren bzw. die zugehörigen Winkel stehen dabei im gleichen Verhältnis zueinander wie die entsprechenden Häufigkeiten. Die Gesamthäufigkeit kann in einem Kreisdiagramm durch die gesamte Fläche des Kreises veranschaulicht werden.

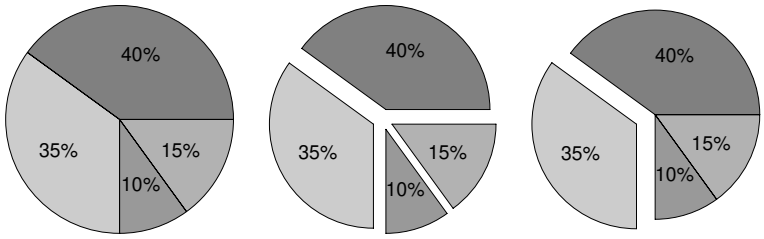


Abbildung 1.5.2: *Verschiedene Varianten von Kreisdiagrammen für relative Häufigkeiten*

Liniendiagramm/ Kurvendiagramm

Ein **Liniendiagramm/Kurvendiagramm** ist eine grafische Darstellung von Messzahlen oder Indexzahlen in einem Koordinatensystem durch Kurven bzw. geradlinig verbundene Punkte.

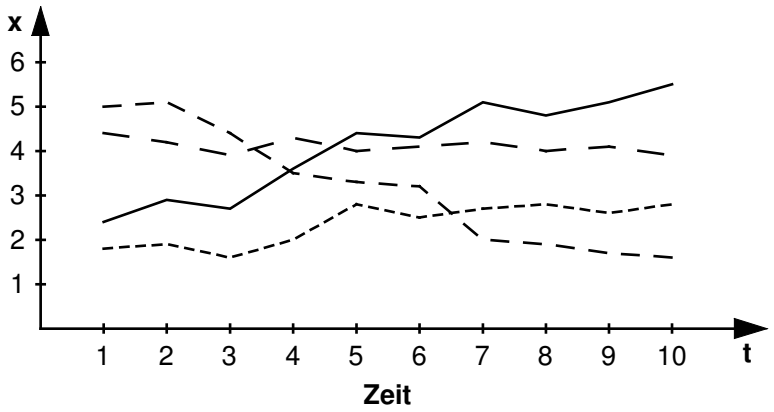


Abbildung 1.5.3: *Liniendiagramm*

Liniendiagramme werden häufig für die Illustration von Zeitreihen verwendet.

Histogramm

Ein **Histogramm** ist eine grafische Darstellung der Häufigkeiten eines klassierten, quantitativen Merkmals durch rechteckige Flächen über den Klassen in einem Koordinatensystem. Es ist zu beachten, dass die einzelnen Rechtecke des Histogramms unmittelbar aneinander schließen und nicht wie beim Säulendiagramm getrennt sind.

Bei einem Histogramm handelt es sich um eine **flächenproportionale** Darstellung der Häufigkeiten. Dies ist besonders dann zu beachten, wenn die Klassen nicht gleich breit sind. Nur bei einem Histogramm mit gleichbreiten Klassen ist die Rechteckhöhe proportional zu den beobachteten Häufigkeiten. Liegen gleichbreite Klassen mit einer Breite von 1 vor, so entspricht die Rechteckhöhe den beobachteten Häufigkeiten.

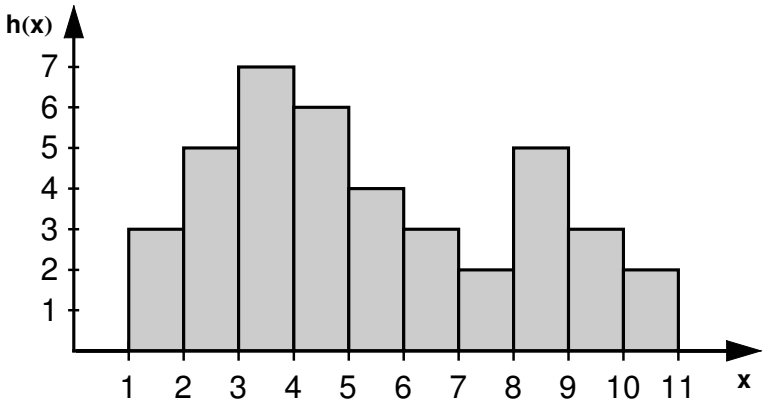


Abbildung 1.5.4: *Beispiel eines Histogramms mit gleichbreiten Klassen (Breite=1)*

Flächendiagramm

Allgemein wird ein Diagramm als **Flächendiagramm** bezeichnet, wenn die Häufigkeiten durch Flächen (sog. flächenproportionale Darstellung) dargestellt werden. Ein Flächendiagramm ist nicht zwingendermaßen ein Histogramm, welches ausschließlich Häufigkeiten für klassierte Merkmale darstellt. Weitere Flächendiagramme sind z.B. Kreisdiagramme.

Grafische Darstellungen können optische Täuschungen erzeugen und beim Betrachter falsche Assoziationen auslösen. Deshalb ist die Skalierung von Grafiken immer mit Bedacht zu wählen.

2 Eindimensionale Häufigkeitsverteilungen

2.1 Darstellung von Häufigkeitsverteilungen

2.1.1 Verteilung absoluter und relativer Häufigkeiten

Häufigkeitsverteilungen (s. Abschnitt 1.4), können in Form einer **Häufigkeitstabelle** oder Grafik angegeben werden. Dabei ist zu beachten, dass nicht alle möglichen, sondern nur die tatsächlich beobachteten Merkmalsausprägungen bzw. Merkmalsklassen berücksichtigt werden.
Die praktische Ermittlung einer Häufigkeitstabelle von Hand erfolgt oft mittels einer Strichliste.

Häufigkeitstabelle

Beispiel 2.1.1:

20 *Personen wurden nach ihrer Steuerklasse befragt.*

I V III I III V II I I III III II IV I IV I I I V III

Merkmalsausprägung x_j	Strichliste	absolute Häufigkeit h_j	relative Häufigkeit f_j
I	𝍸 |	8	0.40
II	||	2	0.10
III	𝍸	5	0.25
IV	||	2	0.10
V	|||	3	0.15
$\sum$	𝍸 𝍸 𝍸 𝍸	20	1.00

Bei der Darstellung einer Häufigkeitsverteilung in Form einer **Grafik** muss das vorliegende Skalenniveau berücksichtigt werden.

Wird eine **Nominalskala** vorausgesetzt, bei der keine natürliche oder vorgegebene Ordnung vorliegt, bieten sich Flächendiagramme in Form eines Kreisdiagramms an. Hierbei lässt sich am besten illustrieren, wie die Gesamtzahl auf die einzelnen Ausprägungen aufgeteilt ist. Auch Säulen-, Stab- und Balkendiagramme sind geeignete Darstellungsmöglichkeiten.

Nominalskala

Liegt eine **Ordinalskala** vor, wird in der Regel auf Säulen-, Stab- und Balkendiagramme zurückgegriffen, wobei auch hier Flächendiagramme sinnvoll verwendet werden können.

Ordinalskala

Beispiel 2.1.2:

Untersucht wurden 30 *Hotels einer Großstadt nach ihrer Güteklasse.*

Hotelgüteklasse		A	B	C	D	$\sum$
Häufigkeit	absolut	9	12	6	3	30
	relativ	30%	40%	20%	10%	100%

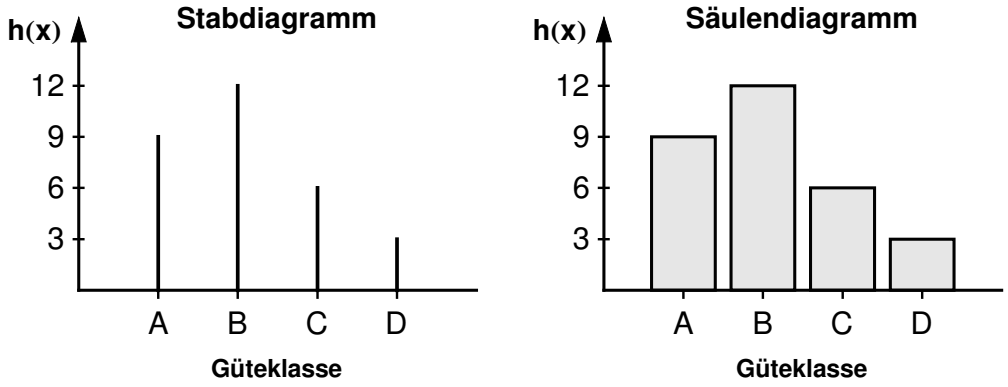


Abbildung 2.1.1: *Stab- und Säulendiagramm*

metrische Skala

Metrische Merkmale werden zwischen diskret und stetig unterschieden. Ein diskretes Merkmal kann wie ein ordinales Merkmal dargestellt werden. Häufigkeitsverteilungen stetiger Merkmale werden mittels eines **Histogramms** veranschaulicht, wobei zunächst eine Klassenbildung erfolgen muss.

Angabe der Rechteckhöhe

Liegen **ungleiche Klassenbreiten** vor, so gibt die *Höhe der Rechtecke* eines Histogramms *keine Auskunft* über die Klassenhäufigkeiten, welche **proportional zur Fläche** der Rechtecke sind.
Die Angabe der **Rechteckhöhe rh_j** der Klasse j hängt davon ab, welche Häufigkeit zugrundegelegt wird (relative oder absolute Häufigkeit). Für absolute Häufigkeiten folgt $rh_j = \frac{h_j}{b_j}$, während für relative Häufigkeiten $rh_j = \frac{f_j}{b_j}$ gilt, wobei h_j die Anzahl der Beobachtungen in der Klasse j ist, f_j der relativen Häufigkeit entspricht und b_j die Klassenbreite bezeichnet.

Wird dies nicht beachtet besteht die Gefahr, dass ein Histogramm falsch interpretiert wird. Dieses Problem entsteht nicht, wenn alle Klassen gleich breit sind.

Die **Höhe der Rechtecke** eines Histogramms ist bei **gleichen Klassenbreiten** proportional zu den Klassenhäufigkeiten.

Beispiel 2.1.3:

Es wurden 20 *nicht selbständige Personen nach ihren Schulden im letzten Jahr befragt (ohne Berücksichtigung von Immobilienschulden). Zugrundegelegt wurde nachstehende geordnete Reihe (Schulden in Tausend €):*

1 1 2 2 2 4 4 5 6 6 7 9 12 15 17 18 20 22 26 29

Zum Vergleich werden zwei Tabellen mit den dazugehörigen Histogrammen erstellt; einmal mit gleicher Klassenbreite, einmal mit ungleicher Klassenbreite ($rh_j = \frac{f_j}{b_j}$).

Schulden	h_j	f_j	b_j	rh_j
(0;5]	8	0.4	5	0.08
(5;10]	4	0.2	5	0.04
(10;15]	2	0.1	5	0.02
(15;20]	3	0.15	5	0.03
(20;25]	1	0.05	5	0.01
(25;30]	2	0.1	5	0.02

Schulden	h_j	f_j	b_j	rh_j
(0;5]	8	0.4	5	0.08
(5;15]	6	0.3	10	0.03
(15;30]	6	0.3	15	0.02

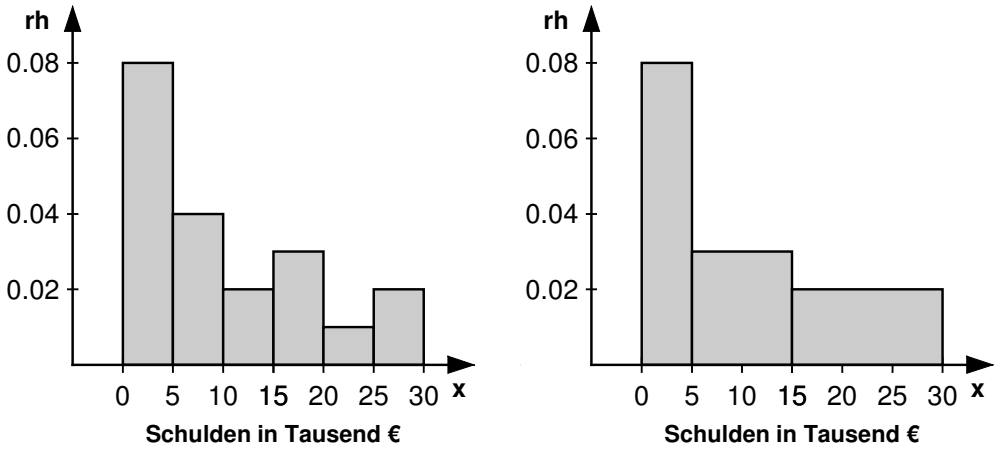


Abbildung 2.1.2: *Histogramme mit unterschiedlicher Klassenbildung*

Im linken Histogramm ist zu sehen, dass die Rechteckhöhe proportional zur relativen Häufigkeit ist. Dies ist im rechten Histogramm nicht der Fall.

Ein Nachteil von Histogrammen bei der Schätzung einer Häufigkeitsverteilung ist darin zu sehen, dass diese von der Wahl der Klassenbreite und des Anfangspunktes abhängt. Bei einer geringen Klassenbreite können einzelne Beobachtungen einen erheblichen Einfluss auf die Glättung der geschätzten Häufigkeitsverteilung haben.
Ein weiterer Nachteil bei der Schätzung einer stetigen Häufigkeitsverteilung durch Histogramme ist, dass hier eine nicht stetige Schätzung vorliegt.

Kerndichteschätzer

Eine stetige Schätzung der unbekannten Häufigkeitsverteilung kann mittels **Kerndichteschätzer** durchgeführt werden, wobei die Wahl der dort verwendeten Bandbreite (Glättungsparameter) die Qualität der Schätzung bestimmt.[2] Die Kerndichteschätzung verwendet eine beliebige von einer Bandbreite abhängige Kernfunktion, die über jede Beobachtung gelegt und anschließend gemittelt wird. Die Kernfunktionen müssen den Eigenschaften einer Dichtefunktion entsprechen und sind i.d.R. um Null symmetrisch und unimodal (z.B. Gauß-Kern, Epanechnikov-Kern, Cauchy-Kern, Dreieckskern und Rechteckskern). Dabei ist zu beachten, dass es sich bei dem Rechteckskern um eine unstetige Funktion handelt, was sich bei der Schätzung durch eine rauere Struktur äußert (vgl.[4]).

2.1.2 Summenhäufigkeiten

Ist es von Interesse festzustellen, wie viele Merkmalswerte insgesamt unterhalb oder oberhalb einer bestimmten Merkmalsausprägung liegen, wird auf die **Summenhäufigkeitsverteilung** zurückgegriffen.

Beispiel 2.1.4:

Interessierende Fragestellung: Wie viele Einwohner der Bundesrepublik haben ein monatliches Einkommen von z.B. höchstens 3000 €*?*

[2] siehe interaktive Mathematica-Applets auf der Homepage des Lehrstuhls http://www.fernuni-hagen.de/ls_statistik/forschung/multimedia/

In solch einem Fall werden die **kumulierten absoluten oder relativen Häufigkeiten** bestimmt. Dabei werden für jede Merkmalsausprägung alle Häufigkeiten der Merkmalsausprägungen addiert, die diese Ausprägung oder einen kleineren Wert annehmen.

Summenhäufigkeit

Die **Summenhäufigkeit** einer Merkmalsausprägung oder einer oberen Klassengrenze eines wenigstens ordinal messbaren Merkmals ist die zugeordnete Häufigkeit aller Beobachtungswerte, die diese Merkmalsausprägung bzw. diese Klassengrenze nicht überschreiten.

Bei der Berechnung der Summenhäufigkeit müssen die Beobachtungswerte **aufsteigend geordnet** vorliegen.

absolute Summenhäufigkeit

Für die **absolute Summenhäufigkeit $H(x)$** gilt:

$$H_j = H(x_j) = \sum_{x_{j'} \leq x_j} h_{j'} = \sum_{j'=1}^{j} h_{j'}$$

relative Summenhäufigkeit

Für die **relative Summenhäufigkeit $F(x)$** gilt:

$$F_j = F(x_j) = \frac{H_j}{n} = \sum_{x_{j'} \leq x_j} f_{j'} = \sum_{j'=1}^{j} f_{j'}$$

Anhand der Definition von H_j bzw. F_j ist zu erkennen, dass bei Klassenbildung die Klassen nach der Regel „über ... bis einschließlich“ gebildet werden müssen.

Summenhäufigkeitsverteilung

Als **Summenhäufigkeitsverteilung** wird die tabellarische oder auch grafische Darstellung der geordneten Merkmalsausprägungen bzw. Merkmalsklassen und der zugehörigen Summenhäufigkeiten bezeichnet.

Liegen diskrete metrische Merkmale oder ordinale Merkmale vor, ergibt die grafische Darstellung der Summenhäufigkeitsverteilung eine Treppenfunktion.

***Beispiel 2.1.5**:*

50 *Studenten nahmen an einer Prüfung teil, bei der maximal* 10 *Punkte erreicht werden konnten.*

Erreichte Punktzahl x_j	absolute Häufigkeit h_j	absolute Summenhäufigkeit H_j	relative Häufigkeit f_j in %	relative Summenhäufigkeit F_j in %
0	1	1	2	2
1	3	4	6	8
2	4	8	8	16
3	2	10	4	20
4	5	15	10	30
5	6	21	12	42
6	8	29	16	58
7	10	39	20	78
8	4	43	8	86
9	5	48	10	96
10	2	50	4	100

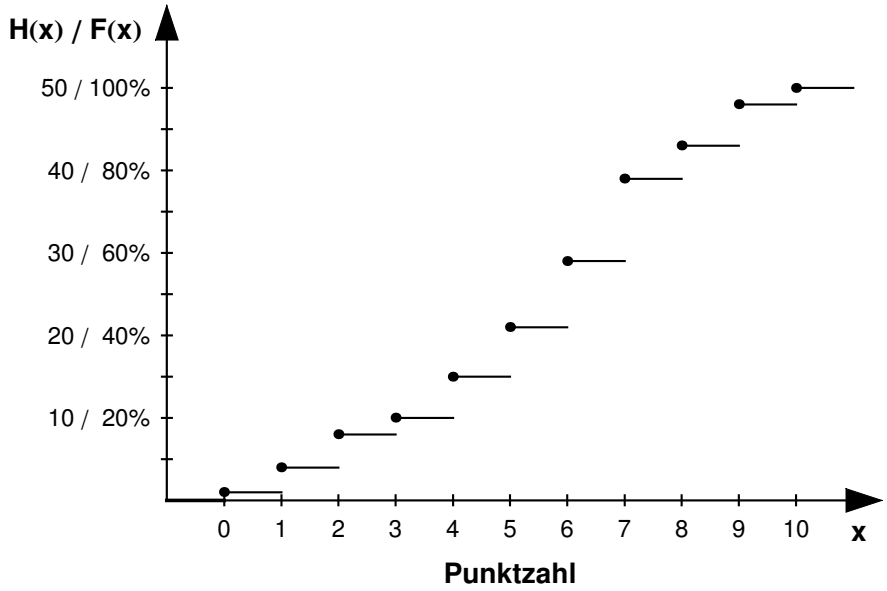


***Abbildung 2.1.3:** Diskrete Summenhäufigkeitsverteilung zu Beispiel 2.1.5*

Bei Klassenbildung kann die Summenhäufigkeitsverteilung nur für die obere Klassengrenze exakt angeben werden. Die Summenhäufigkeiten geben an, wie viele Werte (bzw. wie viel % der Werte) unterhalb der jeweiligen oberen Klassengrenze liegen. Die grafische

Darstellung der Summenhäufigkeitsverteilung eines stetigen Merkmals ergibt eine stückweise lineare Funktion. Dabei wird unterstellt, dass die Werte innerhalb einer Klasse gleichmäßig verteilt (gleichverteilt) sind.

***Beispiel 2.1.6**:*

Eine Untersuchung über Mietpreise bei 200 *ausgewählten Wohnungen hat folgende Daten ergeben ($rh_j = \frac{f_j}{b_j}$).*

Klasse	Mietpreis in €/m^2	Klassenmitte in €	h_j	f_j	b_j	rh_j
I	(3.50;5.50]	4.50	40	0.2	2	0.1
II	(5.50;6.50]	6.00	40	0.2	1	0.2
III	(6.50;7.00]	6.75	60	0.3	0.5	0.6
IV	(7.00;7.50]	7.25	40	0.2	0.5	0.4
V	(7.50;9.50]	8.50	20	0.1	2	0.05

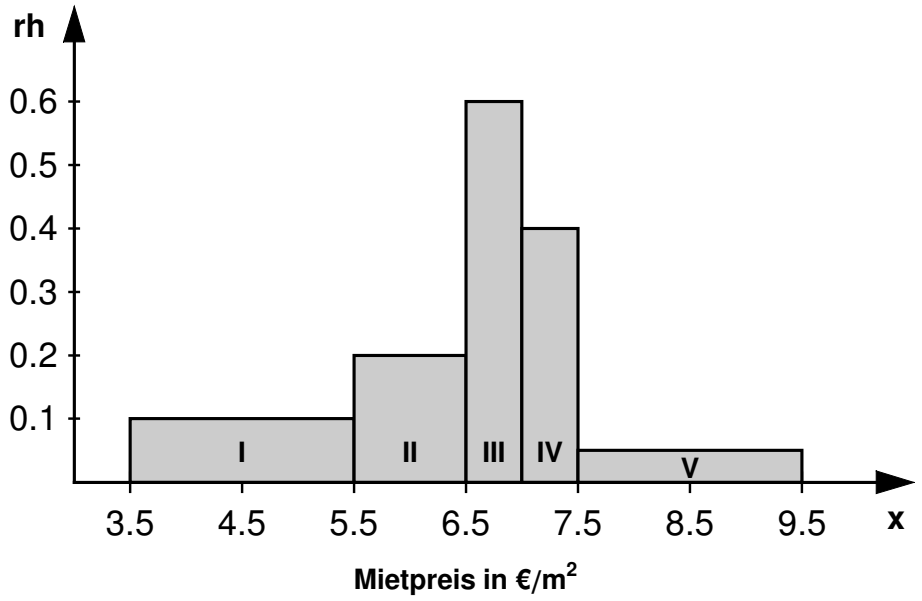


***Abbildung 2.1.4:** Histogramm zu Beispiel 2.1.6*

Die vorhergehende Tabelle gibt die einfachen (absoluten und relativen) Häufigkeiten wieder, während aus der folgenden Tabelle die Summenhäufigkeiten hervorgehen. Z.B. wird bei 140 *Wohnungen oder* 70% *aller Wohnungen ein Mietpreis von* 7 € *pro* m^2 *nicht überschritten.*

Mietpreis €/m^2	Summenhäufigkeit H_j	Summenhäufigkeit F_j in %
≤ 5.5	40	20
≤ 6.5	80	40
≤ 7	140	70
≤ 7.5	180	90
≤ 9.5	200	100

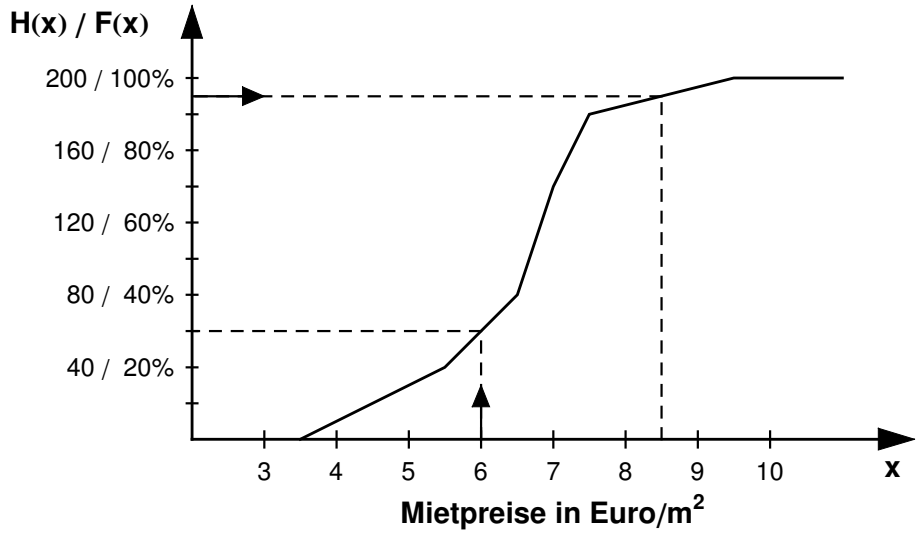


Abbildung 2.1.5: *Summenhäufigkeitsverteilung zu Beispiel 2.1.6*

Die Summenhäufigkeit gibt zu jedem Mietpreis an, wie viele Mietpreise nicht größer sind als dieser Wert. Aufgrund der Annahme der Gleichverteilung innerhalb der Klassen, können somit auch die Zwischenwerte angegeben werden. Aus der Abb. 2.1.5 kann direkt abgelesen werden, dass bis zu einem Mietpreis von 6 € *pro* m^2 30% *der betrachteten Mietpreise entfallen.*

Mit der Angabe einer bestimmten Gesamthäufigkeit kann umgekehrt auch diejenige Merkmalsausprägung bestimmt werden, unterhalb der dieser vorgegebene Anteil der Merkmalswerte liegt. So kann aus der Abb. 2.1.5 auch abgelesen werden, dass 95% *der Mietpreise pro* m^2 *nicht größer als* 8.5 € *sind.*

Für manche Anwendungen ist es zusätzlich nützlich zu wissen, wie viele Beobachtungen bzw. wie viel Prozent der Beobachtungen über einer bestimmten Ausprägung oder Klassengrenze liegen.

***Beispiel 2.1.7**:*

a) Interessierende Fragestellung: Wie viel Prozent der Erwerbstätigen haben ein monatliches Nettoeinkommen von mehr als 1000 *Euro?*

b) Interessierende Fragestellung: Wie viel Prozent der untersuchten Kraftfahrzeuge überschreiten eine bestimmte Geschwindigkeit?

Resthäufigkeiten

In diesem Fall werden die sogenannten **Resthäufigkeiten $\boldsymbol{HR(x)}$** bzw. **$\boldsymbol{FR(x)}$** berechnet, welche aus den Summenhäufigkeiten hergeleitet werden können. Wie bei der Berechnung der Summenhäufigkeiten müssen auch hier die Beobachtungswerte **aufsteigend geordnet** vorliegen.

$$HR_j = HR(x_j) = n - H_j = \sum_{x_{j'} > x_j} h_{j'} = \sum_{j'=j+1}^{n} h_{j'}$$

$$FR_j = FR(x_j) = 1 - F_j = \sum_{x_{j'} > x_j} f_{j'} = \sum_{j'=j+1}^{n} f_{j'}$$

2.2 Lagemaße eindimensionaler Verteilungen

Tabellarische oder grafische Häufigkeitsverteilungen geben einen ersten Eindruck über die zugrundeliegenden Daten wieder. Zur weiteren Charakterisierung werden jedoch zusätzlich Kenngrößen benötigt. Die bekanntesten Kenngrößen sind die Lagemaße.

Lagemaße (Mittelwerte)

Lagemaße (Mittelwerte) geben die zentrale Tendenz einer Beobachtungsreihe mittels einer einzigen charakteristischen Größe wieder, welche die beobachteten Merkmalswerte möglichst gut repräsentieren soll.

Zu beachten ist, dass Mittelwerte über die Lage, jedoch nicht über die Gestalt (Form) einer Verteilung Auskunft geben.

2.2.1 Modalwert

Das einfachste Lagemaß ist der Modalwert, der auch als Modus bezeichnet wird.

Modalwert (Modus)

Der **Modalwert** $\boldsymbol{x_{mod}}$ einer Häufigkeitsverteilung ist jene Merkmalsausprägung, die am häufigsten vorkommt. Es gilt somit $h(x_{mod}) = \max_j h(x_j)$ (Maximum über alle x_j).

Mehrere Modalwerte liegen vor, wenn die größte Häufigkeit mehrmals vorkommt. Im Fall von nominal skalierten Merkmalen kann nur der Modalwert als sinnvolles Lagemaß eingesetzt werden.

Beispiel 2.2.1:

300 *Aktionäre wurden nach der Anzahl ihrer Stammaktien gefragt.*

Anzahl der Aktien	5	6	7	8	9	10	11	12	13	14	15	mehr als 15
Häufigkeit	**86**	16	35	47	21	36	72	**86**	23	39	22	17

Die häufigsten Werte sind $x_{mod,1} = 5$ *und* $x_{mod,2} = 12$.

Modalklasse

Liegen gruppierte Daten vor, kann der Modus in dieser Form nicht bestimmt werden. Hier liegt es nahe, die sogenannte **Modalklasse** zu bestimmen, welche die am dichtesten besetzte Klasse bezeichnet. Oft wird

die Klassenmitte als Modalwert angenommen. Zu beachten ist, dass die Modalklasse nicht unbedingt den Modalwert der ursprünglichen Beobachtungswerte enthält. Auf den feinberechneten Modalwert, der unterschiedliche Klassenbreiten berücksichtigt, wird hier nicht näher eingegangen (vgl. dazu [6]).

2.2.2 Median

Der Median setzt voraus, dass die Merkmalsausprägungen wenigstens nach einer Ordinalskala geordnet werden können.

Median

Der **Median** $\boldsymbol{x_{med}}$ zerlegt eine geordnete Reihe von Beobachtungswerten $x_{(1)}, x_{(2)}, ..., x_{(n)}$ in zwei gleiche Teile, so dass unterhalb und oberhalb des Medians gleich viele Beobachtungswerte liegen.

$\boldsymbol{n}$ **ungerade**: $x_{med} = x_{(\frac{n+1}{2})}$

Ist n gerade, liegt der Median zwischen zwei Werten, nämlich zwischen dem $(\frac{n}{2})$-ten und dem $(\frac{n}{2}+1)$-ten Wert. Definitionsgemäß teilt der Median die Stichprobe in zwei gleiche Teile, so dass theoretisch jede Ausprägung zwischen diesen beiden Werten möglich ist.

Für Merkmale, bei denen ein Mittelwert der Form $\frac{x_i + x_{i'}}{2}$ (s. arithmetisches Mittel) gebildet werden kann, gilt:

$\boldsymbol{n}$ **gerade**: $x_{med} = \frac{1}{2}(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)})$

Die obige Definition gibt in diesem Fall eine eindeutige Bestimmung des Medians an.

Beispiel 2.2.2:

a) Ein Test mit 15 *Studenten ergab folgendes Ergebnis (nach Punktzahlen geordnet):*

5 7 8 8 9 10 11 **12** 12 13 13 13 14 14 15

Der Median ist der $(\frac{n+1}{2})$*-te, also der 8. Wert.*
Es ist $x_{mod} = 13$ *und* $x_{med} = 12$.

b) Von 10 *Aktien wurde die Tagessteigerung ermittelt:*

3.80 3.87 3.93 4.02 **4.07** **4.13** 4.16 4.20 4.21 4.29

Die beiden mittleren Reihenwerte sind 4.07 *und* 4.13*.*
Es gilt also: $x_{med} = \frac{4.07+4.13}{2} = 4.1$*.*

Ist die Häufigkeitsverteilung eines Merkmals gegeben, dann muss bei ordinalen und diskreten metrischen Merkmalen die Ausprägung bestimmt werden, auf die der mittlere Wert fällt bzw. die mittleren Werte fallen. Sind Summenhäufigkeitsverteilungen gegeben, kann der Median mittels folgender Eigenschaft bestimmt werden:

$$F(x_{med}) = 0.5 \text{ bzw. } F(x_{med}) = 50\%.$$

Einfallsklasse

Bei klassierten Merkmalswerten fällt der Zentralwert meistens *in* die sogenannte **Einfallsklasse**, also nicht genau *auf* eine Klassengrenze. In diesem Fall kann der Zentralwert mittels des sogenannten feinberechneten Medians ermittelt werden. Darauf wird an dieser Stelle nicht näher eingegangen (vgl. dazu [6]).

Beispiel 2.2.3:

200 *Studenten wurden nach ihrer Körpergröße befragt:*

Körpergröße in cm	h_j	f_j	H_j	F_j
(140;160]	30	0.15	30	0.15
(160;170]	80	0.40	$110 > \mathbf{101}$	$0.55 > \mathbf{50}\%$
(170;175]	50	0.25	160	0.80
(175;180]	20	0.10	180	0.90
(180;200]	20	0.10	200	1.00

Der Median entspricht dem Wert $x_{med} = \frac{1}{2}(x_{(100)} + x_{(101)})$ *und fällt somit in die Klasse* $(160; 170]$ *(Einfallsklasse). In den beiden ersten Klassen liegen zusammen mehr als die Hälfte der Beobachtungen.*

Besonders einfach lässt sich der Median grafisch bestimmen, wenn die Summenhäufigkeitsverteilung gezeichnet wird. Der Median entspricht der zu $F(x) = 0.5$ gehörigen Merkmalsausprägung. Das Vorgehen ist in Abbildung 2.2.1 für klassierte und unklassierte Werte skizziert. Bei den klassierten Werten wird dabei Gleichverteilung innerhalb der Klassen angenommen.

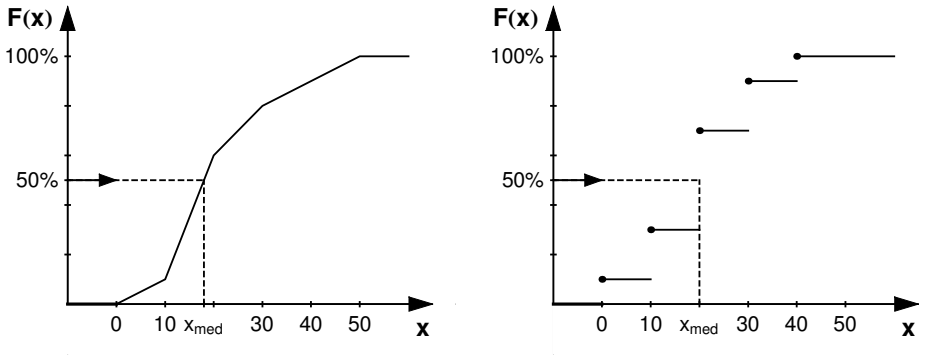


Abbildung 2.2.1: *Grafische Bestimmung des Medians*

Bei quantitativen Merkmalen besitzt der Median folgende Eigenschaft:

Die Summe der absoluten Abweichungen von einem beliebigen Mittelwert wird am kleinsten, wenn sie auf den Median bezogen wird, d.h. die **mittlere absolute Abweichung** d mit $d = \frac{1}{n}\sum_{i=1}^{n}|x_i - c|$ wird **minimal** für $c = x_{med}$.

mittlere absolute Abweichung

Eine weitere wichtige Eigenschaft des Medians ist seine Robustheit gegenüber Ausreißern, d.h. der Median reagiert nicht auf Veränderungen der Werte, die am Rande der Verteilung liegen.

Beispiel 2.2.4:

Die folgenden Reihen von je 3 *Beobachtungswerten haben den gleichen Median* $x_{med} = 20$*:*

1 20 30 10 20 30 10 20 300

Der Median entspricht dem sogenannten 0.5-Quantil $x_{0.5}$, dem Wert der Beobachtungsreihe bis zu dem 50% der Beobachtungen liegen. Allgemein wird als p-Quantil $x_p (0 < p < 1)$ der Wert x_i der geordneten Reihe $x_{(1)}, ..., x_{(n)}$ bezeichnet bis zu dem p% der Beobachtungen liegen.

p-Quantil

$$x_p = \begin{cases} x_{(i)}, & i \text{ ist die kleinste ganze Zahl größer als } n \cdot p, \text{ falls } n \cdot p \text{ keine ganze Zahl ist.} \\ \frac{1}{2}(x_{(i)} + x_{(i+1)}), & i = n \cdot p, \text{ falls } n \cdot p \text{ eine ganze Zahl ist und ein Durchschnittswert berechnet werden kann.} \end{cases}$$

Ist $n \cdot p$ eine ganze Zahl, so wäre prinzipiell jede Zahl zwischen $x_{(i)}$ und $x_{(i+1)}$ als p-Quantil geeignet. Falls sinnvoll ein Durchschnittswert gebildet werden kann, wird hier als spezielles p-Quantil das arithmetische Mittel dieser beiden Werte gewählt. Bei vielen ordinalskalierten Merkmalen wie z.B. beim Rating von Fonds kann kein Durchschnittswert gebildet werden. Weitere Quantile, die einen speziellen Namen tragen, sind die

Dezile

0.1-, 0.2-, ... bzw. 0.9-Quantile $x_{0.1}, x_{0.2}, ..., x_{0.9}$, die auch als **Dezile** bezeichnet werden, und zwar als 10%-Dezil, 20%-Dezil,...,90%-Dezil, welche häufig mit 1.Dezil, 2.Dezil, ..., 9.Dezil abgekürzt werden. Der Median entspricht somit dem 50%-Dezil. Zu erwähnen sind weiter das

untere/obere Quartil

untere Quartil $x_{0.25}$ und das **obere Quartil** $x_{0.75}$.

***Beispiel 2.2.5**:*

Reihe 1*:* 1 1 2 **3** 3 4 5 6 7 7 8 **9** 9 9 10

Es ist $n \cdot 0.25 = 15 \cdot 0.25 = 3.75$*, d.h. das untere Quartil entspricht dem* 4*. geordnetem Wert. Mit* $n \cdot 0.75 = 15 \cdot 0.75 = 11.25$ *lässt sich das obere Quartil mittels des* 12*. geordneten Wertes bestimmen.*

Reihe 2*:* 2 3 **3** **4** 5 6 7 8 **8** **9** 10 10

Hier ist $n \cdot 0.25 = 12 \cdot 0.25 = 3$*, d.h. das untere Quartil entspricht dem arithmetischen Mittel aus dem* 3*. und* 4*. geordnetem Wert* $(3+4)/2 = 3.5$*. Mit* $n \cdot 0.75 = 12 \cdot 0.75 = 9$ *entspricht das obere Quartil dem Wert* $(8+9)/2 = 8.5$*.*

In vielen Analysen reicht es aus, eine empirische Verteilung mittels mehrerer Charakteristika in Form des sogenannten **Box-Plots** grafisch darzustellen. In der einfachsten Variante werden die fünf Kennzahlen $x_{0.25}$, x_{med}, $x_{0.75}$, x_{min} und x_{max} herangezogen.

Box-Plot

In einem einfachen **Box-Plot** werden die Quartile $x_{0.25}$ und $x_{0.75}$ durch eine Box dargestellt, in deren Inneren der Median als Punkt oder als Linie dargestellt ist. Die Extremwerte x_{min} und x_{max} werden mit der Box durch Striche („whisker“) verbunden.

Durch die Lage des Medians innerhalb der Box kann ein Eindruck von der Schiefe der zugrundeliegenden Verteilung vermittelt werden.

Folgende Grafik zeigt eine Variante des einfachen Box-Plots, bei der nicht die Extremwerte mit der Box verbunden sind, sondern der größte und der kleinste „normale“ Wert, der noch nicht als Ausreißer angesehen wird. Ausreißer werden hier mittels eines Kreises dargestellt. Als Ausreißer gelten an dieser Stelle Werte, die weiter als 1.5 Boxlängen unterhalb bzw. oberhalb der Box liegen. Zugrundegelegt wurde ein fiktiver Datensatz von Arbeitslosenzahlen aus 33 Agenturen für Arbeit.

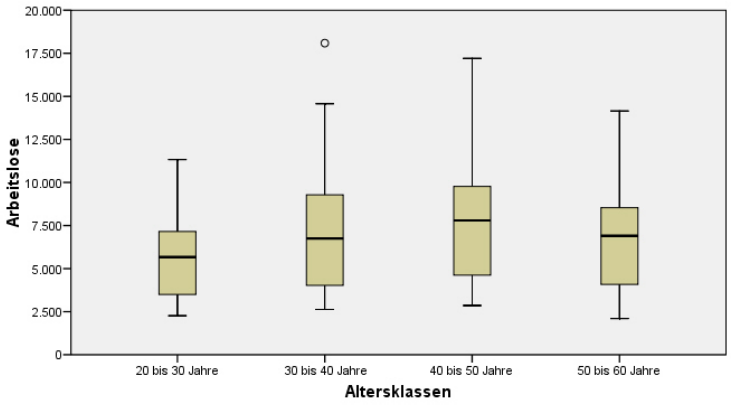


Abbildung 2.2.2: *Arbeitslosenzahlen nach Altersklassen sortiert, Box-Plot erstellt mit SPSS 15.0*

2.2.3 Arithmetisches Mittel

Das **arithmetische Mittel** $\overline{x}$ ist der bekannteste Mittelwert und entspricht dem Wert, der häufig als „Durchschnitt“ bezeichnet wird. Anwendbar ist das arithmetische Mittel nur bei metrischen Merkmalen.

arithmetische Mittel

Liegen einzelne Beobachtungswerte vor, berechnet sich das arithmetische Mittel wie folgt:

Gegeben sind die n Beobachtungswerte x_i $(i = 1, 2, ..., n)$, dann ergibt sich das **arithmetische Mittel** $\overline{x}$ zu

$$\overline{x} = \frac{x_1 + x_2 + ... + x_n}{n} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

Das so berechnete arithmetische Mittel wird auch als **ungewogenes arithmetisches Mittel** bezeichnet.

ungewogene arithmetische Mittel

***Beispiel 2.2.6**:*

Von einer Zeitschrift werden in einem Quartal in den einzelnen Wochen folgende Stückzahlen verkauft (in Millionen):

1.4 1.6 1.8 1.7 1.5 1.7 1.5 1.6 1.8 2.0 1.8 1.8 1.6

Der durchschnittliche Absatz beträgt dann:

$$\begin{aligned}\overline{x} &= \frac{1}{13}(1.4+1.6+1.8+1.7+1.5+1.7+1.5+1.6 \\ &\quad +1.8+2.0+1.8+1.8+1.6) \\ &= \frac{21.8}{13} = 1.677.\end{aligned}$$

Liegen keine einzelnen Beobachtungswerte vor, sondern eine (diskrete oder klassierte) Häufigkeitsverteilung, dann berechnet sich das arithmetische Mittel anders.

gewogene arithmetische Mittel

Für jede vorkommende Merkmalsausprägung bzw. Klasse x_j $(j = 1, ..., m)$ sind die absoluten bzw. relativen Häufigkeiten h_j bzw. f_j gegeben, wobei die Gesamtzahl der Beobachtungen n entspricht. In diesem Fall berechnet sich das arithmetische Mittel $\overline{x}$ zu:

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{m} x_j h_j = \sum_{j=1}^{m} x_j f_j.$$

Dieser Wert, bei dem die Merkmalsausprägungen mit den beobachteten Häufigkeiten gewichtet werden, wird als **gewogenes arithmetisches Mittel** bezeichnet.

***Beispiel 2.2.7**:*

Während eines halben Jahres (120 *Arbeitstage*) *wird täglich im Rahmen einer Untersuchung über den Publikumsverkehr beim Sozialamt einer Großstadt die Anzahl der persönlich vorsprechenden Antragsteller festgehalten.*

Anzahl der Antragsteller	0	1	2	3	4	5	6	7	8	9	10
Anzahl der Tage	5	4	10	12	20	18	18	12	15	2	4

Als arithmetisches Mittel ergibt sich daraus ein durchschnittlicher Publikumsverkehr von

$$\begin{aligned}\overline{x} &= \frac{0\cdot 5+1\cdot 4+2\cdot 10+3\cdot 12+4\cdot 20+5\cdot 18}{120}\\ &\quad +\frac{6\cdot 18+7\cdot 12+8\cdot 15+9\cdot 2+10\cdot 4}{120}\\ &= \frac{4+20+36+80+90+108+84+120+18+40}{120}\\ &= \frac{600}{120}=5.\end{aligned}$$

Liegen klassierte Daten vor, ist zu beachten, dass bei der Berechnung des gewogenen arithmetischen Mittels Verzerrungen im Vergleich zu dem ungewogenen arithmetischen Mittel, welches aus den Ursprungswerten berechnet wird, auftreten können, z.B. wenn die Werte innerhalb der Klassen nicht gleichverteilt sind.

***Beispiel 2.2.8**:*

Eine Befragung über das Alter von 30 *Angehörigen eines Betriebes hat folgendes Ergebnis geliefert:*

24	22	40	21	32	51	62	22	41	42	43	51	22	32	33
63	19	22	21	50	50	33	60	17	20	50	42	30	20	41

Zur Ermittlung der Altersstruktur werden Altersklassen zu je 10 *Jahren gebildet. Für die offenen Randklassen wurden die Klassenmitten* $x_1 = 15$ *und* $x_6 = 65$ *festgelegt.*

j	Altersklasse	x_j	h_j
1	≤ 20	15	4
2	(20;30]	25	8
3	(30;40]	35	5
4	(40;50]	45	8
5	(50;60]	55	3
6	>60	65	2

Aus den Ursprungswerten ergibt das arithmetische Mittel folgendes Durchschnittsalter:

$$\overline{x} = \frac{24+22+40+...}{30} = 35.9.$$

Für die klassierten Werte ergibt sich dagegen

$$\overline{x} = \frac{1}{30}(4 \cdot 15 + 8 \cdot 25 + 5 \cdot 35 + 8 \cdot 45 + 3 \cdot 55 + 2 \cdot 65) = 36.\overline{3}.$$

Der Unterschied zwischen den beiden Ergebnissen beruht auf der ungleichmäßigen Verteilung der Werte innerhalb der Klassen. Die Werte liegen jeweils in der unteren Hälfte der Klasse. Das Ausmaß solcher Verzerrungen hängt somit von der Wahl der Klassengrenzen ab.

Das arithmetische Mittel besitzt folgende wichtige Eigenschaften:

> Die Summe der einfachen Abweichungen der Beobachtungswerte vom arithmetischen Mittel ist 0.

Beweis:

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

□

> Die Summe der quadratischen Abweichungen der Beobachtungswerte von einem beliebigen Mittelwert M wird dann ein Minimum, wenn M dem arithmetischen Mittel entspricht.

Beweis:

Es sei M ein beliebiger Mittelwert. Dann wird die Summe S der quadratischen Abweichungen der Beobachtungswerte von diesem Mittelwert $S = \sum_{i=1}^{n}(x_i - M)^2$ ein Minimum, wenn gilt:

$$\frac{dS}{dM} = \sum_{i=1}^{n} -2(x_i - M) = -2\sum_{i=1}^{n} x_i + 2nM \stackrel{!}{=} 0.$$

Daraus folgt:

$$\sum_{i=1}^{n} x_i = nM \text{ oder } M = \frac{\sum x_i}{n} = \overline{x}.$$

Da außerdem gilt $\frac{d^2S}{dM^2} = 2n > 0$, liegt tatsächlich ein Minimum vor.

□

Das arithmetische Mittel ist das am meisten verwendete Lagemaß. Manchmal wird es jedoch missbräuchlich angewandt. So kann z.B. aus ordinalskalierten Zensuren zwar rechnerisch durchaus eine Durchschnittsnote als arithmetisches Mittel bestimmt werden, aber sachlich ist das nicht sinnvoll, denn der Durchschnitt aus einer 1 und einer 5 muss nicht unbedingt eine 3 sein.

Die Anwendung des arithmetischen Mittels kann aber auch bei einer metrischen Skala problematisch werden.

***Beispiel 2.2.9**:*

Eine Firma mit 9 *Angestellten und einem Chef erzielt in einem bestimmten Monat einen Überschuss, wobei die Angestellten* 200 € *als Prämie erhalten, während der Chef* 2000 € *einbehält. Das arithmetische Mittel der Gewinnausschüttung ergibt hier:*

$$\overline{x} = \frac{9 \cdot 200 + 1 \cdot 2000}{10} = 380.$$

In diesem Beispiel führt die Berechnung des arithmetischen Mittels zu einer irreführenden Aussage. Hier liegt es nahe, den Median zur Berechnung heranzuziehen, der gegenüber Ausreißern robust ist. Das gilt für die meisten Durchschnitts- bzw. Mittelwertbestimmungen bei Einkommensuntersuchungen.

Generell lässt sich sagen, dass die Verwendung des arithmetischen Mittels immer dann Probleme aufwirft, wenn die Verteilung nach einer Seite sehr weit ausläuft, bzw. Ausreißer vorliegen. Bei allen ausgeprägt schiefen Verteilungen wird also die Verwendung des arithmetischen Mittels problematisch. Entsprechendes gilt bei mehrgipfligen Verteilungen.

Das **arithmetische Mittel** ist für die Beschreibung der „durchschnittlichen Lage“ einer Verteilung um so **weniger geeignet**, je stärker eine Verteilung von den Eigenschaften Eingipfligkeit und Symmetrie abweicht.

2.2.4 Geometrisches Mittel

Wird eine statistische Untersuchung durchgeführt, bei der durchschnittliche prozentuale Veränderungen von Interesse sind, so kann das arithmetische Mittel nicht angewendet werden.

***Beispiel 2.2.10**:*

Nachstehende Tabelle gibt die Kapitalentwicklung eines Anfangskapitals von 1000 € *bei steigendem Zins an. Von Interesse ist die durchschnittliche Verzinsung.*

Jahr	Zinssatz	Kapital	
		1000	Anfangskapital
1	3%	1030.00	
2	3%	1060.90	
3	5%	1113.95	
4	7%	1191.92	
5	7%	1275.36	Endkapital

Die Berechnung des Endkapitals mittels des durchschnittlichen Zinssatzes von 5% $(5 = (3 + 3 + 5 + 7 + 7)/5)$ *führt zu einem falschen Ergebnis.*

Jahr	Zinssatz	Kapital	
		1000	Anfangskapital
1	5%	1050.00	
2	5%	1102.50	
3	5%	1157.63	
4	5%	1215.51	
5	5%	1276.28	Endkapital

Für die schrittweise Bestimmung des Endkapitals K_5 sei das Kapital am Ende der ersten vier Jahre mit $K_1, ..., K_4$ bezeichnet. Das Anfangskapital wird mit K_0 angegeben.

1. Jahr: $(1 + 0.03) \cdot K_0 = K_1$
2. Jahr: $(1 + 0.03) \cdot K_1 = (1 + 0.03)^2 \cdot K_0 = K_2$
3. Jahr: $(1 + 0.05) \cdot K_2 = (1 + 0.05) \cdot (1 + 0.03)^2 \cdot K_0 = K_3$
4. Jahr: $(1 + 0.07) \cdot K_3 = (1 + 0.07) \cdot (1 + 0.05) \cdot (1 + 0.03)^2 \cdot K_0 = K_4$
5. Jahr: $(1 + 0.07) \cdot K_4 = (1 + 0.07)^2 \cdot (1 + 0.05) \cdot (1 + 0.03)^2 \cdot K_0 = K_5$

Das Endkapital lässt sich als Produkt aus dem Anfangskapital K_0 und den Zinsfaktoren $1 + z_t$ darstellen, wobei z_t den Zinssatz des Jahres t bezeichnet, $t = 1, ..., 5$. Der durchschnittliche Zinsfaktor $1 + z$ berechnet sich dann als 5. Wurzel aus dem Produkt der Zinsfaktoren.

$$\begin{aligned} 1 + z &= \sqrt[5]{1.03^2 \cdot 1.05 \cdot 1.07^2} = 1.0498476 \\ K_5 &= (1 + z)^5 \cdot K_0 = 1.27536 \cdot K_0 = 1275.36 \end{aligned}$$

Der in dem Beispiel berechnete Durchschnittswert $1 + z$ entspricht dem gewogenen arithmetischen Mittel der Zinsfaktoren und wird als **geometrisches Mittel** bezeichnet. Die Verallgemeinerung lautet:

ungewogene

Das **ungewogene geometrische Mittel** entspricht der n-ten Wurzel aus dem Produkt der Zahlen $x_1 \dots x_n$:

$$\overline{x}_{geom} = \sqrt[n]{x_1 \cdot x_2 \cdot ... \cdot x_n} = \sqrt[n]{\prod_{i=1}^{n} x_i}.$$

gewogene geometrische Mittel

Liegen einige Zahlen mehrmals vor, so ergibt sich das **gewogene geometrische Mittel**:

$$\overline{x}_{geom} = \sqrt[n]{x_1^{h_1} \cdot x_2^{h_2} ... \cdot x_m^{h_m}} = \sqrt[n]{\prod_{j=1}^{m} x_j^{h_j}} = \prod_{j=1}^{m} x_j^{f_j}.$$

Das geometrische Mittel wird immer dann benötigt, wenn es um die Analyse durchschnittlicher prozentualer Veränderungen geht.

Besonders wichtig ist zu beachten, dass das geometrische Mittel nicht von Zinssätzen (Zuwachsraten und dergleichen) berechnet wird, sondern von den Zinsfaktoren (Wachstumsfaktoren und dergleichen).

***Beispiel 2.2.11**:*

Wird die Kapitalentwicklung des Anfangskapitals von 1000 € *anhand des geometrischen Mittels der Zinssätze berechnet (Zinssätze s. Beispiel 2.2.10), so ergibt sich:*

$$\begin{aligned} \overline{x}_{geom} &= \sqrt[5]{3 \cdot 3 \cdot 5 \cdot 7 \cdot 7} = 4.66318, \\ K_5 &= 1.0466318^5 \cdot K_0 = 1.25594 \cdot K_5 = 1255.94. \end{aligned}$$

Dies entspricht einer falschen Berechnung des Endkapitals.

2.2.5 Zusammenfassung zu den Mittelwerten

Die Wahl des geeigneten Mittelwertes hängt sowohl von der Struktur des Datensatzes als auch von der zugrundeliegenden Fragestellung ab.

Der **Modus x_{mod}** ist bei nominalen Merkmalen der einzige sinnvoll zu bestimmende „Mittelwert“. Er ist ein **sehr grobes** Lagemaß.

Der **Median x_{med}** ist bei ordinalen Merkmalen das wichtigste Lagemaß. Da der Median robust gegenüber Ausreißern ist, wird er unter Umständen auch bei metrisch skalierten Merkmalen angewendet.

Das **arithmetische Mittel $\overline{x}$** ist bei metrischen Merkmalen das wichtigste Lagemaß. Für $\overline{x}$ gilt:

$$n\overline{x} = \sum_{i=1}^{n} x_i \quad \text{bzw.} \quad n\overline{x} = \sum_{j=1}^{m} h_j x_j .$$

Die Summe aller Beobachtungswerte liefert also gerade das n-fache des arithmetischen Mittels. Besäßen alle statistischen Einheiten die Merkmalsausprägung des arithmetischen Mittels, dann ergäbe die Summe für alle Einheiten denselben Wert wie die Summe der tatsächlichen Beobachtungswerte. Das arithmetische Mittel wird deshalb immer angewendet, **wenn eine Addition der Beobachtungswerte sinnvoll ist**.

Probleme bei der Anwendung des arithmetischen Mittels treten auf, wenn die Verteilung nach einer Seite sehr weit ausläuft (schiefe Verteilung), bzw. Ausreißer vorliegen. Entsprechendes gilt für mehrgipflige Verteilungen.

Das **geometrische Mittel $\overline{x}_{geom}$** ist ebenfalls nur bei metrischen Merkmalen anwendbar. Es gilt:

$$\overline{x}^n_{geom} = \prod_{i=1}^{n} x_i \quad \text{bzw.} \quad \overline{x}^n_{geom} = \prod_{j=1}^{m} x_j^{h_j} .$$

Das geometrische Mittel wird dann angewendet, **wenn die Beobachtungswerte sinnvoll durch Multiplikation verknüpft werden können**, wie vor allem bei Zuwachs- oder Wachstumsfaktoren eines Merkmals im Zeitablauf.

2.3 Streuungsmaße eindimensionaler Verteilungen

Im vorherigen Abschnitt wurde gezeigt, dass die Lagemaße die zentrale Tendenz einer Verteilung beschreiben. Zur weiteren Charakterisierung einer Verteilung werden zusätzlich Kenngrößen benötigt, welche Aussagen über die **Streuung** einer Verteilung machen. Von Interesse ist dabei abzuschätzen, inwieweit die Stichprobenwerte um den Mittelwert verteilt sind bzw. streuen.

Streuung

Beispiel 2.3.1:

Nachstehende Abbildung verdeutlicht, dass Verteilungen mit gleicher Lage eine unterschiedliche Gestalt besitzen können. Die Verteilung C zeigt einen viel engeren Wertebereich als die Verteilungen A und B, d.h. die Werte der Verteilung C streuen weniger bzw. die Verteilung C hat eine geringere Streuung als die Verteilungen A und B.

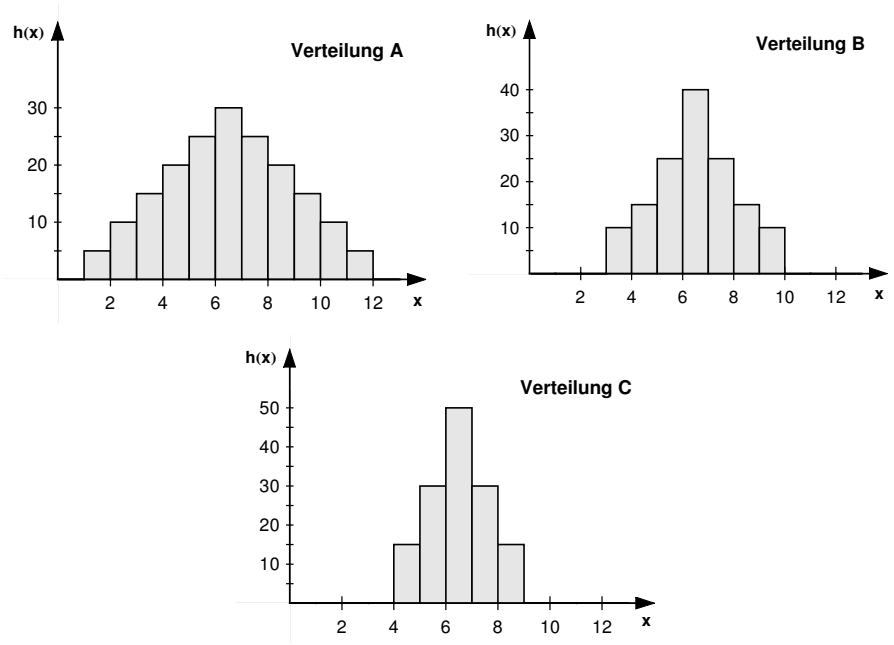


Abbildung 2.3.1: *Verteilungen mit gleicher Lage und unterschiedlicher Streuung*

Im Folgenden werden die wichtigsten **Streuungsmaße**, wie die Spannweite, die Varianz, die Standardabweichung und der Variationskoeffizient, eingeführt. Ein weiteres Streuungmaß, die mittlere absolute Abweichung, wurde bereits im vorherigen Kapitel kurz erwähnt.

Streuungsmaße

2.3.1 Spannweite

Von den Streuungsmaßen ist die Spannweite am einfachsten zu berechnen.

Spannweite

Die **Spannweite w** ist als Differenz der beiden Extremwerte, dem kleinsten und dem größten vorkommenden Beobachtugswert, definiert:

$$w = \max_i x_i - \min_i x_i.$$

$\max_i x_i$: größter x_i-Wert für alle i (Maximum über alle x_i),
$\min_i x_i$: kleinster x_i-Wert für alle i (Minimum über alle x_i).

Beispiel 2.3.2:

Gegeben sind folgende Beobachtungswerte:

27 4 8 3 12 10 26 6 19 16

Die Spannweite beträgt $27 - 3 = 24$.

Quartilsabstand

Da die Spannweite nur den kleinsten und den größten Wert der Verteilung berücksichtigt, ist dieses Maß nicht besonders aussagekräftig. Abbildung 2.3.2 verdeutlicht, dass die Spannweite nicht robust gegenüber Ausreißern ist. Alternativ kann der **Quartilsabstand $x_{0.75} - x_{0.25}$** als Streuungsmaß gewählt werden.

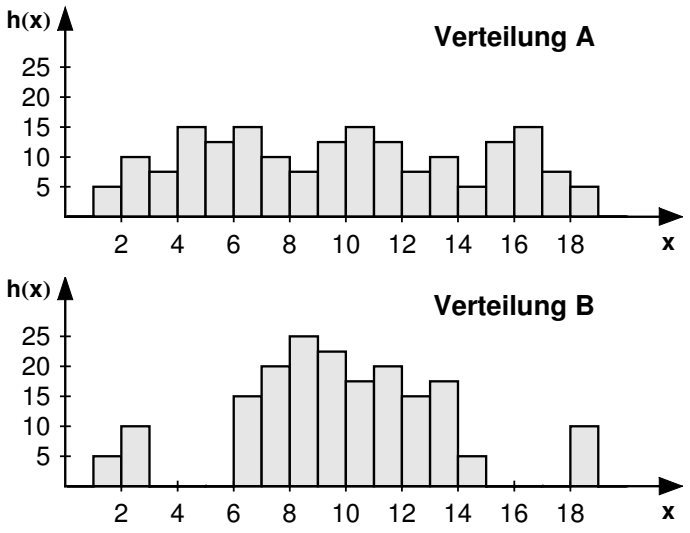


Abbildung 2.3.2: *Unterschiedliche Verteilungen mit gleicher Spannweite*

2.3.2 Varianz und Standardabweichung

Das am häufigsten verwendete Streuungsmaß ist die sogenannte Varianz $\tilde{s}^2$ bzw. die Quadratwurzel daraus, die Standardabweichung $\tilde{s}$.

Diese Maße berücksichtigen die quadratischen Abweichungen aller Beobachtungswerte vom arithmetischen Mittel, so dass größere Abstände zum Mittelwert stärker berücksichtigt werden.

Die empirische **Varianz $\tilde{s}^2$** ist die **mittlere quadratische Abweichung** der Beobachtungswerte $x_i (i = 1, ..., n)$ vom arithmetischen Mittel $\overline{x}$.

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

Für eine Häufigkeitsverteilung gilt:

$$\begin{aligned}\tilde{s}^2 &= \frac{1}{n}\sum_{j=1}^{m}(x_j - \overline{x})^2 h_j = \frac{1}{n}\sum_{j=1}^{m} x_j^2 h_j - \overline{x}^2 \\ &= \sum_{j=1}^{m}(x_j - \overline{x})^2 f_j = \sum_{j=1}^{m} x_j^2 f_j - \overline{x}^2.\end{aligned}$$

Bei klassierten Werten sind die x_j die Klassenmitten.

Varianz

Das arithmetische Mittel besitzt hier die gleiche Eigenschaft wie der Median bei der mittleren absoluten Abweichung, d.h. $\sum_{i=1}^{n}(x_i - c)^2$ wird für $c = \overline{x}$ minimal (Beweis s. Abschnitt 2.2.3).

Die empirische **Standardabweichung $\tilde{s}$** ist die Quadratwurzel aus der Varianz $\tilde{s}^2$.

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

$$\tilde{s} = \sqrt{\frac{1}{n}\sum_{j=1}^{m}(x_j - \overline{x})^2 h_j} = \sqrt{\sum_{j=1}^{m}(x_j - \overline{x})^2 f_j}$$

Standard-abweichung

Werden für verschiedene Variablen (X, Y, ...) die Varianzen bzw. Standardabweichungen berechnet, werden diese mit $\tilde{s}_x^2$, $\tilde{s}_y^2$, ... bzw. $\tilde{s}_x$, $\tilde{s}_y$, ... bezeichnet.

Beispiel 2.3.3:

Für die Werte 3, 7, 8, 9, 13 *mit* $\overline{x} = 8$ *ergibt sich für die Varianz*

$$\begin{aligned}\tilde{s}^2 &= \frac{(3-8)^2+(7-8)^2+(8-8)^2+(9-8)^2+(13-8)^2}{5} \\ &= \frac{(-5)^2+(-1)^2+0^2+1^2+5^2}{5} \\ &= \frac{25+1+1+25}{5} = \frac{52}{5} = 10.4\end{aligned}$$

und für die Standardabweichung

$$\tilde{s} = \sqrt{\tilde{s}^2} = \sqrt{10.4} = 3.22.$$

Für eine gegebene Häufigkeitsverteilung lässt sich die Varianz mittels einer Arbeitstabelle einfach berechnen.

Beispiel 2.3.4:

Innerhalb eines Betriebes wurden 30 *Mitarbeiter nach ihrem Alter befragt.*

Alter	Klassen-mitte x_j	h_j	$x_j h_j$	$x_j - \overline{x}$	$(x_j - \overline{x})^2$	$(x_j - \overline{x})^2 \cdot h_j$
≤ 20	15	2	30	-24	576	1152
(20;30]	25	9	225	-14	196	1764
(30;40]	35	5	175	-4	16	80
(40;50]	45	6	270	6	36	216
(50;60]	55	5	275	16	256	1280
>60	65	3	195	26	676	2028
		$\sum x_j h_j =$	1170	$\sum (x_j - \overline{x})^2 \cdot h_j =$		6520

$$\begin{aligned}\overline{x} &= \frac{1}{n}\sum_{j=1}^{m} x_j h_j = \frac{1}{30} \cdot 1170 = 39 \\ \tilde{s}^2 &= \frac{1}{n}\sum_{j=1}^{m} (x_j - \overline{x})^2 h_j = \frac{1}{30} \cdot 6520 = 217.33 \\ \tilde{s} &= \sqrt{217.33} = 14.74\end{aligned}$$

2.3.3 Variationskoeffizient

Die Varianz bezieht in die Berechnung das arithmetische Mittel ein, jedoch wird dabei nicht berücksichtigt, in welcher Relation die Varianz und das arithmetische Mittel zueinander stehen.

***Beispiel 2.3.5**:*

Es liegen die Merkmale X und Y mit den Mittelwerten $\overline{x} = 1000$ und $\overline{y} = 1$ und den Varianzen $\tilde{s}_x^2 = \tilde{s}_y^2 = 10$ vor. Im Fall des Merkmals X ist die Varianz recht klein, während die Varianz für Y recht groß erscheint.

Gesucht ist ein dimensionsloses Streuungsmaß, welches Streuung und arithmetische Mittel ins Verhältnis setzt, und somit für Vergleichszwecke besonders gut geeignet ist.

Der **Variationskoeffizient v** ist eine relative Größe, welche das Verhältnis von Standardabweichung und arithmetischem Mittel darstellt.

$$v = \frac{\tilde{s}}{\overline{x}}$$

Variations-koeffizient

In einigen Fällen wird auch die mittlere absolute Abweichung zum Median in Beziehung gesetzt ($v = \frac{d}{x_{med}}$).

***Beispiel 2.3.6**:*

Für eine Flasche Parfüm wurden in „Duty-free-shops“ auf verschiedenen Flughäfen folgende Preise in Euro und in US-Dollar verlangt:

Flughafen	1	2	3	4	5	6
Preis in \$	39	43.88	37.05	40.95	44.85	39.98
Preis in €	30	33.75	28.50	31.50	34.50	30.75

Für die Preise in US-Dollar und Euro gelten:

$$\begin{array}{lccc} \textit{US Dollar:} & \overline{x}_1 = 40.95 & \tilde{s}_1 = 2.700 & v_1 = 0.065934 \\ \textit{Euro:} & \overline{x}_2 = 31.50 & \tilde{s}_2 = 2.077 & v_2 = 0.065937 \end{array}$$

Die Variationskoeffizienten v_1 und v_2 stimmen nahezu überein. Die Differenz ist auf Rundungsabweichungen zurückzuführen.

2.3.4 Standardisierung von Daten

Liegen Beobachtungen vor, die unterschiedliche Maßeinheiten besitzen bzw. die aus verschiedenen Stichproben mit unterschiedlichem Erwartungswert und/oder Varianz stammen, so kann eine Vergleichbarkeit der Daten mittels der Standardisierung erzielt werden.

Standardisierung von Daten

Gegeben sei eine Stichprobe mit dem arithmetischen Mittel $\bar{x}$ und der Varianz $\tilde{s}^2$, dann wird bei der **Standardisierung** von Daten, auch ***z*-Transformation** genannt, die Beobachtung x_i in den Wert

$$z_i = \frac{x_i - \bar{x}}{\tilde{s}}$$

transformiert. Nach der Standardisierung liegen Daten mit dem Mittelwert 0 und der Varianz 1 vor.

***Beispiel 2.3.7**:*

Ein Student möchte seine Ergebnisse der Statistikklausur (S) und Mathematikklausur (M), unter Berücksichtigung des Gesamtergebnisses, miteinander vergleichen. In der Statistikklausur erzielte er von 150 *möglichen Punkten* 82*, während er in der Mathematikklausur* 45 *von* 100 *Punkten erreichte. Die Mittelwerte und Standardabweichungen wurden mit $\bar{x}_S = 76$, $\bar{x}_M = 40$, $\tilde{s}_S = 30$, $\tilde{s}_M = 20$ angegeben.*

$$z_S = \frac{82 - 76}{30} = 0.2 \qquad z_M = \frac{45 - 40}{20} = 0.25$$

Der Student hat somit in der Mathematikklausur besser abgeschnitten.

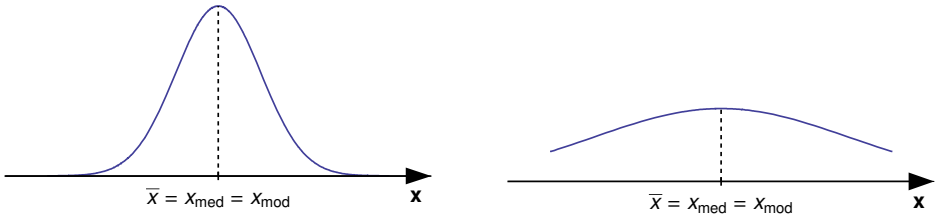
2.4 Schiefe und Wölbung einer Verteilung

Zur Beschreibung von Verteilungen können auch die Begriffe **Symmetrie, Schiefe**, **Steilheit** und **Wölbung** herangezogen werden.

Symmetrie, Schiefe, Steilheit, Wölbung

Die Schiefe gibt an, wie stark eine eingipflige Verteilung nach rechts bzw. nach links geneigt ist. Eine Verteilung ist symmetrisch, wenn sie in Bezug auf das arithmetische Mittel symmetrisch ist. Merkmalsausprägungen, die um den gleichen Betrag nach unten bzw. oben vom arithmetischen Mittel abweichen, haben dann die gleiche absolute bzw. relative Häufigkeit. Des Weiteren kann zwischen flachen und steilen Verteilungen unterschieden werden.

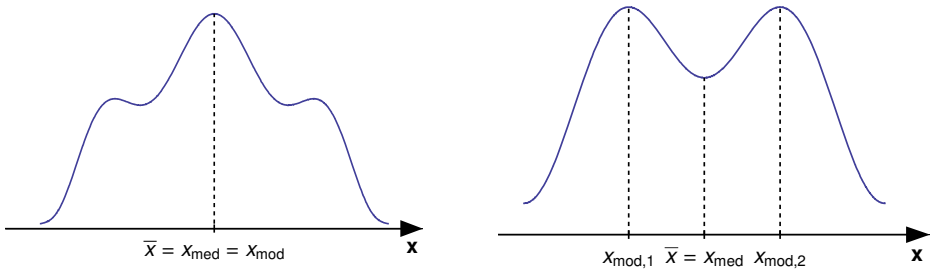
Arithmetisches Mittel, Median und Modalwert stimmen bei einer **eingipfligen, symmetrischen Verteilung** überein.



steile Verteilung flache Verteilung

Abbildung 2.4.1: *Eingipflige, symmetrische Verteilungen*

Auch bei **mehrgipfligen, symmetrischen Verteilung** stimmen das arithmetische Mittel und der Median stets überein. Aufgrund der Mehrgipfligkeit können jedoch mehrere Modalwerte auftreten.

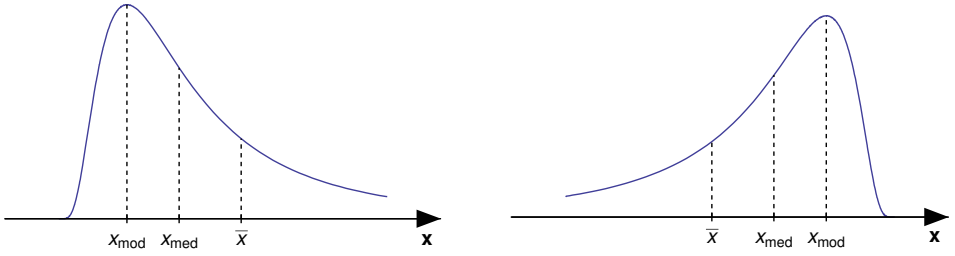


unimodale Verteilung bimodale Verteilung

Abbildung 2.4.2: *Mehrgipflige, symmetrische Verteilungen*

rechtsschief (linkssteil)
linksschief (rechtssteil)

Eingipflige Verteilungen können nach ihrer Schiefe beurteilt werden, wobei zwischen rechtsschief und linksschief unterschieden werden kann. **Rechtsschiefe Verteilungen** steigen von links nach rechts steil an und fallen dann nach rechts flach ab. **Linksschiefe Verteilungen** steigen dagegen von rechts nach links steil an und fallen dann nach links flach ab. Daher werden diese auch als **linkssteile** bzw. **rechtssteile Verteilungen** bezeichnet.



linkssteile bzw. rechtsschiefe Verteilung

linksschiefe bzw. rechtssteile Verteilung

Abbildung 2.4.3: *Schiefe Verteilungen*

Fechnersche Lageregel

Mit Hilfe der **Fechnerschen Lageregel** können Verteilungen hinsichtlich der Schiefe betrachtet werden, ohne eine Grafik hinzuzuziehen.

Fechnersche Lageregel:

$$\begin{aligned} x_{mod} = \overline{x} = x_{med} &\quad \text{für symmetrische Verteilungen,} \\ x_{mod} < x_{med} < \overline{x} &\quad \text{für rechtsschiefe Verteilungen,} \\ \overline{x} < x_{med} < x_{mod} &\quad \text{für linksschiefe Verteilungen.} \end{aligned}$$

Als Maßzahl zur Beurteilung der Schiefe einer eingipfligen Verteilung metrischer Merkmale kann der Momentenkoeffizient berechnet werden.

Momentenkoeffizient der Schiefe

Der **Momentenkoeffizient der Schiefe g_3** lautet:

$$g_3 = \frac{m_3}{\tilde{s}^3} = \frac{\frac{1}{n}\sum_i^n (x_i - \overline{x})^3}{\left[\frac{1}{n}\sum_i^n (x_i - \overline{x})^2\right]^{\frac{3}{2}}}.$$

Aufgrund der dritten Potenz im Zähler, bleiben die Vorzeichen der Abweichungen $(x_i - \overline{x})$ erhalten. Da bei linksschiefen (rechtsschiefen) Verteilungen mehr negative (positive) Abweichungen vorliegen, die aufgrund des größeren Abstandes $(x_i - \overline{x})$ beim Potenzieren mehr ins Gewicht fallen, gilt allgemein für den Momentenkoeffizient der Schiefe:

$g_3 = 0$ für symmetrische Verteilungen,
$g_3 < 0$ für linksschiefe Verteilungen,
$g_3 > 0$ für rechtsschiefe Verteilungen.

Maßzahlen der Wölbung beurteilen eingipflige Verteilungen metrischer Merkmale bezüglich ihrer „Flachheit". Momentenkoeffizienten der Wölbung sind die **Kurtosis** und der **Kurtosis-Exzess** (auch Wölbungsmaß von Fisher genannt).

Die **Kurtosis g_4** berechnet sich zu:

$$g_4 = \frac{m_4}{\tilde{s}^4} = \frac{\frac{1}{n}\sum_i^n (x_i - \overline{x})^4}{\left[\frac{1}{n}\sum_i^n (x_i - \overline{x})^2\right]^2}.$$

Kurtosis

Der **Kurtosis-Exzess $\tilde{g}_4$** berechnet sich zu:

$$\tilde{g}_4 = g_4 - 3.$$

Kurtosis-Exzess

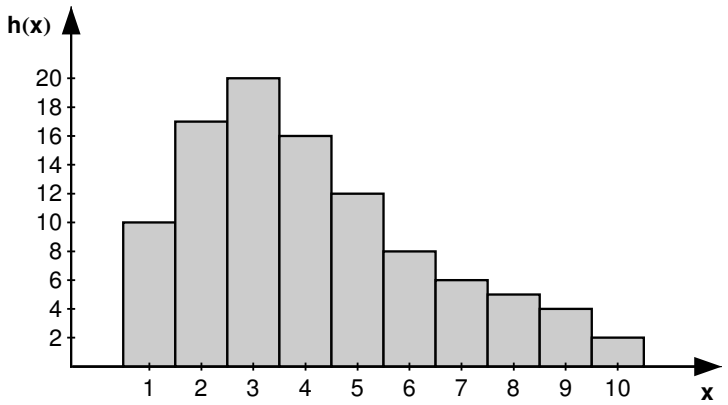
Da die Kurtosis einer Normalverteilung, die wichtigste Verteilung in der Statistik (Einführung der Normalverteilung s. Kurseinheit 2), den Wert 3 annimmt, gibt der Kurtosis-Exzess an, inwieweit sich die Wölbung der vorliegenden Verteilung von der einer Normalverteilung unterscheidet. Es gilt:

$\tilde{g}_4 = 0$ für normalgipflige (mesokurtische) Verteilungen,
$\tilde{g}_4 < 0$ für flachgipflige (platykurtische) Verteilungen,
$\tilde{g}_4 > 0$ für hochgipflige (leptokurtische) Verteilungen.

Zu beachten ist, dass Verteilungen mit gleicher Streuung nicht unbedingt dieselbe Wölbung aufweisen. Hier können Unterschiede in den Randbereichen vorliegen.

***Beispiel 2.4.1**:*

a) *Gegeben ist folgende linkssteile Häufigkeitsverteilung für das Merkmal X:*



Zur Berechnung der Kenngrößen Momentenkoeffizient der Schiefe, Kurtosis bzw. Kurtosis-Exzess wird folgende Tabelle aufgestellt mit $\overline{x} = 4.14$:

x_j	h_j	$(x_j - \overline{x})$	$h_j \cdot (x_j - \overline{x})^2$	$h_j \cdot (x_j - \overline{x})^3$	$h_j \cdot (x_j - \overline{x})^4$
1	10	-3.14	98.596	-309.591	972.117
2	17	-2.14	77.853	-166.606	356.537
3	20	-1.14	25.992	-29.631	33.779
4	16	-0.14	0.314	-0.044	0.006
5	12	0.86	8.875	7.633	6.564
6	8	1.86	27.677	51.479	95.751
7	6	2.86	49.078	140.362	401.435
8	5	3.86	74.498	287.562	1109.990
9	4	4.86	94.478	459.165	2231.542
10	2	5.86	68.679	402.460	2358.416
$\sum$	100		526.04	842.789	7566.138

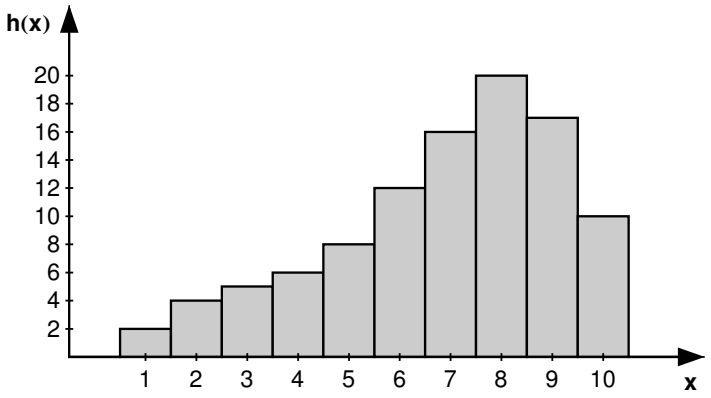
Somit lauten:

$$g_3 = \frac{\frac{1}{n}\sum_j^m h_j(x_j - \overline{x})^3}{\left[\frac{1}{n}\sum_j^m h_j(x_j - \overline{x})^2\right]^{\frac{3}{2}}} = \frac{8.4279}{5.2604^{\frac{3}{2}}} = \frac{8.4279}{12.065} = 0.699,$$

$$g_4 = \frac{\frac{1}{n}\sum_j^m h_j(x_j - \overline{x})^4}{\left[\frac{1}{n}\sum_j^m h_j(x_j - \overline{x})^2\right]^2} = \frac{75.6614}{5.2604^2} = \frac{75.6614}{27.6718} = 2.734,$$

$$\tilde{g}_4 = g_4 - 3 = -0.266.$$

b) *Gegeben ist folgende rechtssteile Häufigkeitsverteilung für das Merkmal X:*



Zur Berechnung der Kenngrößen Momentenkoeffizient der Schiefe, Kurtosis bzw. Kurtosis-Exzess wird folgende Tabelle aufgestellt mit $\overline{x} = 6.86$*:*

x_j	h_j	$(x_j - \overline{x})$	$h_j \cdot (x_j - \overline{x})^2$	$h_j \cdot (x_j - \overline{x})^3$	$h_j \cdot (x_j - \overline{x})^4$
1	2	-5.86	68.679	-402.460	2358.416
2	4	-4.86	94.478	-459.165	2231.542
3	5	-3.86	74.498	-287.562	1109.990
4	6	-2.86	49.078	-140.362	401.435
5	8	-1.86	27.677	-51.479	95.751
6	12	-0.86	8.875	-7.633	6.564
7	16	0.14	0.314	0.044	0.006
8	20	1.14	25.992	29.631	33.779
9	17	2.14	77.853	166.606	356.537
10	10	3.14	98.596	309.592	972.117
$\sum$	100		526.04	-842.789	7566.138

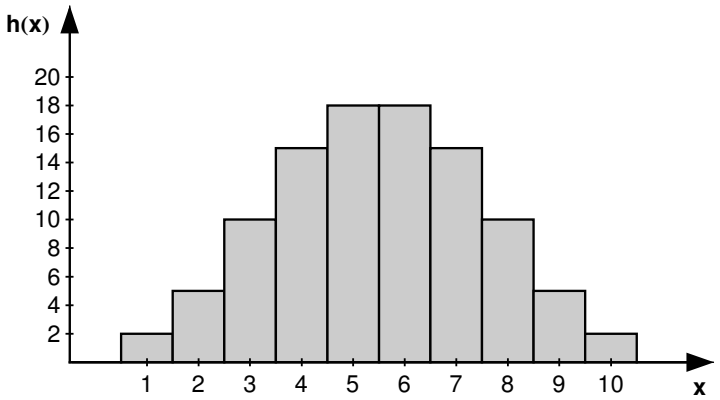
Somit lauten:

$$g_3 = \frac{\frac{1}{n}\sum_j^m h_j(x_j - \overline{x})^3}{\left[\frac{1}{n}\sum_j^m h_j(x_j - \overline{x})^2\right]^{\frac{3}{2}}} = \frac{-8.4279}{5.2604^{\frac{3}{2}}} = \frac{-8.4279}{12.065} = -0.699,$$

$$g_4 = \frac{\frac{1}{n}\sum_j^m h_j(x_j - \overline{x})^4}{\left[\frac{1}{n}\sum_j^m h_j(x_j - \overline{x})^2\right]^2} = \frac{75.6614}{5.2604^2} = \frac{75.6614}{27.6718} = 2.734,$$

$$\tilde{g}_4 = g_4 - 3 = -0.266.$$

c) *Gegeben ist folgende symmetrische Häufigkeitsverteilung für das Merkmal X mit $\overline{x} = 5.5$:*



x_j	h_j	$(x_j - \overline{x})$	$h_j \cdot (x_j - \overline{x})^2$	$h_j \cdot (x_j - \overline{x})^3$	$h_j \cdot (x_j - \overline{x})^4$
1	2	-4.5	40.50	-182.25	820.125
2	5	-3.5	61.25	-214.38	750.313
3	10	-2.5	62.50	-156.25	390.625
4	15	-1.5	33.75	-50.63	75.938
5	18	-0.5	4.50	-2.25	1.125
6	18	0.5	4.50	2.25	1.125
7	15	1.5	33.75	50.63	75.938
8	10	2.5	62.50	156.25	390.625
9	5	3.5	61.25	214.38	750.313
10	2	4.5	40.50	182.25	820.125
$\sum$	100		405	0	4076.252

$$g_3 = \frac{\frac{1}{n}\sum_j^m h_j(x_j - \overline{x})^3}{\left[\frac{1}{n}\sum_j^m h_j(x_j - \overline{x})^2\right]^{\frac{3}{2}}} = 0$$

$$g_4 = \frac{\frac{1}{n}\sum_j^m h_j(x_j - \overline{x})^4}{\left[\frac{1}{n}\sum_j^m h_j(x_j - \overline{x})^2\right]^2} = \frac{40.7625}{4.05^2} = \frac{40.7625}{16.4025} = 2.485$$

$$\tilde{g}_4 = g_4 - 3 = -0.515$$

Die Verteilungen in a) und b) sind spiegelverkehrt, so dass die Momentenkoeffizienten der Schiefe vom Betrag her gleich sind, jedoch verschiedene Vorzeichen aufweisen. Kurtosis bzw. Kurtosis-Exzess nehmen aufgrund der gleichen Wölbung bei beiden Verteilungen denselben Wert an. Die Verteilung in c) ist symmetrisch, so dass der Momentenkoeffizient der Schiefe dem Wert 0 *entspricht.*

2.5 Konzentrationsmessung

Nach der Charakterisierung der Verteilung eines metrischen Merkmals durch Mittelwert und Streuung ist auch von Interesse, ob sich die Summe der Merkmalswerte **gleichmäßig** auf die Merkmalsträger verteilt oder ob eine **Konzentration** vorliegt.

***Beispiel 2.5.1**:*

In einer Großstadt werden für 10 *Baubetriebe sowie für* 20 *Einzelhandelsgeschäfte die Jahresumsätze erfasst. Für die nach Umsatzstärke geordneten Baubetriebe ergibt sich:*

Betrieb Nr. i	1	2	3	4	5	6	7	8	9	10
Umsatz in Mill. €	0.7	0.7	0.7	0.95	0.95	1.5	1.5	8	10	15

Der Gesamtumsatz (= Summe der Merkmalswerte) der 10 *Baubetriebe beträgt* 40 *Mill. €.*

Für jedes Geschäft der Einzelhandelsgeschäften wird ein Umsatz von 2 *Mill. € ermittelt. Der Gesamtumsatz aller Geschäfte beträgt somit ebenfalls* 40 *Mill. €.*

Fast der gesamte Umsatz bei den Baubetrieben konzentriert sich auf drei große Baubetriebe, d.h. auf 30% *der Betriebe entfallen* 82.5% *des Gesamtumsatzes. Auf einen verhältnismäßig kleinen Anteil der Merkmalsträger entfällt ein relativ großer Anteil an der Summe der Merkmalswerte. Die übrigen* 70% *der (kleineren) Betriebe erbringen hingegen nur* 17.5% *des Gesamtumsatzes.*

Der Umsatz im Einzelhandel ist gleichmäßig auf die Betriebe „verteilt". Hier liegt keine Konzentration vor.

2.5.1 Die Lorenzkurve

Eine grafische Darstellung, anhand der die Stärke der Konzentration direkt abgelesen werden kann, ist die **Lorenzkurve**.

Gegeben sei ein nicht negatives metrisches Merkmal X mit den aufsteigend geordneten Merkmalsausprägungen oder Klassen x_j $(j = 1, ..., m)$ mit den absoluten Häufigkeiten h_j bzw. den relativen Häufigkeiten f_j.

Lorenzkurve

Für eine geordnete Reihe $x_1 \leq x_2 \leq ... \leq x_m$ der beobachteten Merkmalsausprägungen bzw. Klassen wird die **Lorenzkurve** als Polygonzug durch die Punkte

$$(0,0), (F_1, G_1), ..., (F_j, G_j), ..., (1,1)$$

angegeben ($j = 1, ..., m$). Dabei entspricht (F_m, G_m) dem Punkt (1,1).

$F_j = \sum_{j'=1}^{j} f_{j'}$ relative Summenhäufigkeit der j-ten Ausprägung,

$G_j = \sum_{j'=1}^{j} g_{j'}$ relative Merkmalssumme der j-ten Ausprägung mit $g_{j'} = \frac{x_{j'} h_{j'}}{\sum_{k=1}^{m} x_k h_k}$.

Mit F_j wird der Anteil der Merkmalsausprägungen bis zur Ausprägung bzw. Klasse j dargestellt, während G_j den Anteil der Summe der Beobachtungswerte bis zu dieser Ausprägung bzw. Klasse an der Gesamtsumme aller Beobachtungswerte bestimmt. Ein Punkt (F_j, G_j) gibt also an, dass auf $F_j \cdot 100\%$ der Untersuchungseinheiten $G_j \cdot 100\%$ des Gesamtbetrages aller Beobachtungswerte entfallen.

Beispiel 2.5.2:

Mit den Daten der Baubetrieben aus Beispiel 2.5.1 lässt sich folgende Arbeitstabelle aufstellen:

Umsatz in Mill.€ x_j	Anzahl der Baubetriebe h_j	Prozentsatz der Baubetriebe f_j	anteilige Umsatzsumme in Mill. € $x_j h_j$	relativer Umsatz $g_j = \frac{x_j h_j}{\sum_{k=1}^{m} x_k h_k}$	aufsumm. Betriebsanteile F_j	aufsumm. Umsatzanteile G_j
0.7	3	0.3	2.1	0.0525	0.3	0.0525
0.95	2	0.2	1.9	0.0475	0.5	0.1
1.5	2	0.2	3	0.075	0.7	0.175
8	1	0.1	8	0.2	0.8	0.375
10	1	0.1	10	0.25	0.9	0.625
15	1	0.1	15	0.375	1.0	1.0
$\sum$	10	1.0	40	1.0		

Die zugehörige Lorenzkurve gibt an, wie viel % des Gesamtumsatzes ($G_j \cdot 100\%$) auf wie viel % der Baubetriebe ($F_j \cdot 100\%$) entfallen.

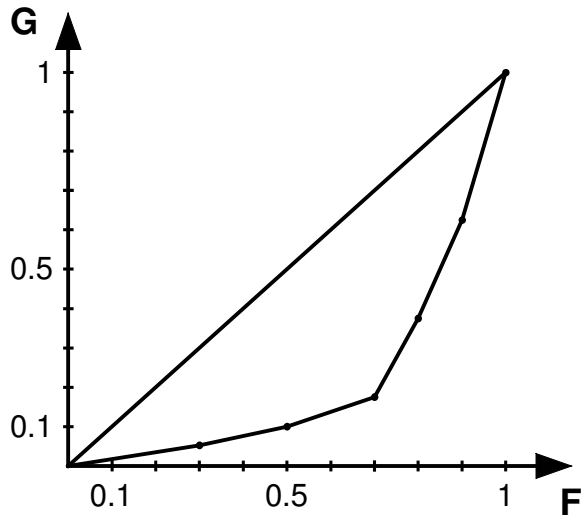


Abbildung 2.5.1: *Lorenzkurve zu Beispiel 2.5.1 (Daten der Baubetriebe)*

Für die Daten der Einzelhandelsbetriebe aus Beispiel 2.5.1 gilt $F_j = G_j$ für alle j. Es liegt somit keine Konzentration vor und die Lorenzkurve ergibt sich als Ursprungsgerade mit der Steigung 1.

Ausgehend von dem Beispiel kann eine Lorenzkurve wie folgt interpretiert werden: Liegt keine Konzentration vor, so ergibt sich als Lorenzkurve eine Ursprungsgerade mit der Steigung 1. Je weiter eine Lorenzkurve von dieser Geraden nach unten abweicht (d.h. je weiter sie „durchhängt“), desto stärker ist die Konzentration. Die Lorenzkurve wächst monoton und ist konvex, d.h. es liegt eine Wölbung nach unten vor.

Der statistische Konzentrationsbegriff erfasst keine wirtschaftliche Konzentrationsphänomene, die sich auf die Verringerung der Anzahl der Einheiten beziehen. Wenn z.B. in einer Branche 200 Betriebe existieren, die alle den gleichen Jahresumsatz tätigen, und 10 Jahre später nur noch 5 Betriebe existieren, die wiederum alle den gleichen (entsprechend größeren) Jahresumsatz haben, dann ergibt sich in beiden Fällen keine statistische Konzentration, obwohl volkswirtschaftlich durchaus Konzentration vorliegt.

2.5.2 Das Lorenzsche Konzentrationsmaß (Gini-Koeffizient)

Im vorherigen Abschnitt wurde gezeigt, dass die Konzentration umso höher ist, je weiter die Lorenzkurve nach unten durchhängt. Als Maß der Konzentration bietet sich daher an, die Fläche zwischen der Diagonalen und der Lorenzkurve zu der Gesamtfläche zwischen Diagonalen und Abszisse in Relation zu setzen. Diese Maßzahl wird als **Lorenzsches Konzentrationsmaß (LKM)** bzw. **Gini-Koeffizient** bezeichnet. Zu beachten ist, dass die Fläche zwischen der Diagonalen und der Abszisse den Wert 0.5 annimmt.

$$\begin{aligned} LKM &= \frac{\text{Fläche zwischen Diagonale und Lorenzkurve}}{\text{Fläche zwischen Diagonale und Abszisse}} \\ &= 2 \cdot \text{Fläche zwischen Diagonale und Lorenzkurve} \end{aligned}$$

Anhand der Abbildung 2.5.**??** wird verdeutlicht, dass zur Berechnung des Lorenzschen Konzentrationsmaßes, dass dem Wert $2K$ entspricht, die Flächen der Trapeze K_j herangezogen werden können.

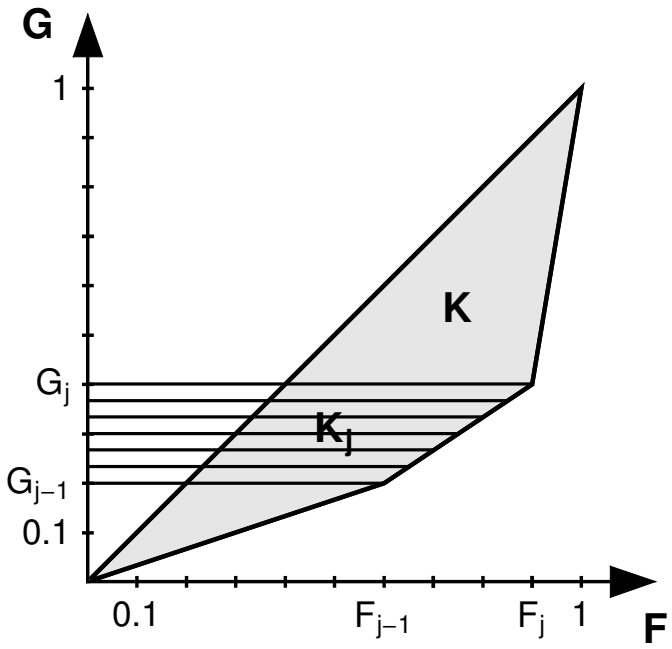


Abbildung 2.5.2: *Berechnung des Konzentrationsmaßes mittels Trapeze*

Die Fläche K setzt sich somit aus der Summe der Trapezflächen K_j abzüglich der Fläche 0.5 des oberen Dreiecks zusammen.[3]

$$K = \sum_{j=1}^{m} K_j - 0.5$$

mit

$$\begin{aligned} K_j &= \frac{F_{j-1} + F_j}{2} \cdot (G_j - G_{j-1}) \\ &= \frac{F_{j-1} + F_j}{2} \cdot \frac{1}{\sum_{k=1}^{m} x_k h_k} \cdot \Big(\sum_{j'=1}^{j} x_{j'} h_{j'} - \sum_{j'=1}^{j-1} x_{j'} h_{j'} \Big) \\ &= \frac{F_{j-1} + F_j}{2} \cdot \frac{x_j h_j}{\sum_{k=1}^{m} x_k h_k} \ . \end{aligned}$$

Lorenzsche Konzentrationsmaß (Gini-Koeffizient)

Das **Lorenzsche Konzentrationsmaß (LKM)**, welches auch **Gini-Koeffizient** genannt wird, nimmt Werte in dem Bereich $[0; \frac{n-1}{n}]$ an und berechnet sich zu

$$\text{LKM} = 2K = \Big[\sum_{j=1}^{m} (F_{j-1} + F_j) \cdot g_j \Big] - 1$$

mit

$$\begin{aligned} g_j &= \frac{x_j h_j}{\sum_{k=1}^{m} x_k h_k} \ , \\ F_j &= \sum_{j'=1}^{j} f_{j'} \text{ für } j = 1, ..., m \text{ und } F_0 = 0. \end{aligned}$$

Bei fehlender Konzentration nimmt das Lorenzsche Konzentrationsmaß den Wert 0 an und bei vollständiger statistischer Konzentration, d.h. eine einzige statistische Einheit trägt die gesamte Merkmalssumme, den Wert $\frac{n-1}{n}$. Die maximale Konzentration hängt somit von der Anzahl der Beobachtungen ab. Um ein Maß zu erhalten, das im Falle vollständiger statistischer Konzentration den Wert 1 annimmt, wird das Lorenzsche Konzentrationsmaß normiert.

[3] Die Fläche eines Trapezes mit den parallelen Seiten a und b und der Höhe h berechnet sich zu $\frac{a+b}{2} \cdot h$.

normierte Lorenzsche Konzentrationsmaß

Das **normierte Lorenzsche Konzentrationsmaß** wird mit $\mathbf{LKM}_{norm}$ bezeichnet und berechnet sich zu:

$$\text{LKM}_{norm} = \frac{n}{n-1}\text{LKM}.$$

Beispiel 2.5.3:

Berechnung des Lorenzschen Konzentrationsmaßes der Daten aus Beispiel 2.5.1.

$$\begin{aligned}
\text{LKM} &= [(0+0.3)\cdot 0.0525 + (0.3+0.5)\cdot 0.0475 \\
&\quad +(0.5+0.7)\cdot 0.075 + (0.7+0.8)\cdot 0.2 \\
&\quad +(0.8+0.9)\cdot 0.25 + (0.9+1)\cdot 0.375] - 1 \\
&= [0.01575 + 0.038 + 0.09 + 0.3 + 0.425 + 0.7125] - 1 \\
&= 1.58123\text{-}1 \\
&= 0.58
\end{aligned}$$

$$\text{LKM}_{norm} = \tfrac{10}{9}\cdot 0.58123 = 0.65$$

Bei der Interpretation des Lorenzschen Konzentrationsmaßes muss beachtet werden, dass verschiedene Konzentrationsstrukturen zum gleichen LKM führen können.

Beispiel 2.5.4:

In zwei Regionen A und B werden jeweils n Bauernhöfe bezüglich ihrer landwirtschaftlich genutzten Fläche untersucht. Folgende Sachverhalte liegen vor:

1. *In* 50% *der Bauernhöfe* (G_j) *der Region A werden nur* 1% *der Nutzfläche* (F_j) *bewirtschaftet. Die anderen* 50% *haben alle die gleiche Fläche; zusammen* 99% *der Nutzfläche.*

2. *In* 1% *der Bauernhöfe werden* 50% *der Nutzfläche bewirtschaftet. Die übrigen* 99% *der Bauernhöfe haben alle die gleiche Fläche zusammen; die anderen* 50% *der Nutzfläche.*

Das Lorenzsche Konzentrationsmaß beträgt in beiden Fällen

$$\text{LKM} = 0.49.$$

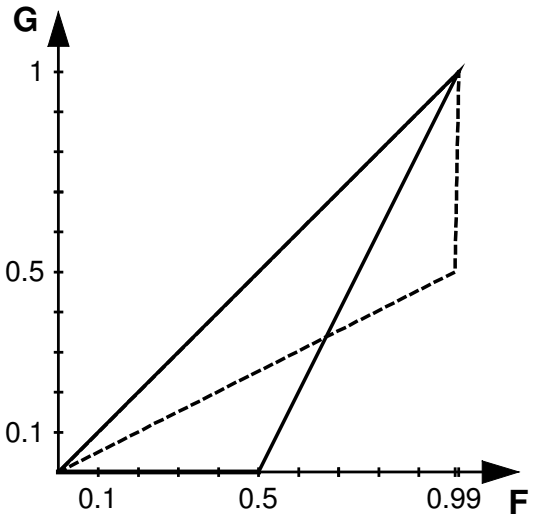


Abbildung 2.5.3: *Lorenzkurven zu Beispiel 2.5.4*

3 Zusammenhänge zwischen Merkmalen

3.1 Häufigkeitsverteilungen zweier Merkmale

Bisher wurde bei der statistischen Analyse nur ein einzelnes Merkmal betrachtet. Oft werden in einer Erhebung jedoch mehrere Merkmale erfasst.

Beispiel 3.1.1:

Für Volkszählungen werden u.a. die Merkmale Geschlecht, Alter, Beruf und Familienstand erhoben.

Eine Untersuchung des Bremsverhaltens von Kraftfahrzeugen betrachtet die Merkmale Geschwindigkeit und Länge des Bremsweges.

Für die Ermittlung des Body-Mass-Index (BMI) werden die gemeinsam auftretenden Merkmale Körpergröße und Körpergewicht benötigt.

Die für eine Erhebung erfassten Merkmale können entweder einzeln oder partiell gemeinsam weiterverarbeitet werden. Dabei stellt sich die Frage:

Gibt es zwischen gemeinsam auftretenden Merkmalen einen Zusammenhang?

Beispiel 3.1.2:

Das Wirtschaftswachstum eines Landes wird von der Arbeitslosenrate beeinträchtigt.

Die Bevölkerungsentwicklung einer Region wird vom vorherrschenden Arbeitsmarkt beeinflusst.

Der Ernteertrag je ha landwirtschaftliche Nutzfläche hängt mit dem Düngemitteleinsatz zusammen.

Die gemeinsame Aufbereitung und Auswertung von zwei (oder mehr) Merkmalen dient vor allem dem Zweck, festzustellen, ob ein Zusammenhang zwischen den Merkmalen besteht, wie ausgeprägt und von welcher Art dieser ist.

Im Bereich der Naturwissenschaften können Zusammenhänge zwischen Merkmalen in den meisten Fällen durch eine (eindeutige) Funktion ausgedrückt werden, da sich in ihnen Naturgesetze widerspiegeln.

Beispiel 3.1.3:

Zwischen den Größen

s = zurückgelegte Entfernung bzw. Wegstrecke,
v = Geschwindigkeit, die während der Messzeit konstant sein soll,
t = Zeit

besteht die Beziehung $s = v \cdot t$.

Werden für verschiedene Vorgänge s,v und t gemessen, dann muss zwischen den Werten immer diese Beziehung bestehen. Abweichungen können nur durch Mess- bzw. Beobachtungsfehler vorkommen.

In den Wirtschaftswissenschaften und anderen Anwendungsgebieten der Statistik (z.B. der Medizin oder der Biologie) lassen sich Zusammenhänge nicht immer eindeutig durch eine Funktion beschreiben.

Mittels sogenannter Streudiagramme lässt sich oft ein erster Eindruck über die Art und die Stärke des Zusammenhangs gewinnen.

Streudiagramm

Eine grafische Darstellung von Wertepaaren metrischer Merkmale (x_i, y_i) als Punkte in einem kartesischen Koordinatensystem heißt **Streuungsdiagramm** oder kurz **Streudiagramm**.

Beispiel 3.1.4:

Von 40 *Personen wurde die Körpergröße X (in cm) und das Körpergewicht Y (in kg) gemessen. Für jede Person ergibt sich ein Beobachtungspaar* (x_i, y_i). *Die Beobachtung* x_i *gibt die Körpergröße und die Beobachtung* y_i *das Körpergewicht der Person Nummer i an.*

Körpergröße x_i in cm				Körpergewicht y_i in kg			
x_i	y_i	x_i	y_i	x_i	y_i	x_i	y_i
150	48	160	63	170	80	180	75
150	51	160	68	170	85	180	85
150	55	160	75	172	68	181	89
150	58	165	55	175	64	185	80
152	63	165	62	175	68	186	74
153	57	165	73	175	84	186	85
155	50	166	79	176	73	186	90
155	63	167	65	176	77	189	77
157	70	170	60	178	93	189	85
160	58	170	75	180	65	189	96

Die Wertepaare (x_i, y_i) können in einem Streudiagramm grafisch dargestellt werden.

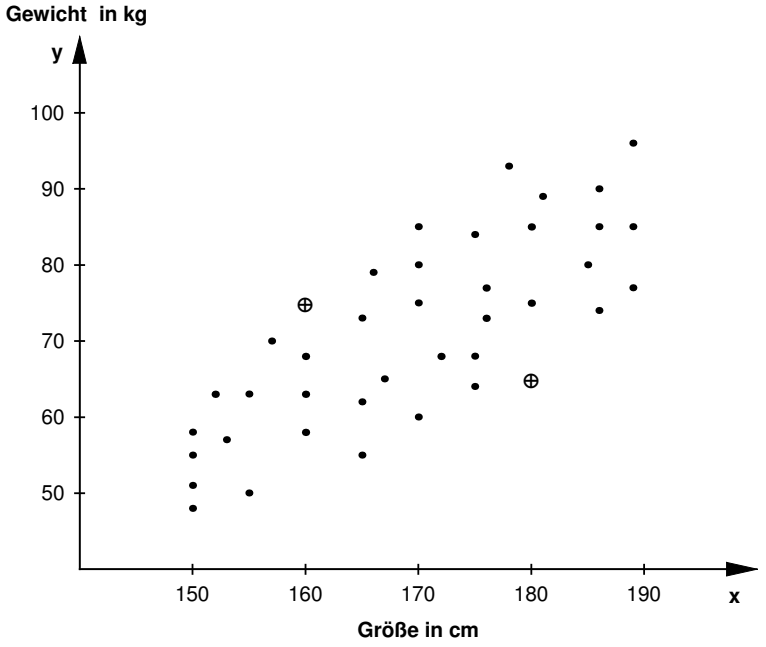


Abbildung 3.1.1: *Streudiagramm der Beobachtungspaare (Körpergröße,Körpergewicht)*

Abbildung 3.1.1 macht deutlich, dass zwischen Körpergröße und Körpergewicht der untersuchten Personen kein eindeutiger Zusammenhang besteht. Der Zusammenhang kann nicht direkt mittels einer einfachen Funktion dargestellt werden. Andererseits ist zu erkennen, dass größere

Personen im Schnitt auch schwerer sind. Körpergröße und Körpergewicht hängen offensichtlich voneinander ab, wobei diese Beziehung aber nur tendenziell gilt. Bei großen Personen wird im Durchschnitt ein höheres Gewicht gemessen als bei kleineren. Im Einzelfall muss diese Aussage nicht zutreffen, wie ein Vergleich der beiden in Abbildung 3.1.1 besonders kenntlich gemachten Punkte zeigt. Der eine Punkt kennzeichnet eine Person, die 180 *cm groß und* 65 *kg schwer ist und der andere eine, die nur* 160 *cm groß, aber* 75 *kg schwer ist.*

Das Beispiel zeigt deutlich das wichtigste Problem bei der Betrachtung des gemeinsamen Auftretens mehrerer Merkmale. Zwischen den Ausprägungen der Merkmale besteht ein tendenzieller Zusammenhang. Dieser lässt sich aber für die Beobachtungswerte nicht auf eine eindeutige Form bringen. Folgende Fragestellungen sind von Interesse:

- Wie ausgeprägt ist ein Zusammenhang? Tritt er sehr deutlich hervor, ist er nur schwach oder ist gar kein Zusammenhang vorhanden?
- Von welchem Typ ist ein Zusammenhang oder die durchschnittliche Tendenz eines Zusammenhangs? Ist er linear, quadratisch oder von einer anderen Form?

Wie sich Unterschiede in der Intensität oder Ausgeprägtheit eines Zusammenhangs äußern, machen die beiden folgenden Zeichnungen deutlich.

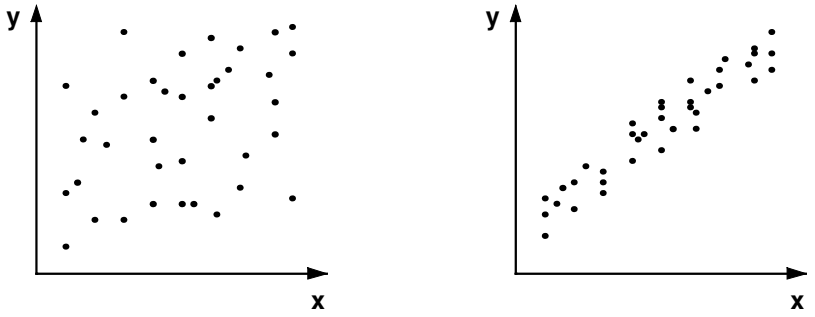


Abbildung 3.1.2: *Streudiagramme*

Die rechte Darstellung in Abbildung 3.1.2 deutet auf einen ziemlich ausgeprägten Zusammenhang zwischen den beiden Merkmalen X und Y hin. Die linke Zeichnung lässt dagegen keine Anzeichen eines Zusammenhangs erkennen.

3.1.1 Zweidimensionale Häufigkeitstabellen

Es sei X ein Merkmal mit den Ausprägungen x_j $(j = 1, ..., m)$ und Y ein Merkmal mit den Ausprägungen y_k $(k = 1, ..., r)$. Beide Merkmale können gemeinsam auftreten und es interessieren die Merkmalskombinationen (x_j, y_k). Da jede Ausprägung von X mit jeder Ausprägung von Y kombiniert auftreten kann, gibt es insgesamt $m \cdot r$ mögliche Kombinationen der Merkmalsausprägungen.

Werden in einer statistischen Analyse n statistische Einheiten untersucht, dann ergeben sich n Beobachtungspaare (x_i, y_i) $i = 1, ..., n$ in Form einer ungeordneten Reihe. Eine Ordnung dieser Reihe ist problematisch, da eine gleichzeitige Ordnung nach beiden Merkmalen nicht möglich ist. Daher werden die Paare zunächst nach den Ausprägungen eines Merkmals und anschließend für übereinstimmende Ausprägungen dieses Merkmals nach den Ausprägungen des anderen Merkmals geordnet.

lexikografische Ordnung

Ein derartiges Ordnungsprinzip heißt **lexikografische Ordnung**.

***Beispiel 3.1.5**:*

Von 10 *Studenten wurde die Mathematiknote und die Statistiknote erfasst (Mathematiknote, Statistiknote).*

(4,3) (2,3) (2,1) (4,5) (1,2) (2,4) (3,3) (4,3) (4,4) (3,5)

Mittels der lexikografischen Ordnung wird hier zuerst nach der Mathematiknote und anschließend nach der Statistiknote geordnet.

(1,2) (2,1) (2,3) (2,4) (3,3) (3,5) (4,3) (4,3) (4,4) (4,5)

Die absoluten Häufigkeiten, mit denen die Merkmalskombinationen auftreten, werden mit $h_{jk} = h(x_j, y_k)$ und die relativen Häufigkeiten mit $f_{jk} = f(x_j, y_k)$ bezeichnet. Werden n Merkmalskombinationen untersucht, so gilt

$$f_{jk} = \frac{h_{jk}}{n} \quad \text{und} \quad \sum_{j=1}^{m} \sum_{k=1}^{r} h_{jk} = n.$$

zweidimensionale Häufigkeitsverteilung

Die Gesamtheit aller Kombinationen von Merkmalsausprägungen und der dazugehörigen absoluten und relative Häufigkeiten ergibt die **zweidimensionale Häufigkeitsverteilung**. Ihre tabellarische Darstellung

heißt **zweidimensionale Häufigkeitstabelle** und hat allgemein folgende Form, wobei statt der absoluten Häufigkeiten auch relative Häufigkeiten verwendet werden können:

zweidimensionale Häufigkeitstabelle

	y_1	y_2	y_3	...	y_r
x_1	h_{11}	h_{12}	h_{13}	...	h_{1r}
x_2	h_{21}	h_{22}	h_{23}	...	h_{2r}
...	...	...	...	...	...
x_m	h_{m1}	h_{m2}	h_{m3}	...	h_{mr}

An dieser Stelle sind die Ausprägungen des Merkmals X den Zeilen und die Ausprägungen des Merkmals Y den Spalten zugeordnet.

Die Häufigkeitstabelle zweier metrischer oder ordinalskalierter Merkmale wird auch **Korrelationstabelle** genannt.

Korrelationstabelle

Die tabellarische Darstellung der Häufigkeitsverteilung zweier nominalskalierter Merkmale heisst auch **Kontingenztabelle**.

Kontingenztabelle

***Beispiel 3.1.6**:*

Die folgende Häufigkeitstabelle (Korrelationstabelle) enthält das Ergebnis einer Erhebung der Englisch- und Mathematiknote von 70 *Studienanfängern.*

Mathematiknote	Englischnote 1	2	3	4	5	$\sum$
1	1	2	3	-	-	6
2	3	5	8	2	1	19
3	2	4	12	9	2	29
4	2	2	2	6	1	13
5	-	-	2	1	-	3
$\sum$	8	13	27	18	4	70

3.1.2 Grafische Darstellung zweidimensionaler Verteilungen

Ein wichtiger Gesichtspunkt, der bei der grafischen Darstellung zweidimensionaler Verteilungen zu beachten ist, ist die Übersichtlichkeit. Inwieweit die Übersichtlichkeit gewährleistet werden kann, hängt nicht nur von der Darstellungsform sondern auch von der Struktur des Datensatzes ab. Des Weiteren wird die Wahl der Darstellungsform von der zugrundeliegenden Zielsetzung bestimmt.
Im Folgenden sind für das Beispiel 3.1.6 und für die Aufgabe 3.2 zwei verschiedene Darstellungsformen gewählt. Eine dreidimensionale Grafik und ein zweidimensionales Säulendiagramm, in dem die Säulen nach dem zweiten Merkmal unterteilt sind.

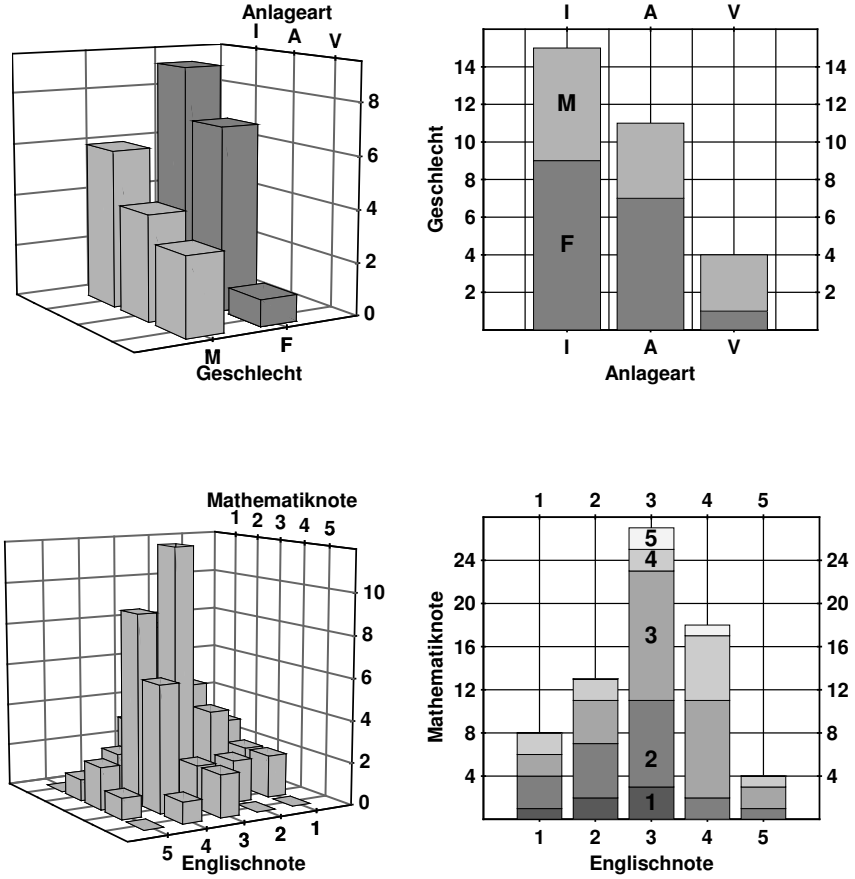


Abbildung 3.1.3: *Häufigkeitsverteilungen der Daten aus Beispiel 3.1.6 und Aufgabe 3.2*

Anhand der Abbildung 3.1.3 ist zu erkennen, dass 3D-Säulendiagramme unübersichtlich sein können bzw. aufgrund der vielen Merkmalsausprä-

gungen nicht alle Häufigkeiten von dem Diagramm abgelesen werden können, da die Säulen sich gegenseitig verdecken (Beispiel Noten). In diesem Fall bietet sich ein zweidimensionales Säulendiagramm an, wobei auch hier zu erkennen ist, dass aufgrund der vielen Unterteilungen direkte Vergleiche der einzelnen Segmente schwer durchzuführen sind. Liegen für beide Merkmale viele Ausprägungen vor, wird oft für jede Merkmalsausprägung von einem der beiden Merkmale ein separates Säulendiagramm erstellt.

Eine wichtige Darstellungsmöglichkeit gemeinsam auftretender Merkmale sind Streudiagramme (s. Abbildungen 3.1.1 und 3.1.2).

3.1.3 Randverteilungen

Liegt eine zweidimensionale Häufigkeitsverteilung vor, kann neben der gemeinsamen Verteilung der beiden Merkmale auch die Verteilung nur eines Merkmals von Interesse sein. Das jeweils andere Merkmal wird dabei nicht berücksichtigt.

Randverteilung (marginale Verteilung)

Die Verteilung nur eines Merkmals einer zweidimensionalen Häufigkeitsverteilung, wobei das andere Merkmal unberücksichtigt bleibt, heißt **Randverteilung** oder **marginale Verteilung**. Durch Berechnung der Zeilen- bzw. Spaltensummen ergeben sich die **absoluten** bzw. **relativen Randverteilungen** der beiden Merkmale X und Y, $\boldsymbol{h_{j.}}$, $\boldsymbol{h_{.k}}$ bzw. $\boldsymbol{f_{j.}}$, $\boldsymbol{f_{.k}}$, wobei der Punkt kennzeichnet, über welches Merkmal aufsummiert wird.

absolute Randverteilung

Für die **absoluten Randverteilungen** gilt:

$$h_{j.} = \sum_{k=1}^{r} h_{jk} \quad \text{Randverteilung für X,}$$

$$h_{.k} = \sum_{j=1}^{m} h_{jk} \quad \text{Randverteilung für Y.}$$

relative Randverteilung

Für die **relativen Randverteilungen** gilt:

$$f_{j.} = \sum_{k=1}^{r} f_{jk} \quad \text{Randverteilung für X,}$$

$$f_{.k} = \sum_{j=1}^{m} f_{jk} \quad \text{Randverteilung für Y.}$$

Zweidimensionale Häufigkeitstabellen können somit wie folgt angegeben werden:

	y_1	y_2	...	y_r	$\sum$
x_1	h_{11}	h_{12}	...	h_{1r}	$h_{1.}$
x_2	h_{21}	h_{22}	...	h_{2r}	$h_{2.}$
...	...	...	...	...	...
x_m	h_{m1}	h_{m2}	...	h_{mr}	$h_{m.}$
$\sum$	$h_{.1}$	$h_{.2}$	...	$h_{.r}$	n

	y_1	y_2	...	y_r	$\sum$
x_1	f_{11}	f_{12}	...	f_{1r}	$f_{1.}$
x_2	f_{21}	f_{22}	...	f_{2r}	$f_{2.}$
...	...	...	...	...	...
x_m	f_{m1}	f_{m2}	...	f_{mr}	$f_{m.}$
$\sum$	$f_{.1}$	$f_{.2}$	...	$f_{.r}$	1

***Beispiel 3.1.7**:*

Von 200 *Personen wurde die Körpergröße und das Körpergewicht gemessen. Durch Addition aller Werte der Zeilen bzw. der Spalten ergeben sich die angegebenen relativen Randverteilungen für das Körpergewicht bzw. die Körpergröße.*

Körpergewicht in kg	Körpergröße in cm					$\sum$
	(150;160]	(160;170]	(170;180]	(180;190]	(190;200]	
(50;60]	0.015	0.025	0.04	0.015	0.005	0.1
(60;70]	0.02	0.125	0.2	0.05	0.005	0.4
(70;80]	0.01	0.05	0.1	0.03	0.01	0.2
(80;90]	0.005	0.04	0.05	0.08	0.025	0.2
(90;100]	0	0.01	0.01	0.025	0.055	0.1
$\sum$	0.05	0.25	0.4	0.2	0.1	1

3.1.4 Bedingte Verteilungen

Eine bei zweidimensionalen Häufigkeitsverteilungen oft auftretende Frage ist die nach der Verteilung eines Merkmals für einen gegebenen Wert des anderen Merkmals. Aus allen gegebenen Merkmalskombinationen (x_j, y_k) werden nur noch die betrachtet, bei denen X oder Y einen vorgegebenen Wert annimmt. Für gegebenes x_j bzw. y_k wird dann die Häufigkeitsverteilung von Y bzw. X betrachtet.

bedingte Verteilung (konditionale Verteilung)

Gegeben sei die zweidimensionale relative Häufigkeitsverteilung der Merkmale X und Y. Die relative Häufigkeitsverteilung des Merkmals X, die sich für eine gegebene Ausprägung y_k des Merkmals Y ergibt, heißt **bedingte Verteilung** oder **konditionale Verteilung** von X für gegebenes y_k. Die Häufigkeiten der bedingten Verteilungen werden mit $f(x_j|Y = y_k)$ oder kurz $f(x_j|y_k)$ bezeichnet $(j = 1, ..., m)$. Die bedingte Verteilung von Y für gegebenes x_j ist analog definiert durch $f(y_k|x_j)$ $(k = 1, ..., r)$.

An dieser Stelle wird die bedingte Verteilung nur für die relativen Häufigkeiten betrachtet, da die bedingte absolute Häufigkeit $h(x_j|y_k)$ mit $h(x_j|y_k) = h_{jk}$ direkt aus der Tabelle abgelesen werden kann. Die bedingten (relativen) Häufigkeiten ergeben sich, indem die relativen Häufigkeiten der entsprechenden Zeilen bzw. Spalten der zweidimensionalen Häufigkeitsverteilung durch den zugehörigen Wert der Randverteilung dividiert werden.

$$f(x_j|y_k) = \frac{f_{jk}}{f_{.k}} = \frac{h_{jk}}{h_{.k}} \qquad f(y_k|x_j) = \frac{f_{jk}}{f_{j.}} = \frac{h_{jk}}{h_{j.}}$$

***Beispiel 3.1.8**:*

Für die zweidimensionale Verteilung in Beispiel 3.1.7 ergibt sich als bedingte Häufigkeitsverteilung für die Körpergröße der Personen mit einem Gewicht von 80 *bis* 90 *kg folgende Verteilung:*

Körpergröße in cm	(150;160]	(160;170]	(170;180]	(180;190]	(190;200]
$f(y_k\|x \in (80;90])$	0.025	0.2	0.25	0.4	0.125

Die bedingte Verteilung des Körpergewichtes für Personen, die von 170 bis 180 cm groß sind, lautet:

Körpergewicht in kg	(50;60]	(60;70]	(70;80]	(80;90]	(90;100]
$f(x_j\|y \in (170;180])$	0.1	0.5	0.25	0.125	0.025

3.1.5 Kenngrößen zweidimensionaler Häufigkeitsverteilungen

bedingte Mittelwerte
bedingte Streuungen

Für zweidimensionale Verteilungen können Lage- und Streuungsmaße für die Randverteilungen und für die bedingten Verteilungen bestimmt werden. Im letzteren Fall wird von **bedingten Mittelwerten** und **bedingten Streuungen** gesprochen.

***Beispiel 3.1.9**:*

Es wird auf die zweidimensionale Verteilung aus Beispiel 3.1.7 zurückgegriffen. Die bedingten Mittelwerte für das Körpergewicht lauten:

Körpergröße in cm	(150;160]	(160;170]	(170;180]	(180;190]	(190;200]
bedingter Mittelwert $\overline{x}_j$	66	70.4	69.75	77.5	87

Der bedingte Mittelwert für die Klasse $(150; 160]$ *berechnet sich wie folgt:*

$$\overline{x} = \frac{3 \cdot 55 + 4 \cdot 65 + 2 \cdot 75 + 1 \cdot 85 + 0 \cdot 95}{10} = 66.$$

Außer dem arithmetischen Mittel können auch der Median und der Modus für die bedingten Verteilungen ermittelt werden. Auf Einzelheiten dazu wird hier nicht weiter eingegangen, da sich keine neuen Aspekte ergeben.

Ein spezielles Maß für zweidimensionale Verteilungen quantitativer Merkmale ist die **Kovarianz**.

Kovarianz

Die **Kovarianz** $\mathrm{Cov}(X, Y)$ der gemeinsamen Verteilung der quantitativen Merkmale X und Y ist ein Maß für die gemeinsame Streuung der beiden Merkmale. Das Vorzeichen der Kovarianz gibt die Richtung des vorliegenden Zusammenhangs an.

Je nachdem, ob für die Berechnung der Kovarianz die absoluten oder die relativen Häufigkeiten verwendet werden, ergeben sich für die Bestimmung der Kovarianz folgende Formeln:

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \frac{1}{n}\sum_{j=1}^{m}\sum_{k=1}^{r}(x_j - \overline{x})(y_k - \overline{y}) \cdot h_{jk}, \\ \mathrm{Cov}(X,Y) &= \sum_{j=1}^{m}\sum_{k=1}^{r}(x_j - \overline{x})(y_k - \overline{y}) \cdot f_{jk}. \end{aligned}$$

Werden in den Summanden die Produkte $(x_j - \overline{x})(y_k - \overline{y})$ ausmultipliziert, dann ergeben sich folgende Formeln, welche die Berechnung der Kovarianz in vielen Fällen erleichtern:

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \frac{1}{n}\sum_{j=1}^{m}\sum_{k=1}^{r} x_j y_k h_{jk} - \overline{x}\ \overline{y}, \\ \mathrm{Cov}(X,Y) &= \sum_{j=1}^{m}\sum_{k=1}^{r} x_j y_k f_{jk} - \overline{x}\ \overline{y}. \end{aligned}$$

Für die Berechnung der Kovarianz unter Verwendung relativer Häufigkeiten ist die Umformung der Formel ausführlich dargestellt.

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \sum_{j=1}^{m}\sum_{k=1}^{r}(x_j - \overline{x})(y_k - \overline{y}) \cdot f_{jk} \\ &= \sum_{j=1}^{m}\sum_{k=1}^{r}[x_j y_k - \overline{x} y_k - x_j \overline{y} + \overline{x}\ \overline{y}] \cdot f_{jk} \\ &= \sum_{j}\sum_{k} x_j y_k f_{jk} - \overline{x}\sum_{k} y_k \sum_{j} f_{jk} - \overline{y}\sum_{j} x_j \sum_{k} f_{jk} + \overline{x}\ \overline{y}\sum_{j}\sum_{k} f_{jk} \\ &= \sum_{j}\sum_{k} x_j y_k f_{jk} - \overline{x}\sum_{k} y_k f_k - \overline{y}\sum_{j} x_j f_j + \overline{x}\ \overline{y} \\ &= \sum_{j}\sum_{k} x_j y_k f_{jk} - \overline{x}\ \overline{y} - \overline{x}\ \overline{y} + \overline{x}\ \overline{y} \\ &= \sum_{j}\sum_{k} x_j y_k f_{jk} - \overline{x}\ \overline{y} \end{aligned}$$

Entsprechendes ergibt sich bei Umformung der Formel mit den absoluten Häufigkeiten.

***Beispiel 3.1.10**:*

Gegeben ist die folgende zweidimensionale Verteilung:

	$y_1 = 1$	$y_2 = 2$	$y_3 = 3$	$y_4 = 4$	$\sum$
$x_1 = 2$	3	9	2	1	15
$x_2 = 4$	8	7	2	3	20
$x_3 = 6$	4	9	1	1	15
$\sum$	15	25	5	5	50

Es ist

$$\overline{x} = \frac{2 \cdot 15 + 4 \cdot 20 + 6 \cdot 15}{50} = 4 \text{ und } \overline{y} = 2.$$

Für die Kovarianz ergibt sich:

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= \frac{1}{n}\sum_{j=1}^{3}\sum_{k=1}^{4}(x_j - \overline{x})(y_k - \overline{y}) \cdot h_{jk} \\
&= \frac{1}{50}[(-2)\cdot(-1)\cdot 3 + (-2)\cdot 0 \cdot 9 + (-2)\cdot 1 \cdot 2 + (-2)\cdot 2 \cdot 1 \\
&\quad + 0\cdot(-1)\cdot 8 + 0\cdot 0 \cdot 7 + 0 \cdot 1 \cdot 2 + 0 \cdot 2 \cdot 3 + 2\cdot(-1)\cdot 4 + 2 \cdot 0 \cdot 9 \\
&\quad + 2 \cdot 1 \cdot 1 + 2 \cdot 2 \cdot 1] \\
&= \frac{1}{50}[6 - 4 - 4 - 8 + 2 + 4] = \frac{1}{50} \cdot (-4) = -\frac{4}{50} = -0.08.
\end{aligned}$$

Unter Anwendung der zweiten angegebenen Formel gilt:

$$\begin{aligned}
\mathrm{Cov}(X,Y) &= \frac{1}{n}\sum_{j=1}^{3}\sum_{k=1}^{4} x_j y_k h_{jk} - \overline{x}\,\overline{y} \\
&= \frac{1}{50}[(2\cdot 1\cdot 3 + 2\cdot 2\cdot 9 + 2\cdot 3\cdot 2 + 2\cdot 4\cdot 1 + 4\cdot 1\cdot 8 + 4\cdot 2\cdot 7 + 4\cdot 3\cdot 2 \\
&\quad + 4\cdot 4\cdot 3 + 6\cdot 1\cdot 4 + 6\cdot 2\cdot 9 + 6\cdot 3\cdot 1 + 6\cdot 4\cdot 1)] - 4\cdot 2 \\
&= \frac{1}{50}(6 + 36 + 12 + 8 + 32 + 56 + 24 + 48 + 24 + 108 + 18 + 24) - 8 \\
&= \frac{1}{50}\cdot 396 - 8 = 7.92 - 8 = -0.08.
\end{aligned}$$

3.2 Unabhängige und abhängige Merkmale

3.2.1 Empirische Unabhängigkeit von Merkmalen

abhängige Merkmale

unabhängige Merkmale

Für **abhängige Merkmale** gilt, dass die Werte bzw. die Gestalt der bedingten Verteilung des einen Merkmals davon abhängt, welchen Wert bzw. welche Ausprägung das andere Merkmal annimmt. Für **unabhängige Merkmale** gilt, dass für alle Ausprägungen des einen Merkmals die gleiche bedingte Verteilung des anderen Merkmals vorliegt. Die bedingten Verteilungen eines Merkmals dürfen nicht davon abhängen, welchen Wert das andere Merkmal annimmt.

Beispiel 3.2.1:

Von 100 *Studenten der Wirtschaftswissenschaften wurden die Statistiknote und das Geschlecht erhoben. Zum Vergleich wurde sowohl die absolute als auch die relative Häufigkeitstabelle aufgestellt.*

	Statistiknote					
	1	2	3	4	5	$\sum$
w	6	12	24	12	6	60
m	4	8	16	8	4	40
$\sum$	10	20	40	20	10	100

	Statistiknote					
	1	2	3	4	5	$\sum$
w	0.06	0.12	0.24	0.12	0.06	0.6
m	0.04	0.08	0.16	0.08	0.04	0.4
$\sum$	0.1	0.2	0.4	0.2	0.1	1

Anhand der obigen Tabellen ist nicht unmittelbar zu erkennen, ob Abhängigkeit vorliegt. Werden nun die bedingten Verteilungen bei gegebenem Geschlecht berechnet, ergibt sich die untere linke Tabelle. In der unteren rechten Tabelle sind die bedingten Verteilungen bei gegebener Statistiknote aufgeführt.

	Statistiknote					
	1	2	3	4	5	$\sum$
w	0.1	0.2	0.4	0.2	0.1	1
m	0.1	0.2	0.4	0.2	0.1	1

	Statistiknote				
	1	2	3	4	5
w	0.6	0.6	0.6	0.6	0.6
m	0.4	0.4	0.4	0.4	0.4
$\sum$	1	1	1	1	1

Die bedingten Verteilungen (hier: gleich den Randsummen der relativen Häufigkeitsverteilung) stimmen innerhalb jeder Tabelle überein, d.h. die Verteilung der Statistiknote hängt nicht von der Ausprägung des Merkmals „Geschlecht“ ab bzw. die Verteilung des Geschlechts hängt nicht von der Ausprägung des Merkmals „Statistiknote“ ab. Die beiden Merkmale sind somit voneinander unabhängig.

Unabhängigkeit kann auch erkannt werden, wenn alle relativen oder absoluten Häufigkeitsverteilungen, nach den einzelnen Ausprägungen eines Merkmals getrennt betrachtet, durch Division oder Multiplikation in dieselbe Verteilung transformiert werden können (d.h. die Zeilen und Spalten sind untereinander ein Vielfaches). Die Häufigkeitstabelle besteht somit aus linear abhängigen Zeilen bzw. Spalten. Hier kann z.B. die Zeile des Geschlechts m der oberen Häufigkeitstabellen durch Multiplikation mit $\frac{3}{2}$ in die Zeile des Geschlechts w transformiert werden.

Das Beispiel macht deutlich, dass es bei einer großen Häufigkeitstabelle einfacher ist, für die Untersuchung der Unabhängigkeit die bedingten Verteilungen der **relativen** Häufigkeiten heranzuziehen, da hier die Unabhängigkeit schneller zu erkennen ist.

empirisch unabhängig, empirisch abhängig

Gegeben sei die zweidimensionale relative Häufigkeitsverteilung der beiden Merkmale X und Y. Stimmen alle bedingten Verteilungen überein, d.h. gilt $\boldsymbol{f(x_j|y_k) = f(x_j|y_l)}$ für alle $k, l = 1, ..., r$ und für alle $j = 1, ..., m$ bzw. $\boldsymbol{f(y_k|x_j) = f(y_k|x_h)}$ für alle $j, h = 1, ..., m$ und für alle $k = 1, ..., r$, dann heißen $\boldsymbol{X}$ und $\boldsymbol{Y}$ **empirisch unabhängig**, andernfalls heißen sie **empirisch abhängig**.

Für die Interpretation und Anwendung der Überlegungen dieses Abschnittes ist Folgendes unbedingt zu beachten:

Alle Ergebnisse und Aussagen im Bereich der beschreibenden Statistik beziehen sich immer nur auf die jeweils erfasste Masse. Verallgemeinerungen auf irgendwelche übergeordnete Massen sind unzulässig. Das gilt für Häufigkeitsverteilungen, für die Kenngrößen von Häufigkeitsverteilungen und auch für die Unabhängigkeit von Merkmalen.

Wird für eine gegebene zweidimensionale Häufigkeitsverteilung Unabhängigkeit (oder Abhängigkeit) der beiden Merkmale festgestellt, dann gilt diese Unabhängigkeit nur bezüglich der erfassten Beobachtungswerte, aber nicht allgemein.

Beispiel 3.2.2:

200 *von* 500 *Mitarbeitern eines Betriebes werden nach Alter und Einkommen befragt. Für jede Altersklasse ergibt sich die gleiche bedingte Einkommensverteilung. Bei den* 200 *erfassten Mitarbeitern sind somit Einkommen und Alter unabhängige Merkmale. Für den Betrieb insgesamt muss das nicht zutreffen, da über Alter und Einkommen der übrigen* 300 *Mitarbeiter nichts bekannt ist.*

Hinsichtlich der Interpretation einer festgestellten Unabhängigkeit ist weiterhin Folgendes zu beachten:

Die Unabhängigkeit wird für eine gegebene Verteilung ermittelt. Sind für die Merkmalsausprägungen Klassen gebildet worden, so kann es sein, dass sich bei anderer Klasseneinteilung keine Unabhängigkeit mehr ergibt.

Eine festgestellte Unabhängigkeit bezieht sich im Falle klassierter Daten auf eine gegebene Klasseneinteilung.

Der eingeführte Unabhängigkeitsbegriff gilt für beliebige Merkmale, d.h. die Messbarkeitseigenschaften der Merkmale haben keinen Einfluss.

3.2.2 Zweidimensionale Verteilung empirisch unabhängiger Merkmale

Für empirisch unabhängige Merkmale stimmen alle Verteilungen der bedingten relativen Häufigkeiten untereinander und mit der entsprechenden Randverteilung überein.

$$f(x_j|y_k) = f_{j.} \quad \text{bzw.} \quad f(y_k|x_j) = f_{.k}$$

Durch Einsetzen der Definition der bedingten relativen Häufigkeit $f(x_j|y_k) = \dfrac{f_{jk}}{f_{.k}}$ folgt dann $\dfrac{f_{jk}}{f_{.k}} = f_{j.}$ und somit gilt:

$$f_{jk} = f_{j.} \cdot f_{.k} \text{ bzw. } f(x_j, y_k) = f(x_j) \cdot f(x_k).$$

Die relative Häufigkeit für das gemeinsame Auftreten der Ausprägungen x_j und y_k der Merkmale X und Y ergibt sich somit bei unabhängigen Merkmalen als Produkt der entsprechenden relativen Häufigkeiten der Randverteilungen.

Beispiel 3.2.3:

Für zwei unabhängige Merkmale X und Y werden folgende relative Häufigkeitsverteilungen angegeben:

X:

x_1	x_2	x_3
0.2	0.6	0.2

Y:

y_1	y_2	y_3
0.1	0.7	0.2

Die gemeinsame Verteilung der (unabhängigen) Merkmale lautet:

	x_1	x_2	x_3	$\sum$
y_1	$0.2 \cdot 0.1 = 0.02$	$0.6 \cdot 0.1 = 0.06$	$0.2 \cdot 0.1 = 0.02$	0.1
y_2	$0.2 \cdot 0.7 = 0.14$	$0.6 \cdot 0.7 = 0.42$	$0.2 \cdot 0.7 = 0.14$	0.7
y_3	$0.2 \cdot 0.2 = 0.04$	$0.6 \cdot 0.2 = 0.12$	$0.2 \cdot 0.2 = 0.04$	0.2
$\sum$	0.2	0.6	0.2	1

Sind für zwei unabhängige Merkmale die absoluten Häufigkeiten gegeben, dann gilt für die absolute Häufigkeit der gemeinsamen zweidimensionalen Verteilung:

$$h_{jk} = \frac{1}{n} h_{j.} \cdot h_{.k}$$

Die Berechnung der absoluten Häufigkeiten einer zweidimensionalen Verteilung unabhängiger Merkmale aus den beiden Randverteilungen spielt bei der Bestimmung von Zusammenhangsmaßen und bei der Prüfung von Zusammenhängen mit Hilfe von Stichprobenverfahren eine Rolle.

3.3 Arten von Zusammenhängen

Die Aufstellung und Analyse zweidimensionaler Verteilungen hat insbesondere den Zweck folgende Fragestellungen zu untersuchen:

- Besteht eine Abhängigkeit zwischen den betrachteten Merkmalen? Wenn ja, in welcher Ausprägung und von welcher Art liegt diese vor?
- Wie kann eine vorhandene Abhängigkeit quantitativ beschrieben werden?

Die Frage, ob überhaupt Abhängigkeit vorliegt, kann bei quantitativen Merkmalen in vielen Fällen durch Streudiagramme geklärt werden.

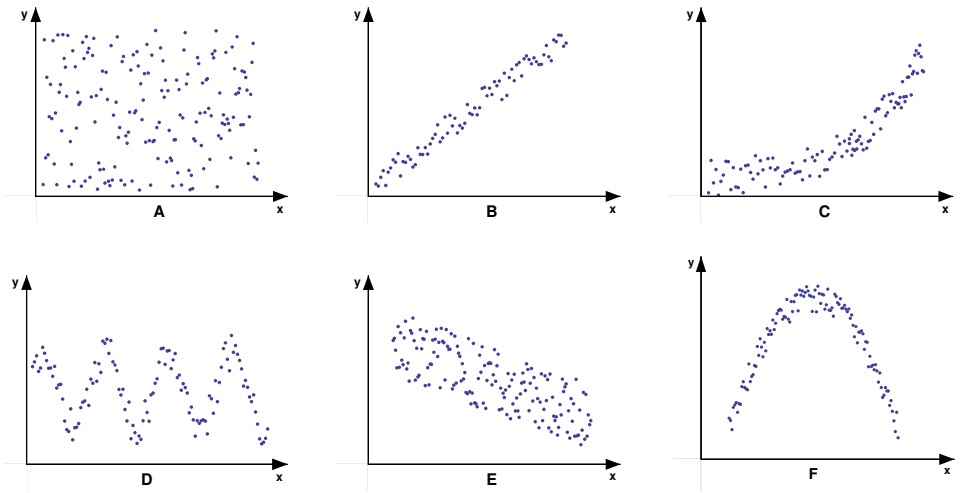


Abbildung 3.3.1: *Streudiagramme mit unterschiedlichen Abhängigkeiten der Merkmale*

In der Darstellung A sind keine Anzeichen von Abhängigkeit zwischen den Merkmalen X und Y zu erkennen. Die anderen Darstellungen weisen auf eine Abhängigkeit hin, welche jedoch nicht exakt im Sinne der Eindeutigkeit ist, wie sie bei Funktionen vorliegt. Die Abhängigkeit besteht tendenziell.

In B hat das Merkmal Y im Durchschnitt um so größere Werte, je größer die Werte von X sind, während in der Darstellung E die Werte von Y im Durchschnitt um so kleiner sind, je größer die Werte von X sind.

Beide Grafiken veranschaulichen einen linearen Zusammenhang. Grafik D deutet auf einen periodischen Zusammenhang hin, während in C die Werte von Y für zunehmende Werte des Merkmals X exponentiell steigen. Die Darstellung F verdeutlicht einen quadratischen Zusammenhang.

Allgemein ist es anhand einer grafischen Darstellung i.d.R. nicht möglich, Aussagen folgender Form zu treffen: „Wenn X den Wert x annimmt, dann nimmt Y immer genau den Wert y an."

Die Überlegungen zu Bild 3.3.1 zeigen, dass bei quantitativen Merkmalen die grafische Darstellung der zweidimensionalen Verteilung in einem Streudiagramm bereits ausreichen kann, um grundlegende Aussagen über eventuelle Abhängigkeiten zu machen. Dabei ist oft auch bereits eine Aussage über den Grad (die Ausgeprägtheit) eines Zusammenhangs möglich.

Für Merkmale, die nach einer Nominalskala oder einer Ordinalskala klassifiziert werden können, ist eine grafische Darstellung einer zweidimensionalen Verteilung in einem Streudiagramm nicht möglich. Hier können nicht so schnell Urteile über eventuelle Abhängigkeiten gemacht werden.

Als Aufgabenstellung für die Untersuchung von Zusammenhängen zwischen Merkmalen kann festgehalten werden:

- Bestimmung von Maßzahlen, die angeben, ob überhaupt ein Zusammenhang vorliegt und wie ausgeprägt dieser ist. Die Maßzahlen werden im Allgemeinen so definiert, dass sie bei Unabhängigkeit den Wert 0 annehmen. Bei einem vorhandenen Zusammenhang soll aus dem Wert der Maßzahl deutlich werden, wie ausgeprägt der Zusammenhang ist.

- Bestimmung von Funktionen, die die durchschnittliche Tendenz eines Zusammenhangs wiedergeben. Das ist nur für metrisch messbare Merkmale möglich. Diese Funktionen sind so zu bestimmen, dass sie die in einem Zusammenhang steckende Tendenz möglichst gut beschreiben.

Zur Beschreibung des Zusammenhangs können für die Streudiagramme in Abbildung 3.3.1 folgende Funktionen verwendet werden: Für C bietet sich eine Exponentialfunktion, für D eine Sinusfunktion und für F eine Parabel an, während die Tendenz des Zusammenhangs in B und E am besten durch eine steigende bzw. fallende Gerade beschrieben wird.

3.4 Korrelationsrechnung

Ist lediglich von Interesse zu untersuchen, ob Abhängigkeit vorliegt oder nicht, ohne die Art der Abhängigkeit oder eine Gesetzmäßigkeit dafür anzugeben, so wird von Korrelation oder Kontingenz gesprochen.

Kontingenz

Korrelation

Kontingenz bezeichnet den Zusammenhang zwischen qualitativen Merkmalen und von **Korrelation** wird bei einem Zusammenhang zwischen quantitativen bzw. mindestens ordinalskalierten Merkmalen gesprochen.

Korrelationskoeffizient

Kontingenzkoeffizient

Die Maßzahlen, die für den Grad bzw. die Ausgeprägtheit eines Zusammenhangs berechnet werden, heißen **Korrelationskoeffizient** bzw. **Kontingenzkoeffizient**. An diese Maßzahlen werden folgende Forderungen gestellt:
Existiert kein Zusammenhang, d.h. sind die beiden untersuchten Merkmale unabhängig voneinander, dann soll das Zusammenhangsmaß den Wert 0 annehmen.

Liegt ein eindeutiger Zusammenhang vor, dann soll die Maßzahl den Wert 1 bzw. $+1$ oder -1 annehmen.

Ein Zusammenhangsmaß, welches diese Eigenschaften erfüllt, nimmt somit Werte in dem Intervall $[-1, 1]$ an.

Die Messbarkeitseigenschaften der betrachteten Merkmale spielen bei der Bestimmung der Korrelations- bzw. Kontingenzkoeffizienten eine wichtige Rolle.

3.4.1 Korrelationskoeffizient nach Bravais-Pearson

In Abschnitt 3.1.5 wurde die Kovarianz als Maßzahl für die gemeinsame Streuung zweier Merkmale X und Y eingeführt. Wird die Kovarianz $\mathrm{Cov}(X, Y)$ durch das Produkt der Standardabweichungen der Randverteilungen der beiden Merkmale ($\tilde{s}_x$ bzw. $\tilde{s}_y$) dividiert, so ergibt sich der Korrelationskoeffizient, der nach den englischen Statistikern Bravais und Pearson benannt wurde. Vorausgesetzt wird dabei, dass eine lineare Beziehung zwischen den Merkmalen X und Y vorliegt und beide Variablen mindestens auf einer Intervallskala gemessen werden können.

Korrelationskoeffizient r

Der **Korrelationskoeffizient nach Bravais-Pearson**

$$\begin{aligned} r &= \frac{\mathrm{Cov}(X,Y)}{\tilde{s}_x \cdot \tilde{s}_y} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^n (x_i - \overline{x})^2 \sum_{i=1}^n (y_i - \overline{y})^2}} \\ &= \frac{\sum_{i=1}^n x_i y_i - n\overline{x}\ \overline{y}}{\sqrt{(\sum_{i=1}^n x_i^2 - n\overline{x}^2)(\sum_{i=1}^n y_i^2 - n\overline{y}^2)}} \end{aligned}$$

ist ein Maß für den Grad des **linearen** Zusammenhangs zweier quantitativer Merkmale.

***Beispiel 3.4.1**:*

Für die Häufigkeitsverteilung in Beispiel 3.1.10, wurde die Kovarianz $\mathrm{Cov}(X,Y) = -0.08$ *berechnet. Bei Vorliegen einer zweidimensionalen Häufigkeitsverteilung erfolgt die Berechnung der einzelnen Standardabweichungen für* X *und* Y *über die entsprechende Randverteilung.*

$$\tilde{s}_x = \sqrt{\frac{1}{50}(4 \cdot 15 + 0 \cdot 20 + 4 \cdot 15)} = \sqrt{2.4} = 1.55$$

und

$$\tilde{s}_y = \sqrt{\frac{1}{50}(1 \cdot 15 + 0 \cdot 25 + 1 \cdot 5 + 4 \cdot 5)} = \sqrt{0.8} = 0.89.$$

Für den Korrelationskoeffizienten nach Bravais-Pearson ergibt sich damit

$$r = \frac{-0.08}{1.55 \cdot 0.89} = -0.05799.$$

Der Korrelationskoeffizient nach Bravais-Pearson kann Werte im Bereich von -1 bis $+1$ annehmen, d.h. es gilt

$$-1 \leq r \leq 1.$$

Liegt überhaupt **kein linearer Zusammenhang** vor, so gilt

$$r = 0.$$

Liegen alle Beobachtungswerte auf einer steigenden Geraden, so gilt

$$r = 1.$$

Liegen alle Wertepaare auf einer fallenden Geraden, so gilt

$$r = -1.$$

Je enger sich die Beobachtungswerte um eine Gerade scharen, desto näher kommt der Korrelationkoeffizient dem Wert +1 oder -1.

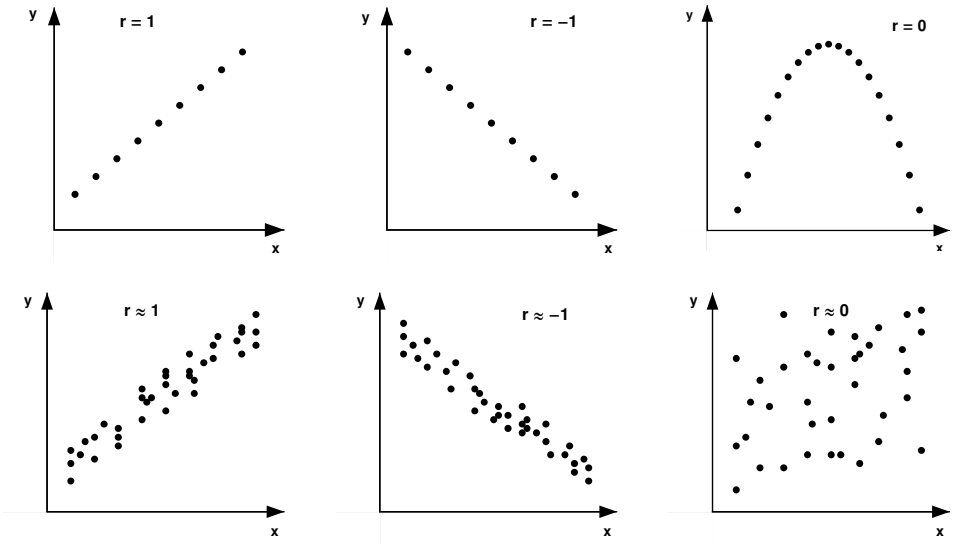


Abbildung 3.4.1: *Streudiagramme mit unterschiedlichen Korrelationskoeffizienten*

Im Fall des Studiums eines nichtlinearen Zusammenhangs kann der Korrelationskoeffizient nach Bravais-Pearson alle möglichen Werte in dem Bereich $-1 < r < 1$ annehmen, unabhängig davon, wie eng der Zusammenhang ist. Anhand der rechten oberen Grafik der Abbildung 3.4.1 ist zu erkennen, dass ein eindeutiger quadratischer Zusammenhang vorliegt. Das *lineare* Zusammenhangsmaß r ergibt in dem Fall 0.

Der Korrelationskoeffizient nach Bravais-Pearson kann somit nur sinnvoll für Merkmale berechnet werden, die einen linearen Zusammenhang aufweisen.

Beispiel 3.4.2:

In Abbildung 3.4.2 sind in drei Streudiagrammen Beobachtungswerte dargestellt, die jeweils alle dieselbe Funktionsgleichung erfüllen. Obwohl in allen drei Fällen ein eindeutiger Zusammenhang vorliegt, nimmt der Korrelationskoeffizient nach Bravais-Pearson verschiedene Werte an.

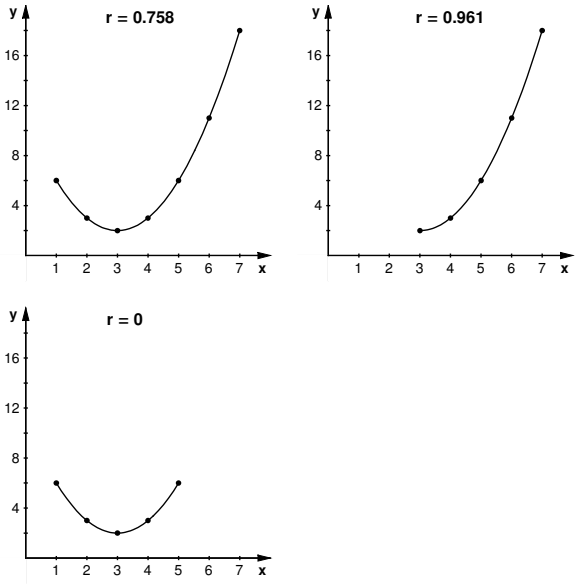


Abbildung 3.4.2: *Nichtlineare Zusammenhänge und zugehörige Korrelationskoeffizienten*

Werden in der Praxis solche Zusammenhänge untersucht, bei denen nicht ausgeschlossen ist, dass eine nichtlineare Abhängigkeit vorliegt, sollte dies bei der Interpretation des Korrelationskoeffizienten nach Bravais-Pearson berücksichtigt werden.

Der Korrelationskoeffizient r nimmt die Werte $+1$ oder -1 nur genau dann an, wenn alle Beobachtungswerte auf einer Geraden liegen. Die Beispiele zeigen, dass bei Nichtbeachtung dieser Tatsachen leicht Fehlinterpretationen vorkommen können.

Der Korrelationskoeffizient nach Bravais-Pearson wird daher auch als Maß für die Linearität eines Zusammenhangs bezeichnet.

In der Literatur wird an einigen Stellen von starker, mittelmäßiger oder schwacher Korrelation gesprochen und jeweilige Grenzen des Korrelationskoeffizienten angegeben. Die Verwendung derartiger Begriffe und Grenzen ist jedoch problematisch.

3.4.2 Korrelationskoeffizient nach Spearman

Der Korrelationskoeffizient nach Spearman wird angewendet, wenn ordinalskalierte Merkmale vorliegen. Im Vergleich zu r wird kein linearer sondern ein monotoner Zusammenhang betrachtet. Für die Berechnung des Korrelationskoeffizient nach Spearman werden Rangdaten benötigt. Dabei werden die Beobachtungswerte jedes Merkmals für sich geordnet und den geordneten Werten wird entsprechend der Reihenfolge eine Rangziffer zugeordnet.

Beispiel 3.4.3:

Gegeben sind zwei ordinalskalierte Merkmale X und Y. Die Ausprägungen beider Merkmale sind mit den Ziffern A-G *kodiert.*

$$(x_i, y_i): \quad (\text{B,D}) \quad (\text{C,E}) \quad (\text{A,F}) \quad (\text{D,B}) \quad (\text{E,G})$$

Für beide Merkmale lautet die Ordnung der Beobachtungswerte mit den dazugehörenden Rangziffern $rg(x_i)$ und $rg(y_i)$:

x_i	A	B	C	D	E
$rg(x_i)$	1	2	3	4	5

y_i	B	D	E	F	G
$rg(y_i)$	1	2	3	4	5

Anstelle der ursprünglichen Beobachtungswerte werden die Rangziffern betrachtet, so dass sich die Beobachtungspaare $(rg(x_i), rg(y_i))$ ergeben.

$$(rg(x_i), rg(y_i)): \quad (2,2) \quad (3,3) \quad (1,4) \quad (4,1) \quad (5,5)$$

Da die Rangziffern genau die Reihenfolge der Beobachtungswerte wiedergeben, kann ein eventueller Zusammenhang zwischen zwei ordinal messbaren Merkmalen unter der Verwendung der Rangziffern untersucht werden. Die Rangziffern selbst sind natürliche Zahlen und haben alle Eigenschaften eines quantitativen Merkmals.

Korrelationskoeffizient nach Spearman (Rangkorrelationskoeffizient)

Der **Korrelationskoeffizient nach Spearman** auch **Rangkorrelationskoeffizient** genannt, ergibt sich aus dem Korrelationskoeffizienten nach Bravais-Pearson, angewendet auf die Rangpaare $(rg(x_i), rg(y_i))$ für $i = 1, ..., n$:

$$\begin{aligned} r_s &= \frac{\sum_{i=1}^n (rg(x_i) - \overline{rg}_X)(rg(y_i) - \overline{rg}_Y)}{\sqrt{\sum_{i=1}^n (rg(x_i) - \overline{rg}_X)^2 \sum_{i=1}^n (rg(y_i) - \overline{rg}_Y)^2}} \\ &= \frac{\sum_{i=1}^n rg(x_i) rg(y_i) - n\overline{rg}_X \overline{rg}_Y}{\sqrt{\sum_{i=1}^n rg(x_i)^2 - n\overline{rg}_X^2} \sqrt{\sum_{i=1}^n rg(y_i)^2 - n\overline{rg}_Y^2}} \end{aligned}$$

mit $\overline{rg}_X = \frac{1}{n}\sum_{i=1}^n rg(x_i) = \frac{1}{n}\sum_{i=1}^n i = (n+1)/2 = \overline{rg}_Y$. Er ist ein Maß für den monotonen Zusammenhang zwischen Merkmalen, die nach einer Rangskala geordnet werden können. Es gilt $-1 \leq r_s \leq 1$.

Gilt $r_s = 1$, liegt ein gleichsinniger monotoner Zusammenhang vor. d.h. kleine/große x-Werte entsprechen kleinen/großen y-Werten bei **Betrachtung der Rangwerte**. Gilt $r_s = -1$, liegt ein gegensinniger monotoner Zusammenhang in der Art große/kleine x-Rangwerte entsprechen kleinen/großen y-Rangwerten vor. Werden die Originaldaten betrachtet, kann diese Interpretation nur mit Vorsicht verwendet werden.

***Beispiel 3.4.4**:*

Von zehn Studenten werden die Ergebnisse der Klausuren Mathematik (X) und Statistik (Y) betrachtet. In beiden Klausuren konnten maximal 100 *Punkte erreicht werden.*

x_i	24	37	41	45	59	63	78	82	89	96
$rg(x_i)$	1	2	3	4	5	6	7	8	9	10
y_i	28	31	32	33	34	36	37	38	39	41
$rg(y_i)$	1	2	3	4	5	6	7	8	9	10

Der Rangkorrelationskoeffizient enstspricht dem Wert 1, doch kann hier nicht gesagt werden, dass eine hohe Punktzahl in der Mathematikklausur auch einer hohen Punktzahl in der Statistikklausur entspricht. Bei 41 *von* 100 *Punkten kann nicht von einer hohen Punktzahl gesprochen werden. Diese Interpretation gilt lediglich bei Betrachtung der Rangzahlen. Der Student mit der höchsten Punktzahl in der Mathematikklausur hat auch in der Statistikklausur von allen die höchste der vorliegenden Punktzahlen erreicht, ungeachtet des Grössenverhältnisses beider Zahlen.*

Kommen bei den Beobachtungswerten x_i bzw. y_i identische Merkmalswerte vor, wird von **Bindungen** (Ties) gesprochen. In dem Fall können die Ränge nicht direkt vergeben werden und es werden sogenannte Durchschnittsränge gebildet, in dem den identischen Werten das arithmetische Mittel der entsprechenden Rangziffern zugeordnet wird.

Bindungen

***Beispiel 3.4.5**:*

In der folgenden Tabelle wird der Wert 3.6 *dreimal aufgeführt, so dass aus den zu vergebenden Rängen* 3, 4, 5 *der Durchschnittsrang* $rg(x_i) = (3+4+5)/3 = 4$ *für* $i = 3, 4, 5$ *gebildet wird.*

x_i	1.2	2.4	3.6	3.6	3.6	4.8
$rg(x_i)$	1	2	4	4	4	6

Treten keine Bindungen auf, d.h. jede Rangzahl tritt genau einmal auf, kann eine einfachere Formel zur Berechnung des Rangkorrelationskoeffizienten r_s herangezogen werden.

Unter der Voraussetzung, dass jede Beobachtung nur einmal auftritt, gilt ausgehend von der Formel des Korrelationskoeffizienten nach Bravais-Pearson, wobei die Beobachtungen x_i und y_i den **Rangzahlen** entsprechen:

$$r = \frac{\frac{1}{n}\sum_{i=1}^{n} x_i y_i - \overline{x}\;\overline{y}}{\sqrt{\frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x}^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \overline{y}^2}}.$$

Bei n Beobachtungspaaren durchlaufen x_i und y_i somit alle Rangziffern von 1 bis n und es gilt $\overline{x} = \overline{y}$, $\tilde{s}_x^2 = \tilde{s}_y^2$, $\sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i = \frac{1}{2}n(n+1)$ sowie $\sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i^2 = \frac{1}{6}n(n+1)(2n+1)$.

Werden die Ausdrücke in die Formel des Korrelationskoeffizienten r eingesetzt, ergibt sich:

$$r = \frac{\frac{1}{12}(n^2-1) - \frac{1}{2n}\sum_{i=1}^{n}(x_i - y_i)^2}{\frac{1}{12}(n^2-1)} \quad \text{oder} \quad r = 1 - \frac{6\sum_{i=1}^{n}(x_i - y_i)^2}{n(n^2-1)}.$$

Korrelationskoeffizient nach Spearman (ohne Bindungen)

Liegen keine Bindungen vor, kann mit Hilfe der Rangdifferenzen $d_i = rg(x_i) - rg(y_i)$ eine einfache Formel zur Berechnung des **Korrelationskoeffizienten nach Spearman** angegeben werden:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \qquad -1 \le r_s \le 1.$$

***Beispiel 3.4.6**:*

Sechs Studenten nehmen während ihres Winterurlaubs an einem Skiwettbewerb teil. Gewertet wurden ein Abfahrts- und ein Slalomrennen.

Student	1	2	3	4	5	6	$\sum$
Abfahrt	2	1	3	4	5	6	
Slalom	2	3	1	5	4	6	
d_i	0	-2	2	-1	1	0	
d_i^2	0	4	4	1	1	0	10

Für den Rangkorrelationskoeffizienten r_s ergibt sich:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 10}{6 \cdot (36-1)} = 1 - \frac{10}{35} = \frac{5}{7} = 0.7143.$$

Da die Berechnung des exakten Korrelationskoeffizienten nach Spearman im Fall von Bindungen aufwendiger ist als die obige Formel, wird oft auf die einfache Formel (Fall ohne Bindungen) mit Berücksichtigung der Durchschnittsränge zurückgegriffen. Je weniger identische Werte vorliegen, desto geringer ist die Abweichung. Dieses Problem tritt allerdings nicht auf, sobald entsprechende Software für statistische Analysen verwendet wird (z.B. Mathematica, SAS).

***Beispiel 3.4.7**:*

Die Äpfel und Birnen, die 11 Obsthändler auf einem Wochenmarkt anbieten, wurden nach den Güteklassen A, B, C, D, E, F eingestuft. Dabei beschreibt die Zufallsvariable X das Merkmal „Apfelsorte“ und das Merkmal „Birnensorte“ wird durch Y dargestellt. Für die Berechnung des exakten Rangkorrelationskoeffizienten r_s wird folgende Tabelle benötigt ($\overline{rg}_X = \overline{rg}_Y = (11+1)/2 = 6$ bzw. $n\overline{rg}_X^2 = n\overline{rg}_Y^2 = 11 \cdot 36 = 396$):

i	x_i	$rg(x_i)$	$rg(x_i)^2$	y_i	$rg(y_i)$	$rg(y_i)^2$	$rg(x_i)rg(y_i)$
1	B	2.5	6.25	B	3	9	7.5
2	A	1	1	A	1	1	1
3	F	11	121	F	10.5	110.25	115.5
4	E	9.5	90.25	F	10.5	110.25	99.75
5	E	9.5	90.25	E	9	81	85.5
6	D	7.5	56.25	D	7	49	52.5
7	D	7.5	56.25	D	7	49	52.5
8	C	5	25	D	7	49	35
9	B	2.5	6.25	B	3	9	7.5
10	C	5	25	B	3	9	15
11	C	5	25	C	5	25	625
$\sum$			502.5			501.5	496.75

$$r_s = \frac{\sum_{i=1}^{n} rg(x_i)rg(y_i) - n\overline{rg}_X\overline{rg}_Y}{\sqrt{\sum_{i=1}^{n} rg(x_i)^2 - n\overline{rg}_X^2}\sqrt{\sum_{i=1}^{n} rg(y_i)^2 - n\overline{rg}_Y^2}}$$
$$= \frac{496.75 - 396}{\sqrt{502.5 - 396}\sqrt{501.5 - 396}} = \frac{100.75}{\sqrt{106.5}\sqrt{105.5}} = \frac{100.75}{105.9988} = 0.9505$$

Für die Berechnung des Rangkorrelationskoeffizienten r_s (Formel ohne Bindungen) wird dagegen folgende Tabelle benötigt

Obst-händler	Äpfel x_i	Rang-ziffer	Birnen y_i	Rang-ziffer	d_i	d_i^2
1	B	2.5	B	3	-0.5	0.25
2	A	1	A	1	0	0
3	F	11	F	10.5	0.5	0.25
4	E	9.5	F	10.5	-1	1
5	E	9.5	E	9	0.5	0.25
6	D	7.5	D	7	0.5	0.25
7	D	7.5	D	7	0.5	0.25
8	C	5	D	7	-2	4
9	B	2.5	B	3	-0.5	0.25
10	C	5	B	3	2	4
11	C	5	C	5	0	0
					$\sum$	10.5

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 10.5}{11 \cdot (121-1)} = 1 - \frac{63}{1320} = 0.9523$$

Die Abweichung beider Korrelationskoeffizienten untereinander ist hier gering.

3.4.3 Kontingenztabelle und Kontingenzkoeffizient

Die Abhängigkeit zwischen zwei nominalen Merkmalen wird anhand des **Kontingenzkoeffizienten** untersucht. Dabei wird von der zweidimensionalen absoluten Häufigkeitstabelle ausgegangen, welche in diesem Zusammenhang **Kontingenztabelle** genannt wird.

Kontingenztabelle

	y_1	y_2	...	y_r	$\sum$ Randverteilung für X
x_1	h_{11}	h_{12}	...	h_{1r}	$h_{1.} = \sum_{k=1}^{r} h_{1k}$
x_2	h_{21}	h_{22}	...	h_{2r}	$h_{2.} = \sum_{k=1}^{r} h_{2k}$
...	...	...	...	...	...
x_m	h_{m1}	h_{m2}	...	h_{mr}	$h_{m.} = \sum_{k=1}^{r} h_{mk}$
$\sum$ Randverteilung für Y	$h_{.1} = \sum_{j=1}^{m} h_{j1}$	$h_{.2} = \sum_{j=1}^{m} h_{j2}$	...	$h_{.r} = \sum_{j=1}^{m} h_{jr}$	n

Um eine Aussage über den Grad einer möglichen Abhängigkeit zwischen den Merkmalen X und Y zu gewinnen, wird zunächst berechnet, welche absoluten Häufigkeiten sich für die gemeinsame Verteilung ergeben, wenn beide Merkmale unabhängig sind. Für unabhängige Merkmale entspricht

$$\frac{h_{j.}h_{.k}}{n} = nf_{j.}f_{.k}$$

der absoluten Häufigkeit der Merkmalskombination (x_j, y_k). Mit den tatsächlich beobachteten absoluten Häufigkeiten h_{jk} und den sich bei Unabhängigkeit ergebenden absoluten Häufigkeiten wird die **Hilfsgröße χ^2(Chi-Quadrat)** berechnet.

Chi-Quadrat

$$\chi^2 = \sum_{j=1}^{m}\sum_{k=1}^{r} \frac{\left(h_{jk} - \frac{h_{j.}h_{.k}}{n}\right)^2}{\frac{h_{j.}h_{.k}}{n}}$$

Die Summanden bestehen aus der quadrierten Abweichung der beobachteten absoluten Häufigkeiten von den Häufigkeiten, die sich bei Unabhängigkeit ergeben. Die Division durch die absolute Häufigkeit bei Unabhängigkeit der Merkmale bedeutet, dass die relativen quadratischen Abweichungen betrachtet werden.

Die Hilfsgröße Chi-Quadrat wird auch in der schließenden Statistik für die Untersuchung von Zusammenhängen mittels statistischer Testverfahren benötigt.

Unter Verwendung dieser Hilfsgröße wird nun als Zusammenhangsmaß der sogenannte Kontingenzkoeffizient berechnet. Der **Kontingenzkoeffizient**

$$C = \sqrt{\frac{\chi^2}{n + \chi^2}}$$

ist ein Maß für die Ausgeprägtheit eines Zusammenhangs zwischen nominal skalierten Merkmalen.

Kontingenzkoeffizient

Beispiel 3.4.8:

Eine Befragung von 100 *Männern nach ihrer bevorzugten Zahlungsmodalität und ihrem Berufsstand ergab folgende Kontingenztabelle:*

	Kreditkarte	EC-Karte	Barzahlung	$\sum$
Selbständig	10	25	5	40
Angestellter	8	40	2	50
in Ausbildung	2	5	3	10
$\sum$	20	70	10	100

$$\begin{aligned}
\chi^2 &= \frac{(10 - \frac{20\cdot 40}{100})^2}{\frac{20\cdot 40}{100}} + \frac{(25 - \frac{70\cdot 40}{100})^2}{\frac{70\cdot 40}{100}} + \frac{(5 - \frac{10\cdot 40}{100})^2}{\frac{10\cdot 40}{100}} \\
&\quad + \frac{(8 - \frac{20\cdot 50}{100})^2}{\frac{20\cdot 50}{100}} + \frac{(40 - \frac{70\cdot 50}{100})^2}{\frac{70\cdot 50}{100}} + \frac{(2 - \frac{10\cdot 50}{100})^2}{\frac{10\cdot 50}{100}} \\
&\quad + \frac{(2 - \frac{20\cdot 10}{100})^2}{\frac{20\cdot 10}{100}} + \frac{(5 - \frac{70\cdot 10}{100})^2}{\frac{70\cdot 10}{100}} + \frac{(3 - \frac{10\cdot 10}{100})^2}{\frac{10\cdot 10}{100}} \\
&= \frac{4}{8} + \frac{9}{28} + \frac{1}{4} + \frac{4}{10} + \frac{25}{35} + \frac{9}{5} + 0 + \frac{4}{7} + 4 = 8.56 \\
C &= \sqrt{\frac{8.56}{100 + 8.56}} = 0.28
\end{aligned}$$

Zur Vereinfachung empfiehlt sich eine Erweiterung der einzelnen Felder einer Kontingenztabelle in der Form

$\boldsymbol{h_{jk}}$	$\left(h_{jk} - \frac{h_{j.}h_{.k}}{n}\right)^2$
$h_{jk} - \frac{h_{j.}h_{.k}}{n}$	$\frac{h_{j.}h_{.k}}{n}$

anzugeben. Die erweiterte Kontingentabelle für das vorherige Beispiel lautet:

	Kreditkarte	EC-Karte	Barzahlung	$\sum$
Selbständig	**10** 4 2 8	**25** 9 −3 28	**5** 1 1 4	40
Angestellter	**8** 4 −2 10	**40** 25 5 35	**2** 9 −3 5	50
in Ausbildung	**2** 0 0 2	**5** 4 −2 7	**3** 4 2 1	10
$\sum$	20	70	10	100

Somit kann die Hilfsgröße χ^2 sofort zu

$$\chi^2 = \frac{4}{8} + \frac{9}{28} + \frac{1}{4} + \frac{4}{10} + \frac{25}{35} + \frac{9}{5} + 0 + \frac{4}{7} + 4 = 8.56$$

berechnet werden.

Der **Kontingenzkoeffizient** C nimmt Werte von 0 bis $\sqrt{\frac{C^*-1}{C^*}}$ an, wobei C^* der kleinere Wert von Zeilenzahl und Spaltenzahl ist, d.h.

$$0 \leq C \leq \sqrt{\frac{C^* - 1}{C^*}} \text{ mit } C^* = \min(m, r).$$

Der maximale Wert, den der Kontingenzkoeffizient annehmen kann, hängt somit von der Größe der Kontingenztabelle ab.

***Beispiel 3.4.9**:*

Die Zahlen aus Beispiel 3.4.8 werden an dieser Stelle so modifiziert, dass zwischen den beiden Merkmalen ein eindeutiger Zusammenhang besteht.

40	0	0	40
0	50	0	50
0	0	10	10
40	50	10	100

$$\chi^2 = 36 + 20 + 4 + 4 + 5 + 81 + 20 + 25 + 5 = 200$$

$$C = \sqrt{\frac{200}{200+100}} = \sqrt{\frac{2}{3}}$$

Obwohl ein eindeutiger Zusammenhang vorliegt, wird C nicht 1.

Am Anfang des Abschnittes 3.4 wurde gefordert, dass ein Zusammenhangsmaß im Bereich $[-1; 1]$ liegen soll.

Um Kontingenzkoeffizienten verschieden großer Tabellen besser vergleichen zu können, wird der Kontingenzkoeffizient zu

$$C_{korr} = C\sqrt{\frac{C^*}{C^* - 1}}$$

normiert. Für diesen **korrigierten Kontingenzkoeffizienten** gilt dann

$$0 \leq C_{korr} \leq 1.$$

korrigierter Kontingenzkoeffizient

***Beispiel 3.4.10**:*

Für das Beispiel 3.4.8 ist $C^* = 3$ und

$$C_{korr} = 0.28\sqrt{\frac{3}{2}} = 0.34.$$

Für das Beispiel 3.4.9 folgt wie gefordert

$$C_{korr} = \sqrt{\frac{2}{3}}\sqrt{\frac{3}{2}} = 1.$$

Im Vergleich zu den vorherigen Koeffizienten, kann der Kontingenzkoeffizient lediglich die Stärke des Zusammenhangs messen. Welche Struktur

die Abhängigkeit besitzt kann nicht direkt gefolgert werden.

Beispiel 3.4.11:

In zwei Städten wurden 100 *Schüler nach ihren monatlichen Ausgaben für Lebensmittel (Süßigkeiten, Fastfood,...) befragt, die sie von ihrem monatlichem Taschengeld bestreiten.*

	Stadt A				Stadt B			
	Ausgaben				Ausgaben			
Taschengeld	[0;10]	(10;20]	(20;30]	$\sum$	[0;10]	(10;20]	(20;30]	$\sum$
[20;30]	30	0	0	30	30	0	0	30
(30;40]	0	50	0	50	0	0	50	50
(40;50]	0	0	20	20	0	20	0	20
$\sum$	30	20	50	100	30	20	50	100

$$\chi^2_A = 49 + 6 + 15 + 15 + 25 + 10 + 6 + 10 + 64 = 200$$

$$\chi^2_B = 49 + 6 + 15 + 15 + 10 + 25 + 6 + 64 + 10 = 200$$

$$C_{A,B} = \sqrt{\frac{200}{200+100}} = \sqrt{\frac{2}{3}} \qquad C_{korr_{A,B}} = \sqrt{\frac{2}{3}}\sqrt{\frac{3}{2}} = 1$$

In der linken Hälfte der Tabelle ist ein eindeutiger linearer Zusammenhang der Form „Je höher das Taschengeld, desto höher die Ausgaben“ zu erkennen. In der rechten Hälfte liegt dagegen ein eindeutiger nichtlinearer Zusammenhang vor. In beiden Fällen nimmt der Kontingenzkoeffizient den Wert 1 *an.*

Der Kontingenzkoeffizient besitzt die Eigenschaft der **Invarianz**.

Invarianz-eigenschaft

Werden in einer Kontingenztabelle Zeilen bzw. Spalten miteinander vertauscht, so ändert sich der Kontingenzkoeffizient nicht.

Aufgrund der Invarianzeigenschaft kann mittels des Kontingenzkoeffizienten die Struktur des Zusammenhangs nicht gemessen werden. Da die Größe χ^2 mit zunehmenden Stichprobenumfang steigt, sollte ein Vergleich von Kontingenztabellen mit stark abweichenden Stichprobenumfängen vermieden werden.

3.4.4 Ergänzende Bemerkungen zur Berechnung der Zusammenhangsmaße

Die in den vorhergehenden Abschnitten behandelten Zusammenhangsmaße (Korrelationskoeffizient nach Bravais-Pearson, Rangkorrelationskoeffizient, Kontingenzkoeffizient) wurden für unterschiedliche Skalen definiert.

Der Korrelationskoeffizient nach Bravais-Pearson darf nur für metrisch messbare Merkmale berechnet werden, da er nur für diese ein sinnvolles Maß für den Grad des Zusammenhangs liefert.

Der Rangkorrelationskoeffizient (Korrelationskoeffizient nach Spearman) ist für Merkmale geeignet, die wenigstens auf einer Ordinalskala gemessen werden können. Seiner Berechnung liegen die Rangziffern der geordneten Merkmalswerte zugrunde. Der Rangkorrelationskoeffizient ist der Korrelationskoeffizient nach Bravais-Pearson dieser Rangziffern. Aus diesem Zusammenhang ergeben sich dann auch entsprechende Interpretationsmöglichkeiten für diesen Korrelationskoeffizienten.

Der Kontingenzkoeffizient kann für nominal messbare Merkmale berechnet werden. Seine Bestimmung ist unabhängig davon, in welcher Reihenfolge die einzelnen Merkmalsausprägungen in der Kontingenztabelle aufgeführt werden. Wird die Reihenfolge der Merkmalsausprägungen geändert, so entspricht das einer Vertauschung von Zeilen und/oder Spalten, wobei der Wert des Kontingenzkoeffizienten gleich bleibt (Invarianzeigenschaft).

Die Beziehungen zwischen Zusammenhangsmaßen und Messbarkeitseigenschaften der betrachteten Merkmale sind bei der Untersuchung von Abhängigkeiten zwischen Merkmalen besonders zu beachten. Werden Merkmale mit unterschiedlichen Messbarkeitseigenschaften auf Abhängigkeiten untersucht, dann ist die jeweils schwächste Skala maßgebend. Das bedeutet z.B., dass ein ordinal messbares und ein metrisch messbares Merkmal nicht mit dem Korrelationskoeffizienten nach Bravais-Pearson, sondern mit dem Rangkorrelationskoeffizienten auf einen Zusammenhang hin untersucht werden können.

Da eine Ordinalskala auch alle Eigenschaften einer Nominalskala und eine metrische Skala auch alle Eigenschaften einer Ordinal- und einer Nominalskala hat, können für Merkmale, die nach einer Ordinalskala gemessen

werden, auch der Kontingenzkoeffizient und für Merkmale, die nach einer metrischen Skala geordnet werden, auch der Rangkorrelationskoeffizient und der Kontingenzkoeffizient berechnet werden. Die damit ermittelten Aussagen sind jedoch aufgrund des Informationsverlustes im Allgemeinen schwächer im Vergleich zum Korrelationskoeffizienten nach Spearman bzw. nach Bravais-Pearson.

3.4.5 Diskrepanz zwischen mathematischer Korrelation und Kausalität

Ein mathematisch nachgewiesener korrelativer Zusammenhang zwischen zwei Merkmalen X und Y besagt lediglich, dass eine Beziehung zwischen den beiden Merkmalen besteht. Die Richtung der Beziehung ist dadurch nicht nachgewiesen. Das Merkmal Y kann von X abhängen und umgekehrt. Möglich ist auch, dass beide Merkmale durch ein drittes Merkmal beeinflusst werden und ein direkter kausaler Zusammenhang zwischen X und Y nicht besteht.

Beispiel 3.4.12:

Die Überprüfung des Zusammenhangs zwischen Weizenpollenallergien und Weizenpreis ergab eine negative Korrelation, die jedoch auf das Wetter bzw. das Wachstum des Weizens zurückzuführen ist. Bei optimalem Wetter wächst und blüht der Weizen gut. Aufgrund des hohen Angebots sinken die Preise. Die Merkmale Weizenpollenallergie und Weizenpreis hängen von einem dritten Merkmal, dem Wachstum des Weizens, ab.

Bei erwachsenen Männern kann eine negative Korrelation zwischen dem Einkommen und der Zahl der Haare auf dem Kopf nachgewiesen werden. Auch hier liegt die Ursache an einem dritten Merkmal, dem Alter. Mit zunehmendem Alter nimmt das Einkommen zu, die Anzahl der Haare allerdings ab.

Eine mathematisch positive Korrelation kann zwischen dem Benzinverbrauch eines Motorrades und der Größe des Tanks festgestellt werden. Hier beruht die Korrelation auf der Größe des Hubraums. Größere Motorräder mit größerem Hubraum weisen in der Regel einen größeren Tank auf. Der Zusammenhang zwischen Hubraum und Benzinverbrauch ist technisch bedingt, während der Zusammenhang zwischen Benzinverbrauch und Größe des Tanks nicht auf Kausalität beruht.

Die oben angeführten Beispiele stellen Scheinkorrelationen dar, d.h. Korrelationen ohne Kausalität, die auf einer versteckten Variablen beruhen. Scheinkorrelationen können auch durch Inhomogenität des vorliegenden Datensatzes entstehen, z.B. wenn der Datensatz in zwei oder mehreren Gruppen aufgeteilt werden kann. Ist in jeder Untergruppe keine Korrelation zwischen zwei interessierenden Merkmalen nachweisbar, so kann im gesamten Datensatz trotzdem eine Korrelation festgestellt werden. Andererseits kann durch Nichtbeachtung eines Merkmals vorhandene Korrelation übersehen werden. Dies ist der Fall, wenn in den Untergruppen entgegengesetzte Beziehungen zu beobachten sind.

Beispiel 3.4.13:

Anhand eines fiktiven Datensatzes wird untersucht, inwieweit die monatlichen Ausgaben für Kosmetik von der Körpergröße abhängen. Werden die Ausgaben nach Geschlecht getrennt betrachtet, so ergibt sich für die Frauen ein Korrelationskoeffizient von $r = -0.03$. *Für die Männer nimmt* r *den Wert* -0.06 *an. Bleibt das Merkmal Geschlecht unberücksichtigt, ergibt sich bei Betrachtung des gesamten Datensatzes ein Korrelationskoeffizient von* 0.68. *Dieser Zusammenhang wird durch nachstehende Abbildung verdeutlicht.*

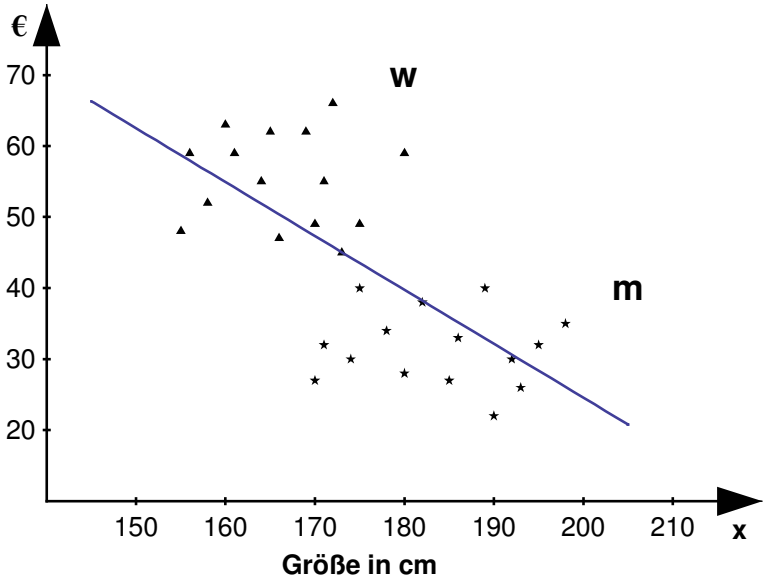


Abbildung 3.4.3: *Scheinkorrelation durch Inhomogenität des Datensatzes*

***Beispiel 3.4.14**:*

Anhand eines fiktiven Datensatzes werden nun die durchschnittlichen Ausgaben für Kosmetik der letzten fünf Jahre nach Geschlecht getrennt betrachtet.

Jahr	1	2	3	4	5	r
m	20.2	22.6	25.4	28	33.6	0.98
w	63	58.6	55.8	50.2	48.6	-0.99

Die Frauen haben über die Jahre hinweg weniger für Kosmetik ausgegeben, während der Konsum bei den Männern über die Jahre hinweg gestiegen ist. Hier ist eindeutig eine Korrelation zu erkennen. Wird der gesamte Datensatz betrachtet, ergibt sich lediglich eine sehr schwache negative Korrelation ($r = -0.02$).

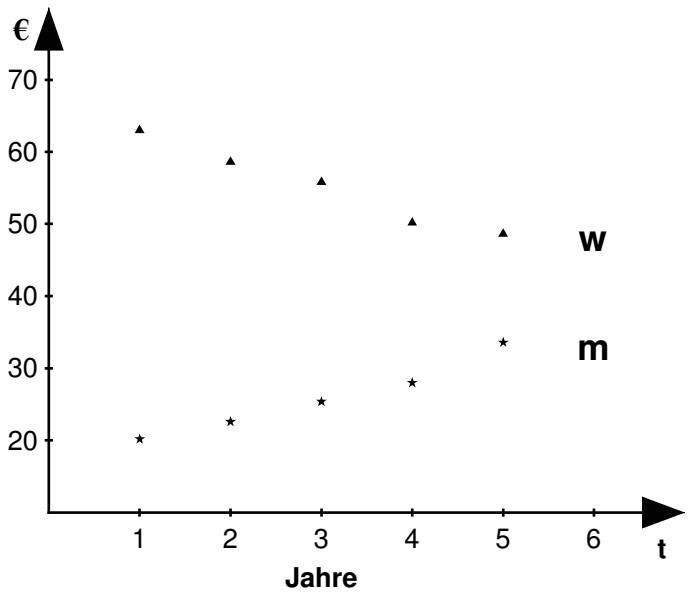


***Abbildung 3.4.4:** Aufhebung der Korrelation durch gegenläufige Entwicklung*

Wie die Beispiele zeigen, ist das Vorhandensein mathematischer Korrelation noch lange kein Nachweis für Kausalität. Ebenso kann bei mangelnder mathematischer Korrelation eine Kausalität generell nicht ausgeschlossen werden. Ohne Zusatzinformationen sollte die kausale Interpretation gut bedacht sein. In allen Fällen ist es ratsam mittels Streudiagramme die Datenstruktur zu betrachten.

3.5 Regressionsanalyse

In der Analyse der Abhängigkeit zwischen quantitativen Merkmalen wird eine mathematische Regel gesucht, die es gestattet, aus gegebenen Merkmalswerten des einen Merkmals die zugehörige **durchschnittliche** Merkmalsausprägung des anderen (abhängigen) Merkmals zu errechnen.

Regression

Der Begriff **Regression** bezeichnet die Untersuchung der Abhängigkeit der Veränderungen eines quantitativen Merkmals von Änderungen eines anderen quantitativen Merkmals (=einfache Regression) oder von Änderungen mehrerer quantitativer Merkmale (=mehrfache Regression).
Zur Beschreibung dieser Abhängigkeit wird die **Regressionsfunktion** verwendet.

Regressionsfunktion

Für zwei Merkmale X und Y hat die Regressionsfunktion die allgemeine Form:

$$\hat{y} = f(x).$$

Regressionsfunktionen beschreiben für gewöhnlich keinen eindeutigen Zusammenhang. Die einzelnen Paare von Beobachtungswerten (x_j, y_j) werden im allgemeinen nicht auf der Regressionsfunktion liegen, sondern um die Funktion herum streuen. Es wird ein Zusammenhang zwischen den Ausprägungen des Merkmals X und den zugehörigen durchschnittlichen Werten des Merkmals Y beschrieben, d.h. mit $\hat{y}$ wird ein Schätzwert für y angegeben.

Regressionsrechnung
Regressor
Regressand

Für eine allgemeine Funktion der Form $y = f(x)$ heißt x unabhängige und y abhängige Variable. In der **Regressionsrechnung** wird x **Regressor**, exogene Variable oder erklärende Variable genannt und y wird als **Regressand**, endogene Variable oder erklärte Variable bezeichnet.

Soll eine Regressionsfunktion bestimmt werden, wird zunächst der Typ der Regressionsfunktion vorgegeben. Mögliche Funktionstypen sind:

Gerade:	$\hat{y} = a + bx$
Parabel:	$\hat{y} = a + bx + cx^2$
Potenzfunktion:	$\hat{y} = ax^b$
Exponentialfunktion:	$\hat{y} = ab^x$

Ist der Funktionstyp festgelegt, müssen noch die Koeffizienten $a, b, ...$ so bestimmt werden, dass die Funktion den Zusammenhang möglichst gut beschreibt.

Kriterium der Kleinsten Quadrate

Die Koeffizienten der Regressionsfunktion werden so bestimmt, dass die **Summe der quadrierten Abweichungen** der Beobachtungswerte y_i von den Regressionsfunktionswerten $f(x_i)$ **ein Minimum** wird (**Kriterium der Kleinsten-Quadrate**).[4]

3.5.1 Lineare Regression

lineare Regressionsfunktion

Für den Fall einer **linearen Regressionsfunktion $\hat{y} = a + bx$** besagt das Kriterium der Kleinsten-Quadrate, dass die Koeffizienten a und b so zu bestimmen sind, dass die Summe der Quadrate der Abweichungen $u_i = y_i - \hat{y}_i$ ein Minimum wird. Wenn insgesamt n Wertepaare vorliegen, soll die Funktion

$$f(a,b) = \sum_{i=1}^{n} u_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

minimal sein. Abbildung 3.5.1 verdeutlicht den Zusammenhang.

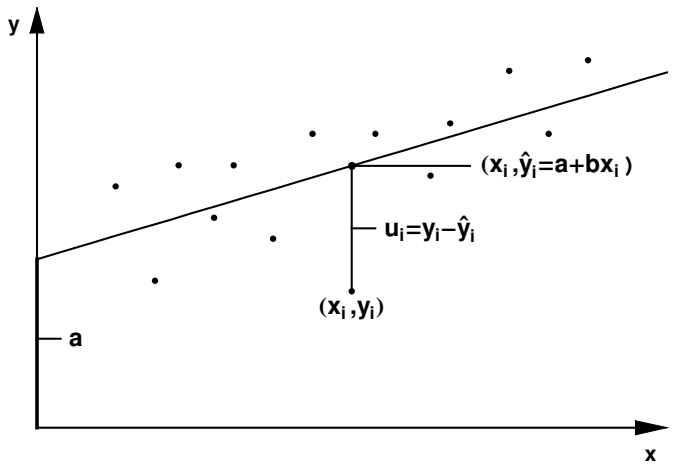


Abbildung 3.5.1: *Streudiagramm mit Regressionsgerade*

Eine notwendige (nicht hinreichende) Bedingung für ein Minimum der Summe der quadratischen Abweichungen ist das Verschwinden der partiellen Ableitungen 1. Ordnung nach a und b. Die partiellen Ableitungen lauten:

[4] siehe interaktive Mathematica-Applets auf der Homepage des Lehrstuhls http://www.fernuni-hagen.de/ls_statistik/forschung/multimedia/

$$\frac{\partial f(a,b)}{\partial a} = \sum_{i=1}^{n} 2(y_i - a - bx_i) \cdot (-1),$$
$$\frac{\partial f(a,b)}{\partial b} = \sum_{i=1}^{n} 2(y_i - a - bx_i) \cdot (-x_i).$$

Durch Nullsetzen der Ableitungen und nach einigen einfachen Umformungen ergeben sich die sogenannten **Normalgleichungen** zur Bestimmung von a und b:

Normalgleichungen

I. Normalgleichung: $\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i$

II. Normalgleichung: $\sum_{i=1}^{n} y_i x_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2$

Für a und b gilt nach Auflösen des Gleichungssystems:[5]

$$a = \frac{\sum x_i^2 \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^2 - (\sum x_i)^2},$$
$$b = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{n\sum x_i^2 - (\sum x_i)^2} = \frac{\sum x_i y_i - n\overline{x}\,\overline{y}}{\sum x_i^2 - n\overline{x}^2}.$$

Werden beide Seiten der I. Normalgleichung durch n dividiert, ergibt sich die vereinfachte Gleichung

$$\overline{y} = a + b\overline{x}.$$

Wird zunächst der Koeffizient b berechnet, kann a mit Hilfe der obigen Gleichung leicht berechnet werden.

Die Koeffizienten a und b heißen auch **Regressionskoeffizienten**:

$$b = \frac{\sum x_i y_i - n\overline{x}\,\overline{y}}{\sum x_i^2 - n\overline{x}^2}, \qquad a = \overline{y} - b\overline{x}.$$

Regressionskoeffizienten

Für die Berechnung der Regressionskoeffizienten werden folgende Summen benötigt:

$$\sum x_i, \quad \sum x_i^2, \quad \sum y_i, \quad \sum x_i y_i.$$

[5] Die Lösung des Gleichungssystems kann kompliziert durch Ausrechnen oder einfacher mittels der Cramerschen Regel bestimmt werden.

***Beispiel 3.5.1**:*

Gegeben seien folgende Wertepaare:

x_i	1	1	3	3
y_i	1	3	2	4

Für die Berechnung der Hilfssummen ergibt sich:

i	x_i	y_i	x_i^2	$x_i y_i$
1	1	1	1	1
2	1	3	1	3
3	3	2	9	6
4	3	4	9	12
$\sum$	8	10	20	22

Aus der Tabelle ergeben sich $\overline{x} = 2$ *und* $\overline{y} = 2.5$*, so dass die Regressionskoeffizienten wie folgt berechnet werden:*

$$b = \frac{22 - 4 \cdot 2 \cdot 2.5}{20 - 4 \cdot 2^2} = \frac{22 - 20}{20 - 16} = \frac{2}{4} = 0.5,$$

$$a = 2.5 - 0.5 \cdot 2 = 2.5 - 1 = 1.5.$$

Die Regressionsgerade lautet somit $\hat{y} = 1.5 + 0.5x$*. In Abbildung 3.5.**??** sind die Beobachtungswerte und die Regressionsgerade grafisch dargestellt.*

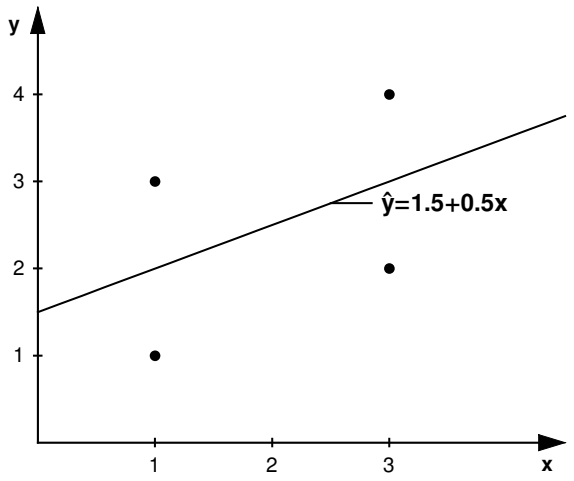


***Abbildung 3.5.2:** Streudiagramm mit Regressionsgerade*

Ist der kausale Zusammenhang von Merkmalen ungewiss, so ist auch unbekannt, welche Variable Regressor und welche Regressand ist. In dem Fall ist es sinnvoll, beide Regressionsfunktionen zu bestimmen.

Beispiel 3.5.2:

Es soll der Zusammenhang zwischen Körpergröße und Körpergewicht mit den Daten aus Beispiel 3.1.4 bestimmt werden. Betrachtet wird hier die Abhängigkeit des Körpergewichts Y von der Körpergröße X.

$$\hat{y} = -60.69 + 0.78x$$

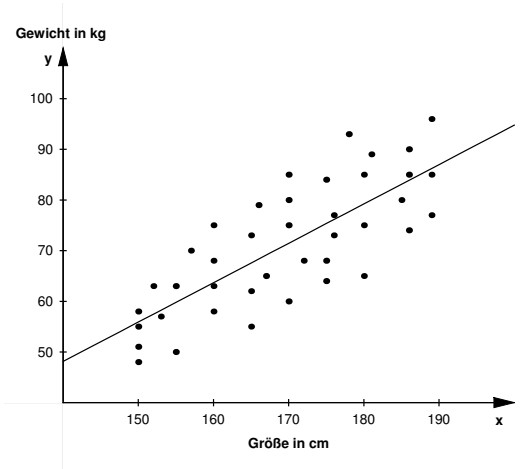


Abbildung 3.5.3: *Streudiagramm mit Regressionsgerade*

Abbildung 3.5.2 zeigt noch einmal das Streudiagramm aus Abbildung 3.1.1 und die eingezeichnete Regressionsgerade.

Der errechnete Wert des Regressionskoeffizienten b, nämlich $b = 0.78$, ist wie folgt zu interpretieren: Bei den untersuchten 40 Personen steigt (fällt) unter sonst gleichen Bedingungen (= ceteris paribus) das Körpergewicht durchschnittlich um 0.78 kg, wenn sich die Körpergröße um 1 cm vergrößert (verkleinert). Für $x_i = 175$ cm und $x_{i'} = 176$ gilt:

$$\begin{aligned}
\hat{y}_i &= -60.69 + 0.78 \cdot 175 = -60.69 + 13650 = 75.81, \\
\hat{y}_{i'} &= -60.69 + 0.78 \cdot 176 = -60.69 + 137.28 = 76.59, \\
\hat{y}_{i'} - \hat{y}_i &= 76.59 - 75.81 = 0.78.
\end{aligned}$$

Interpretation der Regressionskoeffizienten

Die Gleichung $\overline{y} = a + b\overline{x}$ (**I. Normalgleichung**) besagt, dass eine lineare Regressionsfunktion durch den Punkt $(\overline{x}, \overline{y})$ geht.

Der Wert von b gibt an, um wie viel sich der Wert des Merkmals Y durchschnittlich verändert, wenn der Wert des Merkmals X um 1 Einheit geändert wird.

Der Wert von a, dem Absolutglied oder Achsenabschnitt der linearen Regressionsfunktion $\hat{y} = a + bx$, gibt den durchschnittlichen Wert des erklärten Merkmals Y an, wenn das erklärende Merkmal X den Wert $x = 0$ angenommen hat. Diese Interpretation ist zwar formal korrekt; fachwissenschaftlich ist sie aber nur sinnvoll, wenn der Wert Null im Bereich der beobachteten Werte des erklärenden Merkmals X liegt.

Die Regressionsfunktion $\hat{y} = f(x)$ kann auch zur Prognose verwendet werden. Dazu wird in die Regressionsfunktion der betreffende x-Wert eingesetzt und der zugehörige $\hat{y}$-Wert berechnet.

Beispiel 3.5.3:

Für die gemeinsame Verteilung von Körpergröße X und Körpergewicht Y aus dem vorherigen Beispiel wurde die Regressionsfunktion $\hat{y} = -60.69 + 0.78x$ ermittelt. Für $x = 171$ ergibt sich dann:

$$\hat{y} = -60.69 + 0.78 \cdot 171 = 72.69.$$

Für das gegebene Datenmaterial kann also aufgrund der Regressionsfunktion prognostiziert werden, dass Personen mit 171 *cm Körpergröße durchschnittlich* 72.69 *kg schwer sind.*

In dieser Weise kann mit jeder Kleinsten-Quadrate-Regressionsfunktion $\hat{y} = f(x)$ für einen vorgegebenen Wert des Merkmals X der zugehörige Durchschnittswert des Merkmals Y berechnet werden.

Es muss allerdings davor gewarnt werden, diese Schätzung zu verallgemeinern. In dem hier behandelten Rahmen können sich Aussagen immer nur auf das gegebene Datenmaterial beziehen. Problematisch ist es, eine berechnete Regressionsfunktion beliebig zu verlängern, denn die Gültigkeit kann sich immer nur auf den Bereich der Beobachtungswerte beziehen.

3.6 Zusammenhang zwischen Regression und Korrelation - das Bestimmtheitsmaß

Der Korrelationskoeffizient r und der Koeffizient b einer linearen Regressionsfunktion $\hat{y} = a + bx$ stehen in einem engen Zusammenhang. Beide Koeffizienten können mittels des anderen berechnet werden.

$$r = b \cdot \frac{\tilde{s}_x}{\tilde{s}_y}$$

Liegt ein positiver linearer Zusammenhang vor, so kann dieser mittels einer steigenden Geraden beschrieben werden und der Regressionskoeffizient b ist daher auch positiv.

Eine Umformung der obigen Darstellung von r führt zu einem Gütemaß in der Regressionsanalyse, welches die Güte der Anpassung, die mittels der vorliegenden Regressionsfunktion erzielt wird, betrachtet.

$$\begin{aligned} r^2 &= b^2 \cdot \frac{\frac{1}{n}\sum(x_i - \overline{x})^2}{\tilde{s}_y^2} = \frac{\sum(bx_i - b\overline{x})^2}{\sum(y_i - \overline{y})^2} \\ &= \frac{\sum(a + bx_i - a - b\overline{x})^2}{\sum(y_i - \overline{y})^2} \\ &= \frac{\sum(\hat{y}_i - \overline{y})^2}{\sum(y_i - \overline{y})^2} = \frac{\tilde{s}_{\hat{y}}^2}{\tilde{s}_y^2} \end{aligned}$$

In dieser Darstellung wird die Varianz der geschätzten Werte, welche mit $\tilde{s}_{\hat{y}}^2$ bezeichnet wird, in Relation zu der Varianz der beobachteten Werte gesetzt. Naturgemäß ist $\tilde{s}_{\hat{y}}^2$ kleiner als $\tilde{s}_y^2$, wie die **Streuungszerlegung** zeigt.

Streuungszerlegung

Für die Gesamtstreuung $n\tilde{s}_y^2$ gilt die **Streuungszerlegung**

$$\begin{aligned} \sum_{i=1}^{n}(y_i - \overline{y})^2 &= \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2, \\ \text{SQT} &= \text{SQE} + \text{SQR}. \end{aligned}$$

Dabei bezeichnet:

SQT:	Sum of Squares Total (Gesamtstreuung, $n\tilde{s}_y^2$)
SQE:	Sum of Squares Explained (erklärte Streuung, $n\tilde{s}_{\hat{y}}^2$)
SQR:	Sum of Squares Residual (Reststreuung)

Die Gesamtvarianz der vorliegenden Daten setzt sich somit aus der erklärten, d.h. die auf den Zusammenhang zwischen X und Y zurückführbare Variation und einer Restvariation (z.B. Messfehler) zusammen. Je besser die Regressionsfunktion die vorliegenden Daten beschreibt, je besser somit die Anpassung ist, desto größer ist der Anteil der erklärten Varianz an der Gesamtvarianz.

Bestimmtheitsmaß

Im Falle linearer Regression heißt das Quadrat des Korrelationskoeffizienten nach Bravais-Pearson **Bestimmtheitsmaß**. Das Bestimmtheitsmaß

$$R^2 = r^2 = \frac{\tilde{s}_{\hat{y}}^2}{\tilde{s}_y^2}$$

gibt den Anteil der durch das Merkmal X erklärten Varianz $\tilde{s}_{\hat{y}}^2$ an der Gesamtvarianz $\tilde{s}_y^2$ an. Das Bestimmtheitsmaß kann Werte zwischen Null und Eins annehmen.

***Beispiel 3.6.1**:*

In Bild 3.6.1 ist ein Streudiagramm mit der zugehörigen linearen Regressionsfunktion $\hat{y} = 1.75 + 0.25x$ dargestellt.

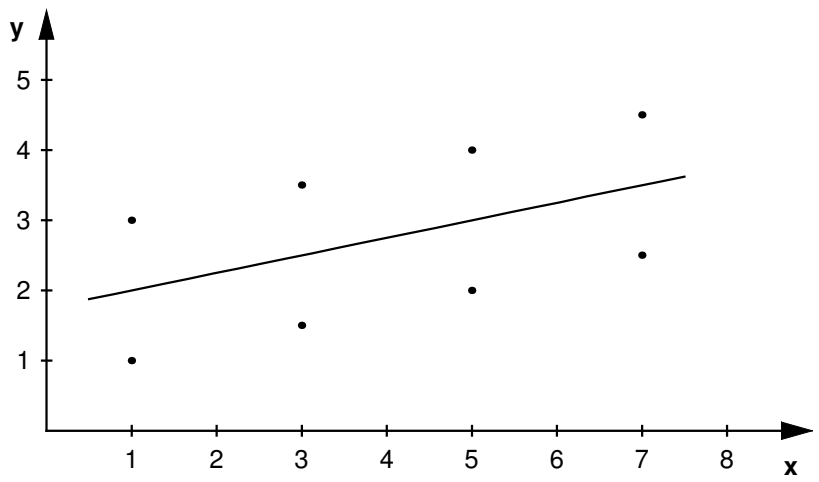


***Abbildung 3.6.1:** Streudiagramm und Regressionsgerade zu Beispiel 3.6.1*

Die Beobachtungswerte und die zu den gegebenen x-Werten auf der Regressionsgeraden liegenden $\hat{y}$-Werte sind in der folgenden Tabelle dargestellt.

x_i	1	1	3	3	5	5	7	7
y_i	1	3	1.5	3.5	2	4	2.5	4.5
$\hat{y}_i$	2	2	2.5	2.5	3	3	3.5	3.5

Es ist $\tilde{s}_y^2 = 1.3125$ *und* $\tilde{s}_{\hat{y}}^2 = 0.3125$. *Damit ergibt sich*

$$R^2 = \frac{0.3125}{1.3125} = 0.2381$$

und

$$r = 0.48795.$$

Das Bestimmtheitsmaß von 0.2381 *bedeutet, dass* 23.81% *der Varianz der Y-Werte durch die lineare Regression erklärt werden, wobei hier „erklären“ nicht so zu verstehen ist, dass hier ein ursächlicher Zusammenhang vorliegt.*

Übungsaufgaben

Aufgaben zu Kapitel 1

Übungsaufgabe 1.1:

Entscheiden Sie ob es sich um diskrete oder stetige Merkmale handelt.

		diskret	*stetig*
a)	*Geschwindigkeit von Pkws*	☐	☐
b)	*Hörerzahl einer Vorlesung*	☐	☐
c)	*Anzahl der Mitarbeiter eines Betriebes*	☐	☐
d)	*Einkommen*	☐	☐
e)	*Zeit für die Beschleunigung eines Pkws von 0 auf 100 km/h*	☐	☐
f)	*Punkte in einer Klausur*	☐	☐
g)	*Bücherbestand in einer Bibliothek*	☐	☐
h)	*Stromverbrauch*	☐	☐
i)	*Treibstoffverbrauch eines Orbiters in Abhängigkeit von der Flughöhe*	☐	☐

Übungsaufgabe 1.2:

Geben Sie zu den folgenden Merkmalen an, wie diese skaliert werden können (Nominal-, Ordinal-, Intervall-, Verhältnis-, Absolutskala).

a)	*Militärdienstgrad:*
b)	*Alter:*
c)	*Geschlecht:*
d)	*Nationalität:*
e)	*Intelligenzquotient:*
f)	*Semesterzahl:*
g)	*Wassergüte:*
h)	*Anzahl von Verkehrsunfällen:*
i)	*Tarifklasse der Kfz-Haftpflicht:*
j)	*Studienfach:*
k)	*Fahrtkosten:*

Übungsaufgabe 1.3:

Bei einer Betriebszählung wird von 30 *Betrieben u.a. die Anzahl der Beschäftigten ermittelt.*

12 438 623 187 216 25 98 100 617 367 560 116 270 304 36
87 54 124 517 410 160 125 44 76 62 260 342 570 520 234

Für die Beschäftigtenzahl der Betriebe sollen Klassen gebildet werden, und zwar 1 *bis* 100, 101 *bis* 200 *usw. Bestimmen Sie die Verteilung der absoluten und relativen Häufigkeiten.*

Klasse Nr.	Klasse	Strichliste	absolute Häufigkeit	relative Häufigkeit
1	1 bis 100			
2	101 bis 200			
3				

Aufgaben zu Kapitel 2

Übungsaufgabe 2.1:

40 *Personen werden danach gefragt, ob sie Arbeiter* (A)*, Angestellter* (K)*, Beamter* (B) *oder selbständig* (S) *sind.*

B A A K S A A B K B S S A B B A A A B A
B B A K K S K B A K B A A K A K K A K A

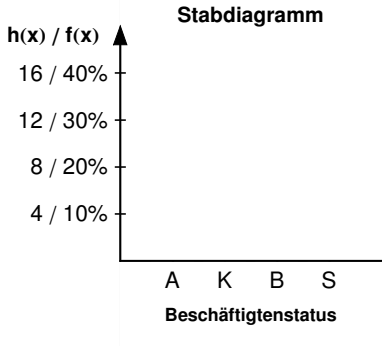
a) Was sind bei dieser Aufgabe

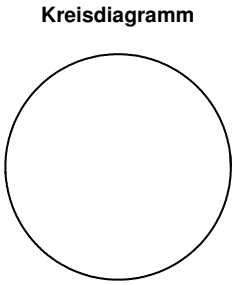
Merkmalsträger:
Merkmal:
Merkmalsausprägungen:

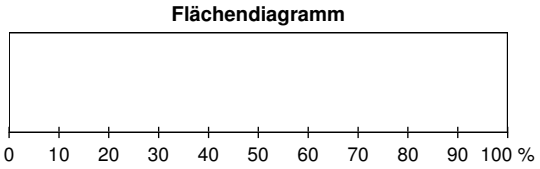
b) Bestimmen Sie die absoluten und die relativen Häufigkeiten der Merkmalsausprägungen.

Merkmals-ausprägung	Strichliste	absolute Häufigkeit	relative Häufigkeit in %

c) Stellen Sie die Häufigkeitsverteilung als Stabdiagramm, als Flächendiagramm und als Kreisdiagramm grafisch dar.





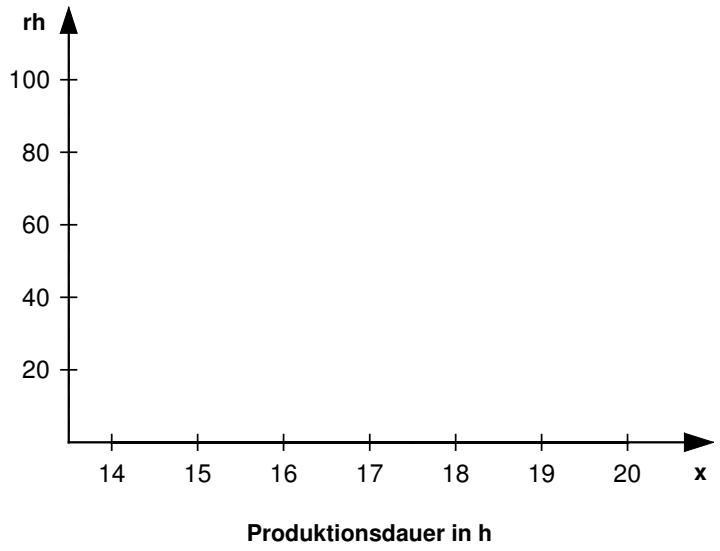


Übungsaufgabe 2.2:

Die Produktionsdauer von 200 *Werkstücken sei wie folgt verteilt:*

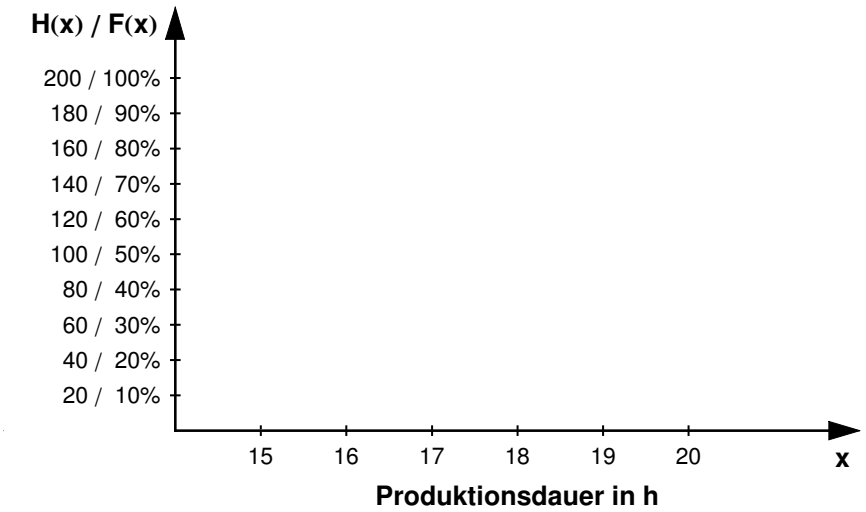
Klasse	Produktionsdauer in h	absolute Häufigkeit h_j	relative Häufigkeit f_j
I	(14;16]	30	0.15
II	(16;17]	80	0.40
III	(17;17.5]	50	0.25
IV	(17.5;18]	20	0.10
V	(18;20]	20	0.10

a) *Geben sie die grafische Darstellung der Häufigkeitsverteilung in Form eines Histogramms an mit* $rh_j = \frac{h_j}{b_j}$.



b) *Bestimmen Sie die tabellarische und grafische Summenhäufigkeitsverteilung und zeigen Sie anhand der grafischen Summenhäufigkeitsverteilung, bei wie viel % der Werkstücke die Produktionsdauer nicht länger als* 16.5 *h ist und welche Produktionsdauer von* 85% *der Werkstücke nicht überschritten wird.*

Produktionsdauer in h	Summenhäufigkeit H_j	Summenhäufigkeit F_j in %
≤ 16		
≤ 17		
≤ 17.5		
≤ 18		
≤ 20		



Übungsaufgabe 2.3:

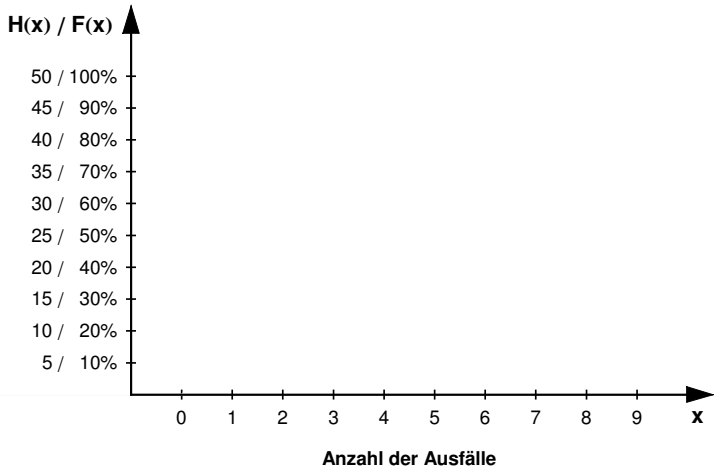
In einer Fabrik wurden innerhalb eines Jahres an 50 *Maschinen die Ausfallzeiten notiert. Es wurde untersucht, wie oft die Maschinen im vergangenen Jahr ausgefallen sind.*

0 1 0 5 4 3 1 7 9 3 2 1 0 4 2 6 7 5 0 1 4 1 1 3 4
8 6 2 6 1 0 0 1 4 3 1 2 6 3 5 4 7 4 2 3 3 1 1 5 6

a) Bestimmen Sie die absoluten, relativen Häufigkeiten und Summenhäufigkeiten.

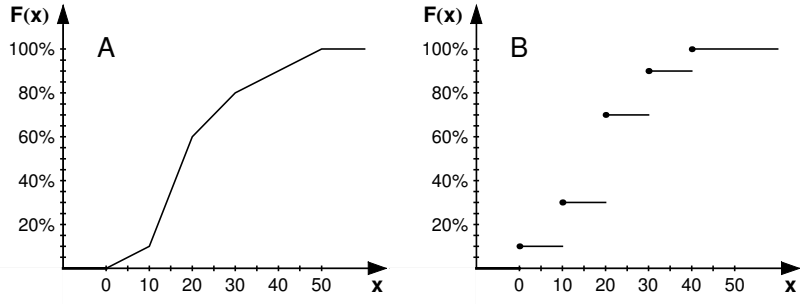
Anzahl der Ausfälle	Strichliste	h_j	f_j	H_j	F_j
0					
1					
2					
3					
4					
5					
6					
7					
8					
9					

b) Stellen Sie die Summenhäufigkeitsverteilung grafisch dar.



Übungsaufgabe 2.4:

Gegeben sind die beiden grafischen Darstellungen von Summenhäufigkeitsverteilungen.



Bestimmen Sie die relativen Häufigkeiten für die Fälle a) - e):

	A	B
a) $x \leqq 20$		
b) $x > 10$		
c) $x \geqq 10$		
d) $10 < x \leqq 30$		
e) $10 < x < 30$		

Für welchen Wert nimmt die Summenhäufigkeitsverteilung A bzw. B den Wert $F(x) = 70\%$ an ?

Übungsaufgabe 2.5:

Für ein Fernsehgerät eines bestimmten Typs werden in 12 *Fachgeschäften die folgenden Verkaufspreise festgestellt:*

199 195 219 229 249 199 229 199 299 195 209 149

Berechnen Sie den Modalwert x_{mod}*, den Median* x_{med}*, das untere Quartil* $x_{0.25}$*, das 8.Dezil und das arithmetische Mittel* $\overline{x}$*.*

Übungsaufgabe 2.6:

Von n *Teilzeitkräften einer Firma wurde der Stundenlohn ermittelt. Bestimmen Sie unter Benutzung der Tabelle die Modalklasse, die Einfallsklasse des Medians und das arithmetische Mittel* $\overline{x}$ *des Stundenlohns.*

Stundenlohn in €	Klassenmitte x_j	Häufigkeit h_j	$x_j \cdot h_j$	H_j	F_j
(7;8]	7.5	14			
(8;9]	8.5	40			
(9;10]	9.5	38			
(10;11]	10.5	32			
(11;12]	11.5	26			
		$n =$	$\sum =$		

Übungsaufgabe 2.7:

Sie legen 1000 € *am* 1.1. *eines Jahres mit Zinseszinsen an. In den ersten beiden Jahren bekommen Sie* 4% *Zinsen, in den folgenden drei Jahren* 5.5% *Zinsen, im* 6. *Jahr* 6% *Zinsen. Auf wie viel* € *ist Ihr Kapital nach* 6 *Jahren angewachsen? Bei welcher (konstanten) Durchschnittsverzinsung* z *erhalten Sie das gleiche Endkapital?*

Übungsaufgabe 2.8:

In 10 Bäckereien wurden folgende Preise für ein belegtes Brötchen ermittelt (in €):

1.40 1.60 1.70 1.50 1.40 1.80 1.70 1.60 1.50 1.80

Berechnen Sie das arithmetische Mittel, die Varianz und die Standardabweichung.

$$\begin{aligned} \overline{x} &= \frac{1}{n}\sum_{i=1}^{n} x_i = \\ \tilde{s}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 = \\ \tilde{s} &= \sqrt{\tilde{s}^2} = \end{aligned}$$

Übungsaufgabe 2.9:

Untersucht wurde die Lebensdauer von n Glühbirnen. Berechnen Sie die Varianz und die Standardabweichung.

Stunden	x_j	h_j	$x_j h_j$	$x_j - \overline{x}$	$(x_j - \overline{x})^2$	$(x_j - \overline{x})^2 \cdot h_j$
(600;700]	650	5				
(700;800]	750	11				
(800;900]	850	16				
(900;1000]	950	10				
(1000;1100]	1050	8				
$\sum$						

$$\begin{aligned} \overline{x} &= \frac{1}{n}\sum_{j=1}^{m} x_j h_j = \\ \tilde{s}^2 &= \frac{1}{n}\sum_{j=1}^{m}(x_j - \overline{x})^2 h_j = \\ \tilde{s} &= \sqrt{\tilde{s}^2} = \end{aligned}$$

Übungsaufgabe 2.10:

Studenten wurden nach ihrem Einkommen aus Nebentätigkeiten befragt. Das Ergebnis ist in der Tabelle enthalten. Berechnen Sie das arithmetisches Mittel, die Varianz und die Standardabweichung.

verfügbares Einkommen	Klassenmitte x_j	f_j	$x_j f_j$	x_j^2	$x_j^2 f_j$
$\leq$400	350	0.2			
(400;500]	450	0.4			
(500;600]	550	0.2			
(600;700]	650	0.1			
>700	750	0.1			
$\sum$					

$$\overline{x} = \sum_{j=1}^{m} x_j f_j \quad =$$

$$\tilde{s}^2 = \sum_{j=1}^{m} x_j^2 f_j - \overline{x}^2 \quad =$$

$$\tilde{s} = \sqrt{\tilde{s}^2} \quad =$$

Übungsaufgabe 2.11:

Die folgende Tabelle gibt die Körpergrößen von 5 *Kindern in Zoll und cm an (es wird der Einfachheit halber* 1 *Zoll* = 2.5 *cm gesetzt).*

x_i cm	120	130	125	130	135
y_i Zoll	48	52	50	52	54

Berechnen Sie für die beiden Messreihen

a) das arithmetische Mittel,

b) die Standardabweichung,

c) den Variationskoeffizienten.

Übungsaufgabe 2.12:

Der Jahresumsatz eines Baugewerbes der Bundesrepublik betrug 67.5 *Milliarden* €. *Der Umsatz verteilte sich auf* 14500 *Betriebe wie folgt:*

Umsatz in Mio. €	x_j	Anzahl der Unternehmen h_j
≤ 1	0.5	4000
(1;5]	3	8300
(5;10]	7.5	1300
(10;50]	30	800
(50;90]	70	100

Zeichnen Sie die Lorenzkurve und bestimmen Sie das Lorenzsche Konzentrationsmaß.

Aufgaben zu Kapitel 3

Übungsaufgabe 3.1:

Von 28 *Studienanfängern wurden Mathematiknoten und Englischnoten beim Abitur erfasst (Mathematiknote, Englischnote):*

(4,2)	(3,1)	(3,3)	(2,3)	(4,4)	(3,4)	(3,3)	(1,3)	(3,2)	(5,3)
(3,3)	(3,4)	(3,3)	(3,4)	(2,3)	(2,1)	(2,2)	(3,4)	(3,3)	(3,3)
(1,1)	(4,5)	(5,4)	(2,5)	(2,2)	(2,3)	(2,3)	(3,4)		

Ordnen Sie die Wertepaare zuerst nach der Mathematiknote und bei gleicher Mathematiknote nach der Englischnote (lexikografische Ordnung).

Übungsaufgabe 3.2:

Befragt wurden 30 *Männer und Frauen in welche Anlagemöglichkeit sie zur Zeit investieren würden. Zur Auswahl stehen Aktien, Immobilien und Versicherungen.*

Person	Geschlecht*	Anlage**	Person	Geschlecht	Anlage
1	M	I	16	F	A
2	M	A	17	F	I
3	F	A	18	M	A
4	F	I	19	M	I
5	M	V	20	F	I
6	F	I	21	M	A
7	F	A	22	F	A
8	M	I	23	M	I
9	M	V	24	F	A
10	F	I	25	F	I
11	F	A	26	M	A
12	F	I	27	F	V
13	M	V	28	F	A
14	M	I	29	M	I
15	F	I	30	F	I

* M=Mann, F=Frau
** A=Aktie, I=Immobilie, V=Versicherung

Stellen Sie eine zweidimensionale Häufigkeitstabelle auf.

Übungsaufgabe 3.3:

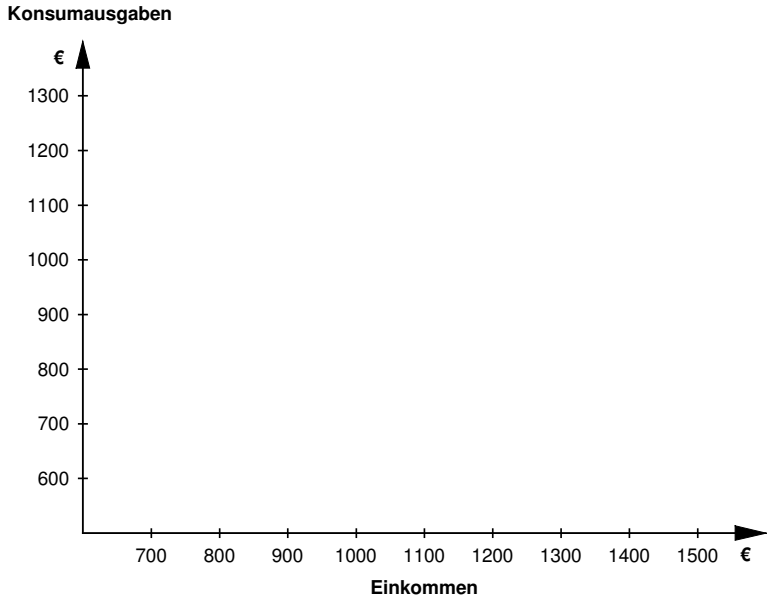
20 *Haushalte werden nach ihrem monatlichen Einkommen X und den monatlichen Konsumausgaben Y befragt.*

Haushalt	1	2	3	4	5	6	7	8	9	10
Einkommen	800	1200	1100	1480	1300	900	1000	1200	800	925
Konsum-ausgaben	700	820	930	1270	1160	840	620	970	680	750
Haushalt	11	12	13	14	15	16	17	18	19	20
Einkommen	1150	870	1420	950	1350	1280	1040	1470	1220	1120
Konsum-ausgaben	870	870	1050	920	1250	1010	820	1310	1200	980

a) Stellen Sie die Beobachtungswerte in einer Korrelationstabelle mit der angegebenen Klasseneinteilung dar.

Konsum-ausgaben	Einkommen			
	(700;900]	(900;1100]	(1100;1300]	(1300;1500]
(500;700]				
(700;900]				
(900;1100]				
(1100;1300]				
(1300;1500]				

b) Zeichnen Sie ein Streudiagramm.



Übungsaufgabe 3.4:

Gegeben sei die folgende zweidimensionale Verteilung absoluter Häufigkeiten:

	X					
Y	1	2	3	4	5	$\sum$
1	3	5	10	8	4	
2	5	8	20	20	7	
6	9	15	50	40	6	
8	3	12	20	12	3	
$\sum$						

Bestimmen Sie die Randverteilungen.

Übungsaufgabe 3.5:

50 *Personen wurden nach Alter und Einkommen befragt.*

	Alter				
Einkommen	(0;30]	(30;40]	(40;50]	(50;60]	(60;70]
[0;1000]	1	2	1	1	1
(1000;1500]	2	4	4	3	1
(1500;2000]	3	6	6	3	2
>2000	1	3	2	2	2

Bestimmen Sie für jede Altersklasse die bedingten Verteilungen des Einkommens mit Angabe der Häufigkeiten in %.

	Alter				
Einkommen	(0;30]	(30;40]	(40;50]	(50;60]	(60;70]
[0;1000]					
(1000;1500]					
(1500;2000]					
>2000					

Übungsaufgabe 3.6:

Berechnen Sie für die Verteilung in Aufgabe 3.5 das bedingte arithmetische Mittel der Einkommen für jede Altersklasse. Für Einkommen über 2000 € *verwenden Sie als repräsentativen Wert* 2500 €.

Alter	Durchschnittseinkommen
(20;30]	
(30;40]	
(40;50]	
(50;60]	
(60;70]	

Übungsaufgabe 3.7:

Berechnen Sie für die folgende zweidimensionale Häufigkeitsverteilung die Kovarianz

	$x_1 = 1$	$x_2 = 3$	$x_3 = 6$	$\sum$
$y_1 = 4$	0	0	16	
$y_2 = 5$	0	10	0	
$y_3 = 9$	10	0	4	
$\sum$				

Die arithmetischen Mittelwerte der Randverteilungen und die Kovarianz lauten:

$$\begin{aligned} \overline{x} &= \\ \overline{y} &= \\ \mathrm{Cov}(X,Y) &= \frac{1}{n}\sum_{j=1}^{3}\sum_{k=1}^{3} x_j y_k h_{jk} - \overline{x}\,\overline{y} = \end{aligned}$$

Übungsaufgabe 3.8:

Prüfen Sie durch Bestimmung der bedingten Verteilungen, bei welcher der beiden angegebenen Verteilungen die Merkmale abhängig sind.

a)

	X			
Y	1	3	9	$\sum$
2	4	6	10	20
10	5	7	8	20
12	11	7	2	20
$\sum$	20	20	20	60

b)

	X				
Y	1	2	4	8	$\sum$
2	20	12	8	4	44
4	10	6	4	2	22
20	5	3	2	1	11
$\sum$	35	21	14	7	77

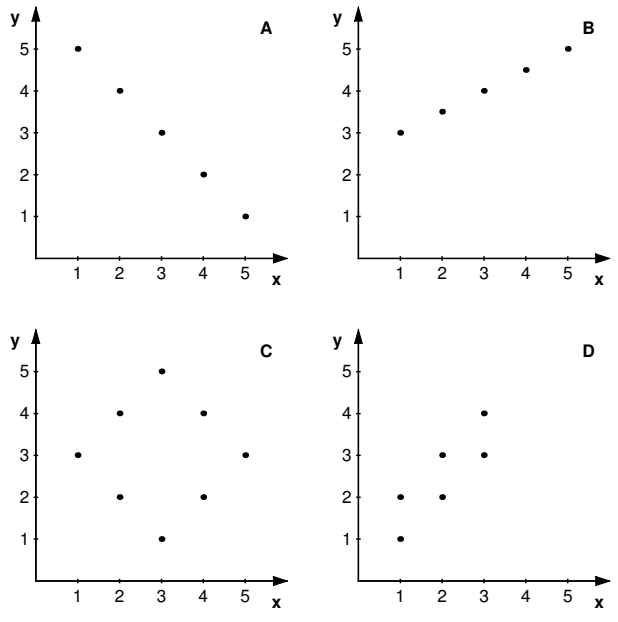
Übungsaufgabe 3.9:

Berechnen Sie für die Verteilung aus Aufgabe 3.7 den Korrelationskoeffizienten nach Bravais-Pearson.

$$\tilde{s}_x = \qquad \tilde{s}_y = \qquad r = \frac{\mathrm{Cov}(X,Y)}{\tilde{s}_x \cdot \tilde{s}_y} = \frac{-3}{\quad} =$$

Übungsaufgabe 3.10:

Berechnen Sie für die in den folgenden Zeichnungen dargestellten Verteilungen jeweils den Korrelationskoeffizienten nach Bravais-Pearson.



Für die Berechnung ist zu beachten, dass alle absoluten Häufigkeiten den Wert 1 *annehmen. Bei den zu bestimmenden Summen für die Kovarianz werden nur die von Null verschiedenen Summanden aufgeführt.*

$A: \overline{x} =$ $\tilde{s}_x =$ $\overline{y} =$ $\tilde{s}_y =$

$$\mathrm{Cov}(X,Y) = \frac{1}{5}\sum_{j=1}^{5}\sum_{k=1}^{5} x_j y_k h_{jk} - \overline{x}\,\overline{y} =$$

$$r = \frac{\mathrm{Cov}(X,Y)}{\tilde{s}_x \cdot \tilde{s}_y} =$$

$B: \overline{x} =$ $\tilde{s}_x =$ $\overline{y} =$ $\tilde{s}_y =$

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \\ r &= \end{aligned}$$

$C: \overline{x} =$ $\tilde{s}_x =$ $\overline{y} =$ $\tilde{s}_y =$

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \\ r &= \end{aligned}$$

$D: \overline{x} =$ $\tilde{s}_x =$ $\overline{y} =$ $\tilde{s}_y =$

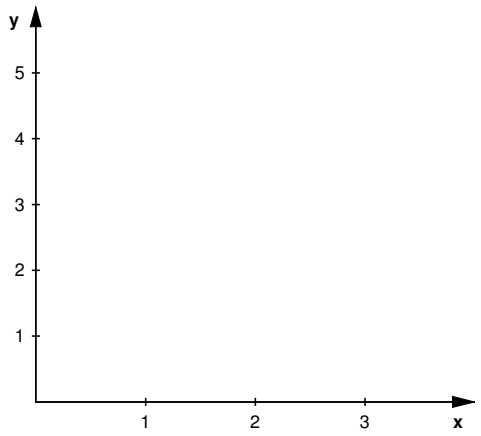
$$\begin{aligned} \mathrm{Cov}(X,Y) &= \\ r &= \end{aligned}$$

Übungsaufgabe 3.11:

Gegeben sind folgende Wertepaare:

x_i	1	1	2	2	3	3
y_i	1	3	2	4	3	5

a) *Stellen Sie die Wertepaare grafisch dar.*



b) *Bestimmen Sie und zeichnen Sie eine lineare y-x-Regressionsfunktion nach dem Kriterium der Kleinsten-Quadrate.*

i	x_i	y_i	x_i^2	$x_i y_i$
1				
2				
3				
4				
5				
6				
$\sum$				

$$b = \frac{\sum x_i y_i - n\overline{xy}}{\sum x_i^2 - n\overline{x}^2}$$

$$= $$

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

$$= $$

Regressionsfunktion:

c) *Zeichnen Sie die Regressionsgerade in das Diagramm aus a) ein.*

Übungsaufgabe 3.12:

Folgende Tabelle gibt Aufschluss über die Ersparnisse von privaten Haushalten sowie deren verfügbares Einkommen der letzten zehn Jahre (in Mill. €). Es wird vermutet, dass die Ersparnisse annähernd linear vom verfügbaren Einkommen abhängen. Bestimmen Sie die lineare Regressionsfunktion mit Hilfe des Verfahrens der „Kleinsten Quadrate".

Jahr i	Einkommen x_i	Ersparnisse y_i
1	34.2	2.8
2	40.8	4.1
3	42.5	4.5
4	47.3	4.3
5	50.1	4.9
6	52.6	5.8
7	56.9	7.0
8	61.4	7.7
9	73.5	8.1
10	76.7	8.8

Übungsaufgabe 3.13:

Zeigen Sie, dass gilt:

$$r \;=\; b \cdot \frac{\tilde{s}_x}{\tilde{s}_y}$$

Übungsaufgabe 3.14:

Gegeben sind die folgenden Beobachtungswerte für zwei quantitative Merkmale X und Y.

Berechnen Sie unter Benutzung der Tabelle eine lineare Regressionsfunktion.

i	x_i	y_i	x_i^2	$x_i y_i$	y_i^2	$\hat{y}_i$
1	2	5				
2	2	7				
3	4	4				
4	4	6				
5	5	4.5				
6	6	3				
7	6	5				
8	8	3				
$\sum$						

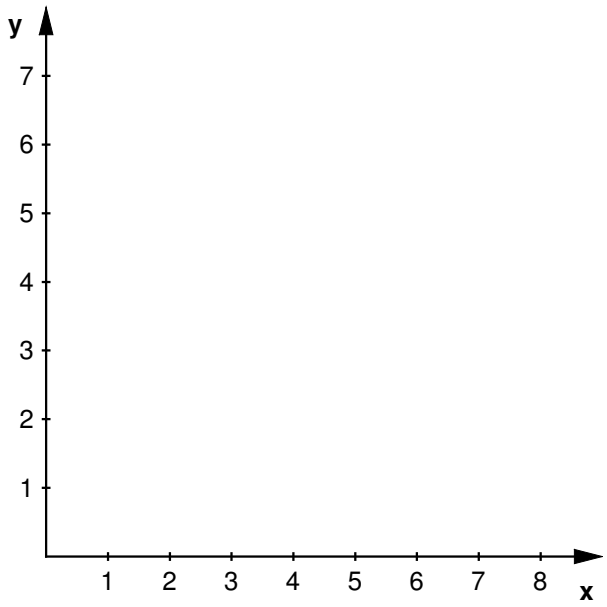
$\overline{x} =$ $\overline{y} =$ $\tilde{s}_x^2 =$ $\tilde{s}_y^2 =$

$$b = \frac{\sum x_i y_i - n\overline{x}\,\overline{y}}{\sum x_i^2 - n\overline{x}^2} =$$

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

Die Regressionsgerade lautet:

Stellen Sie die Beobachtungswerte und die Regressionsgerade grafisch dar.



Berechnen Sie die Kovarianz.

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum x_i y_i - \overline{x}\ \overline{y} =$$

Berechnen Sie den Korrelationskoeffizienten nach Bravais-Pearson.

$$r = \frac{\mathrm{Cov}(X,Y)}{\tilde{s}_x \tilde{s}_y} =$$

Berechnen Sie die Varianz der $\hat{y}$-Werte.

$$\tilde{s}^2_{\hat{y}} =$$

Bestimmen Sie das Bestimmtheitsmaß R^2.

$$R^2 = \frac{\tilde{s}^2_{\hat{y}}}{\tilde{s}^2_y} =$$

Interpretieren Sie den Wert von R^2.

Übungsaufgabe 3.15:

In einer Fußballliga starten 12 *Vereine, die ihren Spielern unterschiedlich hohe Prämien für den Gewinn der Meisterschaft in Aussicht stellen. Die folgende Übersicht zeigt Endstand und Höhe der Prämie. Berechnen Sie den Rangkorrelationskoeffizienten r_s.*

Verein	K	B	E	A	C	D	L	M	G	H	F	J
Platz	1	2	3	4	5	6	7	8	9	10	11	12
Prämie pro Spieler in 1000 €	10	180	150	200	120	50	100	80	60	40	30	20
d_i												
d_i^2												

$r_s =$

Übungsaufgabe 3.16:

Bei 10 *Studenten wurde der durchschnittliche tägliche Kaffeekonsum X während des letzten Semesters und die in der Statistik-Klausur erreichte Punktzahl Y festgehalten.*

Student	1	2	3	4	5	6	7	8	9	10
Kaffeekonsum in l	0.6	0.6	0.8	1.0	1.2	1.4	1.8	1.8	2.0	2.2
Punktzahl	74	86	66	78	58	70	50	66	42	50

Berechnen Sie den Korrelationskoeffizienten nach Spearman (exakter Koeffizient und Koeffizient nach der Formel ohne Bindungen). Ordnen Sie dabei nach „zunehmendem Kaffeekonsum“ bei steigender Rangzahl.

i	x_i	$rg(x_i)$	$rg(x_i)^2$	y_i	$rg(y_i)$	$rg(y_i)^2$	$rg(x_i)rg(y_i)$
1							
2							
3							
4							
5							
6							
7							
8							
9							
10							
			$\sum$			$\sum$	$\sum$

$r_s =$

i	x_i	$rg(x_i)$	y_i	$rg(y_i)$	d_i	d_i^2
1						
2						
3						
4						
5						
6						
7						
8						
9						
10						
						$\sum$

$r_s =$

Übungsaufgabe 3.17:

200 *Männer wurden nach ihrem Berufsstand und dem ihres Vaters befragt:*

	Vater			
Sohn	Arbeiter	Angestellter	Beamter	Selbständig
Arbeiter	40	10	0	0
Angestellter	40	25	5	10
Beamter	10	25	25	0
Selbständig	0	0	0	10

Berechnen Sie nachstehende Größen.

$$\begin{aligned} \chi^2 &= \\ C &= \\ C^* &= \\ C_{korr} &= \end{aligned}$$

Lösung der Übungsaufgaben

Lösung der Aufgaben zu Kapitel 1

Lösung 1.1:

		diskret	*stetig*
a)	*Geschwindigkeit von Pkws*	☐	☒
b)	*Hörerzahl einer Vorlesung*	☒	☐
c)	*Anzahl der Mitarbeiter eines Betriebes*	☒	☐
d)	*Einkommen*	☐	☒
e)	*Zeit für die Beschleunigung eines Pkw´s von 0 auf 100 km/h*	☐	☒
f)	*Punkte in einer Klausur*	☒	☐
g)	*Bücherbestand in einer Bibliothek*	☒	☐
h)	*Stromverbrauch*	☐	☒
i)	*Treibstoffverbrauch eines Orbiters in Abhängigkeit von der Flughöhe*	☐	☒

Lösung 1.2:

a)	*Militärdienstgrad:*	*Ordinalskala*
b)	*Alter:*	*Verhältnisskala*
c)	*Geschlecht:*	*Nominalskala*
d)	*Nationalität:*	*Nominalskala*
e)	*Intelligenzquotient:*	*Intervallskala*
f)	*Semesterzahl:*	*Absolutskala*
g)	*Wassergüte:*	*Ordinalskala*
h)	*Anzahl von Verkehrsunfällen:*	*Absolutskala*
i)	*Tarifklasse der Kfz-Haftpflicht:*	*Ordinalskala*
j)	*Studienfach:*	*Nominalskala*
k)	*Fahrtkosten:*	*Verhältnisskala*

Lösung 1.3:

Klasse Nr.	Klasse	Strichliste	absolute Häufigkeit	relative Häufigkeit
1	1 bis 100	𝍸 𝍸	10	33.33%
2	101 bis 200	𝍸	5	16.67%
3	201 bis 300	𝍷	4	13.33%
4	301 bis 400	𝍶	3	10.00%
5	401 bis 500	𝍵	2	6.67%
6	501 bis 600	𝍷	4	13.33%
7	601 bis 700	𝍵	2	6.67%

Lösung der Aufgaben zu Kapitel 2

Lösung 2.1:

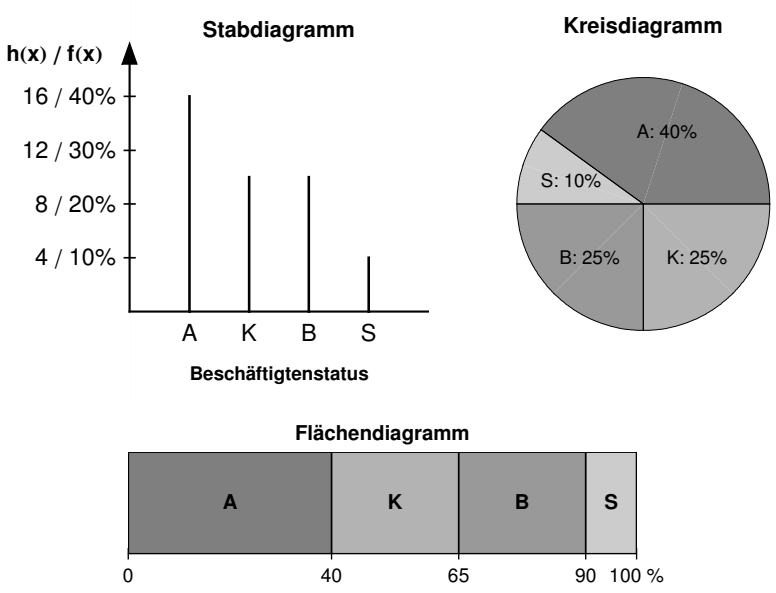
a)

Merkmalsträger:	Personen
Merkmal:	Berufsgruppen
Merkmalsausprägungen:	A,K,B,S

b)

Merkmals-ausprägungen	Strichliste	absolute Häufigkeit	relative Häufigkeit in %
A	𝍸 𝍸 |	16	40
K	𝍸 𝍸	10	25
B	𝍸 𝍸	10	25
S	||||	4	10

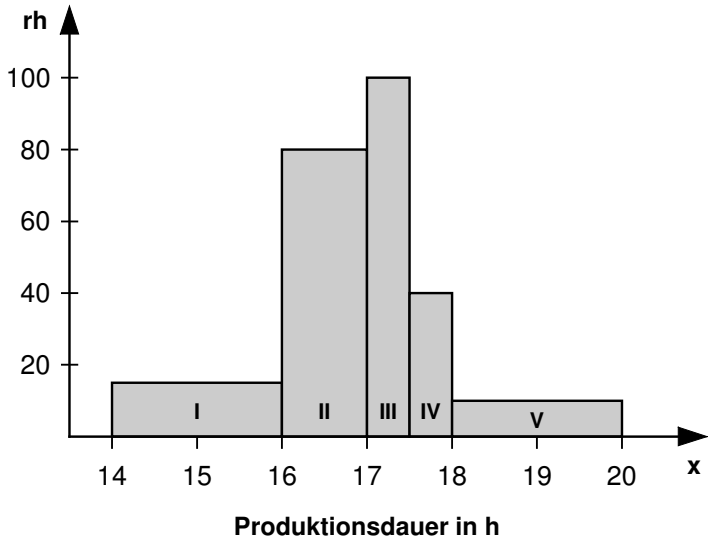
c)



Lösung 2.2:

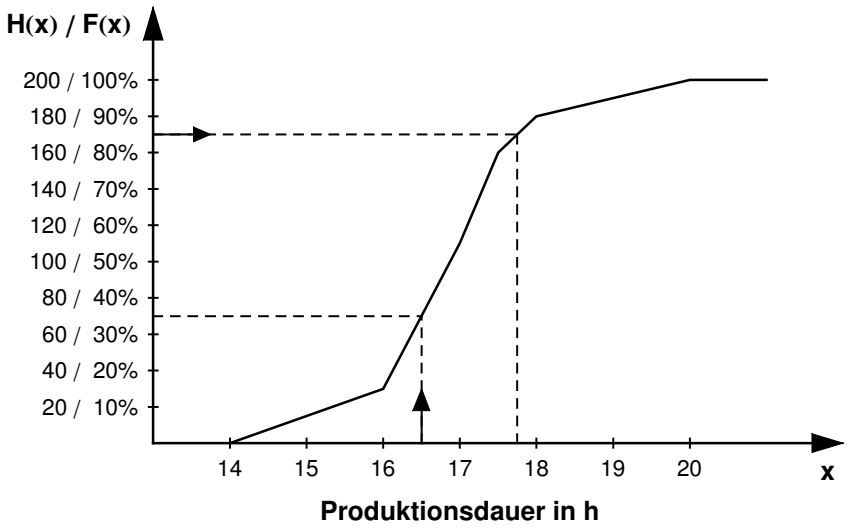
a) Geben sie die grafische Darstellung der Häufigkeitsverteilung in Form eines Histogramms an mit $rh_j = \frac{h_j}{b_j}$.

Klasse	Produktions-dauer in h	absolute Häufigkeit h_j	relative Häufigkeit f_j	$rh_j = \frac{h_j}{b_j}$
I	(14;16]	30	0.15	15
II	(16;17]	80	0.40	80
III	(17;17.5]	50	0.25	100
IV	(17.5;18]	20	0.10	40
V	(18;20]	20	0.10	10



b) Bestimmen Sie die tabellarische und grafische Summenhäufigkeits-verteilung und zeigen Sie anhand der grafischen Summenhäufig-keitsverteilung, bei wie viel % der Werkstücke die Produktionsdauer nicht länger als 16.5 *h ist und welche Produktionsdauer von* 85% *der Werkstücke nicht überschritten wird.*

Produktionsdauer in h	Summenhäufigkeit absolut H_j	Summenhäufigkeit in % F_j
≤ 16	30	15
≤ 17	110	55
≤ 17.5	160	80
≤ 18	180	90
≤ 20	200	100



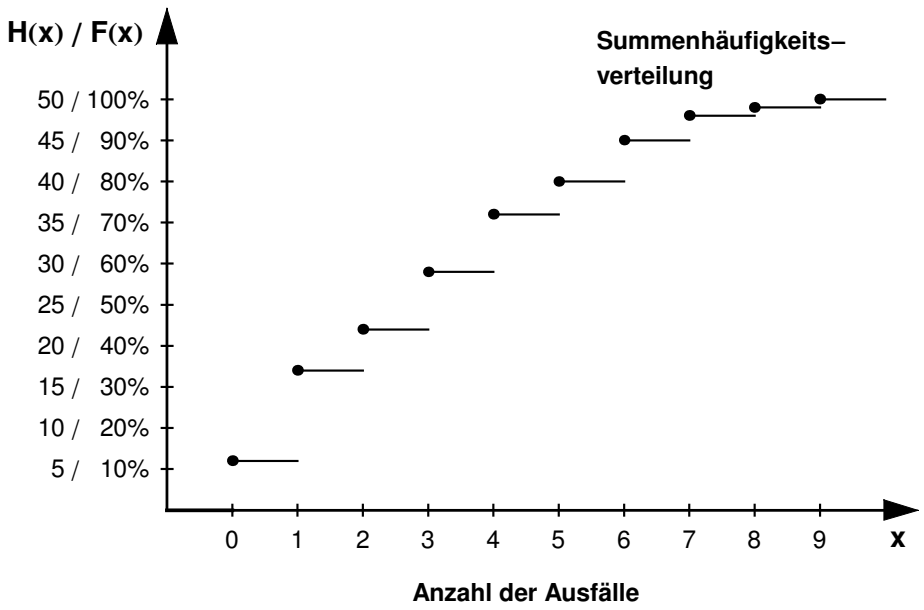
Anhand der grafischen Summenhäufigkeitsverteillung lässt sich erkennen, dass die Produktionsdauer bei 35% *der Werkstücke nicht länger als* 16.5 *h ist und dass eine Produktionsdauer von* 17.75 *h von* 85% *der Werkstücke nicht überschritten wird.*

Lösung 2.3:

a)

Anzahl der Ausfälle	Strichliste	h_j	f_j	H_j	F_j
0	𝍸 |	6	12%	6	12%
1	𝍸𝍸 |	11	22%	17	34%
2	𝍸	5	10%	22	44%
3	𝍸 ||	7	14%	29	58%
4	𝍸 ||	7	14%	36	72%
5	||||	4	8%	40	80%
6	𝍸	5	10%	45	90%
7	|||	3	6%	48	96%
8	|	1	2%	49	98%
9	|	1	2%	50	100%

b)



Lösung 2.4:

		A	B
a)	$x \leqq 20$	60%	70%
b)	$x > 10$	90%	70%
c)	$x \geqq 10$	90%	90%
d)	$10 < x \leqq 30$	70%	60%
e)	$10 < x < 30$	70%	40%

Für $F(x) = 70\%$ *gilt bei A:* $x = 25$ *und bei B:* $x = 20$.

Lösung 2.5:

149 195 **195** **199** 199 **199** **209** 219 229 **229** 249 299

$$x_{mod} = 199$$

$$x_{med} = \frac{199 + 209}{2} = 204$$

$n \cdot p = 12 \cdot 0.5 = 6$, *d.h.* x_{med} *liegt zwischen dem 6. und 7. Wert.*

$$x_{0.25} = \frac{195 + 199}{2} = 197$$

$n \cdot p = 12 \cdot 0.25 = 3$, *d.h.* $x_{0.25}$ *liegt zwischen dem 3. und 4. Wert.*

$$x_{0.8} = 229$$

$n \cdot p = 12 \cdot 0.8 = 9.6$, *d.h.* $x_{0.8}$ *entspricht dem 10. Wert.*

$$\overline{x} = (149 + 2 \cdot 195 + 3 \cdot 199 + 209 + 219 + 2 \cdot 229 + 249 + 299)/12$$
$$= 2570/12 = 214.17$$

Lösung 2.6:

Stundenlohn in €	Klassenmitte x_j	Häufigkeit h_j	$x_j \cdot h_j$	H_j	F_j
(7;8]	7.5	14	105	14	0.093
(8;9]	8.5	40	340	54	0.36
(9;10]	9.5	38	361	92	0.613
(10;11]	10.5	32	336	124	0.827
(11;12]	11.5	26	299	150	1
		$n = 150$	$\sum = 1441$		

Modalklasse: $(8; 9]$

Einfallsklasse des Medians: $(9; 10]$

$$\overline{x} = \frac{1441}{150} = 9.61$$

Lösung 2.7:

t	Kapital am Anfang des Jahres t	Zins z_t	Kapital am Ende des Jahres t
1	1000	4%	$1000 \cdot (1 + 0.04) = 1000 \cdot 1.04$
2	$1000 \cdot 1.04$	4%	$1000 \cdot 1.04 \cdot 1.04$
3	$1000 \cdot (1.04)^2$	5.5%	$1000 \cdot 1.04^2 \cdot 1.055$
4	$1000 \cdot (1.04)^2 \cdot (1.055)$	5.5%	$1000 \cdot 1.04^2 \cdot 1.055^2$
5	$1000 \cdot (1.04)^2 \cdot (1.055)^2$	5.5%	$1000 \cdot 1.04^2 \cdot 1.055^3$
6	$1000 \cdot (1.04)^2 \cdot (1.055)^3$	6%	$1000 \cdot 1.04^2 \cdot 1.055^3 \cdot 1.06$

a) $K_6 = 1000 \cdot 1.04^2 \cdot 1.055^3 \cdot 1.06 = 1346.26$

b) $z = \overline{x}_{geom} - 1 = \sqrt[6]{1.04^2 \cdot 1.055^3 \cdot 1.06} - 1 = 0.051 = 5.1\%$

Mit z als durchschnittlichen Zinssatz und $\overline{x}_{geom}$ als geometrisches Mittel der Zinsfaktoren $1 + z_t$.

Lösung 2.8:

$$\begin{aligned}
\overline{x} &= \frac{1}{10}\sum_{i=1}^{n} x_i \\
&= \frac{1}{10}(1.4 + 1.6 + 1.7 + 1.5 + 1.4 + 1.8 + 1.7 + 1.6 + 1.5 + 1.8) \\
&= \frac{16}{10} = 1.6 \\
\tilde{s}^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\
&= \frac{1}{10}[(-0.2)^2 + 0^2 + 0.1^2 + (-0.1)^2 + (-0.2)^2 + 0.2^2 + 0.1^2 + 0^2 \\
&\qquad + (-0.1)^2 + 0.2^2] \\
&= \frac{1}{10}(0.04 + 0.01 + 0.01 + 0.04 + 0.04 + 0.01 + 0.01 + 0.04) \\
&= \frac{0.2}{10} = 0.02 \\
\tilde{s} &= \sqrt{\tilde{s}^2} = 0.14
\end{aligned}$$

Lösung 2.9:

Stunden	x_j	h_j	$x_j h_j$	$x_j - \overline{x}$	$(x_j - \overline{x})^2$	$(x_j - \overline{x})^2 \cdot h_j$
(600;700]	650	5	3250	-210	44100	220500
(700;800]	750	11	8250	-110	12100	133100
(800;900]	850	16	13600	-10	100	1600
(900;1000]	950	10	9500	90	8100	81000
(1000;1100]	1050	8	8400	190	36100	288800
$\sum$			43000			725000

$$\begin{aligned}
\overline{x} &= \frac{1}{n}\sum_{j=1}^{m} x_j \cdot h_j = \frac{1}{50} \cdot 43000 = 860 \\
\tilde{s}^2 &= \frac{1}{n}\sum_{j=1}^{m} (x_j - \overline{x})^2 \cdot h_j = \frac{725000}{50} = 14500 \\
\tilde{s} &= \sqrt{\tilde{s}^2} = 120.42
\end{aligned}$$

Lösung 2.10:

verfügbares Einkommen	Klassen-mitte x_j	f_j	$x_j f_j$	x_j^2	$x_j^2 f_j$
$\leq$400	350	0.2	70	122500	24500
(400;500]	450	0.4	180	202500	81000
(500;600]	550	0.2	110	302500	60500
(600;700]	650	0.1	65	422500	42250
>700	750	0.1	75	562500	56250
$\sum$			500		264500

$$\begin{aligned}
\overline{x} &= \sum_{j=1}^{m} x_j \cdot f_j = 500 \\
\tilde{s}^2 &= \sum_{j=1}^{m} x_j^2 f_j - \overline{x}^2 = 264500 - 250000 = 14500 \\
\tilde{s} &= \sqrt{\tilde{s}^2} = 120.42
\end{aligned}$$

Lösung 2.11:

x_i cm	120	130	125	130	135
y_i Zoll	48	52	50	52	54

a)

$$\overline{x} = \frac{1}{5}(120 + 130 + 125 + 130 + 135) = \frac{640}{5} = 128$$

$$\overline{y} = \frac{1}{5}(48 + 52 + 50 + 52 + 54) = \frac{256}{5} = 51.2$$

b)

$$\tilde{s}_x^2 = \frac{1}{5}(64 + 4 + 9 + 4 + 49) = \frac{130}{5} = 26$$

$$\tilde{s}_x = \sqrt{26} = 5.1$$

$$\tilde{s}_y^2 = \frac{1}{5}(10.24 + 0.64 + 1.44 + 0.64 + 7.84) = \frac{20.8}{5} = 4.16$$

$$\tilde{s}_y = \sqrt{4.16} = 2.04$$

c)

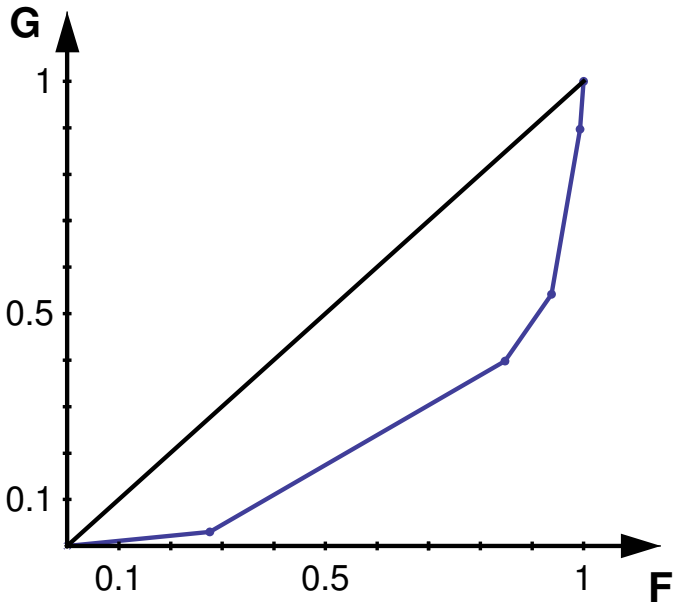
$$v_x = \frac{\tilde{s}_x}{\overline{x}} = \frac{5.1}{128} = 0.03984$$

$$v_y = \frac{\tilde{s}_y}{\overline{y}} = \frac{2.04}{51.2} = 0.03984$$

Lösung 2.12:

x_j	h_j	f_j	$x_j h_j$	$g_j = \frac{x_j h_j}{\sum_{j'=1}^{m} x_{j'} h_{j'}}$	F_j	G_j
0.5	4000	0.276	2000	0.03	0.276	0.03
3	8300	0.572	24900	0.368	0.848	0.398
7.5	1300	0.09	9750	0.144	0.938	0.542
30	800	0.055	24000	0.355	0.993	0.897
70	100	0.007	7000	0.103	1	1
$\sum$	14500	1	67650	1		

$$\begin{aligned}
\text{LKM} &= [(0 + 0.276) \cdot 0.03 + (0.276 + 0.848) \cdot 0.368 + (0.848 + 0.938) \cdot 0.144 \\
&\quad + (0.938 + 0.993) \cdot 0.355 + (0.993 + 1) \cdot 0.103] - 1 \\
&= [0.008028 + 0.413632 + 0.257184 + 0.685505 + 0.205279] - 1 \\
&= 1.56988 - 1 \\
&= 0.57
\end{aligned}$$



Lösung der Aufgaben zu Kapitel 3

Lösung 3.1:

(1,1) (1,3)
(2,1) (2,2) (2,2) (2,3) (2,3) (2,3)
(2,3) (2,5)
(3,1) (3,2) (3,3) (3,3) (3,3) (3,3) (3,3)
(3,3) (3,4) (3,4) (3,4)
(3,4) (3,4)
(4,2) (4,4) (4,5)
(5,3) (5,4)

Lösung 3.2:

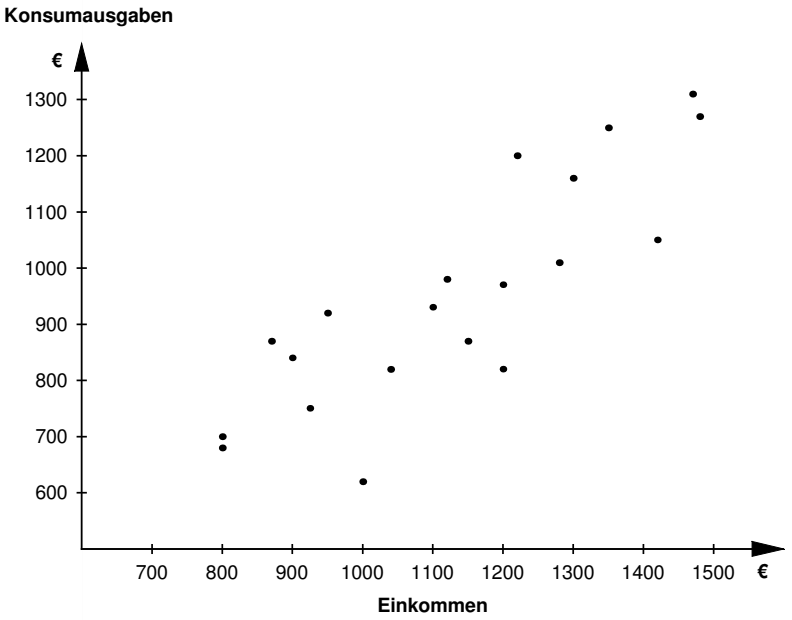
Geschlecht	Immobilien	Aktien	Versicherungen	$\sum$
M	6	4	3	13
F	9	7	1	17
$\sum$	15	11	4	30

Lösung 3.3:

a)

Konsum-ausgaben	Einkommen				$\sum$
	(700;900]	900;1100]	(1100;1300]	(1300;1500]	
(500;700]	2	1	0	0	3
(700;900]	2	2	2	0	6
(900;1100]	0	2	3	1	6
(1100;1300]	0	0	2	2	4
(1300;1500]	0	0	0	1	1
$\sum$	4	5	7	4	20

b)



Lösung 3.4:

	X					
Y	1	2	3	4	5	$\sum$
1	3	5	10	8	4	30
2	5	8	20	20	7	60
6	9	15	50	40	6	120
8	3	12	20	12	3	50
$\sum$	20	40	100	80	20	260

Lösung 3.5:

	Alter				
Einkommen	(0;30]	(30;40]	(40;50]	(50;60]	(60;70]
[0;1000]	14.29%	13.33%	7.69%	11.11%	16.67%
(1000;1500]	28.57%	26.67%	30.77%	33.33%	16.67%
(1500;2000]	42.86%	40%	46.15%	33.33%	33.33%
>2000	14.29%	20%	15.38%	22.22%	33.33%

Lösung 3.6:

Alter	Durchschnittseinkommen
(20;30]	$\frac{1}{7}(1 \cdot 500 + 2 \cdot 1250 + 3 \cdot 1750 + 1 \cdot 2500) = 1535.71$
(30;40]	$\frac{1}{15}(2 \cdot 500 + 4 \cdot 1250 + 6 \cdot 1750 + 3 \cdot 2500) = 1600$
(40;50]	$\frac{1}{13}(1 \cdot 500 + 4 \cdot 1250 + 6 \cdot 1750 + 2 \cdot 2500) = 1615.38$
(50;60]	$\frac{1}{9}(1 \cdot 500 + 3 \cdot 1250 + 3 \cdot 1750 + 2 \cdot 2500) = 1611.11$
(60;70]	$\frac{1}{6}(1 \cdot 500 + 1 \cdot 1250 + 2 \cdot 1750 + 2 \cdot 2500) = 1708.33$

Lösung 3.7:

	$x_1 = 1$	$x_2 = 3$	$x_3 = 6$	$\sum$
$y_1 = 4$	0	0	16	16
$y_2 = 5$	0	10	0	10
$y_3 = 9$	10	0	4	14
$\sum$	10	10	20	40

Die arithmetischen Mittelwerte der Randverteilungen berechnen sich zu:

$$\begin{aligned} \overline{x} &= \frac{1}{40}(1 \cdot 10 + 3 \cdot 10 + 6 \cdot 20) = \frac{160}{40} = 4, \\ \overline{y} &= \frac{1}{40}(4 \cdot 16 + 5 \cdot 10 + 9 \cdot 14) = \frac{240}{40} = 6. \end{aligned}$$

Für die Kovarianz ergibt sich damit:

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \frac{1}{n}\sum\sum x_j y_k h_{jk} - \overline{x}\,\overline{y} \\ &= \frac{1}{40}(1 \cdot 4 \cdot 0 + 1 \cdot 5 \cdot 0 + 1 \cdot 9 \cdot 10 + 3 \cdot 4 \cdot 0 + 3 \cdot 5 \cdot 10 \\ &\qquad + 3 \cdot 9 \cdot 0 + 6 \cdot 4 \cdot 16 + 6 \cdot 5 \cdot 0 + 6 \cdot 9 \cdot 4) - 4 \cdot 6 \\ &= \frac{1}{40}(90 + 150 + 384 + 216) - 24 \\ &= \frac{840}{40} - 24 = 21 - 24 = -3. \end{aligned}$$

Lösung 3.8:

a) $f(x|y)$: $f(y|x)$:

	X			
Y	1	3	9	$\sum$
2	0.2	0.3	0.5	1
10	0.25	0.35	0.4	1
12	0.55	0.35	0.1	1

	X		
Y	1	3	9
2	0.2	0.3	0.5
10	0.25	0.35	0.4
12	0.55	0.35	0.1
$\sum$	1	1	1

Die Merkmale X und Y sind abhängig.

b) $f(x|y)$: $f(y|x)$:

	X				
Y	1	2	4	8	$\sum$
2	0.45	0.27	0.18	0.09	1
4	0.45	0.27	0.18	0.09	1
20	0.45	0.27	0.18	0.09	1

	X			
Yy	1	2	4	8
2	0.57	0.57	0.57	0.57
4	0.29	0.29	0.29	0.29
20	0.14	0.14	0.14	0.14
$\sum$	1	1	1	1

Die Merkmale X und Y sind unabhängig.

Lösung 3.9:

$$\tilde{s}_x = \sqrt{\frac{1}{40}(9 \cdot 10 + 1 \cdot 10 + 4 \cdot 20)} = \sqrt{4.5} = 2.12$$

$$\tilde{s}_y = \sqrt{\frac{1}{40}(4 \cdot 16 + 1 \cdot 10 + 9 \cdot 14)} = \sqrt{5} = 2.24$$

$$r = \frac{\mathrm{Cov}(X,Y)}{\tilde{s}_x \cdot \tilde{s}_y} = \frac{-3}{2.12 \cdot 2.24} = -0.63$$

Lösung 3.10:

$$A: \quad \overline{x} = 3 \quad \tilde{s}_x = \sqrt{2} = 1.414 \quad \overline{y} = 3 \quad \tilde{s}_y = \sqrt{2} = 1.414$$

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \frac{1}{n}\sum_{j=1}^{5}\sum_{k=1}^{5} x_j y_k h_{jk} - \overline{x}\,\overline{y} \\ &= \frac{1}{5}(1\cdot 5 + 2\cdot 4 + 3\cdot 3 + 4\cdot 2 + 5\cdot 1) - 3\cdot 3 = -2 \end{aligned}$$

$$r = \frac{\mathrm{Cov}(X,Y)}{\tilde{s}_x \cdot \tilde{s}_y} = \frac{-2}{1.41 \cdot 1.41} = -1$$

$$B: \quad \overline{x} = 3 \quad \tilde{s}_x = \sqrt{2} = 1.414 \quad \overline{y} = 4 \quad \tilde{s}_y = \sqrt{0.5} = 0.71$$

$$\mathrm{Cov}(X,Y) = \frac{1}{5}(1\cdot 3 + 2\cdot 3.5 + 3\cdot 4 + 4\cdot 4.5 + 5\cdot 5) - 3\cdot 4 = 1$$

$$r = \frac{1}{\sqrt{2}\cdot\sqrt{0.5}} = 1$$

$$C: \quad \overline{x} = 3 \quad \tilde{s}_x = \sqrt{1.5} = 1.225 \quad \overline{y} = 3 \quad \tilde{s}_y = \sqrt{1.5} = 1.23$$

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \frac{1}{8}(1\cdot 3 + 2\cdot 2 + 2\cdot 4 + 3\cdot 1 + 3\cdot 5 + 4\cdot 2 + 4\cdot 4 \\ &+ 5\cdot 3) - 3\cdot 3 = 9 - 9 = 0 \end{aligned}$$

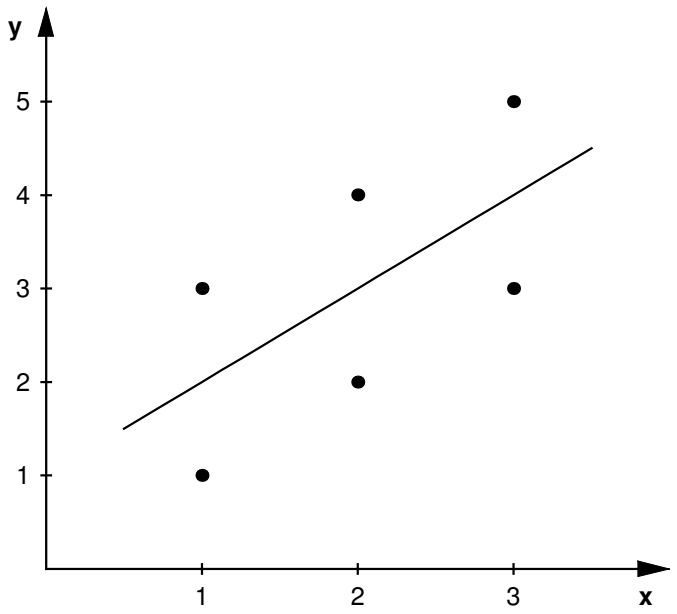
$$r = \frac{0}{1.23 \cdot 1.23} = 0$$

$$D: \quad \overline{x} = 2 \quad \tilde{s}_x = 0.8165 \quad \overline{y} = 2.5 \quad \tilde{s}_y = 0.9574$$

$$\begin{aligned} \mathrm{Cov}(X,Y) &= \frac{1}{6}(1\cdot 1 + 1\cdot 2 + 2\cdot 2 + 2\cdot 3 + 3\cdot 3 + 3\cdot 4) - 2\cdot 2.25 \\ &= \frac{34}{6} - 5 = \frac{4}{6} = 0.6667 \end{aligned}$$

$$r = \frac{0.6667}{0.8165 \cdot 0.9574} = 0.85$$

Lösung 3.11: a), c)



b)

i	x_i	y_i	x_i^2	$x_i y_i$
1	1	1	1	1
2	1	3	1	3
3	2	2	4	4
4	2	4	4	8
5	3	3	9	9
6	3	5	9	15
$\sum$	12	18	28	40

$$\begin{aligned}
\overline{x} &= 2 \\
\overline{y} &= 3 \\
b &= \frac{\sum x_i y_i - n\overline{x}\,\overline{y}}{\sum x_i^2 - n\overline{x}^2} \\
&= \frac{40 - 6 \cdot 2 \cdot 3}{28 - 6 \cdot 2^2} = \frac{4}{4} = 1 \\
a &= \overline{y} - b\overline{x} \\
&= 3 - 1 \cdot 2 = 1
\end{aligned}$$

Regressionsfunktion: $\hat{y} = 1 + x$

Lösung 3.12:

Folgende Tabelle gibt Aufschluss über die Ersparnisse von privaten Haushalten sowie deren verfügbares Einkommen der letzten zehn Jahre (in Mill. €). Es wird vermutet, dass die Ersparnisse annähernd linear vom verfügbaren Einkommen abhängen. Bestimmen Sie die lineare Regressionsfunktion mit Hilfe des Verfahrens der „Kleinsten Quadrate".

Jahr i	Einkommen x_i	Ersparnisse y_i	x_i^2	$x_i \cdot y_i$
1	34.2	2.8	1169.64	95.76
2	40.8	4.1	1664.64	167.28
3	42.5	4.5	1806.25	191.25
4	47.3	4.3	2237.29	203.39
5	50.1	4.9	2510.01	245.49
6	52.6	5.8	2766.76	305.08
7	56.9	7.0	3237.61	398.30
8	61.4	7.7	3769.96	472.78
9	73.5	8.1	5402.25	595.35
10	76.7	8.8	5882.89	674.96
$\sum$	536	58	30447.30	3349.64

$$\sum x_i = 536 \quad \sum x_i^2 = 30447.3 \quad \sum y_i = 58 \quad \sum x_i y_i = 3349.64$$

$$\begin{aligned} b &= \frac{\sum x_i y_i - 10\overline{x}\,\overline{y}}{\sum x_i^2 - 10\overline{x}^2} \\ &= \frac{3349.64 - 10 \cdot 53.6 \cdot 5.8}{30447.3 - 10 \cdot 53.6^2} = \frac{3349.64 - 3108.8}{30447.3 - 28729.6} = \frac{240.84}{1717.7} = 0.14 \end{aligned}$$

$$a = \overline{y} - b\overline{x} = 5.8 - 0.1402 \cdot 53.6 = -1.71$$

Regressionsfunktion: $\hat{y} = -1.71 + 0.14 \cdot x$

Lösung 3.13:

Zeigen Sie, dass gilt:

$$r \;=\; b \cdot \frac{\tilde{s}_x}{\tilde{s}_y}.$$

Die Formel für den quadrierten Korrelationskoeffizienten nach Bravais-Pearson lautet:

$$r^2 = \frac{[\frac{1}{n}\sum(x_i-\overline{x})(y_i-\overline{y})]^2}{[\frac{1}{n}\sum(x_i-\overline{x})^2]\tilde{s}_y^2}.$$

Der Zähler entspricht der Formel für die Kovarianz. Die Varianz $\tilde{s}_y^2$ wird nicht explizit angegeben, da diese in den folgenden Umformungen unverändert bleibt. Eine Erweiterung des Bruches mit $\frac{1}{n}\sum(x_i-\overline{x})^2$ ergibt:

$$r^2 = \left[\frac{\frac{1}{n}\sum(x_i-\overline{x})(y_i-\overline{y})}{\frac{1}{n}\sum(x_i-\overline{x})^2}\right]^2 \cdot \frac{\frac{1}{n}\sum(x_i-\overline{x})^2}{\tilde{s}_y^2}.$$

Für den ersten Faktor gilt:

$$\begin{aligned}
& \left[\frac{\sum x_iy_i - \overline{x}\sum y_i - \overline{y}\sum x_i + n\overline{x}\;\overline{y}}{\sum x_i^2 - 2\overline{x}\sum x_i + n\overline{x}^2}\right]^2 \\
= \;& \left[\frac{\sum x_iy_i - \frac{\sum x_i}{n}\sum y_i - \frac{\sum y_i}{n}\sum x_i + \frac{\sum x_i\sum y_i}{n}}{\sum x_i^2 - 2\frac{\sum x_i\sum x_i}{n} + \frac{(\sum x_i)^2}{n}}\right]^2 \\
= \;& \left[\frac{\sum x_iy_i - \frac{1}{n}\sum x_i\sum y_i}{\sum x_i^2 - \frac{1}{n}(\sum x_i)^2}\right]^2 = \left[\frac{n\sum x_iy_i - \sum x_i\sum y_i}{n\sum x_i^2 - (\sum x_i)^2}\right]^2 = b^2.
\end{aligned}$$

Somit ergibt sich

$$r^2 \;=\; b^2 \cdot \frac{\frac{1}{n}\sum(x_i-\overline{x})^2}{\tilde{s}_y^2} = b^2 \cdot \frac{\tilde{s}_x^2}{\tilde{s}_y^2}.$$

Lösung 3.14:

i	x_i	y_i	x_i^2	$x_i y_i$	y_i^2	$\hat{y}_i$
1	2	5	4	10	25	6
2	2	7	4	14	49	6
3	4	4	16	16	16	5
4	4	6	16	24	36	5
5	5	4.5	25	22.5	20.25	4.5
6	6	3	36	18	9	4
7	6	5	36	30	25	4
8	8	3	64	24	9	3
$\sum$	37	37.5	201	158.5	189.25	

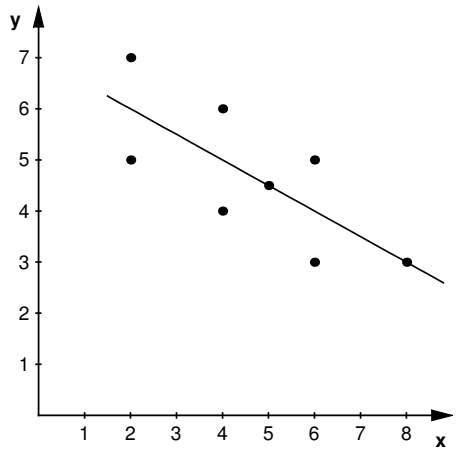
$\overline{x} = 4.625 \qquad \overline{y} = 4.6875 \qquad \tilde{s}_x^2 = 3.7344 \qquad \tilde{s}_y^2 = 1.6836$

$$b = \frac{\sum x_i y_i - n\overline{x}\,\overline{y}}{\sum x_i^2 - n\overline{x}^2} = \frac{158.5 - 8 \cdot 4.625 \cdot 4.6875}{201 - 8 \cdot 4.625^2} = \frac{-14.9375}{29.875} = -0.5$$

$$a = \overline{y} - b\overline{x} = 4.6875 + 0.5 \cdot 4.625 = 7$$

Die Regressionsgerade lautet:

$$\hat{y} = 7 - 0.5x.$$



Berechnen Sie die Kovarianz:

$$\mathrm{Cov}(X,Y) = \frac{1}{n}\sum x_i y_i - \overline{x}\;\overline{y} = \frac{1}{8}158.5 - 4.625 \cdot 4.6875 = -1.8672.$$

Berechnen Sie den Korrelationskoeffizienten nach Bravais-Pearson:

$$r = \frac{\mathrm{Cov}(X,Y)}{\tilde{s}_x \tilde{s}_y} = \frac{-1.8672}{1.932 \cdot 1.298} = -0.7447.$$

Berechnen Sie die Varianz der $\hat{y}$-Werte:

$$\tilde{s}^2_{\hat{y}} = 0.9336.$$

Bestimmen Sie das Bestimmtheitsmaß R^2:

$$R^2 = \frac{\tilde{s}^2_{\hat{y}}}{\tilde{s}^2_y} = \frac{0.9336}{1.6836} = 0.5545.$$

Interpretieren Sie den Wert von R^2:

55.45% *der Varianz der Y-Werte kann durch die lineare Regression erklärt werden.*

Lösung 3.15:

Verein	K	B	E	A	C	D	L	M	G	H	F	J
Platz	1	2	3	4	5	6	7	8	9	10	11	12
Prämie pro Spieler in 1000 €	10 (12)	180 (2)	150 (3)	200 (1)	120 (4)	50 (8)	100 (5)	80 (6)	60 (7)	40 (9)	30 (10)	20 (11)
d_i	-11	0	0	3	1	-2	2	2	2	1	1	1
d_i^2	121	0	0	9	1	4	4	4	4	1	1	1 $\sum$ 150

$$r_s = 1 - \frac{6 \cdot 150}{12(144-1)} = 1 - \frac{900}{1716} = 0.48$$

Lösung 3.16:

i	x_i	$rg(x_i)$	$rg(x_i)^2$	y_i	$rg(y_i)$	$rg(y_i)^2$	$rg(x_i)rg(y_i)$
1	0.6	1.5	2.25	74	3	9	4.5
2	0.6	1.5	2.25	86	1	1	1.5
3	0.8	3	9	66	5.5	30.25	16.5
4	1.0	4	16	78	2	4	8
5	1.2	5	25	58	7	49	35
6	1.4	6	36	70	4	16	24
7	1.8	7.5	56.25	50	8.5	72.25	63.75
8	1.8	7.5	56.25	66	5.5	30.25	41.25
9	2.0	9	81	42	10	100	90
10	2.2	10	100	50	8.5	72.25	85
$\sum$			384			384	369.5

$$\begin{aligned} r_s &= \frac{\sum_{i=1}^{n} rg(x_i)rg(y_i) - n\overline{rg}_X\overline{rg}_Y}{\sqrt{\sum_{i=1}^{n} rg(x_i)^2 - n\overline{rg}_X^2}\sqrt{\sum_{i=1}^{n} rg(y_i)^2 - n\overline{rg}_Y^2}} \\ &= \frac{369.5 - 302.5}{\sqrt{384 - 302.5}\sqrt{384 - 302.5}} = \frac{67}{81.5} = 0.822 \end{aligned}$$

i	x_i	$rg(x_i)$	y_i	$rg(y_i)$	d_i	d_i^2
1	0.6	1.5	74	3	-1.5	2.25
2	0.6	1.5	86	1	0.5	0.25
3	0.8	3	66	5.5	-2.5	6.25
4	1.0	4	78	2	2	4
5	1.2	5	58	7	-2	4
6	1.4	6	70	4	2	4
7	1.8	7.5	50	8.5	-1	1
8	1.8	7.5	66	5.5	2	4
9	2.0	9	42	10	-1	1
10	2.2	10	50	8.5	1.5	2.25
					$\sum$	29

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 29}{10 \cdot (100-1)} = 1 - \frac{174}{990} = 0.824$$

Lösung 3.17:

Sohn	Vater: Arbeiter	Vater: Angestellter	Vater: Beamter	Vater: Selbständig	$\sum$
Arbeiter	**40** 306.25 −17.5 22.5	**10** 25 −5 15	**0** 56.25 −7.5 7.5	**0** 25 −5 5	50
Angestellter	**40** 16 6 36	**25** 1 1 24	**5** 49 −7 12	**10** 4 2 8	80
Beamter	**10** 289 −17 27	**25** 49 7 18	**25** 256 14 9	**0** 36 −6 6	60
Selbständig	**0** 20.25 −4.5 4.5	**0** 9 −3 3	**0** 2.25 −1.5 1.5	**10** 81 9 1	10
$\sum$	90	60	30	20	200

$$\begin{aligned}
\chi^2 &= \frac{306.25}{22.5} + \frac{25}{15} + \frac{56.25}{7.5} + \frac{25}{5} + \frac{16}{36} + \frac{1}{24} + \frac{49}{12} + \frac{4}{8} \\
&\quad + \frac{289}{27} + \frac{49}{18} + \frac{256}{9} + \frac{36}{6} + \frac{20.25}{4.5} + \frac{9}{3} + \frac{2.25}{1.5} + \frac{81}{1} \\
&= 13.61 + 1.67 + 7.5 + 5 + 0.44 + 0.04 + 4.08 + 0.5 \\
&\quad + 10.7 + 2.72 + 28.44 + 6 + 4.5 + 3 + 1.5 + 81 \\
&= 170.7 \\
C &= \sqrt{\frac{170.7}{200 + 170.7}} = 0.6786 \\
C^* &= 4 \\
C_{korr} &= 0.6786\sqrt{\frac{4}{4-1}} = 0.7836
\end{aligned}$$