Statistik

Definition
Statistik umfasst die Entwicklung und Anwendung von Methoden zur Erhebung, Aufbereitung, Analyse und Interpretation von Daten.

Drei Teilgebiete
- **Beschreibende Statistik (Deskriptive Statistik)**
  - In der deskriptiven Statistik geht es um beobachtete Merkmalsausprägungen und deren Verdichtung.
- **Wahrscheinlichkeitsrechnung**
  - In der Wahrscheinlichkeitsrechnung geht es um Resultate von Zufallsvorgängen und Gesetzmässigkeiten von Zufallsvariabeln.
- **Schliessende Statistik (Verteilungen)**
  - Die Schliessende Statistik ermöglicht von Stichproben auf die Grundgesamtheit zu schliessen.

1. Deskriptive Statistik

1.1 Einführung

- **Drei Schritte zur deskriptiven Statistik**

  1. Schritt: Vollständige Erhebung der relevanten Daten des Untersuchungsobjekts.
  2. Schritt: Aufbereitung (Tabelle, Grafik) des gewonnenen Datenmaterials.
  3. Schritt: Analyse der aufbereiteten Daten durch Berechnung von Kennzahlen (Mittelwert, Streuungsmass) oder durch Erkennen von Gesetzmässigkeiten oder durch Berechnen des Abhängigkeitsausmasses

1.1.1 Grundbegriffe

- **Merkmalsträger (1)** (auch: Element, statistische Einheit, Untersuchungseinheit)
  - Gegenstand der statistischen Untersuchung (z.B. Fernsehzuschauer, Kunden, Kleider)
  - Träger der interessierenden statistischen Information
- **Grundgesamtheit (2)** (auch: Kollektiv, statistische Gesamtheit, statistische Masse, Gesamtheit, Masse)
  - Gesamtheit aller Merkmalsträger (z.B. alle Schweizer Bürger, alle Computer einer Produktionsserie etc.)
  - Betrachtet man nur eine Teilmenge der Grundgesamtheit, spricht man von **Stichprobe.** Eine Stichprobe liegt nur dann vor, wenn sie repräsentativ ist, also ein Miniaturbild der Grundgesamtheit.
  - Die Grundgesamtheit muss für eine qualitativ gute statistische Untersuchung exakt abgegrenzt sein:
    - **Inhaltliche Abgrenzung**
      - Festlegung von Abgrenzung- oder Identifikationsmerkmalen
      - Ein Merkmalsträger gehört zur Grundgesamtheit, wenn er alle Abgrenzungsmerkmale besitzt.
    - **Sachliche Abgrenzung**
      - Ist ein Ausgesteuerter ein Arbeitsloser?
      - Wert gilt als Mitarbeiter/Kunde etc?
    - **Räumliche Abgrenzung**
      - Inländer- oder Inlandprinzip bei der volkswirtschaftlichen Gesamtrechnung
    - **Zeitliche Abgrenzung**
      - **Zeitpunkt** oder
        - Wer zu einem bestimmten Zeitpunkt Angestellter einer Unternehmung ist, ist Merkmalsträger und gehört zur Grundgesamtheit.
      - **Zeitraum**
        - Falls Ereignisse zur Grundgesamtheit gehören, ist ein Ereignis genau dann Element der Grundgesamtheit, wenn es innerhalb des definierten Zeitraumes anfällt.

1.1.2 Merkmalsarten

- **Merkmal (3)** (auch: Untersuchungsmerkmale, Prädikatsmerkmal, statistisches Merkmal, Untersuchungsvariable, Variable)
  - Jene Eigenschaft des Merkmalsträgeres, die bei der statistischen Untersuchung von Interesse ist. (z.B. Alter, Geschlecht, Grösse, Parteizugehörigkeit etc.)
  - Als Symbol wird in der Regeln ein Grossbuchstaben (X, Y, Z) verwendet.
- **Merkmalswert (4) (Ausprägung, Wert, Modalität)**
  Jedes Merkmal hat eine bestimmte Ausprägung, einen bestimmten Wert.
  - Ermittlung der Merkmalswerte:
    - Beobachtung
    - Befragung
    - Messung
    - Zählung

Merkmalsträger (1)	Merkmal (3)	Merkmalswert (4)	Art
Personen	Familienstand	ledig, verheiratet, verwitwet, geschieden	qualitativ
Kunden	TV-Eigentum	Eigentümer / Nicht-Eigentümer	qualitativ
Parlamentarier	Parteizugehörigkeit	FDP, SP, CVP, SVP, Grüne etc.	qualitativ
Fernsehzuschauer	Beurteilung e. Sendung	••••• / •••• / ••• / •• / •	qualitativ
Wohnungsmieter	Anzahl Zimmer	1, 1 ½, 2, 2 ½ …	quan.disk
Personen	Körpergrösse	155 – 210 cm	quan.stet
Betriebsangehörige	monatliches Einkommen	Fr. 3'500.00 – Fr. 10'000.00	quan.stet
Motoren	km-Leistung	0 – 200'000 km	quan.stet
Betriebe einer Region	Anzahl Beschäftigte	1, 2, 3… Mitarbeiter	quan.disk
Kleider	Grösse	S, M, L, XL, XXL	qualitativ
Leser von Zeitungen	Zeitungstitel	NZZ, LZ, Blick etc.	qualitativ

- **Merkmalsarten**
  - Qualitative und quantitative Merkmale
    - **Qualitatives Merkmal** (auch kategoriales Merkmal)
      Den Merkmalswerten können lediglich Namen oder Klassenbezeichnungen zugeordnet. Beispiele:
      - Parteizugehörigkeit → FDP, SP, CVP
      - Beurteilung einer Sendung → ••••• / •••• / ••• / •• / •
    - **Quantitatives Merkmal** (auch metrisches Merkmal)
      Ein Merkmal, das eine messbare Dimension besitzt oder in Mengeneinheiten ausgedrückt werden kann.
  - Quantitative (metrische) Merkmale können in Diskrete und stetige Merkmale eingeteilt werden:
    - **Diskretes Merkmal** (diskontinuierliches Merkmal)
      Ein quantitatives Merkmal, das abzählbar (unendlich aber nicht mit unendlichen Zwischenwerten) viele Werte annehmen kann. z.B. Anzahl Mitarbeiter (ein halber Mitarbeiter ist nicht möglich), Anzahl Stück (ein halbes Stück ist nicht möglich)
      - **Dichotome Merkmale** → nur zwei Merkmalswerte
      - **Polytome Merkmale** → mehr als zwei Merkmalswerte
    - **Stetiges Merkmal** (kontinuierliches Merkmal)
      Ein quantitatives Merkmal, das überabzählbar (unendlich mit unendlichen Zwischenwerten) viele Werte annehmen kann. z.B. Körpergrösse, 1.8333m (unendlich)
  - Häufbare und nicht-häufbare Merkmale
    - **Häufbares Merkmal** (immer ein qualitatives Merkmal)
      - Ein Merkmalsträger kann mehr als einen Merkmalswert besitzen.
      - Z.B. Zeitungstitel → Jemand kann 2 Zeitungen lesen!
    - **Nicht-häufbares Merkmal** (qualitatives oder quantitatives Merkmal)
      - Ein Merkmalsträger kann genau einen Merkmalswert besitzen.
      - Z.B. Familienstand → Man kann nur genau einen Familienstand haben

1.1.3 Skalierungen

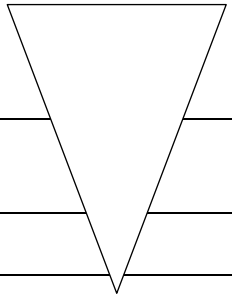
- Instrument zur Ermittlung der Merkmalswerte durch Beobachtung, Befragung, Messung oder Zählung.
- Wahl der Skalierung ist entscheidend für das Informationsniveau, den Aussagegehalt des Merkmalswertes und für das statistische Verfahren, dass eingesetzt werden darf.
- **a) Nominalskala**
  - Namen als Skalenwerte (immer bei qualitativen Merkmalen und häufbare Merkmale)
  - Keine Reihenfolge (Gleichberechtigung)

Merkmal	Merkmalswert
Geschlecht	männlich, weiblich
Familienstand	ledig, verheiratet, geschieden, verwitwet
Rebsorte	Silvaner, Riesling, Traminer etc.

- **b) Ordinalskala**
  - Klassenbezeichnungen als Skalenwerte
  - Keine Gleichberechtigung (auf- od. absteigende Folge) (immer intensitätslässig abgestufte Merkmale)

Merkmal	Merkmalswert
Schulnote	sehr gut, gut, befriedigend, mangelhaft
Wein-Qualität	Tafelwein, Landwein … Eiswein

- **c) metrische Skala oder Kardinalskala**
  - reelle Zahlen (immer bei quantitativen Merkmalen)
  - Je nach Ort des Nullpunktes unterscheidet man:
    - **c1) Intervallskala**
      - Intervall spielt eine Rolle
      - Verhältnisse zwischen den gemessenen Werten haben keine Bedeutung
        - Beispiel: Temperatur an zwei Tagen:
          Tag 1: 12 ° Tag 2: 36 °
          Es ist richtig zu sagen, dass zwischen den Tagen 24° Temperatur-unterschied herrschte. Jedoch ist es falsch zu sagen, dass es am Tag 2 dreimal wärmer war, als am Tag 1. 0° ist ein willkürlich gewählter Nullpunkt.
        - Z.B. Temperatur (um 8 Uhr ist es nicht doppelt so spät wie um 4 Uhr), Uhrzeit, Kalenderzeit
    - **c2) Verhältnisskala**
      - Intervall spielt eine Rolle
      - Verhältnis spielt auch eine Rolle
        - Beispiel: Einkommen
          Haushalt 1: 24'000.00 Haushalt 2: 72'000.00
          Es ist richtig zu sagen, dass zwischen den Haushalten 48'000.00 Unterschied im Einkommen herrscht. Es ist ebenfalls richtig zu sagen, dass der Haushalt 2 das dreifache Einkommen von Haushalt 1 hat. Nullpunkt ist nicht willkürlich gewählt sondern ist bei 0.
        - z.B. Einkommen, Gewicht, Alter
- **Bedeutung der Messskala**
  - Informationsgehalt
    - Das Informationsniveau ist am grössten bei der Verhältnisskala und am kleinsten bei der Nominalskala.

Skala	Informationsgehalt	Was kann festgestellt werden?
Verhältnisskala		Verschiedenartigkeit, Rangfolge, einfache Abstände und verhältnismässige Abstände
Intervallskala		Verschiedenartigkeit Rangfolge und einfache Abstände
Ordinalskala		Verschiedenartigkeit und Rangfolge
Nominalskala		Verschiedenartigkeit

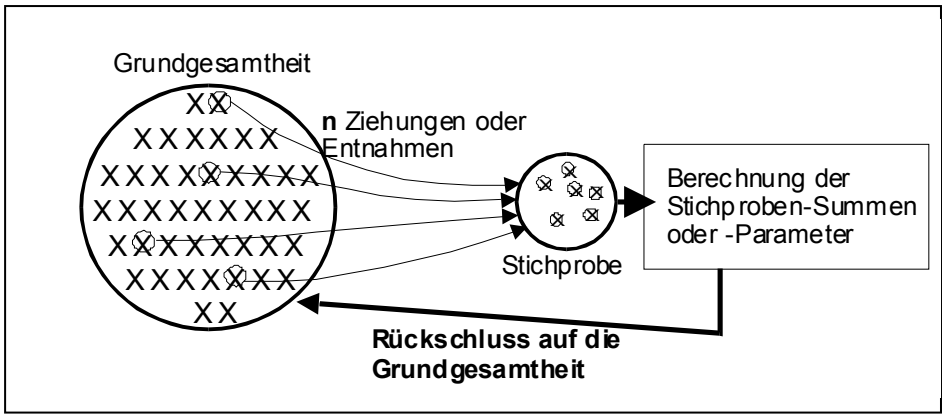
  - Feinheit der Skalierung
    - Höherstehende Skalen (Intervall- und Verhältnisskala) erlauben eine feinere Skalierung. Die Zuordnung der Zahlenwerte bei höherstehenden Skalen ist objektiv möglich und unterliegt nicht subjektiven Einflüssen. Bei qualitativen Merkmalen ist die Zuordnung oft von subjektiven Einflüssen mitgeprägt.
  - Gebrauch statistischer Verfahren
    - Je nach Skalenniveau können unterschiedliche statistische Verfahren zur Anwendung kommen. Beispielsweise kann ein arithmetisches Mittel nur dann angewendet werden, wenn der Abstand zwischen den Merkmalswerten bekannt ist (daher mindestens die Intervallskala)

1.1.4 Statistische Erhebungsarten

- **4 Phasen der Statistischen Untersuchung**
  - **1. Planung**
    - Welche Merkmale sind bei welchen Merkmalsträgern mit welcher Technik zu erheben, welche Aufbereitungsverfahren einzusetzen, welche Formen der Darstellung und welche statistischen Analyseverfahren sind zu wählen?
  - **2. Datenerhebung** *(sh. Erhebungsarten weiter unten)*
    - Aufgabe: die für das Untersuchungsziel relevanten Daten erfassen
      - a) das Untersuchungsziel ist zu konkretisieren und
      - b) die Erhebungstechnik ist festzulegen

- **3. Datenaufbereitung**
  - Daten liegen in Form von Fragebögen, Beobachtungsprotokollen, Interviews oder Versuchsprotokollen vor.
  - Die Daten müssen geprüft, kontrolliert, ausgezählt und in tabellarische und/oder grafische Form gebracht werden. *(sh. 1.2 Bearbeitung und Darstellung von Datenbeständen)*
- **4. Datenanalyse und -interpretation**
  - Die beschreibende Statistik umfasst vier Analyse- und Interpretationsfelder. Die Trennung von der Darstellung der Daten kann nicht immer streng gezogen werden.
    - Niveau der Häufigkeitsverteilung *(sh. 1.3.1 Lageparameter)*
    - Struktur der Häufigkeitsverteilung *(sh. 1.3.3 Streuungsparameter)*
    - Zusammenhang zwischen Merkmalen *(sh. 1.4 Regressions- und Korrelationsrechnung)*
    - Entwicklung von Zeitreihen *(sh. 1.4.2 Methode der kleinsten Quadrate)*

- **Erhebungsarten**
  - Nach der **Herkunft der Daten:**
    - **Primärstatistik** → neue Daten
    - **Sekundärstatistik** → Rückgriff auf bereits vorliegendes Datenmaterial (evtl. nicht mehr aktuell)
  - Nach dem **Erhebungsumfang:**
    - **Vollerhebung**
      - Sämtliche Merkmalsträger der Grundgesamtheit (G) werden erfasst.
      - Womöglich Aktualitätsverlust da hoher Zeitumfang und Kosten.
      - Anzahl Elemente von G: $n = |G|$
    - **Teilerhebung**
      - Nur ein Teil der Merkmalsträger der Grundgesamtheit (G) werden erfasst.
      - Unter einer Stichprobe (S) verstehen wir einen sachadäquat ausgewählten Ausschnitt oder eine Teilmenge, die möglichst wirklichkeitsgetreue qualitative oder quantitative statistische Aussagen über eine Grundgesamtheit zulassen. Umfang der Stichprobe: $n = |S|, S \subset G, \quad |S| < |G|$
        Die Bewertung von Stichproben im Verhältnis zur Grundgesamtheit geschieht in der Wahrscheinlichkeitsrechnung und in der schliessenden Statistik.



  - Nach Art der **Erhebung (Methode der Datenerfassung):**
    - **Beobachtung** (per Augenschein durch das Erhebungspersonal)
    - **Befragung (schriftlich oder mündlich)** (Probleme: Verweigerung, Falschausdrücke)
    - **Experiment**
    - **Automatische Erfassung** (Messgeräte und Aufzeichnungsvorrichtungen)
  - Nach der **Häufigkeit der Datenerfassung**
    - **Einmalige Erhebung**
    - **Laufende Erhebung**
  - Nach der **Zeitfolge der Datenerfassung**
    - **Querschnittserhebung**
      - Zustandsbilder, die sich auf einen bestimmten Beobachtungszeitpunkt oder einen ausgewählten einzelnen Beobachtungszeitraum beziehen.
      - Veränderungen des Zustandes im Zeitablauf interessieren nicht.
    - **Längsschnitterhebung** (Zeitreihe)
      - Betrachtung von Merkmalsausprägungen im Zeitablauf
      - Längsschnittdaten erhält man dadurch, dass man im Zeitablauf immer wieder neu zählt oder misst.

1.2 Bearbeitung und Darstellung von Datenbeständen

1.2.1 Aufbereitung und Klassierung von Daten / Tabellarische Darstellungsarten

- **Urliste**
  - Nach der Erhebung liegen die Daten in Form einer Urliste vor.
  - Die Merkmalswerte und evtl. auch die Merkmalsträger sind zufällig, alphabetisch, in der Reihenfolge der Befragung oder sonst wie aufgereiht.

Nr.	Name	Familienstand	Anzahl Kinder	Tarifgruppe
01	Amberger, Heinz	ledig	0	II
02	Bauer, Regine	verheiratet	2	I
03	Bertram, Günther	geschieden	1	II
04	Dünnes, Rita	ledig	0	I
05	Engel, Erika	verheiratet	1	II
06	Frühauf, Ernst	verwitwet	1	III
07	Frisch, Anton	verheiratet	3	II
08	Gillhuber, Erwin	geschieden	0	III
09	Hell, Marion	ledig	0	II
10	Jahn, Josef	verheiratet	2	II
11	Kaps, Wolfgang	verwitwet	0	III
12	Lechner, Ernst	verheiratet	4	II
13	Maier, Waltraud	ledig	0	II
14	Mayer, Elisabeth	ledig	1	I
15	Pagler, Fritz	ledig	0	I
16	Polzer, Herrmann	verheiratet	2	IV
17	Rabe, Armin	verheiratet	3	III
18	Reiser, Gabriele	geschieden	2	II
19	Schmidt, Heinz	verheiratet	1	IV
20	Zube, Karl	verheiratet	1	IV

- **Streichliste**
  - Um herauszufinden, wie viele Beschäftigte wie viele Kinder haben kann man eine Streichliste anlegen.
  - Auf diese Weise werden Merkmalsträger mit identischen Merkmalen zusammengefasst.
  - Sind die Merkmalswerte bereits vor der Erhebung bekannt, kann die Urliste direkt als Strichliste aufgenommen werden.

Anzahl Kinder	Anzahl der Beschäftigten
0	𝍸 II
1	𝍸 I
2	IIII
3	II
4	I

- **Häufigkeitstabelle**
  - Mit der Streichliste kann man die Häufigkeitstabelle erstellen.
  - Die Häufigkeitstabelle gibt die Häufigkeitsverteilung eines Merkmals wieder.
  - Verglichen mit der Urliste geht die Charakterisierung des einzelnen Merkmalsträgers verloren. Dafür wird aber die Gesamtheit der Merkmalsträger übersichtlich beschrieben.

Anzahl Kinder	Anzahl der Beschäftigten
0	7
1	6
2	4
3	2
4	1

- **A Eindimensionale Häufigkeitsverteilungen**
  - Statistische Untersuchungen die sich auf ***ein einziges Merkmal*** beziehen.
    - **a) Einfache Häufigkeitsverteilungen**

Die einfache Häufigkeit gibt an, wie häufig ein Merkmalswert x_i aufgetreten ist. Die einfache Häufigkeit kann absolut oder relativ ausgedrückt werden.

h_i absolute einfache Häufigkeit
Anzahl Merkmalsträger mit Merkmal x_i (Merkmalswert) (die Summe ergibt die Gesamtanzahl Merkmalsträger)

f_i relative einfache Häufigkeit
Anteil Merkmalsträger mit Merkmal x_i (Merkmalswert) (Quotient; die Summe ergibt 1)

Die relative Häufigkeit f_i ergibt sich als Quotient $\frac{h_i}{n}$ (Prozentwert)
Die Summe aller relativen Häufigkeiten (bei nicht-häufbaren Merkmalen) ist 1.

Serie 1
Aufg. 3 a)

n Gesamtzahl der Merkmalsträger $\sum_{i=1}^{v} h_i = n$

v Anzahl verschiedener Merkmalswerte

Anzahl Kinder	Anzahl Beschäftigte	Anteil der Beschäftigten	Laufindex i
x_i	h_i	f_i	i
0	7	0.35 (7/20) (35%)	1
1	6	0.30 (6/20)	2
2	4	0.20 (4/20)	3
3	2	0.10 (2/20)	4
4	1	0.05 (1/20)	5
	20 n	1	

- **b) Kumulierte Häufigkeitsverteilungen**

Die kumulierte Häufigkeit (Summenhäufigkeit) gibt die Anzahl bzw. den Anteil der Merkmalsträger an, die einen bestimmten Merkmalswert nicht überschreiten.

Serie 1
Aufg. 3 a)

H_i absolute kumulierte Häufigkeit,
Anzahl der Merkmalsträger mit einem Merkmalswert (x_i), der kleiner oder gleich x_i ist.

$$H_i = \sum_{k=1}^{i} h_k$$

F_i relative kumulierte Häufigkeit,
Anteil der Merkmalsträger mit einem Merkmalswert (x_i), der kleiner oder gleich x_i ist. (Quotient)
(z.B. 85 % der Beschäftigten haben weniger oder 2 Kinder.)

$$F_i = \sum_{k=1}^{i} f_k = \frac{H_i}{n}$$

Zahl der Kinder	Anzahl der Beschäftigten	Anteil der Beschäftigten	Absolute kumulierte Häufigkeiten	Relative kumulierte Häufigkeiten	Absolute kumulierte Resthäufig-keiten	Relative kumulierte Resthäufig-keiten
x_i	h_i	f_i	H_i	F_i	HR_i	FR_i
0	7	0.35 (7/20)	7	0.35	13	0.65
1	6	0.30 (6/20)	13 (7+6)	0.65	7	0.35
2	4	0.20 (4/20)	17 (7+6+4)	0.85	3	0.15
3	2	0.10 (2/20)	19 (7+6+4+2)	0.95	1	0.05
4	1	0.05 (1/20)	20 (7+6+4+2+1)	1	0	0.00
	20 **n**	1				

Resthäufigkeit:

HR_i absolute kumulierte Resthäufigkeit
Anzahl der Merkmalsträger mit einem Merkmalswert (x_i), der grösser x_i ist.

$HR_i = n - H_i$ (Gesamtanzahl der Merkmalsträger – absolute kumulierte Häufigkeit)

FR_i relative kumulierte Resthäufigkeit
Anzahl der Merkmalsträger mit einem Merkmalswert (x_i), der grösser x_i ist.

(z.B. 5 % der Merkmalsträger mehr als 3 Kinder.)

$FR_i = 1 - F_i$ (1 – relative kumulierte Häufigkeit)

- **B Mehrdimensionale Häufigkeitsverteilungen**
  - Statistische Untersuchungen können sich aber auch auf mehrere Merkmale beziehen.Eine überschaubare tabellarische Darstellung ist in der Regel nur mit zwei Merkmalswerten möglich.
    - **a) Absolute einfache Häufigkeiten**

Merkmal X: Tarifgruppe, Zeilenindex i (i =1, …., v=4)
Merkmal Y: Zahl der Kinder, Spaltenindex k (k=1, …, w=5)
(gemäss Urliste Seite 5)

Serie 1
Aufg. 4
Aufg. 5

In der Vorspalte stehen die Merkmalswerte für Merkmal X, in der Kopfzeile die Merkmalswerte für Merkmal Y.

Addiert man die Häufigkeiten zeilenweise erhält man in der Summenspalte die eindimensionale Verteilung für das Merkmal X. In der Summenspalte stehen also die Anzahl Beschäftigte mit den Tarifgruppen I bis IV.
Addiert man die Häufigkeiten spaltenweise erhält man in der Summenzeile die eindimensionale Verteilung für das Merkmal Y. In der Summenzeile stehen also die Anzahl Beschäftigten mit 0 bis 4 Kindern.
Im Schnittpunkt von Summenzeile und Summenspalte steht die Gesamtzahl der Merkmalsträger (20).

4 Kinder (Laufindex 5)
0 Kinder (Laufindex 1)

yk / xi	0	1	2	3	4	$\sum_{k=1}^{5} h_{ik}$
I	2	1	1	0	0	4
II	3	2	2	1	1	9
III	2	1	0	1	0	4
IV	0	2	1	0	0	3
$\sum_{k=1}^{4} h_{ik}$	7	6	4	2	1	20

Wie viele Beschäftigte sind in der Tarifgruppe I und haben 0 Kinder?

Summenzeile

Summenspalte

- **b) Absolute einfache (einf) und kumulierte (kum) Häufigkeiten**

y_k / x_i	0		1		2		3		4		$\sum_{k=1}^{5} h_{ik}$
	einf	kum	einf	kum	einf	kum	einf	kum	einf	kum	
I	2	2	1	3 (1+2)	1	4 (1+1+2)	0	4	0	4	4
II	3	5	2	8 (1+2+3+2)	2	11 (1+1+2+2+2+3)	1	12	1	13	9
III	2	7	1	11 (1+2+3+2+1+2)	0	14	1	16	0	17	4
IV	0	7	2	13 (1+2+3+2+1+2+2)	1	17	0	19	0	20	3
$\sum_{k=1}^{4} h_{ik}$	7		6		4		2		1		20

Die Berechnung der absoluten kumulierten Häufigkeiten ergibt sich für die kumulierte absolute Häufigkeit in Zeile i und Spalte k, indem man die absoluten einfachen Häufigkeiten über alle Zeilen bis und mit i und alle Spalten bis und mit k aufsummiert.

Beispiel: Die Zahl 14 setzt sich aus den blau-eingerahmten Zahlen zusammen. Das will heissen dass die kumulierte Häufigkeit 14 die Merkmalsträger mit einem Merkmalswert der kleiner oder gleich misst, auf beide Merkmalswerte bezogen.

- **b) Relative einfache (einf) und kumulierte (kum) Häufigkeiten**
  Die einfachen relativen Häufigkeiten ergeben sich aus den einfachen absoluten Häufigkeiten durch Division durch n (Gesamtzahl der Merkmalsträger: 20) $\sum_{k=1}^{5} h_{ik}$

  Die kumulierten relativen Häufigkeiten ergeben sich aus den einfachen relativen Häufigkeiten durch die gleiche Summation mit der wir oben die absoluten kumulierten Häufigkeiten aus den absoluten einfachen Häufigkeiten errechnet haben. Oder man erhält dies relativen kumulierten Häufigkeiten auch indem man die absoluten kumulierten Häufigkeiten durch n (20) teilt.

  Die Summe der Summenspalte sowie der Summenzeile ergibt 20.

y_k / x_i	0		1		2		3		4		
	einf	kum	einf	kum	einf	kum	einf	kum	einf	kum	
I	0.1	0.1	0.05	0.15	0.05	0.2	0	0.2	0	0.2	0.2
II	0.15	0.25	0.1	0.4	0.1	0.55	0.05	0.6	0.05	0.65	0.45
III	0.1	0.35	0.05	0.55	0	0.7	0.05	0.8	0	0.85	0.2
IV	0	0.35	0.1	0.65	0.05	0.85	0	0.95	0	1	0.15
$\sum_{k=1}^{4} h_{ik}$	0.35		0.3		0.2		0.1		0.05		1

- **C Klassifizierte Häufigkeitsverteilungen**
  Die oben besprochene tabellarische Darstellung von Häufigkeitsverteilungen ist nur möglich, wenn nicht zu viele Merkmalswerte vorkommen. Bei mehr als etwa 10 bis 15 Merkmalswerten muss man ein anderes Vorgehen wählen.

  Angenommen man hätte die Rechnungsbeträge von 140 Kunden. Statt nun eine Liste zu erstellen, die im Maximum 140 Merkmalle umfassen könnte, werden Klassen gebildet und es wird dann gezählt, wie viele Rechnungen in jeder Klasse auftreten.

j	Rechnungsbetrag		hj	Hj	fj	Fj
	von	bis unter				
1	0	20	10	10	7.14 %	7.14 %
2	20	40	20	30	14.29 %	21.43 %
3	40	60	60	90	42.86 %	64.29 %
4	60	80	35	125	25.00 %	89.29 %
5	80	100	10	135	7.14 %	96.43 %
6	100	120	5	140	3.57 %	100.00 %
			140		100.00 %	

j Laufindex für die Klassen (Klassenindex)
$\mathbf{x^u_j}$ Untergrenze der Klasse j, $\mathbf{x^o_j}$ Obergrenze der Klasse j
$\mathbf{h_j}$ Absolute einfache **Klassenhäufigkeit** $x^u_j \le x_i < x^o_j$
Hj Absolute kumulierte **Klassenhäufigkeit** $x_i < x^o_j$

Beispiele:
$h_2 = 20$; d.h. 20 Kunden haben eine Rechnung über einen Betrag von 20 bis unter 40 erhalten.
$H_2 = 30$; d.h. 30 Kunden haben eine Rechnung über einen Betrag von 0 bis unter 40 erhalten.
Einerseits gehen bei der Klassenbildung Informationen verloren, andererseits gewinnt man Übersichtlichkeit.

- **Anzahl Klassen**
  - Je weniger Klassen man hat desto höher ist der Informationsverlust.
  - Je mehr Klassen man hat, desto schlechter ist die Übersichtlichkeit.
- **Klassenbreite**
  - Ideal ist eine für alle Klassen **identische Klassenbreite.** Die Häufigkeiten beziehen sich dann stets auf die gleiche Basis. Sie müssen nicht relativiert werden.
  - Es können jedoch auch **Restklassen** gebildet werden, z.B. sind die normalen Klassenbreiten von 0 – 20, 20 – 40 und die letzte Klasse ist von 40 – 80.
  - Die Festlegung der Klassenbreite soll möglichst so erfolgen, dass der Wert in der Klassenmitte ein typischer Stellvertreter für die Klasse ist.
- **Eindeutige Zuordnung der Merkmalswerte**
  - Die Klassen sollen so gebildet werden, dass immer eindeutig klar ist in welche Klasse ein Merkmalswert gehört.
  - In folgender Darstellung ist unklar ob 10 in die erste oder in die zweite Klasse gehört:

Rechnungsbetrag
0 – 10
10 – 20

Hingegen ist folgende Darstellung besser, da sie präzisiert ist:

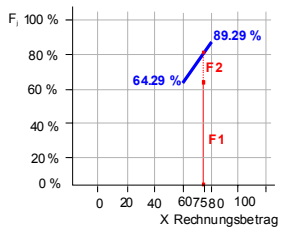
Rechnungsbetrag über … bis unter …
0 – 10
10 – 20

- **Exkurs: Näherungsweise Häufigkeitsberechnungen (Interpolation)**
  - Aus der folgenden Tabelle wollen wir wissen, wie gross der Anteil der Kunden ist, die einen Rechnungsbetrag von weniger als 75 haben:

j	Rechnungsbetrag		hj	Hj	fj	Fj
	von	bis unter				
1	0	20	10	10	7.14 %	7.14 %
2	20	40	20	30	14.29 %	21.43 %
3	40	60	60	90	42.86 %	**64.29 %**
4	**60**	**80**	35	125	25.00 %	**89.29 %**
5	80	100	10	135	7.14 %	96.43 %
6	100	120	5	140	3.57 %	100.00 %
			140		100.00 %	

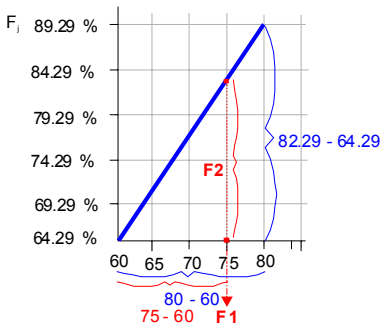
  - Man kann dies nicht herauslesen, da sich der Wert innerhalb der Klasse von 60 – 80 befindet.
  - Was wir wissen ist, dass 64.29 % (=F_3) der Kunden eine Rechnung kleiner als 60 (=x^u_4 und 89.29 % (=F_4) der Kunden eine Rechnung kleiner als 80 (=x^o_4) haben.
  - Falls wir unterstellen, dass die kumulierten relativen Häufigkeiten zwischen den Rechnungsbeträgen von 60 und 80 von 64.29 % auf 89.29 % linear ansteigen, dann ergibt sich folgendes Bild (Abb. 1):

Abb. 1



  - Betrachten wir den Ausschnitt (Abb. 2) zwischen den interessierenden Werten. Gesucht ist der Abstand bei x=75, also f_{75}. Man weiss dass F1 64.29 % beträgt, wie viel aber F2 beträgt weiss man nicht. Wir kennen jedoch die Werte bei x=80 (89.29 %) und bei x=60 (64.29 %). Es entstehen zwei Dreiecke, das bekannte und das gesuchte.
  - Die Seitenverhältnisse ähnlicher Dreiecke sind gleich. Wenn man das blaue Dreieck verkleinert kommt man auf das kleine rote Dreieck. Folgende Rechnung ist anzustellen:

Abb. 2

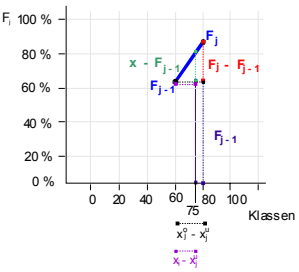


$$\frac{F2}{75 - 60} = \frac{89.29 - 64.29}{80 - 60}$$

$F2 = 18.75\ \%$

18.75 % (F2) + 64.29 % (F1) = 83.04 %

Man nennt diese Rechnung **Interpolation.**

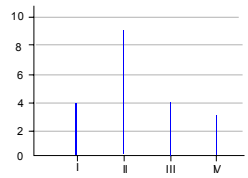


$$F(x_i) = F_{j-1} + \frac{x_i - x^u_j}{x^o_j - x^u_j} \cdot (F_j - F_{j-1}) \quad \text{oder Steigungsdreieck (wie im Bsp)} \quad \frac{x - F_{i-1}}{x_i - x^u_j} = \frac{F_j - F_{j-1}}{x^o_j - x^u_j}$$

1.2.2 Grafische Darstellungsarten

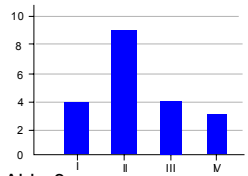
- Grafische Darstellungen werden auf der Basis von Tabellen erstellt.
  - **1.2.2.1 Einfache Häufigkeitsverteilungen**
    - **a) Das Stabdiagramm** (Abb. 1)

Abb. 1



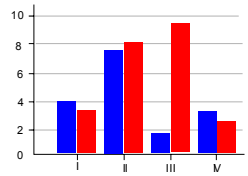
    - **b) Das Säulendiagramm** (Abb. 2)
      - Werden Stäbe im Stabdiagramm verbreitert spricht man vom Säulendiagramm- Wie das Stabdiagramm ist das Säulendiagramm **<u>höhenproportional</u>**.

Abb. 2



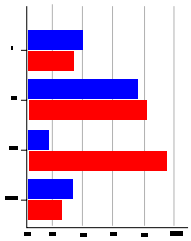
      - Diese grafische Darstellung eignet sich vor allem auch, falls mehrere Gesamtheiten darzustellen sind. (Abb. 3)

Abb. 3



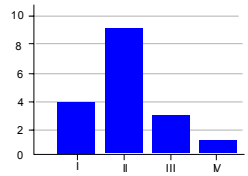
      - Manche Dreidimensionale Säulen von Excel eignen sich nicht, da die Säulen nicht am Raster stehen und man dadurch die Daten nicht aus dem Diagramm ablesen kann.
    - **c) Das Balkendiagramm** (Abb. 4)
      - Das Balkendiagramm ist ein gedrehtes Säulendiagramm.

Abb. 4



    - **d) Das Rechteckdiagramm** (Abb. 5)
      - Das Rechteckdiagramm ist **flächenproportional**.
      - Wählt man die Breite überall gleich ist es auch **höhenproportional**. In diesem Fall unterscheidet es sich nicht im wesentlichen von einem Säulendiagramm.

Abb. 5

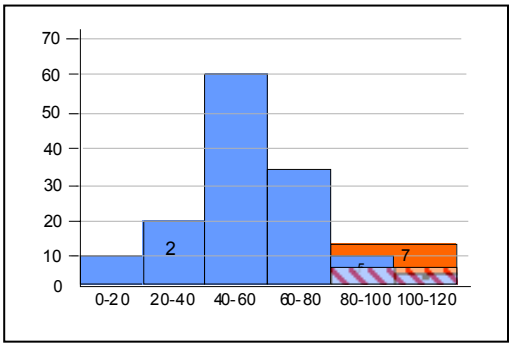


Serie 1
Aufg. 2

    - **e) Das Kreisdiagramm**
      - Beim Kreisdiagramm muss die Kreisfläche **<u>flächenproportional</u>** sein. Das heisst, wenn man beispielsweise verschiedene Jahre miteinander vergleicht, ist der eine Kreis grösser als der andere.
      - Die Fläche wird mit folgender Formel errechnet: $r^2 \cdot \pi$ **(r = radius)** Dabei muss auf den Radius zurück gerechnet werden um die entsprechenden Flächen korrekt miteinander in Proportionalität zu setzen. Nicht der Radius sondern die Flächen sind proportional zueinander.
    - **f) Das Histogramm**
      - Das Histogramm ist geeignet zur grafischen Darstellung **klassifizierter Häufigkeitsverteilungen.** <u>Das Histogramm ist **flächenproportional.**</u>
      - **Beispiel 1:**

j	Rechnungsbetrag		hj	Hj	fj	Fj
	von	bis unter				
1	0	20	10	10	7.14 %	7.14 %
2	20	40	20	30	14.29 %	21.43 %
3	40	60	60	90	42.86 %	64.29 %
4	60	80	35	125	25.00 %	89.29 %
5	80	100	10	135	7.14 %	96.43 %
6	100	120	5	140	3.57 %	100.00 %
			140		100.00 %	

- **Häufigkeitsdichte:**
  Fläche 2: 20 x 20 = 400
  Fläche 5: 10 x 20 = 200
  Fläche 6: 5 x 20 = 100 } 300
- Hätte man eine Restklasse von 80 – 120 gewählt also Nr. 6 einfach auf Nr. 5 heraufgestellt, hätte man die Fläche 7 erhalten,
  Fläche 7 = 15 x 40 = 600; anstatt 300. Dies darf man also nicht tun. Man müsste den Durchschnittswert wählen: Fläche 7 = 7.5 x 40 = 300! 7.5 ist der Durchschnittswert von 10 und 5. Man erhält in diesem Fall den \ \ gestreiften Balken.

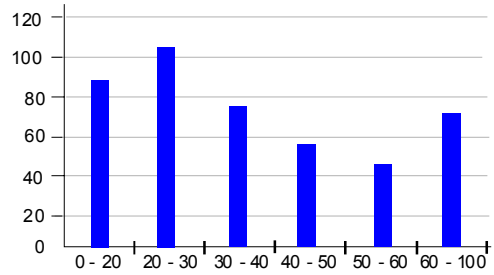


- **Beispiel 2 !:**

Umsatz pro Kunde	0 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 100	**Total**
Anzahl Kunden h_i	88	194	76	56	44	44	220
Häufigkeitsdichte	4.4	10.4	7.6	5.6	4.4	1.8	

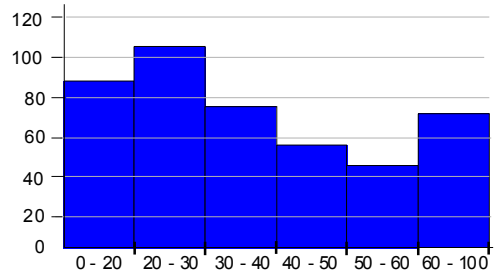
- Die erste Klasse hat eine Breite von 20. Die letzte Klasse hat eine Breite von 40. Die restlichen Klassen haben eine Standardbreite von 10.

Abb. 1 **Säulendiagramm (höhenproportional)**



- Abbildung 1 zeigt ein entsprechendes **Säulendiagramm**, das **höhenproportional** ist. Man erhält durch die Beschriftung den Hinweis, dass die erste und die letzte Klasse von den übrigen Klassen abweichende Breiten haben. Dieses Säulendiagramm ist eher **ungeeignet** für verschiedene Klassenbreiten.

Abb. 2 **Säulendiagramm (höhenproportional)**



- Abbildung 2 zeigt ein **Säulendiagramm** mit breiten Säulen, das **höhenproportional** ist. Das Auge nimmt jetzt die Fläche wahr und die Darstellung täuscht. **Das Auge misst die Fläche.** Man erhält den Eindruck, die erste und letzte Klasse sei sehr stark. Dieses höhenproportionale Säulendiagramm ist **ungeeignet**.

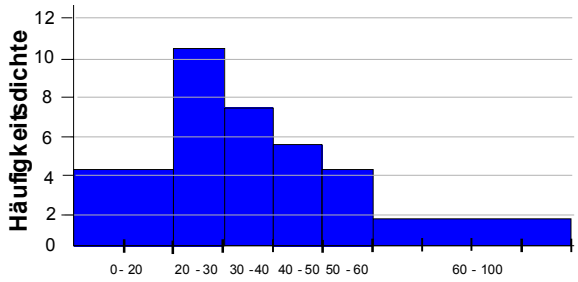
- Wenn man jetzt die Häufigkeitsdichte berechnet und die Klassenbreite entsprechend anpasst, wird die Grafik automatisch richtig interpretiert. Es entsteht ein **flächenproportionales Histogramm.**

Häufigkeitsdichte d = $\frac{\text{Absolute einfache Häufigkeit } h_i}{\text{Klassenbreite}}$

Serie 1
Aufg. 1 a)
Aufg. 3 b)

- Abbildung 3 zeigt das flächenproportionale Histogramm. **Die Breite der Säulen entspricht der Klassenbreite. Die Höhe der Säulen entspricht der Häufigkeitsdichte.** Wenn man die Häufigkeitsdichte mal die Klassenbreite rechnet, erhält man die Anzahl Merkmalsträger (Anzahl Kunden).

Abb. 3 **Histogramm (flächenproportional)**



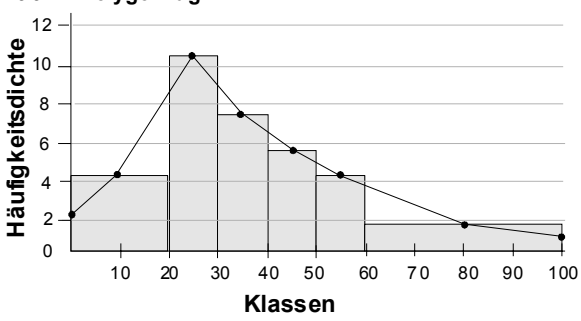
- **g) Der Polygonzug**

Serie 1
Aufg. 1 b)
Aufg. 3 b)

  - Diese Darstellung eignet sich für **klassifizierte Häufigkeitsverteilungen.**
  - Analog zum **Histogramm** werden auf der Abszisse (x-Achse) die Klassen abgetragen und auf der Ordinate (y-Achse) die **Häufigkeitsdichte.** In das Koordinatensystem werden immer in der **Mitte der Klassenbreite** (Klassenmitte x_i) die Häufigkeitsdichten eingetragen. Die so erhaltenen Punkte werden linear verbunden. Damit erhält man unter dem Kurvenzug eine Fläche die gleich der Fläche des Histogramms ist.
  - Am linken Rand der ersten Klasse wird die Hälfte der Häufigkeitsdichte der ersten Klasse aufgetragen. Am rechten Rand der letzen Klasse wird die Hälfte der Häufigkeitsdichte der letzten Klasse aufgetragen. Es muss zwingend die untere Tabelle erstellt werden.
  - **Beispiel 1:** Daten aus Beispiel 2 (Im Hintergrund das Histogramm zum Vergleich)

Klassenmitte x_i	**Häufigkeitsdichte**
0	2.2
10	4.4
25	10.4
35	7.6
45	5.6
55	4.4
80	1.8
100	0.9

Abb. 4 **Polygonzug**

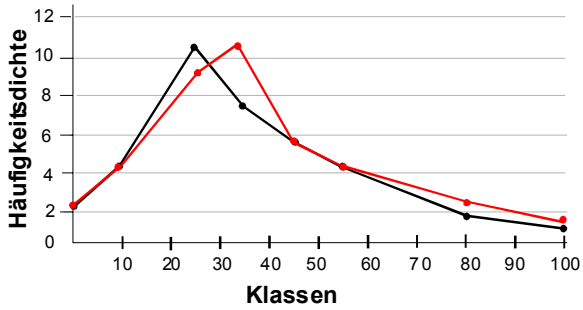


- Der Polygonzug eignet sich sehr gut dazu, Vergleiche anzustellen (sh. Beispiel 2):
- **Beispiel 2:** Angenommen wir hätten die oben ermittelten Umsätze in einer Filiale ermittelt und in einer anderen seien folgende Werte festgestellt worden:

Umsatz pro Kunde	0 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 100
Häufigkeitsdichte Filiale 1	4.4	10.4	7.6	5.6	4.4	1.8
Häufigkeitsdichte Filiale 2	4.4	9	10.5	5.6	4.4	2.8

Klassenmitten	**Häufigkeitsdichte F1**	**Häufigkeitsdichte F2**
0	2.2	2.2
10	4.4	4.4
25	10.4	9
35	7.6	10.5
45	5.6	5.6
55	4.4	4.4
80	1.8	2.8
100	0.9	1.4

Abb. 5 **Polygonzug**



o **1.2.2.2 Kumulierte Häufigkeitsverteilungen**

- **a) Die Treppenfunktion**
  - Die Treppenfunktion benutzen wir bei ordinalskalierten Merkmalen und diskreten, **nicht-klassifizierten Merkmalen.**
  - **Beispiel:**

Tarifgruppe	**Anzahl der Beschäftigten**	**Anzahl der Beschäftigten (kumuliert) H_i**	**Laufindex i**
x_i	h_i	H_i	i
I	4	4	1
II	9	13	2
III	4	17	3
IV	3	20	4

Abb. 6 **Treppenfunktion Variante 1**

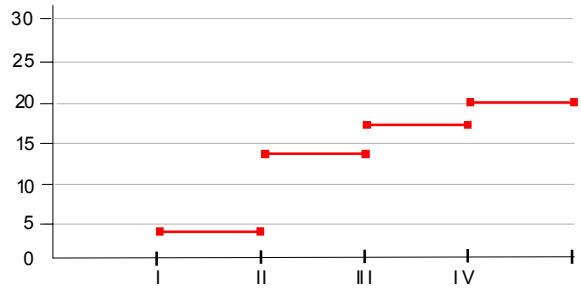
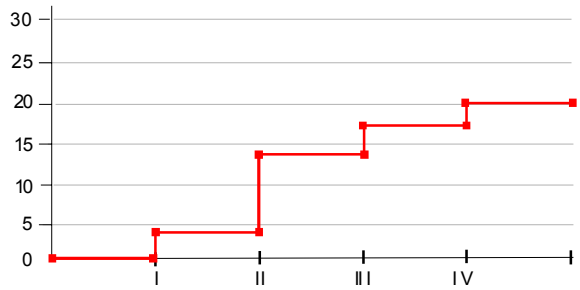


Abb. 7 **Treppenfunktion Variante 2**



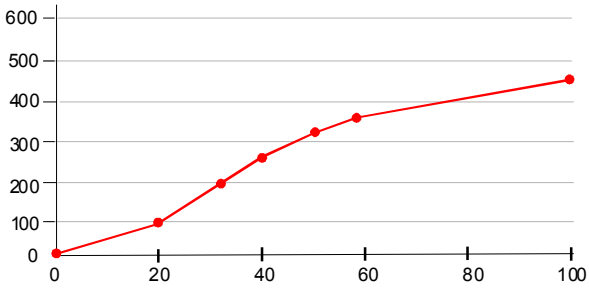
- **b) Das Summenpolygon (Ogive)**
  - Das Summenpolygon ist geeignet zur Darstellung **klassifizierter Häufigkeitsverteilungen.** Zu der Obergrenze jeder Klasse werden die entsprechenden kumulierten Häufigkeiten angegeben. Die Obergrenze wird verwendet da es sich um kumulierte Häufigkeiten handelt.
  - Zusätzlich wird als erster Punkt bei der Untergrenze der ersten Klasse der Wert Null angegeben.

Serie 1
Aufg. 3 c)

Umsatz pro Kunde	0 – 20	20 – 30	30 – 40	40 – 50	50 – 60	60 – 100
Anzahl Kunden	88	104	76	56	44	72

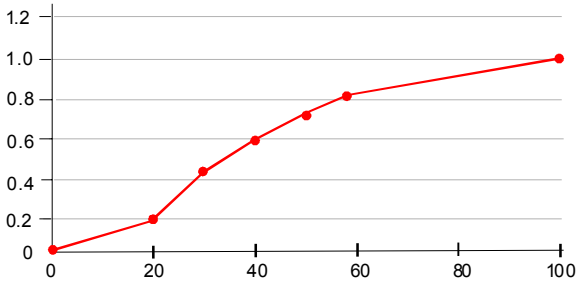
Klassenober-grenzen	**Absolute kumulierte Häufigkeit H_i**	**Relative kumulierte Häufigkeit F_i**
0	0	0.00
20	88	0.20
30	192	0.44
40	268	0.61
50	324	0.74
60	368	0.84
100	440	1.00

Abb. 8 **Absolute kumulierte Häufigkeiten H_i**



- Abbildung 9 zeigt die Darstellung für die relativen kumulierten Häufigkeiten.
- Man nennt das Summenpolygon der relativen kumulierten Häufigkeiten auch **Empirische Verteilungsfunktion.**

Abb. 9 **Relative kumulierte Häufigkeiten F_i**



1.3 Statistische Messzahlen

1.3.1 Lageparameter (Mittelwerte)

- Interpretation einzelner und Vergleich verschiedener Häufigkeitsverteilungen mit Hilfe die Lage der Häufigkeitsverteilungen auf der Merkmalsachse

Serie 1
Aufg. 6

o **1. Modus (Modalwert, häufigster Wert, dichtester Wert)**

Der Modus ist derjenige Merkmalswert x_i, der am häufigsten beobachtet wurde.

- **Voraussetzungen**
  - Der Modus ist bei jeder Verteilung bestimmbar.
- **Beispiel**
  - Der Modus in beiden Häufigkeitsverteilungen beträgt **1** Überstunde (Merkmalswert).

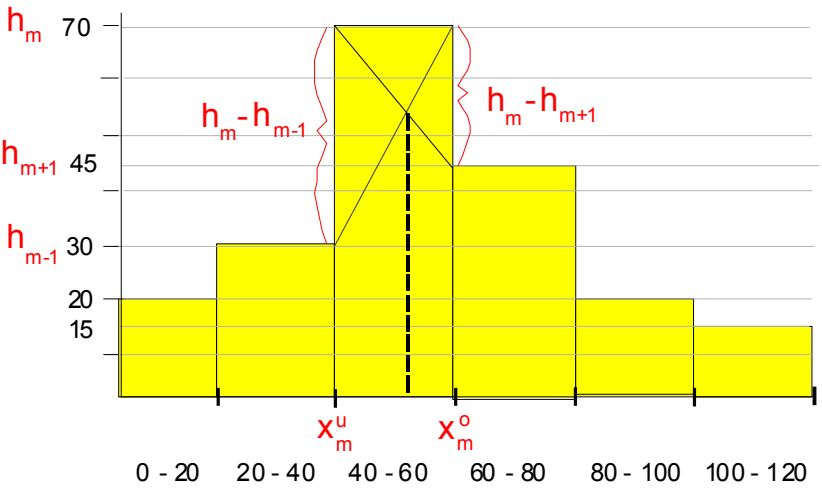
Überstunden der Müller AG

Überstunde x_i	hi
0	3
1	**10**
2	4
3	3
4	2
12	1

Überstunden der Maier AG

Überstunde xi	hi
0	3
1	**5**
2	4
3	4
4	4

- **Beurteilung**
  - Ein Vorteil ist, dass ein Ausreisser wie derjenige Mitarbeiter mit 12 Überstunden bei der Müller AG den Modus nicht beeinflusst. Hier ist die Häufigkeit von zehn deutlich die höchste. Die übrigen Häufigkeiten für die anderen Merkmalsausprägungen sind kleiner oder gleich vier. Das ist bei der Maier AG ganz anders: Hier ist der Modus eigentlich kein geeigneter Wert, da sich der Modus von 5 von den anderen Werten nicht signifikant unterscheidet. Die anderen Werte sind vier und drei.
- **Eignung**
  - Für nominalskalierte Grössen ist der Modus der einzige mögliche Mittelwert.
  - Der Modus ist dann geeignet, wenn seine Häufigkeit heraussticht. Die Häufigkeit muss einen deutlichen Gipfel haben.
  - Bei mehrgipfligen Verteilungen kann man unter Umständen auch von zwei Modi sprechen.
- **Klassifizierte Häufigkeitsverteilung**
  - Aus klassifizierten Häufigkeitsverteilungen kann der Modus nicht mehr direkt abgelesen werden. **Dabei geht man davon aus, dass sich der Modus in der Klasse mit der grössten Klassenhäufigkeit befindet.**
  - Bei schmalen Klassen wird oft die Klassenmitte dieser Klasse als Modus verwendet. Er kann aber auch fein bestimmt werden:
    - **Feinbestimmung des Modus bei Klassifizierten Häufigkeitsverteilungen:**



- Bei dieser Näherungsrechnung wird angenommen, dass der Modus umso näher an der oberen Grenze der Modusklasse liegt, je grösser die Häufigkeit der Klasse m+1 (60 – 80 im Beispiel) gegenüber der Häufigkeit der Klasse m-1 (20 – 40 im Beispiel) ist.

Auftragswert x_i		h_j
von…	bis unter…	
0	20	20
20	40	30
40	60	70
60	80	45
80	100	20
100	120	15

- **Modus (Feinbestimmung bei klassifizierten Häufigkeitsverteilungen)**
  $M_0 = x^u_m + (x^o_m - x^u_m) \cdot \frac{h_m - h_{m-1}}{(h_m - h_{m-1}) + (h_m - h_{m+1})}$

- $M_0 = 40 + (60 - 40) \cdot \frac{70 - 30}{(70 - 30) + (70 - 45)} = \underline{52.31}$

  Tatsächlich ist der Wert 52.31 womöglich gar nie vorgekommen, jedoch konzentrieren sich die meisten Werte um diesen Modus.
- Bei **unterschiedlichen Klassenbreiten ist die Modulklasse h_j die Klasse mit der grössten Häufigkeitsdichte d.** Die Formel ist gleich aufgebaut, nur muss die Klassenhäufigkeit h gegen die Dichte d ausgetauscht werden.

o **2. Median (zentraler Wert)**

Der Median ist derjenige Merkmalswert, dessen Merkmalsträger in der Rangordnung aller Merkmalsträger genau die mittlere Position einnimmt. Die Anzahl Merkmalsträger bzw. Merkmalswerte die vor dem Median liegen ist gleich der Anzahl Merkmalsträger bzw. Merkmalswerte, die hinter ihm liegen.

- **Voraussetzungen**
  - Das Merkmal muss wenigstens ordinalskaliert sein.
- **n (Gesamtzahl der Merkmalsträger) ist <u>ungerade</u>**

Fehltage x_i	0	3	4	7	**8**	9	12	13	59
h_i	3	1	2	3	5	4	2	2	1
H_i	3	4	6	9	**14**	18	20	22	23

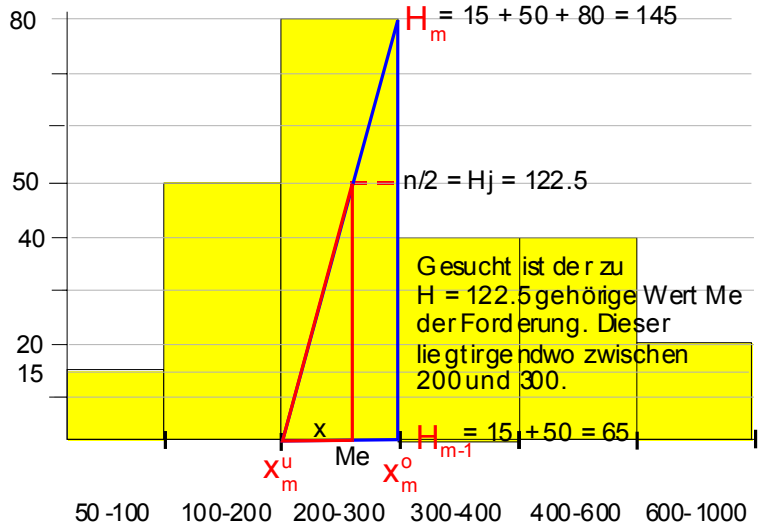
  - **$Me = x_{\left(\frac{n+1}{2}\right)}$ bei ungeradem n (Gesamtanzahl der Merkmalsträger)**
    x_i = Merkmalswert des Merkmalsträgers mit der Positionsziffer i.
  - i = (23 + 1) / 2 = 12 Der Merkmalsträger, der die Mittelposition einnimmt, hat die Positionsziffer (i) = 12. Der Dazugehörige Merkmalswert und zugleich Medium ist <u>8</u>.
- **n (Gesamtzahl der Merkmalsträger) ist <u>gerade</u>**

Fehltage x_i	0	2	5	6	7	11	12	14
h_i	4	2	2	2	4	3	2	1
H_i	4	6	8	10	14	17	19	20

  - **$Me = \frac{1}{2} \cdot \left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right)$ bei geradem n (Gesamtanzahl der Merkmalsträger)**
    x_i = Merkmalswert des Merkmalsträgers mit der Positionsziffer i.
  - Bei geradem n ist die Positionsziffer (n+1)/2 = 10.5 keine gerade Zahl und damit keinem Merkmalsträger zugeordnet. Man setzt deshalb den Median gleich dem Durchschnitt der Merkmalswerte der beiden zentral gelegenen Merkmalsträger.
  - $Me = 0.5 \cdot (6_{(10/2)} + 7_{(10/2+1)}) = \underline{6.5}$
  - <u>Der Median beträgt 6.5 Tage.</u>
    Das bedeutet, dass 50 % der Belegschaft weniger als 50 % und 50 % mehr als 6.5 Fehltage haben.
- **Beurteilung**
  - Ein Vorteil des Medians ist, dass er von Ausreissern unbeeinflusst bleibt.
- **Klassifizierte Häufigkeitsverteilung**
  - Bei klassifizierten Häufigkeitsverteilungen kann der Median nicht mehr exakt abgelesen werden. Er muss **näherungsweise** bestimmt werden.
    - Zuerst muss die **Medianklasse** bestimmt werden, in welcher der Merkmalsträger mit der Positionsziffer (n+1) / 2 liegt. Bei klassifizierten Häufigkeitsverteilungen wird meistens vereinfacht n/2 verwendet.

Forderung		h_j	H_j
von …	bis unter …		
50	100	15	15
100	200	50	65
200	300	80	145
300	400	40	185
400	600	40	225
600	1000	20	245

Gesucht ist der Merkmalswert für die Positionsziffer n/2 = 245/2 = <u>122.5</u>



- Der Merkmalswert, der irgendwo zwischen 200 und 300 liegt, kann mit dem Steigungsdreieck berechnet werden. Dieses lautet wie folgt:

$$\frac{x}{(n/2) - H_{m-1}} = \frac{x^o_m - x^u_m}{H_m - H_{m-1}}$$

Aus dieser Gleichung kann man die Formel für den **Median Me bei klassifizierten Häufigkeitsverteilungen** herleiten. Dabei gibt es drei Lösungsmöglichkeiten:

- **Me** $= x^u_m + x = x^u_m + \frac{(n/2) - H_{m-1}}{H_m - H_{m-1}} \cdot (x^o_m - x^u_m) = x^u_m + \frac{(n/2) - H_{m-1}}{h_m} \cdot (x^o_m - x^u_m)$

  Me = 200 + {[(245/2)-65] : [145 – 65]} · (300 – 200) = 271.875 (Forderungsbetrag)
- Das heisst 50 % der Merkmalsträger haben eine Forderung grösser als der median und 50 % haben eine Forderung die kleiner ist.
- **In obiger Darstellung wird H_i in einem Histogramm dargestellt. Ein Summenpolygon wäre allerdings sinnvoller. H_i kann so direkt abgelesen werden.**

Serie 1
Aufg. 7

- **Quantile**
  - Beim Median wird die Gesamtheit n in zwei Hälften zerlegt.
  - Beim **Perzentile** wird die Gesamtheit n in **hundert Hundertstel** zerlegt.
  - Beim **Dezentile** wird die Gesamtheit n in **zehn Zehntel** zerlegt.
    - Betrachten wir im Beispiel von oben (Forderungen) das neunte Dezentil. Wir wollen wissen unter welchem Wert 90 % aller Merkmalsträger anzutreffen sind. (10 % der Merkmalsträger haben Forderungen über diesem Wert).
    - 90 % von 245 sind 220.5. Damit befindet sich dieses Quantil in der fünften Klasse. Die Formel für den Median wird für das Quantil entsprechend angepasst
    - $D_9 = x^u_m + \frac{(n \cdot 9/10) - H_{m-1}}{H_m - H_{m-1}} \cdot (x^o_m - x^u_m)$

      D_9 = 400 + {[(245 · 9/10) – 185] : [225 – 185]} · (600 – 400) = 577.50 (Forderungsbetrag)
  - Beim **Quartile** wird die Gesamtheit n in **vier Viertel** zerlegt.
    - $Q_1 = x^u_m + \frac{(n \cdot 1/4) - H_{m-1}}{H_m - H_{m-1}} \cdot (x^o_m - x^u_m)$

      Q_1 = 100 + {[(245 · 1/4) – 15] : [65 – 15]} · (200 – 100) = 192.50 (Forderungsbetrag)

o **3. Das arithmetische Mittel** (häufigste Mittelwert, Durchschnitt)

Das arithmetische Mittel ist der Wert, der sich bei gleichmässiger Verteilung der Summe aller beobachteten Merkmalswerte auf alle Merkmalsträger ergibt. Beim arithmetischen Mittel ist die Summe der Abweichungen von Werten kleiner als das arithmetische Mittel zum arithmetischen Mittel gleich der Summe der Abweichungen von Werten grösser als das arithmetische Mittel zum arithmetischen Mittel.
Die Summe aller Produkte des Merkmaltswertes (x_i) mal Häufigkeit (h_i) muss durch die Anzahl aller Merkmalsträger geteilt werden. Das Resultat ist ein Merkmalswert.

$$\bar{x} = \frac{\sum_{i=1}^{k} x_i \cdot h_i}{n}$$

- **Voraussetzungen**
  - Das Merkmal muss mindestens intervallskaliert sein denn die Abstände zwischen den Merkmalswerten müssen messbar sein.
- **Beispiel**

Fehltage x_i	0	3	4	7	8	9	12	13	59
h_i	3	1	2	3	5	4	2	2	1
H_i	3	4	6	9	14	18	20	22	23

[(0 · 2) + (3 · 1) + (4 · 2) + (7 · 3) + (8 · 5) + (9 · 4) + (12 · 2) + (13 · 2) + (59 · 1)] : 23 = 9.43
9.43 entspricht der Mittelwert der Fehltage.

- **Beurteilung**
  - Das arithmetische Mittel beantwortet die Frage, was wäre wenn alle Merkmalsträger gleich gestellt wären. Der Nachteil des arithmetischen Mittels liegt darin, dass Ausreisser den Wert stark beeinflussen. Mediane wären manchmal anstatt arithmetischen Mittelwerten besser.
- **Eignung**
  - Das arithmetische Mittel sollte dort verwendet werden, wo eingipflige, nahezu symmetrische Häufigkeitsverteilungen sowie Verteilungen ohne klar erkennbare Konzentration auf einen Merkmalswert vorliegen. Es ist weniger oder nicht geeignet für schiefe Verteilungen und für kleine Gesamtheiten mit Ausreissern.
- **Klassifizierte Häufigkeitsverteilung**
  - Bei klassifizierten Häufigkeitsverteilungen kann das arithmetische Mittel nur näherungsweise bestimmt werden.
  - **Man rechnet die Klassenmitten mal die entsprechenden Häufigkeiten und summiert diese Ergebnisse. Diese Summe teilt man durch die Anzahl Merkmalsträger.**
  - Die untere Tabelle ist die gleiche wie auf Seite 13 unten. Das arithmetische Mittel beträgt: 78'625 · 245 = 320.92. Der Median (122.50) von Seite 13 ist deutlich kleiner als das arithmetische Mittel. Der Grund liegt darin, dass die hohen Werte in der Klasse von 500 bis 1000 das arithmetische Mittel nach oben ziehen.

Forderung		h_j	x_j^l	$x_j^l \cdot h_j$
von …	bis unter …			
50	100	15	75	1'125
100	200	50	150	7'500
200	300	80	250	20'000
300	400	40	350	14'000
400	600	40	500	20'000
600	1000	20	800	16'000
		245		78'625

o **4. Das geometrische Mittel**

Serie 1
Aufg. 9

Das geometrische Mittel wird dann eingesetzt, wenn es darum geht aus verschiedenen Faktoren (die messen um wie viel grösser oder kleiner ein Merkmalswert im Vergleich zum anderen ist) den durchschnittlichen Vergrösserungsfaktor (Verkleinerungsfaktor) zu ermitteln.
Das geometrische Mittel gibt mehrere aufeinanderfolgende Vervielfachungen als durchschnittliche Vervielfachung F_{GM} wieder.

- Basis ist in der Regel ein Merkmalswert, der zu verschiedenen Zeitpunkten erhoben wird. Z.B. werde immer am Ende eines Jahres das durchschnittliche Einkommen eines Haushaltes festgehalten.

Jahr	Einkommen	Wachstums-faktor	Wachstums-rate
1998	56'000		
1999	67'000	1.20	19.94 %
2000	74'500	1.11	11.19 %
2001	69'000	0.93	-7.38 %
2002	78'500	1.14	13.77 %
2003	83'000	1.06	5.73 %

- Die Jährliche Veränderung ist als Faktor ersichtlich. Dieser errechnet sich:
  Wert zum späteren Zeitpunkt
  Wert zum früheren Zeitpunkt
- **Der Wachstumsfaktor gibt für das Jahr j an, mit wie viel man den Wert im Jahr j-1 multiplizieren muss um den Wert im Jahr j zu erhalten.**
- **Durch Subtraktion von 1 erhält man aus dem Wachstumsfaktor die jährliche prozentuale Veränderung.**

- $56'000 \cdot \underbrace{1.20}_{67'000} \cdot 1.11 \cdot 0.93 \cdot \underbrace{1.14 \cdot 1.06}_{74'000} = 83'000$

- **Berechnung des geometrischen Mittels F_{GM} – Variante 1:**
  $56'000 \cdot F_{GM} \cdot F_{GM} \cdot F_{GM} \cdot F_{GM} \cdot F_{GM} = 83'000$
  oder
  $56'000 \cdot F_{GM}^5 = 83'000 \quad / : 56'0000 \quad = \quad F_{GM} = \sqrt[5]{\frac{83'000}{56'000}} = 1.081877$

$$F_{GM} = \sqrt[n]{\frac{\text{Endwert}}{\text{Anfangswert}}}$$ **wobei n der Anzahl Wachstumsfaktoren entspricht**

- **Berechnung des geometrischen Mittels F_{GM} – Variante 2:**
  $56'000 \cdot 1.20 \cdot 1.11 \cdot 0.93 \cdot 1.14 \cdot 1.06 = 83'000$
  $56'000 \cdot F_{GM} \cdot F_{GM} \cdot F_{GM} \cdot F_{GM} \cdot F_{GM} = 83'000$
  also
  $1.20 \cdot 1.11 \cdot 0.93 \cdot 1.14 \cdot 1.06 = F_{GM} \cdot F_{GM} \cdot F_{GM} \cdot F_{GM} \cdot F_{GM}$
  $1.48 = F_{GM}^{5}$ / $\sqrt{\ }$
  $F_{GM} = \underline{1.0818}$

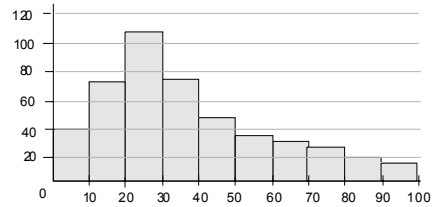
$F_{GM} = \sqrt[n]{F_1 \cdot F_1 \cdot \ldots \cdot F_n}$ wobei n der Anzahl Wachstumsfaktoren entspricht

- **Voraussetzungen**
  - Die zugrundeliegenden Grössen müssen verhältnisskaliert sein.
- **Klassifizierte Häufigkeitsverteilung**
  - Entwicklungs- und Wachstumsprozesse lassen sich nicht durch klassifizierte Häufigkeitsverteilungen beschreiben.

1.3.2 Typen von Verteilungen, Schiefe und Wölbung

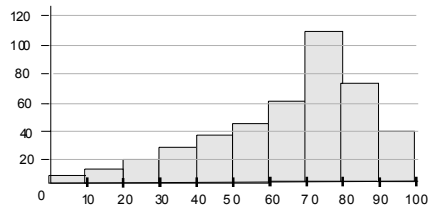
- **Rechtsschiefe (linkssteile) Verteilung** (Abb. 1)

Abb. 1



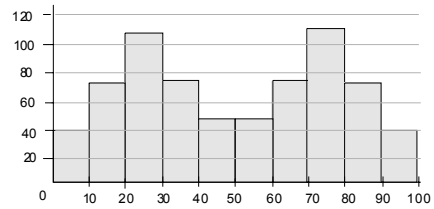
- **Linksschiefe (rechtssteile) Verteilung** (Abb. 2)

Abb. 2



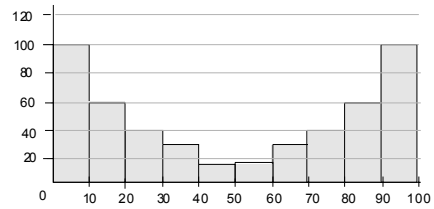
- **Bimodale (zweigipflige) Verteilung)** (Abb. 3)

Abb. 3



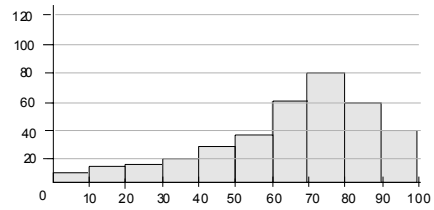
- **U-förmige Verteilung** (Abb. 4)

Abb. 4



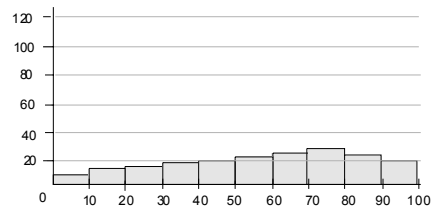
- **Wölbung**
  - **Steile Verteilung** (Abb. 5)

Abb. 5



  - **Flache Verteilung** (Abb. 6)

Abb. 6



- **Symmetrische Verteilungen** (Abb. 7 und 8)

Abb. 7

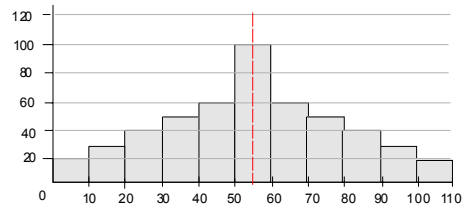
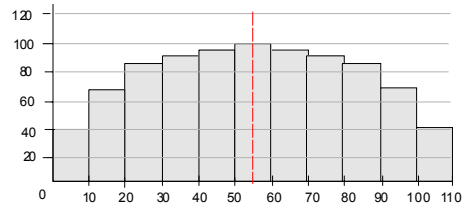
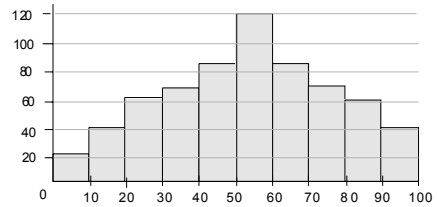


Abb. 8



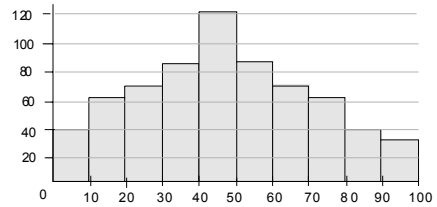
- Zur Verdeutlichung (Gemischte Begriffe)
  - **Eine beinahe symmetrische linksschiefe (rechtssteile) Verteilung** (Abb. 9)

Abb. 9



  - **Eine beinahe symmetrische rechtsschiefe (linkssteile) Verteilung** (Abb. 10)

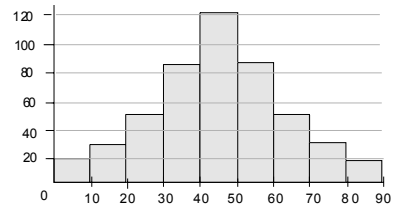
Abb. 10



1.3.3 Streuungsparameter

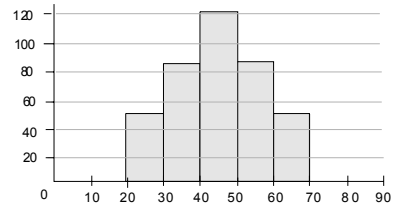
- Neben Mittelwerten (Lageparametern) beschreiben Streuungsparameter die zweite wesentliche Eigenschaft von Häufigkeitsverteilungen-
- **Die Frage ist, ob die Merkmalswerte in einem breiten Bereich auftreten oder ob sie nur in einem engen Bereich streuen.**
- Streuungsparameter sollen die Streuung der Häufigkeitsverteilung in einer einzigen Grösse messen. Zusammen mit dem Mittelwert erhält man so eine informative Beschreibung einer gegebenen Häufigkeitsverteilung.
- Es gibt zwei wichtige Konzepte zur Messung der Streuung:
  - Fall 1: Als Mass für die Streuung wird die Entfernung zwischen zwei ausgewählten Merkmalswerten gemessen.
  - Fall 2: Abweichungen der Merkmalswerte zum Mittelpunkt werden betrachtet.
- **Breite Streuung** (Abb. 11)

Abb. 11



- **Enge Streuung** (Abb. 12)

Abb. 12



o **1. Die Spannweite (Variabilität, Variationsbreite, Range)**

Als Mass für die Streuung wird die Distanz zwischen dem kleinsten und dem grössten Merkmalswert gemessen. Die Spannweite ist die Differenz zwischen dem grössten und dem kleinsten beobachteten Wert.
Für die aufsteigend geordneten Merkmalswerte $x_1, x_2 \dots x_N$ gilt:
Spannweite (Range) $R = x_n - x_1$ (x_n ist der letzte Merkmalswert)

- **Voraussetzungen**
  - Das Merkmal muss wenigstens intervallskaliert sein. In der Praxis wird oft ordinalskaliert als genügend betrachtet.
- **Beispiel**
  - $R = x_n - x_1$
    $= 12 - 0$
    $= 12$
    Die Überstunden streuen in einem Intervall von 12 Stunden.

Überstunde x_i	h_i
0	3
1	10
2	4
3	3
4	2
12	1

- **Eignung**
  - Im obigen Beispiel wäre die Spannweite nur vier wenn der einzige Merkmalsträger mit 12 Überstunden nicht berücksichtigt wird. **Die Spannweite reagiert also empfindlich auf Ausreisser.** Börsenkurse mit Höchst- und Tiefstwert sind ein Beispiel für dieses Streuungsmass.
- **Klassifizierte Häufigkeitsverteilungen**
  - **Bei klassifizierten Häufigkeitsverteilungen wird als kleinster Wert die Untergrenze der ersten Klasse und als grösster Wert die Obergrenze der letzten Klasse verwendet: $R = x^o_v - x^u_1$**

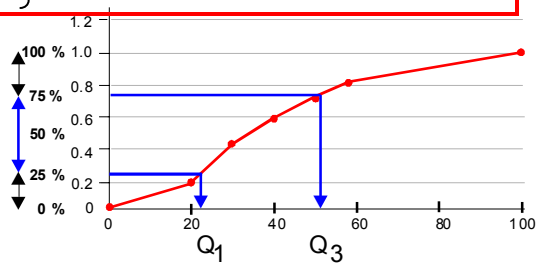
Forderung		h_j	x^l_j	$x^l_j \cdot h_j$
von …	bis unter …			
50	100	15	75	1'125
100	200	50	150	7'500
200	300	80	250	20'000
300	400	40	350	14'000
400	600	40	500	20'000
600	1000	20	800	16'000
		245		78'625

Im obigen Beispiel beträgt die Spannweite: R = 1000 – 50 = 950

o **2. Der Interquartilsabstand (Zentraler Quartilsabstand) und andere Abstände**

Der zentrale Quartilsabstand ist die Entfernung zwischen den beiden Merkmalswerten x, welche die in der Rangordnung zentral gelegenen 50 % der Merkmalsträger (50 % von n welche in der Mitte sind) eingrenzen. Die relative kumulierte Häufigkeit H_i muss vorliegen.
ZQA = $Q_3 - Q_1$ **$Q_1 = x_{\left(\frac{n+1}{4}\right)}$** **$Q_2 = x_{\left(\frac{3(n+1)}{4}\right)}$** **$x_i$ = Merkmalswert des Merkmalsträgers mit der Positionsziffer i.**

- Die nebenstehende Ogive (Summenpolygon) zeigt eine klassifizierte Häufigkeitsverteilung. Auf der y-Achse erscheinen die relativen kumulierten Häufigkeiten. **Der zentrale Quartilsabstand entspricht der Entfernung zwischen Q_1 und Q_2.**
- Analog dazu lassen sich auch andere Quantilsabstände messen. Der zentrale 90 %-Perzentilsabstand schneidet die unteren und oberen 5 % der Häufigkeitsverteilung ab.



- **Voraussetzungen**
  - Das Merkmal muss mindestens intervallskaliert sein. Gibt man nur die Quartilswerte an, ohne einen Abstand zu berechnen, kann man auch ordinalskalierte Merkmalswerte benutzen.
- **Beispiel**

Fehltage **x**	0	2	5	6	7	11	12	14
h_i	4	2	2	2	4	3	2	1
H_i	4	6	8	10	14	17	19	20

  - $Q_1 = x_{((20+1)/4)} = x_{(5.25)} = 2$ $Q_2 = x_{(3(20+1)/4)} = x_{(15.75)} = 11$
    ZQA = 11 – 2 = 9 Fehltage. Die mittleren 50 % der Fehlzeiten streuen mit einem Intervall mit der Länge 9 Tage. Die mittleren 50 % der Beschäftigten haben zwischen 2 und 11 Tagen gefehlt.

- **Beurteilung**
  - Das Problem von Ausreissern, das wir bei der Spannweite hatten, ist hier nicht relevant.
- **Klassifizierte Häufigkeitsverteilung**
  - Das Problem von Ausreissern, das wir bei der Spannweite hatten, ist hier nicht relevant.
  - **Zuerst müssen die Klassen von 25 % und von 75 % bestimmt werden.**
  - **ZQA = $Q_3 - Q_1$** $\quad Q_1 = x^u_m + \frac{(n \cdot 1/4) - H_{m-1}}{H_m - H_{m-1}} \cdot (x^o_m - x^u_m)$

    $Q_3 = x^u_m + \frac{(n \cdot 3/4) - H_{m-1}}{H_m - H_{m-1}} \cdot (x^o_m - x^u_m)$

Forderung **x**		h_j	H_i
von …	bis unter …		
50	100	15	15
100	200	50	**65**
200	300	80	145
300	400	40	**185**
400	600	40	225
600	1000	20	245

  - Relevante Klassen: 25 % von 245 = 61.25 → Klasse 100 – 200
    75 % von 245 = 183.75 → Klasse 300 – 400
  - Q_1 = 100 + {[(245 · 1/4) – 15] : [65 – 15]} · (200 – 100) = 192.50 (Forderungsbetrag)
    Q_3 = 300 + {[(245 · 3/4) – 145] : [185 – 145]} · (400 – 300) = 396.87 (Forderungsbetrag)
    ZQA = 396.875 – 192.5 = 204.375
    Die mittleren 50 % streuen in einem Intervall mit der Länge 204.375 zwischen 192.50 und 396.87.

o **3. Die mittlere absolute Abweichung**

Die betragsmässigen Abstände der Merkmalswerte vom Mittelwert (meist arithmetisches Mittel, aber auch Median) werden aufsummiert und durch die Anzahl Merkmalswerte geteilt. Die mittlere absolute Abweichung ist die durchschnittliche Entfernung aller beobachteten Merkmalswerte vom arithmetischen Mittel.

$$\delta = \frac{1}{n} \cdot \sum_{i=1}^{k} |x_i - \bar{x}| \cdot h_i$$ (Betrag bedeutet Rechnen ohne Vorzeichen)

- **Voraussetzungen**
  - Die Merkmale müssen intervallskaliert sein da die Abstände berechnet werden.
- **Beispiel**

Überstunde x_i	h_i	$x_i \cdot h_i$	$\lvert x_i - \bar{x} \rvert$	$\lvert x_i - \bar{x} \rvert \cdot h_i$
0	3	3	2.4	6.13
1	10	10	1.04	10.43
2	4	8	0.04	0.17
3	3	9	0.96	2.87
4	2	8	1.96	3.91
12	1	12	9.96	9.96
	23	47		33.48

  - Arithmetisches Mittel = Summe von $x_i \cdot h_i$ / n = 47 / 23 = 2.04 h
  - δ = 33.38 / 23 = 1.45 h
  - Die Überstunden weichen durchschnittlich um 1.45 h vom arithmetischen Mittel von 2.04 h ab.
- **Beurteilung**
  - Auch hier haben wir wieder das Problem des Ausreissers.
  - Sehr **geeignetes Mass** in der beschreibenden Statistik.
- **Klassifizierte Häufigkeitsverteilung**
  - **Die Merkmalswerte x_i werden durch die Klassenmitten x^l_j ersetzt.** Innerhalb der Klassen wird Gleichverteilung unterstellt.

$$\delta = \frac{1}{n} \cdot \sum_{i=1}^{k} |x^l_j - \bar{x}| \cdot h_i$$ (Betrag bedeutet Rechnen ohne Vorzeichen)

o 4. Varianz und Standardabweichung

Wiederum werden Abweichungen der Merkmalswerte vom Mittelwert für das Streuungsmass berücksichtigt.

Die <u>Varianz</u> σ^2 ist die Summe der quadrierten Abweichungen der Merkmalswerte vom arithmetischen Mittel, dividiert durch die Anzahl der Merkmalsträger.

Die <u>Standardabweichung</u> σ ist die Quadratwurzel aus der Varianz.

$$\sigma = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{v} (x_i - \bar{x})^2 \cdot h_i} \quad \text{vereinfacht} \rightarrow \quad \sigma^2 = \sum_{i=1}^{v} (x_i)^2 \cdot h_i - \bar{x}^2$$

<u>oder</u> mit relativer einfacher Häufigkeit:

$$\sigma = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{v} (x_i - \bar{x})^2 \cdot f_i} \quad \text{vereinfacht} \rightarrow \quad \sigma^2 = \sum_{i=1}^{v} (x_i)^2 \cdot f_i - \bar{x}^2$$

- **Voraussetzungen**
  - Die Merkmale müssen intervallskaliert sein.
- **Beispiel**

Überstunde x_i	h_i	$x_i \cdot h_i$	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$	$(x_i - \bar{x})^2 \cdot h_i$
0	3	0	-2.04	4.1616	12.48
1	10	10	-1.04	1.0816	10.82
2	4	8	-0.04	0.0016	0.01
3	3	9	0.96	0.9216	2.76
4	2	8	1.96	3.8416	7.68
12	1	12	9.96	99.2016	99.20
	23	47			132.96

  - Arithmetisches Mittel = Summe von $x_i \cdot h_i$ / n = 47 / 23 = <u>2.04 h</u>
  - Varianz = 132.96 / 23 = <u>5.78 h</u>
  - Standardabweichung = $\sqrt{5.78}$ = <u>2.40 h</u>
- **Beurteilung**
  - Diesem Streuungsparameter ist vor allem ihre fehlende Anschaulichkeit vorzuwerfen.
  - Merkmalswerte mit zunehmendem Abstand vom Mittelwert haben einen überproportionalen Einfluss auf das Streuungsmass.
  - Die Bedeutung der Standardabweichung ist auf die schliessende Statistik beschränkt.
- **Klassifizierte Häufigkeitsverteilung**
  - **Die Merkmalswerte x_i werden durch die Klassenmitten x_j^l ersetzt.**

$$\sigma = \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{v} (x_j^l - \bar{x})^2 \cdot h_i}$$

o 5. Der Variationskoeffizient

Der Variationskoeffizient misst nicht die absolute Streuung, sondern er setzt diese in Beziehung zur Lage der Häufigkeitsverteilung. Der Variationskoeffizient ist der Quotient aus Standardabweichung und arithmetischem Mittel. <u>Die Standardabweichung wird als Prozentsatz des arithmetischen Mittels ausgedrückt.</u>

$$VK\,\% = \frac{\sigma}{\bar{x}} \cdot 100 \qquad \text{bzw. falls } x < 0: \quad VK\,\% = \frac{\sigma}{|\bar{x}|} \cdot 100$$

- **Voraussetzungen**
  - Die Merkmale müssen intervallskaliert.
- **Beurteilung**
  - Der VK ist geeignet beim Vergleich von Streuungen von Häufigkeitsverteilungen mit unterschiedlichen Lagen.
- **Beispiel**

Forderung		h_j	x_j^l	$x_j^l \cdot h_j$	$(x_j^l - \bar{x})^2$	$(x_j^l - \bar{x})^2 \cdot h_j$
von …	bis unter …					
50	100	15	75	1'125	60'476.65	907'149.70
100	200	50	150	7'500	29'213.65	1'460'682.32
200	300	80	250	20'000	5'029.65	402'371.71
300	400	40	350	14'000	845.65	33'825.86
400	600	40	500	20'000	32'069.65	1'282'78.86
600	1000	20	800	16'000	229'517.65	4'590.352.93
		245		78'625		

- $\bar{x}$ = 78'625 / 245 = <u>320.91</u>
- Varianz = 35'417.01
- Standardabweichung = <u>188.19</u>
- VK = 188.19 / 320.91 = <u>58.64 %</u>

1.4 Zeitreihenanalyse

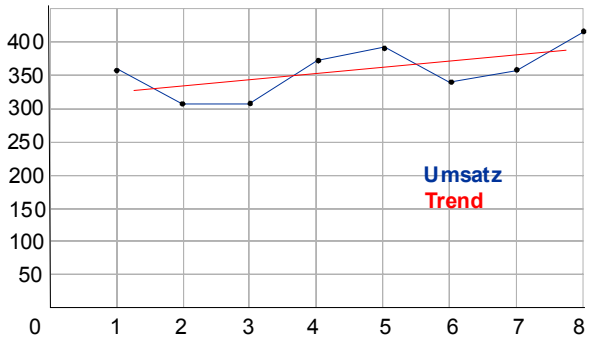
Eine Zeitreihe ist eine zeitlich geordnete Folge von Merkmalswerten.

- Aufgabe der Zeitreihenanalyse ist es, die Struktur und die Gesetzmässigkeiten einer Zeitreihe zu erkennen.
  - *Beispiel:* Ist der Rückgang der Arbeitslosigkeit im letzten Quartal eine Wende auf dem Arbeitsmarkt oder ist der Rückgang vielleicht nur saisonbedingt.

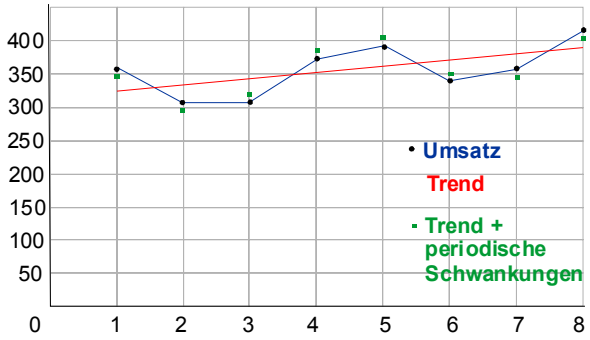
1.4.1 Komponenten der Zeitreihe (Begriffserklärung)

- **Trend**
  - **Als Trend bezeichnet man die Grundrichtung der Entwicklung.**
  - Die **langfristig wirksamen Einflüsse** sollen sichtbar gemacht werden.

Quartal	Umsatz
1	360
2	310
3	310
4	370
5	390
6	340
7	360
8	420



- **Periodische (Saisonale) Schwankunkungen**
  - **Als periodische (Saisonale) Schwankungen bezeichnet man periodisch wiederkehrende Schwankungen um den Trend.**
  - Im Beispiel ist erkennbar, dass das jeweils erste und vierte Quartal eines Jahres über dem Trend liegen während das jeweils zweite und dritte Quartal eines Jahres unter dem Trend liegen.



  - **Ermittelt man die periodischen bzw. saisonalen Schwankungen kann dargestellt werden, welche Werte aufgrund von Trend und saisonaler Schwankung in der Zukunft erwartet werden.**
  - Diese aufgrund von Trend und periodischen Schwankungen erwarteten Umsätze sind noch nicht die effektiven Umsätze. Zwischen dem grünen Punkt und dem Umsatz besteht noch eine Differenz.
  - In der nebenstehenden Abbildung werden die **periodisch wirksamen Einflüsse** dargestellt.

- **Restkomponente**
  - **Die Restkomponente geht auf einmalig wirkende Grössen oder auf unbekannte Grössen die wiederholt aber unregelmässig wirken zurück.**
  - **Die Restkomponente ist in obiger Darstellung die Differenz zwischen dem grünen Punkt (Trend + periodische Schwankungen) und dem effektiven Umsatz.**

Quartal	Umsatz	Trend	Periodische Schwankungen	Restkomponente	Trend + Saisonale Schwankungen + Restkomponente
(1)	(2)	(3)	(4)	(5)	(3) + (4) + (5) = (2)
1	360	325.00	31.43	3.57	360
2	310	334.29	-27.86	3.57	310
3	310	343.57	-27.14	-6.43	310
4	370	352.86	23.57	-6.43	370
5	390	362.14	31.43	-3.57	390
6	340	371.43	-27.86	-3.57	340
7	360	380.71	27.14	6.43	360
8	420	390.00	23.57	6.43	420

- In diesem Beispiel unterstellen wir einen **additiven Zusammenhang** zwischen Trend, saisonaler Schwankung und Restkomponente. **Additiver Zusammenhang** bedeutet, dass die Komponenten unabhängig voneinander wirken.

y_i	**Umsatz im Quartal i**
T_i	**Trend im Quartal i**
S_i	**periodische Schwankung im Quartal i**
R_i	**Restkomponente im Quartal i**

$y_i = T_i + S_i + R_i \qquad (i = 1, \dots n)$

- Der Zusammenhang kann aber auch **multiplikativ** sein. In diesem Fall sind die Komponenten voneinander abhängig:

$y_i = T_i \cdot S_i \cdot R_i \qquad (i = 1, \dots n)$

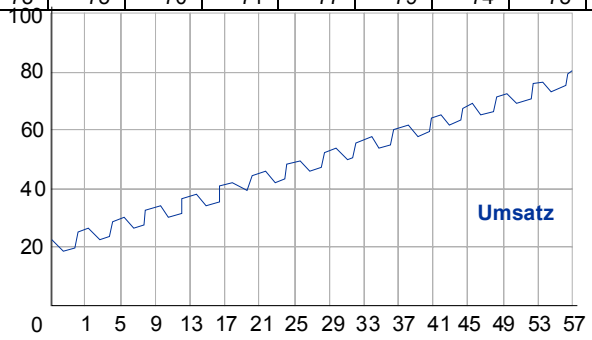
1.4.2 Methoden zur Trendermittlung (Trend)

- **Ziel der Trendermittlung ist es, die Entwicklung darzustellen, die sich ergibt, wenn die den Trend überlagernden Schwankungen eliminiert werden.**
  - 1. Methode der gleitenden Durchschnitte
  - 2. Methode der kleinsten Quadrate

- **1. Methode der gleitenden Durchschnitte**
  - **Beispiel**

Periode	1	2	3	4	5	6	7	8	9	10	11	12
Werte	23	18	19	25	27	22	23	29	31	26	27	33
Periode	13	14	15	16	17	18	19	20	21	22	23	24
Werte	35	30	31	37	39	34	35	41	43	38	39	45
Periode	25	26	27	28	29	30	31	32	33	34	35	36
Werte	47	42	43	49	51	46	47	53	55	50	51	57
Periode	37	38	39	40	41	42	43	44	45	46	47	48
Werte	59	54	55	61	63	58	59	65	67	62	63	69
Periode	49	50	51	52	53	54	55	56	57	58	59	60
Werte	71	66	67	73	75	70	71	77	79	74	75	81

- **Für den gleitenden Durchschnitt erfolgt die Glättung der Kurve dadurch, dass den einzelnen Perioden Durchschnitte zugeordnet werden.**

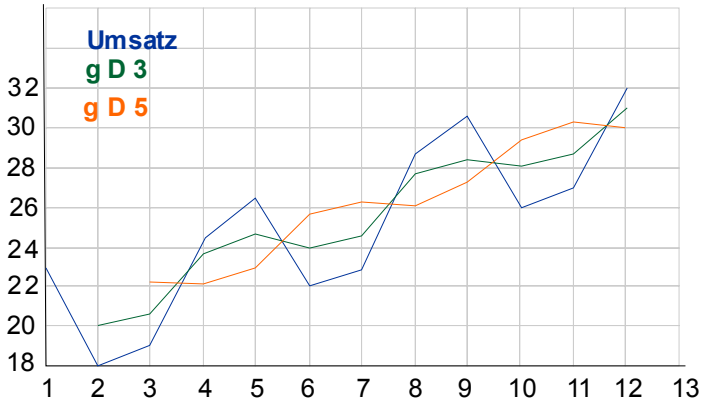


  - a) **Gleitender Durchschnitt ungerader Ordnung**
    - **Soll der gleitende Durchschnitt (g D) ungerader Ordnung ermittelt werden, dann wird der Durchschnitt aus den Zeitreihenwerten (ungerader Anzahl) jeweils der mittleren Periode zugeordnet.**
    - Für die ersten zwölf Perioden erhält man für den gleitenden Durchschnitt der dritten und fünften Ordnung folgende Werte:

Periode	1	2	3	4	5	6	7	8	9	10	11	12
Werte	23	18	19	25	27	22	23	29	31	26	27	33
g D 3 (Arithmetisches Mittel aus 3 Werten)		20.0	20.7	23.7	24.7	24.0	24.7	27.7	28.7	28.0	28.7	31.7
g D 5 (Arithmetisches Mittel aus 5 Werten)			22.4	22.2	23.3	25.2	26.4	26.2	27.2	29.2	30.4	30.2

ungerade

    - Es ist deutlich erkennbar, dass die Werte mit zunehmenden Wert für die Ordnung des gleitenden Durchschnitts glatter werden.



o b) **Gleitender Durchschnitt gerader Ordnung**

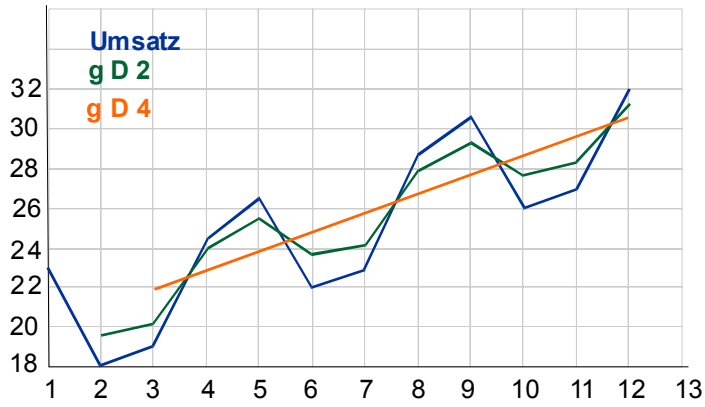
- Soll der gleitende Durchschnitt (g D) gerader Ordnung ermittelt werden, dann muss die Methode modifiziert werden. Bei gleichem Vorgehen müsste man den Durchschnitt aus der geraden Anzahl jeweils einem Zeitpunkt zwischen den Perioden zuordnen. Stattdessen wählt man folgendes Verfahren:
- **Will man den gleitenden Durchschnitt n-ter Ordnung (n gerade) ermitteln, berücksichtigt man n+1 Werte (ungerade Anzahl) auf folgende Art: Der erste Wert und der n+1 Wert werden nur zur Hälfte berücksichtigt, die übrigen Werte werden voll berücksichtigt und die Summe wird durch n geteilt.**

Periode	1	2	3	4	5	6	7	8	9	10	11	12
Werte	23	18	19	25	27	22	23	29	31	26	27	33
g D 2 (Arithmetisches Mittel aus 2 Werten)		19.5	20.3	24.0	25.3	23.5	24.3	28.0	29.3	27.5	28.3	32.0
g D 4 (Arithmetisches Mittel aus 4 Werten)			21.8	22.8	23.8	24.8	25.8	26.8	27.8	28.8	29.8	30.8

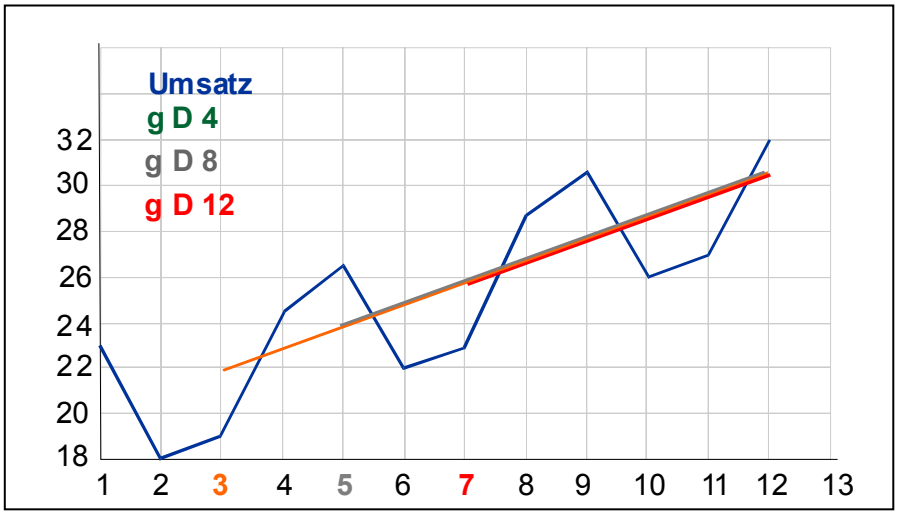
gerade

$$\text{g D 2} = y_{2;\text{g D2}} = \frac{\frac{23}{2} + 18 + \frac{19}{2}}{2} = \underline{19.5} \qquad \text{g D 4} = y_{2;\text{g D2}} = \frac{\frac{22}{2} + 23 + 29 + 31 + \frac{26}{2}}{4} = \underline{26.8}$$

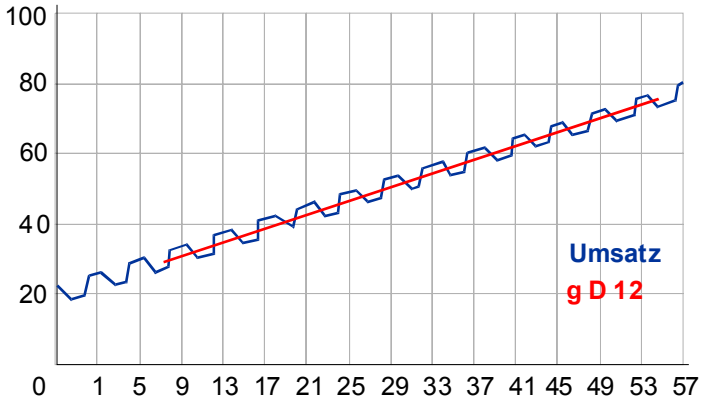
- Man erhält damit die nebenstehende Grafik.



- **Beurteilung**
  - **Wenn man den gleitenden Durchschnitt vierter Ordnung mit dem gleitenden Durchschnitt fünfter Ordnung vergleicht, erkennt man, dass man die vollständige Glättung beim gleitenden Durchschnitt vierter Ordnung erreicht.**
  - **Das ist auch einleuchtend, da die periodischen Schwankungen über vier Perioden** (= 1 Jahr) **laufen. In diesem Fall muss der gleitende Durchschnitt von der Ordnung k mal vier sein (k ∈ N).**
  - Betrachtet man statt des gleitenden Durchschnitts vierter Ordnung die gleitenden Durchschnitte achter und zwölfter Ordnung dann führen diese in unserem Beispiel zu keiner weiteren Veränderung. Allerdings erhalten wir beim gleitenden Durchschnitt achter Ordnung erst Werte ab der Periode fünf und beim gleitenden Durchschnitt zwölfter Ordnung erst Werte ab der Periode sieben. (siehe unten folgende Abbildung)
  - Im Falle einer idealen oder fastidealen vierperiodigen Schwankung wie hier, wird man also mit einem gleitenden Durchschnitt vierter Ordnung arbeiten. Aber in anderen Fällen kann natürlich auch eine acht- oder zwölfperiodige Schwankung auftreten.

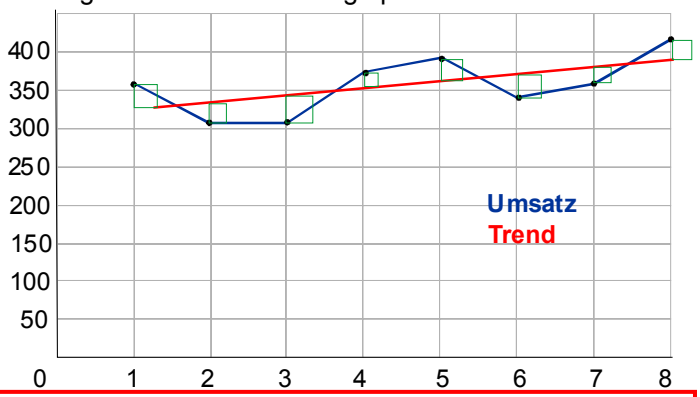


- **Darstellung des gleitenden Durchschnitts zwölfter Ordnung über die ursprünglichen 60 Perioden.**
  - Bei einer idealisierten Schwankung um den Trend, wie in diesem Beispiel, wird die Glättung durch Übergang von der vierten zur achten bzw. von der vierten zur zwölften Ordnung nicht besser.
  - Bei nicht-idealisierten Schwankungen aber kann durch Übergang vom gleitenden Durchschnitt vierter Ordnung auf einen gleitenden Durchschnitt der Ordnung k mal vier eine weitere Glättung erzielt werden. Hier muss dann überlegt werden ob man eine bessere Glättung will (gleitenden Durchschnitt höherer Ordnung wählen) oder ob man mehr Werte will (gleitenden Durchschnitt kleinerer Ordnung wählen).



- **2. Methode der kleinsten Quadrate**
  - **Bei der Methode der kleinsten Quadrate wird eine Funktion gesucht, deren Werte als Trendwerte interpretiert werden.**
    - Es gibt Trendfunktionen die linear verlaufen und solche die nicht linear verlaufen (Exponentialfunktion, Potenzfunktion, logistische Funktion).
    - Bei der Methode der kleinsten Quadrate wird die Trendfunktion mit folgender Eigenschaft gesucht:
    - **Die gesuchte Funktion soll von allen möglichen Funktionen** (Geraden bei linearem Trendverlauf) **die Eigenschaft haben, dass die Summe der Abweichungen zwischen den effektiven Werten und den Trendwerten im Quadrat minimal ist.**
  - a) **Linearer Trendverlauf**
    - **Daher soll die Fläche der unten eingezeichneten Quadrate minimal sein.**
    - **Beispiel:** Umsatzzahlen und Trendgerade mit Abweichungsquadraten

Quartal x_i	Umsatz y_i	$x_i \cdot y_i$	x_i^2
1	360	360	1
2	310	620	4
3	310	930	9
4	370	1480	16
5	390	1950	25
6	340	2040	36
7	360	2520	49
8	420	3360	64
36	2860	13'260	204



- Trendgerade $\hat{y} = b \cdot x + a$

$$a = \bar{y} - b \cdot \bar{x}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

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

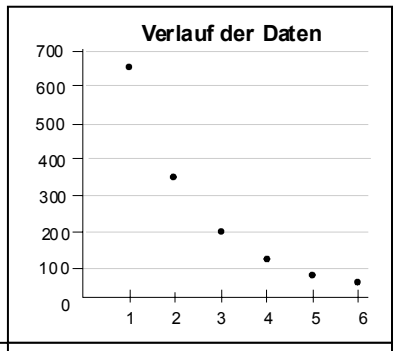
$\bar{x} = \frac{36}{8} = \underline{4.5}$ $\qquad \bar{y} = \frac{2860}{8} = \underline{357.5}$ $\qquad b = \frac{13'260 - 8 \cdot 4.5 \cdot 357.5}{204 - 8 \cdot (4.5)^2} = \underline{9.286}$

$a = 357.5 - 9.286 \cdot 4.5 = \underline{315.714}$ $\qquad \underline{\hat{y} = 9.286x + 315.714}$ → Trendfunktion

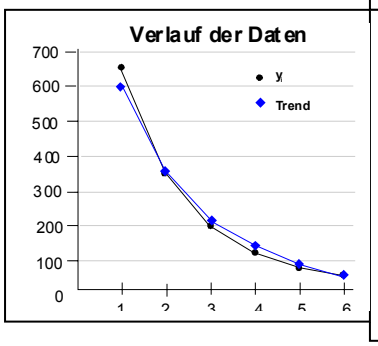
o b) **Nicht-Linearer Trendverlauf**

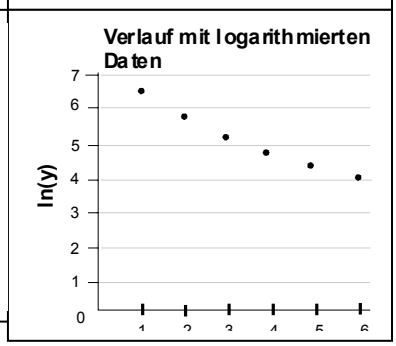
- Dieses Verfahren kann bei Exponentialfunktionen gewählt werden.
- Der Graph zeigt den Verlauf der Daten (x_i und y_i)

x_i	y_i	$\ln(y_i)$	$x_i \cdot \ln(y_i)$	x_i^2
1	650	6.47697	6.48	1
2	350	5.85793	11.71	4
3	200	5.29832	15.90	9
4	125	4.82831	19.31	16
5	85	4.44265	22.21	25
6	65	4.17439	25.05	36
21	1475	31.08	100.66	91



- Aufgrund des exponentiellen Verlaufs wird die Trendgerade als $\hat{y} = a \cdot b^x$ definiert. ($a > 0$ und $b > 0$)
- Logarithmiert (ln natürlicher Logarithmus) erhält man einen nahezu linearen Verlauf (Graph 2). Jetzt kann gleich vorgegangen werden wie im linearen Trendverlauf.





- Trendgerade $\hat{y} = a \cdot b^x$

$\ln(a) = \overline{\ln(y)} - \ln(b) \cdot \bar{x}$ $\qquad \bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ $\qquad \overline{\ln(y)} = \frac{\sum_{i=1}^{n} \ln(y_i)}{n}$

$$\ln(b) = \frac{\left(\sum_{i=1}^{n} x_i \cdot \ln(y_i)\right) + \left(-n \cdot \bar{x} \cdot \overline{\ln(y)}\right)}{\left(\sum_{i=1}^{n} x_i^2\right) + \left(-n \cdot \bar{x}^2\right)}$$

$\ln(\hat{y}) = \ln(a) + x \cdot \ln(b)$

→ **delogarithmieren mit e^x**

$\bar{x} = \frac{21}{6} = \underline{3.5}$ $\qquad \overline{\ln(y)} = \frac{31.08}{6} = \underline{5.18}$ $\qquad \ln(b) = \frac{100.66 + (-6 \cdot 3.5 \cdot 5.18)}{91 + (-6 \cdot 12.25)} = \underline{-0.464}$

$\ln(a) = 5.18 - (-0.464) \cdot 3.5 = \underline{6.803}$

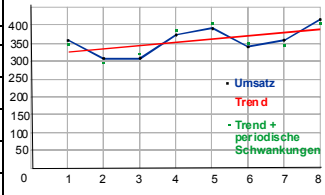
$\ln(a) = 6.803 \rightarrow e^x \; \underline{a = 900.221}$ $\qquad \ln(b) = -0.464 \rightarrow e^x \; \underline{b = 0.629}$

$\underline{\hat{y} = 900.221 \cdot 0.629^x}$ → Trendfunktion

1.4.3 Ermittlung der periodischen Schwankungen

- Man betrachtet alle Werte der **ersten Quartale.** Man bildet den Schnitt aller Abweichungen vom Umsatz und interpretiert dies als periodische Schwankung des ersten Quartals. Dasselbe macht man auch für das zweite, dritte und vierte Quartal. Diese Durchschnitte nennt man **Saisonnormale** für das jeweils erste, zweite, dritte und vierte Quartal.

Quartal	Umsatz	Trend gem. Funktion	Abweichungen vom Umsatz (Quartal)				Saison-normale	Saisonale Schwankungen (Trend + Saisonnormale)
			1. Q	2. Q	3. Q	4. Q		
(1)	(2)	(3)	(4) = (3)- (2)	(5) = (3)- (2)	(6) = (3)- (2)	(7) = (3)- (2)	(8)	(9) = (3) + (8)
1	360	325.0	35.00				31.43	356.43
2	310	334.3		-24.29			-27.86	306.44
3	310	343.6			-33.57		-27.14	316.46
4	370	352.9				17.41	23.57	329.33
5	390	362.1	27.86				31.43	393.53
6	340	371.4		-31.43			-27.86	343.54
7	360	380.7			-20.71		-27.14	353.56
8	420	390.0				30.00	23.57	413.57
Saison-Normale:			31.43	-27.86	-27.14	23.57		



- **Abweichungen vom Umsatz = Umsatz – Trend**
- **Saison-Normale = Durchschnitt der Abweichungen des jeweils 1. Quartals… (1. Q + 5. Q / 2) etc.**
- **Saisonale Schwankungen = Trend + Saisonnormale**
- Bei 8 Quartalen werden zwei Werte benötigt (1. Quartale der beiden Jahre)
  Bei 16 Quartalen werden vier Werte benötigt (1. Quartale der vier Jahre)

1.5 Regressions- und Korrelationsrechnung

- Kennt man zwei Merkmale stellt sich oft die Frage in welchem Zusammenhang die zwei Merkmale zueinander stehen.

Die Regression sucht durch eine Funktion (linear, aber auch exponentiell oder anders) den Zusammenhang zwischen zwei Datensätzen zu beschreiben.

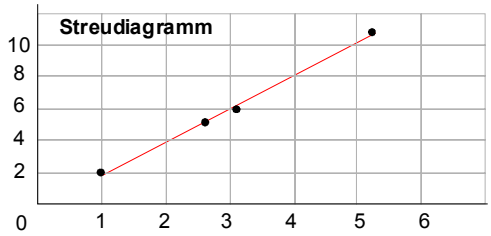
Die Korrelationsrechnung soll angeben, von welcher Stärke ein solcher Zusammenhang ist.

- Wenn man nur ein Merkmal im Zeitablauf betrachtet, handelt es sich um die Zeitreihenanalyse.

1.5.1 Regression

- **Die Regression sucht durch eine Funktion (linear, aber auch exponentiell oder anders) den Zusammenhang zwischen zwei Datensätzen zu beschreiben.**
  - Die Merkmale müssen intervall- oder verhältnisskaliert sein.
  - Der Zusammenhang zwischen den Merkmalswerten kann ein linearer sein, er kann aber auch Exponentiell oder im Sinne einer Potenzfunktion sein. Im Folgenden wird ein linearer Zusammenhang dargestellt, exponentielle Zusammenhänge können wie bei der Zeitreihenanalyse durch geeignetes Logarithmieren ebenfalls betrachtet werden.
- **1. Methode der kleinsten Quadrate**
  - Zur Bestimmung der Regressionsgeraden beziehen wir uns auf die Methode der kleinsten Quadrate.

x_i	y_i	$x_i \cdot y_i$	x_i^2
1	2	2	1
2.5	4.5	11.25	6.25
3	5	15	9
5	9	45	25
11.5	20.5	73.25	41.25



- Regressionsgerade $\hat{y} = a + b \cdot x$

$$a = \bar{y} - b \cdot \bar{x}$$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

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

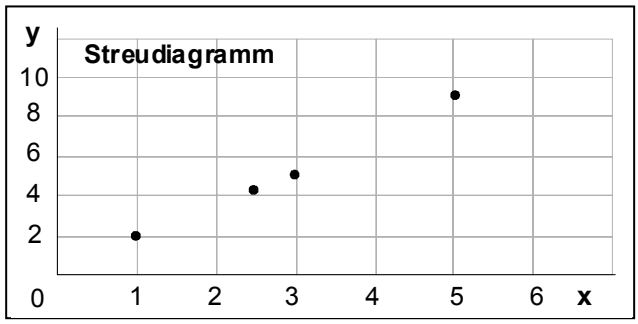
$\bar{x} = \frac{11.5}{4} = \underline{2.875}$ $\quad \bar{y} = \frac{20.5}{4} = \underline{5.125}$ $\quad b = \frac{73.25 - 4 \cdot 5.125 \cdot 2.875}{41.25 - 4 \cdot 8.2656} = \underline{1.748}$

$a = 5.125 - 1.748 \cdot 2.875 = \underline{0.099}$ $\quad \underline{\hat{y} = 0.0099 + 1.748 \cdot x}$ → Regressionsgerade

1.5.2 Streudiagramme und Kovarianz

- Es ist nützlich zuerst die Daten in einem **Streudiagramm** zu zeichnen, um herauszufinden, ob die Daten **zueinander in Beziehung stehen oder nicht.**
- Zur Beurteilung solcher Streudiagramme dient die **Kovarianz.**

x_i	y_i
1	2
2.5	4.5
3	5
5	9
11.5	20.5



- **1. Kovarianz**

Die Kovarianz ist in Analogie zur Varianz definiert. Bei der Varianz mitteln wir die Abweichungen einer Variablen vom arithmetischen Mittel im Quadrat. Die Varianz beider Variablen kann also berechnet werden:

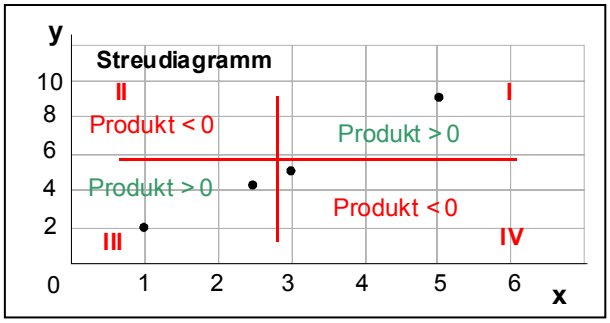
$$\sigma^2_x = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x})^2 \quad \text{(für x-Werte)} = 2.05$$

$$\sigma^2_y = \frac{1}{n} \cdot \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{(für y-Werte)} = 6.30$$

Die Kovarianz ist die gemeinsame Varianz der x- und y-Werte.

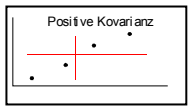
$$\sigma_{xy} = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x}) \cdot (y_i - \bar{y})$$

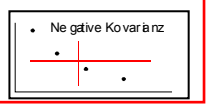
- Die Kovarianz kann positiv oder negativ werden.
- Die nebenstehende Abbildung zeigt die Geraden für das arithmetische Mittel von x und y, $\bar{x}$ und $\bar{y}$ und nummeriert die so entstehenden Quadranten mit I, II, III und IV.



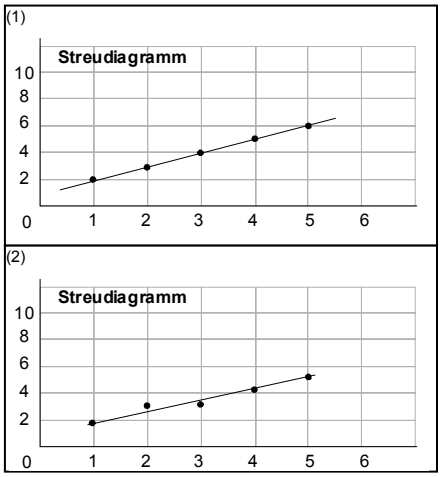
		Abweichung x	Abweichung y	Produkt
x_i	y_i	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
1	2	-1.875	-3.125	5.859375
2.5	4.5	-0.375	-0.625	0.234375
3	5	0.125	-0.125	-0.015635
5	9	2.125	3.875	8.234375
				14.3125

- Die Kovarianz beträgt 14.3125 / 4 = 3.578125
- Falls sich die Punkte nur im ersten und dritten Quadranten befinden, dann erhalten wir für jede einzelne Summanden des Produkts $(x_i - \bar{x}) \cdot (y_i - \bar{y})$ einen positiven Wert und das Produkt bleibt positiv. Im obigen Beispiel ist ein Produkt negativ da einer seiner Summanden positiv ist, der andere negativ. Dies ist der Fall, weil Punkt (3/5) im 4. Quadrant liegt. Dieser Ausreisser hat jedoch nur einen kleinen Einfluss auf die Kovarianz.
- **Erkenntnis:**
  - **Für x zunehmend und y zunehmend besteht eine positive Kovarianz (Zusammenhang).** In diesem Fall befindet sich die Mehrheit der Punkte in den Quadranten III und I. **(gleichläufige x und y)**
  - **Für x zunehmend und y abnehmend besteht eine negative Kovarianz (Zusammenhang).** In diesem Fall befindet sich die Mehrheit der Punkte in den Quadranten II und IV. **(gegenläufige x und y)**





- **Interpretation der Kovarianz**
  - Ist die Kovarianz = 0, so kann man keine Aussage machen, ob mit zunehmenden x-Werten die y-Werte zunehmen oder abnehmen.
  - Die erste (1) Grafik zeigt einen Zusammenhang mit einer Kovarianz von 2. Der zweite (2) Zusammenhang weist auch eine Kovarianz von 2 auf. Man kann also aus dem Betrag der Kovarianz nicht auf die Güte des Zusammenhangs schliessen.
  - Auch das Umgekehrte ist möglich, dass man trotz unterschiedlicher Kovarianz einen gleichen Zusammenhang zwischen den Daten hat. Dies liegt oft daran, dass ein Zusammenhang grundsätzlich identisch ist mit dem ersten, aber einen Datenpunkt mehr hat.
  - **Die Kovarianz sagt also nur mit dem Vorzeichen, ob die beiden Merkmalswerte gleichläufig oder gegenläufig sind. Sie gibt kein Ausmass der Abhängigkeit wieder.**



1.5.3 Korrelationskoeffizient (Masskorrelation, Produkt-Moment-Koeffizient, Korrelationskoeffizient von Bravais-Pearson)

- Mit der Kovarianz kann also nur bestimmt werden, ob ein Zusammenhang gleichläufig oder gegenläufig ist. Um eine Aassage über das **Ausmass der Abhängigkeit** zu machen, muss die Kovarianz normiert werden. Dies ist mit dem Korrelationskoeffizient möglich.
- **Der Korrelationskoeffizient entsteht aus der Kovarianz, indem die Kovarianz durch die Standardabweichung von x mal die Standardabweichung von y dividiert wird. Es können nur mehr Werte zwischen – 1 und 1 entstehen, daher $r \varepsilon$ = [1; -1].**

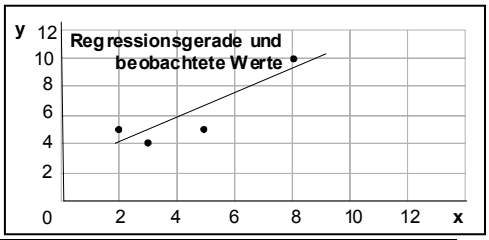
$$r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$ σ = Standardabweichung

- Im obigen Beispiel ergäbe dies also: 3.578125 / (√2.05 · √6.30) = 0.997
- **Das Vorzeichen des Korrelationskoeffizienten gibt an, ob die Merkmalswerte gleichläufig oder gegenläufig sind. Der Betrag des Korrelationskoeffizienten variiert zwischen 0 und 1. Je näher er bei 1 liegt, desto stärker, ausgeprägter ist der Zusammenhang zwischen den Merkmalswerten und der Regressionsgeraden. Je näher er bei 0 liegt, desto schwächer, unausgeprägter ist der Zusammenhang zwischen den Merkmalswerten und der Regressionsgeraden.**

1.5.4 Bestimmtheitsmass (Determinationskoeffizient)

- Die folgenden Erläuterungen gehen einer Regressionsgerade nach der Methode der kleinsten Quadrate von $\hat{y} = 0.905x + 1.923$ aus. Es geht jetzt darum eine Aussage zu machen, ob diese Regression den Zusammenhang zwischen x und y gut oder schlecht abbildet.

x_i	y_i
2	5
3	4
5	5
8	10



- **Fläche der kleinsten Quadrate**
  - Von der Methode der kleinsten Quadrate wissen wir, dass die Abweichungsquadrate zur Regressionsgeraden minimal sind. Im Folgenden wird nun die Fläche dieser Quadrate berechnet.

x_i	y_i	**Regression** $\hat{y} = 0.905x + 1.923$	**Abweichung** $y_i - \hat{y}$	**im Quadrat** Fläche der Quadrate (Abweichung^2)
2	5	3.73	1.26	1.59
3	4	4.64	-0.64	0.41
5	5	6.45	-1.45	2.11
8	10	9.17	0.83	0.69
				4.81

  - Diese Fläche der Quadrate kann auch mit folgender Formel errechnet werden:

$$\sum_{i=1}^{n} (y_i - b \cdot (x_i - \bar{x}) - \bar{y})^2 \qquad \textit{oder} \quad \sum_{i=1}^{n} (y_i - \bar{y})^2 - \frac{\left[\sum_{i=1}^{n} (y_i - \bar{y}) \cdot (x_i - \bar{x})\right]^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

b = gemäss Methode der kl. Quadrate

  - Lösung: $22 - (19^2 / 21) = \underline{4.81}$ (Fläche der kleinsten Quadrate)

- Das **Bestimmtheitsmass R^2** misst die Stärke des Zusammenhangs zwischen zwei Merkmalen x und y, die beide mindestens **intervallskaliert** sind. Die Stärke des Zusammenhangs zwischen zwei Merkmalen x und y kann festgestellt werden, indem über eine Streuungszerlegung (Varianzzerlegung) bestimmt wird, inwieweit die Abweichung der Merkmalswerte y vom durchschnittlichen Merkmalswert $\bar{y}$ durch das Merkmal x verursacht wird. **Das Bestimmtheitsmass informiert darüber, welcher Teil der Varianz durch die Regression** (Methode der kleinsten Quadrate) **bestimmt werden kann. Es können nur Werte zwischen – 1 und 1 entstehen, daher R^2 ε = [1; -1]. Es kann in Prozenten ausgedrückt werden.**
- **Je näher der Wert des Bestimmtheitsmasses bei dem Wert 1 (100 %) liegt, desto stärker bzw. ausgeprägter ist der Zusammenhang. Je näher der Wert des Bestimmtheitsmasses bei dem Wert 0 (0 %) liegt, desto schwächer bzw. weniger ausgeprägt ist der Zusammenhang.**

1. Das Bestimmtheitsmass wird 1 (starker Zusammenhang), wenn die Fläche der kleinsten Quadrate 0 ist:

$$\sum_{i=1}^{n} (y_i - b \cdot (x_i - \bar{x}) - \bar{y})^2 = 0$$

das heisst, wenn die Abweichungen zwischen den berechneten (erklärten) und beobachten Werten Null sind. Je grösser die Abstände, desto kleiner wird R^2. Das Bestimmtheitsmass sagt also aus, ob die Regressionsgerade (Methode der kleinsten Quadrate) den Zusammenhang von x und y schlecht oder gut wiedergibt.

2. Das Bestimmtheitsmass ist eine normierte Kovarianz. Das heisst R^2 kann wie der Korrelationskoeffizient interpretiert werden.

$$\mathbf{R^2 = \frac{б^2_{xy}}{б^2_x \cdot б^2_y}} \quad \text{(in \% · 100)} \qquad \text{oder} \qquad \mathbf{R^2 = r^2}$$

x_i	y_i	$y_i - \bar{y}$	$x_i - \bar{x}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})^2$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
2	5	-1	-2.5	1	6.25	2.5
3	4	-2	-1.5	4	2.25	3
5	5	-1	0 .5	1	0.25	-0.5
8	10	4	3.5	16	12.25	14
4.5	6			22	21	19
Arith. Mittel						

  - $б_{xy} = 19/4 = 4.75$ $\quad б^2_x = 21/4 = 5.25$ $\quad б^2_y = 22/4 = 5.50$
  - $4.75^2 / (5.50 \cdot 5.25) = 0.7814 \rightarrow \underline{\mathbf{R^2 = 78.14\ \%}}$

- **Je näher der Wert des Bestimmtheitsmasses bei dem Wert 1 (100 %) liegt, desto stärker bzw. ausgeprägter ist der Zusammenhang. Je näher der Wert des Bestimmtheitsmasses bei dem Wert 0 (0 %) liegt, desto schwächer bzw. weniger ausgeprägt ist der Zusammenhang.**
- 78.14 % **drückt aus, dass** 78.14 % **der Gesamtvarianz σ^2_y durch die Varianz der Regressionswerte $\sigma^2_{\hat{y}}$ bestimmt werden kann. Das heisst, dass** 78.14 % **von y_i von x_i abhängig ist.**

3. Das Bestimmtheitsmass ist eine normierte Kovarianz. Das heisst R^2 kann wie der Korrelationskoeffizient interpretiert werden (Varianzzerlegung).

$$R^2 = \frac{\sigma^2_{\hat{y}}}{\sigma^2_y}$$

Darüber hinaus gilt, dass die Gesamtstreuung als Summe der Streuung der berechneten (erklärte) Werte und der Streuung der Abweichungen erklärt werden kann:

$\sigma^2_y = \sigma^2_{\hat{y}} + \sigma^2_{y-\hat{y}}$ (Varianzzerlegung)

Beispiel:

	x_i	y_i	Regression $\hat{y} = 0.905x + 1.923$	Abweichung $y_i - \hat{y}$
	2	5	3.73	1.26
	3	4	4.64	-0.64
	5	5	6.45	-1.45
	8	10	9.17	0.83
Arith. Mittel	4.50	6.00	6.00	0.00
Varianz	5.25	5.50	4.30	1.20

4.30 + 1.20 = 5.50

4. Auf diesem Hintergrund kann auch das sogenannte Unbestimmtheitsmass definiert werden. Der Anteil der Abweichung, der durch die Regression unbestimmt bleibt, wird durch das Unbestimmtheitsmass U^2 ausgedrückt

$$U^2 = 1 - R^2$$

Es drückt aus, dass 21.86 % der Varianz nicht von der Regression abhängig sind.

5. Weitere Berechnungsmöglichkeiten des Bestimmtheitsmasses

$$R^2 = r^2$$

$$r = b \cdot \frac{\sigma_x}{\sigma_y} \qquad R^2 = b^2 \cdot \frac{\sigma^2 x}{\sigma^2 y}$$

b = gemäss Methode der kl. Quadrate (Steigung der Regression)

1.5.5 Rangkorrelation

- Soll der Zusammenhang zwischen zwei Merkmalen untersucht werden, von denen das eine nur **ordinalskaliert** ist, während das **andere mindestens ordinalskaliert** ist, dann kann keine Regression erstellt werden, und die bisher beschriebnen Kennzahlen können nicht ohne weiteres angewendet werden. In diesem Fall kann allerdings die Rangkorrelation angewendet werden.
- **Die beiden Merkmalsträger werden hinsichtlich der beiden Merkmale in eine Rangordnung gebracht.** Der Grad des Zusammenhangs zwischen den beiden Merkmalen kann dann festgestellt werden, indem die beiden Rangordnungen auf den Grad ihrer Übereinstimmung verglichen werden.
- **Sind zwei oder mehr Merkmale gleich, dann wird den Merkmalen das arithmetische Mittel der Ränge zugeordnet, die dieser Merkmale erhalten hätten, wenn sie nicht gleich wären.**
  - In nebenstehender Tabelle heisst das, da zweimal (für Rang 2 und 3) gut zugeordnet wurde, dass beide Merkmalsträger den Wert 2.5 erhalten.

Wein	Bewertung x_i	Preis y_i	Rang x_i	Rang y_i
A	ausreichend	13.50	5	5
B	mangelhaft	15.20	6	4
C	gut	16.30	**2.5**	3
D	sehr gut	18.50	1	1
E	befriedigend	17.40	4	2
F	gut	12.90	**2.5**	6

- Aus diesen Rangwerten kann nun der **Korrelationskoeffizient** (in diesem Fall **Rangkorrelation** genannt) berechnet werden

$$\rho \text{ (griech. Rho)} = r = \frac{\sigma_{xy}}{\sigma_x \cdot \sigma_y}$$

x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})^2$	$(x_i - \bar{x}) \cdot (y_i - \bar{y})$
5	5	1.5	1.5	2.25	2.25	2.25
5	4	2.5	0.5	6.25	0.25	1.25
2.5	3	-1	-0.5	1	0.25	0.50
1	1	-2.5	-2.5	6.25	6.25	6.25
4	2	0.5	-1.5	0.25	2.25	-0.75
2.5	6	-1	2.5	1	6.25	-2.50
3.5	**3.5**	**Summe**		**17**	**17.5**	**7**
Mittelwert		**Varianz** σ^2		**2.83**	**2.92**	
		Standardabweichung		**1.68**	**1.71**	
		Kovarianz				**1.17**
		Rangkorrelation ρ = 1.17 / (1.68 · 1.71) = **0.406**				

- Sind die Merkmale **vollständig**, oder wie im Beispiel fast **vollständig unterschiedlich,** so dass keine oder fast **keine arithmetische Mittel für die Rangzuteilung** erfolgen müssen, kann man auch folgende Formel anwenden:

$$\rho \text{ (griech. Rho)} = r = 1 - \frac{6 \cdot \sum_{i=1}^{n} D_i^2}{n^3 - n} \qquad \text{wobei } D_i = \text{Rang}(x_i) - \text{Rang}(y_i)$$

$$1 - \frac{6 \cdot 20.5}{6^3 - 6} = 0.414$$

Wein	Bewertung x_i	Preis y_i	Rang x_i	Rang y_i	D_i	D_i^2
A	ausreichend	13.50	5	5	0	0
B	mangelhaft	15.20	6	4	2	4
C	gut	16.30	2.5	3	-0.5	0.25
D	sehr gut	18.50	1	1	0	0
E	befriedigend	17.40	4	2	2	4
F	gut	12.90	2.5	6	-3.5	12.25
		Mittelwert:	**3.5**	**3.5**		**20.5**

  - Es gibt also zwischen 0.406 und 0.414 eine kleine Abweichung, weil bei den Rängen einmal gemittelt werden musste. Wir sehen aber, dass die Kennzahl auf zwei Kommastellen gerundet in beiden Fällen 0.41 ergibt.
- **Interpretation der Rangkorrelation**
  - Die Kennzahl kann gleich interpretiert werden wie der Korrelationskoeffizient. Allerdings ist darauf zu achten, dass nur der Zusammenhang zwischen den Rängen untersucht wird, und nicht der Zusammenhang zwischen den Merkmalswerten selbst.
  - Das Vorzeichen gibt an, dass bei dieser Zuordnung der Ränge (1: sehr gut, 1: höchster Preis) ein Zusammenhang (wenn auch nur schwach) derart besteht, dass mit höherem Preis mit höherer Qualität gerechnet werden kann. Hätte man folgende Zuordnung gewählt: (1: sehr gut, 1: tiefster Preis), wäre ein negatives Vorzeichen herausgekommen.

2. Wahrscheinlichkeitsrechnung

Aufgabe der Wahrscheinlichkeitsrechnung ist es, das Ausmass der Sicherheit, mit der ein möglicher Ausgang eintritt, zahlenmässig auszudrücken.

2.1 Grundbegriffe

- **Zufallsvorgang (Zufallsexperiment oder Zufallsbeobachtung)**
  - **Ein Zufallsvorgang ist ein Vorgang, dessen Ausgang aufgrund von Unkenntnis oder Unwissenheit nicht vorhergesagt werden kann.**
    - *Beispiel 1:* Ein Würfel wird einmal geworfen
    - *Beispiel 2:* Aus einer Lieferung werden drei Einheiten entnommen und auf ihre Funktionstüchtigkeit geprüft.
    - *Beispiel 3:* Der Benzinverbrauch eines umweltfreundlichen Autos wird für eine Teststrecke von 100 km. gemessen.
- **Elementarereignis und Ereignisraum**
  - **Elementarereignisse heissen die einzelnen, sich gegenseitig ausschliessenden möglichen Ausgänge eines Zufallsvorgangs.**
    - Bei der Durchführung eines Zufallsvorgangs tritt genau ein Elementarereignis ein.
    - *Beispiel 1:* Wird der Würfel einmal geworfen, dann wird eine der Augenzahlen 1, 2, 3, 4, 5 oder 6 erscheinen.
    - *Beispiel 2:* Bei der Prüfung auf Funktionstüchtigkeit einer jeden der drei Einheiten wird das Urteil „ja“ oder „nein“ lauten. Die möglichen Elementarereignisse lauten: (n, n, n), (j, n, n), (n, j, n), (n, n, j), (j, j, n), (j, n, j), (n, j, j) und (j, j, j)
    - *Beispiel 3:* Der Benzinverbrauch des Autos möge, sehr feine Messgenauigkeit vorausgesetzt, jeden Wert zwischen 2.7 und 3.1 annehmen können.
  - **Der Ereignisraum Ω ist die Menge aller möglichen Elementarereignisse.**
    - **Diskrete Ereignisräume**
      - Diskrete Ereignisräume umfassen endlich oder abzählbar viele Elementarereignisse (viele, aber nicht unendlich viele Ereignisräume)
    - **Stetige Ereignisräume**
      - Stetige Ereignisräume umfassen unendlich oder überabzählbar viele Elementarereignisse (unendliche Ereignisräume)
    - *Beispiel 1:* $\Omega = \{1, 2, 3, 4, 5, 6\}$ (diskret)
    - *Beispiel 2:* $\Omega = \{(n, n, n), (j, n, n), (n, j, n), (n, n, j), (j, j, n), (j, n, j), (n, j, j), (j, j, j)\}$ (diskret)
    - *Beispiel 3:* $\Omega = \{\text{Benzinverbrauch } x \mid 2.7 \leq 3.1\}$ (stetig)
      Leseweise: Ω umfasst alle Benzinverbräuche x, die die Bedingung x grösser gleich 2.7 und zusätzlich kleiner gleich 3.1 Liter erfüllen.
- **Ereignis**
  - **Das Ereignis ist der Ausgang des Zufallsvorgangs, also eine bestimmte Menge, die sich aus einem Elementarereignis oder mehreren Elementarereignissen zusammensetzt.** Das Ereignis kann den gewünschten Ausgang oder den effektiv eingetretenen Ausgang bezeichnen. Den Ereignissen werden Grossbuchstaben zugeordnet.
    - *Beispiel 1:*
      - Ereignis A: Werfen der Augenzahl 6
        $A = \{6\}$
      - Ereignis B: Werfen einer geraden Augenzahl.
        $B = \{2, 4, 6\}$
    - *Beispiel 2:*
      - Die Lieferung wird nur bei dem Ereignis C (mindestens zwei funktionstüchtige Einheiten) entgegengenommen.
        $C = \{(j,j,n), (j,n,j), (n,j,j), (j,j,j)\}$
    - *Beispiel 3:*
      - Die Entwicklungsingenieure sind an dem Ereignis D (wenn der Benzinverbrauch unter 3 Litern pro 100 km liegt) interessiert.
        $A = \{\text{Benzinverbrauch } x \mid x < 3\}$
    - **Sicheres Ereignis**
      - Ein Ereignis ist sicher, wenn es alle Elementarereignisse eines Ereignisraumes umfasst.
      - *Beispiel 1:* $E = \{1, 2, 3, 4, 5, 6\}$
    - **Unmögliches Ereignis**
      - Ein Ereignis ist unmöglich, wenn es kein Elementarereignis des Ereignisraums umfasst.
      - *Beispiel 1:* $E = \{\} = \varnothing$

2.2 Direkte Ermittlungen von Wahrscheinlichkeiten

- Bei Zufallsvorgängen stellt sich die Frage, welches Ereignis eintreten wird und wie gross die Chancen und Risiken für dieses Eintreten sind. Diese Wahrscheinlichkeiten können entweder direkt oder indirekt ermittelt werden.
- **Bei der direkten Ermittlung wird der Zufallsvorgang tatsächlich oder gedanklich durchgeführt.**
- **a) Klassische Wahrscheinlichkeitsermittlung** (Laplace-Wahrscheinlichkeit, von Jakob Bernoulli erfunden)
  - **Voraussetzungen**
    - Der Zufallsvorgang besitzt endlich/diskret viele Elementarereignisse und diese sind alle gleichmöglich bzw. gleich wahrscheinlich
  - **Konzept**
    - Der Zufallsvorgang wird nicht tatsächlich durchgeführt, sondern die Wahrscheinlichkeit wird auf rein gedankliche Weise ermittelt. Die Wahrscheinlichkeit wird zum vornherein ermittelt, weshalb man sie auch **a-priori-Wahrscheinlichkeit** nennt.

    1. Bestimmung der für den Eintritt des Ereignisses A **günstigen Elementarereignisse** (sogenannte günstige Elementarereignisse)
    2. Bestimmung der Elementarereignisse, aus denen sich der **Ereignisraum Ω** zusammensetzt.
    3. Berechnung der **Wahrscheinlichkeit für das Eintritten von Ereignis A:** W(A)

    $$W(A) = \frac{\text{Anzahl der für A günstigen Elementarereignisse}}{\text{Anzahl der gleich möglichen Elementarereignisse (gemäss } \Omega)}$$

  - **Beispiel** *(Beispiel 1)*
    - Wie gross ist die Wahrscheinlichkeit beim einmaligen Werfen mit einem Würfel, das Ereignis A, „gerade Augenzahl“, zu erreichen?
    - A (günstige Elementarereignisse) = {2, 4, 6}
      Ω (Ereignisraum) = {1, 2, 3, 4, 5, 6}

      $W(A) = \frac{3}{6} = \frac{1}{2} = 50\ \%$
  - **Probleme**
    - Die Anwendung der klassischen Methode ist oft nicht möglich, da die Elementarereignisse nicht endlich/diskret sind, sondern stetig. Dies ist bei *Beispiel 3* der Fall. Ω (Ereignisraum) = {Benzinverbrauch x | 2.7 ≤ 3.1)
      - Einen Ausweg bietet hier die **geometrische Wahrscheinlichkeitsermittlung:**
        - Der Bereich von 2.7 bis 3.1 wird in 10 Intervalle unterteilt.
        - Falls uns nun die Wahrscheinlichkeit interessiert, in einem dieser 10 Intervalle zu erhalten, dann wäre die Wahrscheinlichkeit hierfür 10 %.
    - Die Anwendung der klassischen Methode ist oft auch nicht möglich, da die Elementarereignisse nicht gleich möglich sind. Dies ist bei *Beispiel 2* der Fall.
      Ω = {(n, n, n), (j, n, n), (n, j, n), (n, n, j), (j, j, n), (j, n, j), (n, j, j), (j, j, j)} Die Chance ganz taugliche Geräte (j, j, j) zu finden, ist grösser.
    - Auch das Auffinden bzw. Abzählen der Elementarereignisse kann problematisch sein.
  - **Bedeutung**
    - Grosse Bedeutung im Glücksspiel
- **b) Statistische Wahrscheinlichkeitsermittlung** (Richard von Mises, von Jakob Bernoulli erfunden)
  - **Voraussetzungen**
    - Der Zufallsvorgang muss unter identischen Bedingungen wiederholt werden. Er wird tatsächlich durchgeführt.
  - **Konzept**
    - Der Zufallsvorgang wird genügend oft tatsächlich durchgeführt. Die Wahrscheinlichkeit des Ereignisses A, W(A), ist dann die **relative Häufigkeit** des Ereignisses A.

      $$W(A) = \frac{\text{Zahl der Zufallsvorgänge } \mathbf{mit\ Ereignis\ A}}{\text{Zahl der Zufallsvorgänge insgesamt } \mathbf{n}}$$

      - Wird der Zufallsvorgang nur wenige Male durchgeführt, dann ist das Risiko gross, dass die festgestellte relative Häufigkeit von der tatsächlichen Wahrscheinlichkeit abweicht.
      - Mit zunehmender Wiederholdung des Zufallsvorgangs nähert sich die relative Häufigkeit der gesuchten Wahrscheinlichkeit asymptotisch an. Es gibt also einen Grenzwert für W(A). Dieser entspricht der relativen Häufigkeit.
      - Der Zufallsvorgang ist daher so lange zu wiederholen, bis sich die relative Häufigkeit stabilisiert hat.
        - Diese statistische Wahrscheinlichkeit findet ihre Berechtigung im **Gesetz der grossen Zahl** von Jakob Bernoulli:

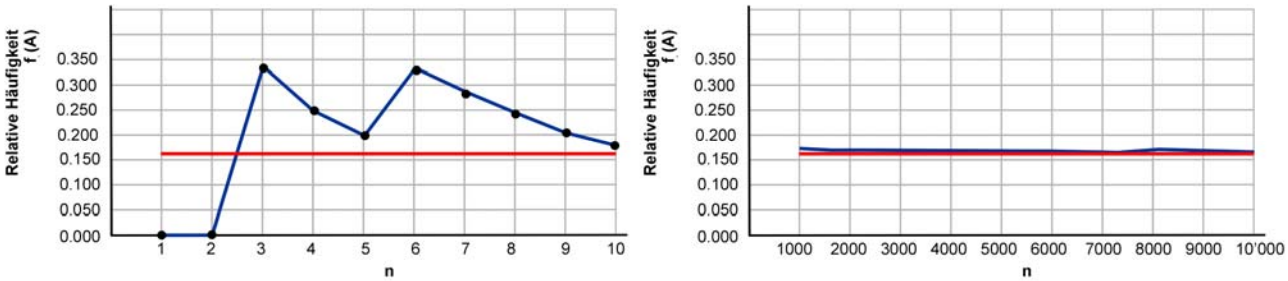
- Es besagt, dass mit wachsender Anzahl der Zufallsvorgänge die Wahrscheinlichkeit gegen Null strebt, dass die absolute Differenz aus der relativen Häufigkeit und der Wahrscheinlichkeit grösser als eine vorgegebene, beliebig kleine positive Zahl ε ist.

  $\lim_{n \to \infty} W(|f_n(A) - W(A)| > \varepsilon) = 0$

  $f_n(A)$: relative Häufigkeit für A bei n Zufallsvorgängen
  n: Anzahl Zufallsvorgänge

- Die Wahrscheinlichkeit wird erst nach Durchführung der Zufallsvorgänge bekannt. Deshalb nennt man sie **a posteriori-Wahrscheinlichkeit.**

o **Beispiel** *(Beispiel 1)*
- Beim einmaligen Werfen mit einem Würfel interessiert der Eintritt des Ereignisses A „Augenzahl 6".
  A = {6}
- Zur Ermittlung der Wahrscheinlichkeit wird nun der Zufallsvorgang „Werfen des Würfels" wiederholt durchgeführt. Dabei wird immer die relative Häufigkeit gerechnet um zu prüfen, ob sich diese schon stabilisiert hat. In einer Computersimulation ist dies 10'000 mal durchgeführt worden:



- Die rote Linie entspricht der Wahrscheinlichkeit wie man sie mit der Formel berechnen kann, wenn entsprechende Versuche angestellt wurden:

  $W(A) = \frac{1}{6}$
- Wie man deutlich sieht, nähert sich der Wert der relativen Häufigkeit immer mehr dieser Wahrscheinlichkeit an, je grösser n ist.
- Wie zeichnet man die Blaue Linie der relativen Häufigkeiten?
  1. Versuch (n=1): Wurf ohne sechs 0/1 (gemäss W(A)-Formel) = 0
  3. Versuch (n=3): Treffer 1/3 = 1/3 = 0.333
  4. Versuch (n=4): Wurf ohne sechs 1/4 = 0.250
  6. Versuch (n=6): Treffer 2/6 = 0.333
  Anschliessend nähert sich das Ergebnis immer 1/6 = 0.166
- Stellt man das rechte Diagramm, wo 10000 Versuche abgebildet ist, beispielsweise im hinteren Bereich zwischen 8000 und 9000 Versuchen vergrössert dar, erkennt man, dass die Wahrscheinlichkeit von 1/6 noch keineswegs erreicht wird. Allerdings hat sich der Wert der relativen Häufigkeiten dann schon sehr stark stabilisiert um 0.170 und 0.169.

o **Probleme**
- In der Praxis können Zufallsvorgänge meist nicht beliebig oft und identisch wiederholt werden. Deshalb kann mit dieser Wahrscheinlichkeit oft nicht gearbeitet werden.

o **Bedeutung**
- In vielen Fällen existieren keine andere Wahrscheinlichkeiten als die statistischen (z.B. Wahrscheinlichkeit für das Erreichen eines bestimmten Alters oder für eine Knabengeburt)
- Auch gewinnt die statistische Wahrscheinlichkeit an Bedeutung, da durch Computersimulation unter Umständen die tatsächliche Wiederholung des Zufallsvorgangs ersetzt werden kann.

▪ **c) Subjektive Wahrscheinlichkeitsermittlung**

o Die Subjektive Wahrscheinlichkeit ist eine **a priori-Wahrscheinlichkeit** (im Vornherein).

o **Sachkundige Personen beurteilen rein gedanklich die Möglichkeit des Eintretens eines Ereignisses zahlenmässig.**

o In der betrieblichen Praxis sehr oft anzutreffen: Beurteilen von Szenarien, Abschätzen von Chancen und Gefahren eines Projektes

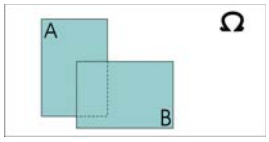
o Da diese Wahrscheinlichkeit nicht an Voraussetzungen wie Gleichmöglichkeit der Elementarereignisse und identische Wiederholbarkeit der Zufallsvorgänge gebunden ist, ist sie oft die einzige Möglichkeit für die Beurteilung von Situationen.

2.3 Indirekte Ermittlungen von Wahrscheinlichkeiten

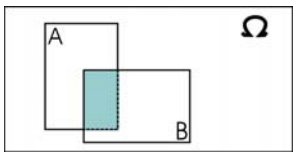
- **Bei der indirekten Ermittlung wird die Wahrscheinlichkeit für ein Ereignis aus den bekannten Wahrscheinlichkeiten anderer Ereignisse abgeleitet.** Der Zufallsvorgang wird nicht tatsächlich oder gedanklich durchgeführt.

- **Einführungsbeispiel**
  - Zufallsvorgang:
    Einmaliges Werfen eines Würfels mit den beiden Ereignissen A = {1} und B = {3, 5}.
  - Berechnung mit der klassischen Methode:
    $W(A) = \frac{1}{6}$ und $W(B) = \frac{2}{6}$
  - Das interessierende Ereignis C = {1, 3, 5} ist offensichtlich eine Relation aus den Ereignissen A und B. Ereignis C vereinigt die Elementarereignisse der Ereignisse A und B. Die Wahrscheinlichkeit für das Ereignis C wird indirekt ermittelt, indem die für die Ereignisse A und B bekannten Wahrscheinlichkeiten addiert werden:
    $W(C) = W(A) + W(B) = \frac{1}{6} + \frac{2}{6} = \frac{1}{2}$

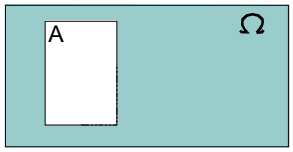
- **a) Relationen von Ereignissen**
  - Interessierende Ereignisse können oft durch eine Relation aus anderen Ereignissen beschrieben werden.
    - **Vereinigung von Ereignissen**
      - **Werden zwei oder mehr Ereignisse zu einem neuen Ereignis vereinigt, dann besteht das neue Ereignis aus genau den Elementarereignissen der vereinigten Ereignisse.**
        A ∪ B
      - **Alle Elementarereignisse, die entweder in A oder in B oder in A und B vorkommen gehören zu A ∪ B.**
      - **Beispiel**
        - A = {1, 2, 3} und B = {2, 4}
          A ∪ C = {1, 2, 3, 4}

      - **Bedeutung**
        - Oft interessiert bei der Durchführung eines Zufallsvorgangs, dass mindestens eines von mehreren Ereignissen eintritt. Das heisst, es interessiert, dass die Vereinigung eintritt.
        - Ein Spieler der auf „ungerade“ und „rot“ setzt gewinnt, wenn „ungerade“ oder „rot“ eintritt.
    - **Durchschnitt von Ereignissen**
      - **Der Durchschnitt der Ereignisse A, B und C umfasst genau die Elementarereignisse, die in jedem der Ereignisse A, B und C enthalten sind.**
        A ∩ B
      - **Alle Elementarereignisse, die sowohl in A als auch in B vorkommen gehören zu A ∩ B.**
      - **Besitzen zwei Mengen kein gemeinsames Elementarereignis, dann ist der Durchschnitt der beiden Mengen leer. Man nennt zwei solche Mengen auch disjunkt. A ∩ B = { } = ø**
      - **Beispiel**
        - A = {1, 2, 3} und B = {2, 3, 4, 5}
          A ∩ B = {2, 3}

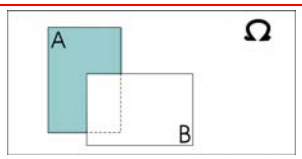
        

      - **Bedeutung**
        - Bei der Durchführung eines Zufallsvorgangs besteht oft ein Interesse daran, dass mehrere Ereignisse zugleich eintreten. Es interessiert der Durchschnitt.
        - Der Spieler ist besonders daran interessiert, dass „ungerade“ und „rot“ eintritt.
    - **Komplementärereignis**
      - **Ein Ereignis ist Komplementärereignis (Gegenereignis) zu einem anderen Ereignis, wenn es genau die Elementarereignisse des Ereignisraums umfasst, die nicht Elementarereignisse des anderen Ereignisses sind.**
      - **Das bedeutet, dass das zu A komplementäre Ereignis $\bar{A}$ genau dann eintritt, wenn das Ereignis A nicht eintritt.**

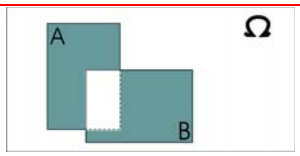


- **Beispiel**
  - A = {1, 2, 3} und $\overline{A}$ = {4, 5, 6}
- **Bedeutung**
  - Die Berechung der Wahrscheinlichkeit des Komplementärereignisses ist oft einfacher als die Berechnung des Ereignisses. Deshalb kommt dieser Relation grosse Bedeutung zu.

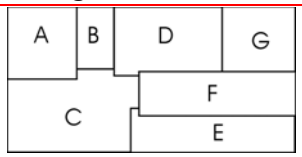
- **Logische Differenz**
  - **Das Ereignis, das aus der Differenz von A und B hervorgeht umfasst alle Elementarereignisse von A, die nicht auch Elementarereignisse von B sind.**
  - **A \ B umfasst also alle Elementarereignisse von A, die nicht Elementarereignisse von B sind.**
  - **Beispiel**
    - A = {1, 2, 3, 4) und B = {3, 4, 5}
      A \ B = {1, 2}



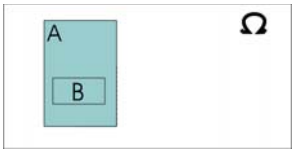
- **Symmetrische Differenz**
  - **Als symmetrische Differenz bezeichnen wir: A°B = A \ B ∪ B \ A**
  - **Die symmetrische Differenz besteht aus den Elementarereignissen der Vereinigung ohne die Elementarereignisse des Durchschnitts.**
  - **Die symmetrische Differenz der Ereignisse A und B ist das Ereignis, das genau aus den Elementarereignissen besteht, die entweder nur zu Ereignis A oder nur zu Ereignis B gehören.**
  - **Beispiel**
    - A = {1, 2, 3, 4} und B = {3, 4, 5)
      A°B = {1, 2, 5}



- **Vollständiges Ereignissystem**
  - **Als vollständiges Ereignissystem wird jede Zerlegung des Ereignisraumes Ω in paarweise disjunkte Ereignisse bezeichnet.**
  - **Ein vollständiges Ereignissystem ist eine Zusammenstellung von Ereignissen derart, dass jedes Elementarereignis des Ereignisraumes Ω in genau einem der Ereignisse enthalten ist.**



- **Teilereignis**
  - **Sind die Elemente eines Ereignisses B alle in A enthalten, so wird das Ereignis B als Teilereignis von A bezeichnet.**
  - **B ist also Teilereignis von A, wenn alle Elementarereignisse von B auch Elementarereignisse von A sind. B ⊆ A.**
  - **Wenn B eintritt, dann tritt auch A ein.**
  - **Beispiel**
    - Für A = {1, 2, 3, 4} und B = {2, 3} gilt B ⊆ A.



- **b) Eigenschaften von Wahrscheinlichkeiten**
  - Wahrscheinlichkeiten müssen bestimmte Grundeigenschaften haben, damit mit Ihnen gerechnet werden kann. Adrej Kolmogoroff hat folgende Axiome aufgestellt: (Ein Axiom ist ein Grundsatz, der nicht bewiesen werden kann.)

- **Axiom 1: Nichtnegativität**
  Jedem Ereignis kann eine Wahrscheinlichkeit zugeordnet werden, die grösser gleich Null ist.
  W(A) ≥ 0

- **Axiom 2: Normierung**
  Die Wahrscheinlichkeit für das sichere Ereignis ist gleich 1 = 100 %.
  W(Ω) = 1 = 100 %

- **Axiom 3: Additivität**
  Sind A und B zwei disjunkte Ereignisse, dann ist die Wahrscheinlichkeit für das Ereignis A ∪ B gleich der Summe der beiden Einzelwahrscheinlichkeiten für A und B.
  W(A ∪ B) = W(A) + W(B) *(siehe Einführungsbeispiel oben)*

- **c) Rechnen mit Wahrscheinlichkeiten**
  - Aufbauend auf dem Axiomensystem können weitere Eigenschaften von Wahrscheinlichkeiten abgeleitet werden.
  - **1. Additionssatz**
    - **Einführungsbeispiel**
      - Die Aufgabenstellung lautet, wie gross ist die Wahrscheinlichkeit, dass mindestens eines von mehreren gegebenen Ereignissen eintreten wird. Das entspricht der Wahrscheinlichkeit der Vereinigung. Es ist also die Eintretens-wahrscheinlichkeit für die Vereinigung von Ereignissen zu ermitteln.
      - Beim Zufallsvorgang „zweimaliges Werfen eines Würfels" interessieren die folgenden Ereignisse:
        - A = {Werfen eines Pasches} = {(1,1), (2,2), (3,3), (4,4), (5,5), (6,6)}
        - B = {Augenzahlsumme ≤ 4} = {(1,1), (1,2), (1,3), (2,1), (2,2), (3,1)}
        - Es gilt:

$$W(A) = \frac{\text{Anzahl der für A günstigen Ereignisse}}{\text{Anzahl der gleich möglichen Elementarereignisse}}$$

$$= \frac{\text{6 günstige Ereignisse (Werfen eines Pasches)}}{\text{6 Augenzahlen, zweimal würfeln} \rightarrow 6^2 = 36 \text{ mögliche Elementarereignisse (gemäss } \Omega)} = \frac{6}{36}$$

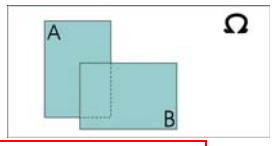
$$W(B) = \frac{6}{36}$$

      - Besteht ein Interesse, dass mindestens eines der beiden Ereignisse A oder B eintritt, so ist die Wahrscheinlichkeit für den Eintritt der Vereinigung von A und B zu bestimmen. Wir sehen, dass eine Addition der beiden Einzelwahrscheinlichkeiten einen zu hohen Wert ergeben würde, da die beiden Elementarereignisse (1,1) und (2,2) sowohl in A als auch in B vorkommen und somit doppelt erfasst würden. Von der Summe der beiden Einzelwahrscheinlichkeiten ist daher die Wahrscheinlichkeit für den Eintritt des Durchschnitts, d.h. der Elementarereignisse (1,1) und (2,2) abzuziehen.

$$W(\text{Durchschnitt A} \cap \text{B}) = \frac{2}{36}$$

$$W(A \cup B) = \frac{6}{36} + \frac{6}{36} - \frac{2}{36} = \frac{5}{18} = 27.78\ \%$$

    - **Additionssatz: Satz / Rechenregel**
      Bei der Addition der Wahrscheinlichkeiten für A und B wird die Wahrscheinlichkeit für die Schnittfläche (=Durchschnitt) doppelt erfasst. Deshalb muss die Wahrscheinlichkeit des Durchschnitts von der Summe der Einzelwahrscheinlichkeiten subtrahiert werden.



      - **Die Wahrscheinlichkeit, dass mindestens eines von zwei Ereignissen A und B eintritt beträgt: W(A ∪ B) = W(A) + W(B) – W(A ∩ B)**
      - **Der Additionssatz für drei Ereignisse lautet:**
        W(A ∪ B ∪ C) = W(A) + W(B) + W(C)
        – W(A ∩ B) – W(A ∩ B) – W(B ∩ C)
        + (A ∩ B ∩ C)
      - **Der Additionssatz für disjunkte Ereignisse lautet:** $\sum_{i=1}^{n} W(A_i)$
        (Einzelwahrscheinlichkeiten aufaddieren)
    - **Beispiele**
      - Klausur
        In der Tabelle sind die relativen Häufigkeiten für das Bestehen bzw. Nicht-Bestehen zweier Prüfungen angegeben. Die relativen Häufigkeiten können als Wahrscheinlichkeiten verwendet werden.
        - Wie gross ist die Wahrscheinlichkeit, dass ein zufällig ausgewählter Student

y	S	$\bar{S}$	Summe
M	60	20	80
$\bar{M}$	5	15	20
Summe	65	35	100

M Mathematik bestanden
$\bar{M}$ Mathematik nicht bestanden
S Statistik bestanden
$\bar{S}$ Statistik nicht bestanden

mindestens eine der beiden Klausuren bestanden hat?

- $W(S \cup M) = W(S) + W(M) - W(S \cap M)$
  $= 0.65 + 0.80 - 0.60 = 0.85 = \underline{85\ \%}$

- Werbeanzeige
  Ein Hersteller von Werkzeugmaschinen hat folgende Angaben:
  70 % seiner Kunden lesen Fachzeitschrift A. W(A)
  40 % seiner Kunden lesen Fachzeitschrift B. W(B)
  35 % seiner Kunden lesen sowohl Fachzeitschrift A wie auch B. $W(A \cap B)$
  - Wie gross ist die Wahrscheinlichkeit, dass ein potentieller Kunde die Anzeige liest, wenn der Werkzeugmaschinenhersteller in beiden Zeitschriften eine Anzeige aufgibt?
  - Merke: Jene 35 % die beide Zeitungen lesen, sind bereits in den 70 resp. 40 % inbegriffen.
  - $W(A \cup B) = W(A) + W(B) - W(A \cap B)$
    $= 0.70 + 0.40 - 0.35 = 0.75 = \underline{75\ \%}$

2. Bedingte Wahrscheinlichkeit

Definition

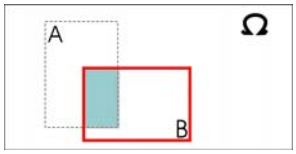
- Oft interessiert die Wahrscheinlichkeit eines Ereignisses unter einer Bedingung. Wie gross ist die Wahrscheinlichkeit für den Eintritt eines Ereignisses A, wenn ein Ereignis B bereits eingetreten ist?
- *Beispiel:* Die Wahrscheinlichkeit, dass ein Auto gestohlen wird, ist kleiner, falls es sich um einen Garagenwagen handelt.
- **Gesucht ist also die Wahrscheinlichkeit des Ereignisses A unter der Bedingung des Ereignisses B.**
  $W(A \mid B)$ (Wahrscheinlichkeit für A, unter der Bedingung B)

Einführungsbeispiel

- Beim Zufallsvorgang „zweimaliges Werfen eines Würfels“ interessiert die Eintrittswahrscheinlichkeit für das Ereignis A = {Augenzahlsumme ≤ 4} = {(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)}
  Die Eintrittswahrscheinlichkeit vor dem ersten Wurf beträgt: $W(A) = \frac{6}{36} = \frac{1}{6}$

  Wenn nun im ersten Wurf eine 1 geworfen wird (Bedingung B), dann reduziert sich die Anzahl der noch möglichen Elementarereignisse von ursprünglich 36 (6^2) auf folgende sechs Elementarereignisse: {(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)}

  Der Ereignisraum für den zweiten Wurf wird nun B:
  B = {(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)}.
  Die günstigen Fälle werden $A \cap B$: $A \cap B$ = {(1,1),(1,2),(1,3)}
  Die Eintrittswahrscheinlichkeit für A beträgt nun $W(A) = \frac{3}{6}$



- Die bedingte Wahrscheinlichkeit für den Eintritt des Ereignisses A, gegeben B, ergibt sich also wie folgt:
  $$W(A \mid B) = \frac{W(A \cap B)}{W(B)} = \frac{3}{6} = \frac{1}{2}$$
  Welche Elemente, die sowohl in A als auch in B vorkommen, kommen in B vor? Den B tritt sowieso ein (Bedingung).

  Ausgeschrieben hiesse dies so:
  $$W(A \mid B) = \frac{\frac{3}{36}}{\frac{6}{36}} = \frac{1}{2}$$
  Weshalb durch 36? Weil die Grundmenge zweimal Werfen ist, also $\Omega = 6^2$.

Bedingte Wahrscheinlichkeit: Satz / Rechenregel

- Haben wir vor dem ersten Wurf 6 günstige Fälle zu 36 gleichmöglichen Fällen, so reduziert sich nach dem ersten Wurf sowohl die Menge der günstigen Fälle (statt 6 nur noch 3) als auch die Menge der gleichmöglichen Fälle (statt 36 nur noch 6). **Die Bedingte Wahrscheinlichkeit ist das Verhältnis der verbleibenden günstigen Fälle (=3) zu den verbleibenden gleichmöglichen Fällen (=6).**
- **Die Wahrscheinlichkeit für das Ereignis A unter der Bedingung des Ereignisses B (W(B) > 0) beträgt:**
  $$W(A \mid B) = \frac{W(A \cap B)}{W(B)}$$
  Welche Elemente, die sowohl in A als auch in B vorkommen, kommen in B vor? Den B tritt sowieso ein (Bedingung).

- **Beispiele**
  - Klausur
    Wie viele der Studenten, die die Mathematikklausur bestanden haben, haben auch die Statistikprüfung bestanden?
    - $W(S \mid M) = \frac{W(S \cap M)}{W(M)} = \frac{60}{80} = \frac{3}{4} = \underline{75\ \%}$

y	S	$\bar{S}$	Summe
M	60	20	80
$\bar{M}$	5	15	20
Summe	65	35	100

M Mathematik bestanden
$\bar{M}$ Mathematik nicht bestanden
S Statistik bestanden
$\bar{S}$ Statistik nicht bestanden

  - Qualitätskontrolle von Tellern
    U Teller mit Unebenheiten, W(U) = 20 %
    D Teller mit Dekorfehlern, W(D) = 12 %
    $U \cap D$ Teller mit Unebenheiten und Dekorfehlern, $W(U \cap D) = 8\ \%$
    - Wie gross ist die Wahrscheinlichkeit, dass bei einem Teller mit Unebenheiten auch ein Dekorfehler auftritt.
    - $W(D \mid U) = \frac{W(D \cap U)}{W(U)} = \frac{0.08}{0.20} = \underline{40\ \%}$

o **3. Unabhängigkeit von Ereignissen**
- Es stellt sich die Frage, ob durch Eintritt eines Ereignisses A der Eintritt eines Ereignisses B beeinflusst wird oder nicht.
- **Einführungsbeispiele**
  - Beispiel mit abhängigen Ereignissen
    Zweimaliges Werfen eines Würfels: Die Wahrscheinlichkeit für das Ereignis A = {Augenzahlsumme ≤ 4} = {(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)} beträgt genau: $\frac{1}{6}$
    Wird im ersten Wurf eine 1 geworfen (Bedingung B), dann steigt die Wahrscheinlichkeit für Ereignis A auf:
    B = {(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)}

    $$W(A \mid B) = \frac{W(A \cap B)}{W(B)} = \frac{\frac{3}{36}}{\frac{6}{36}} = \frac{3}{6}$$

    Weshalb durch 36? Weil die Grundmenge zweimal Werfen ist, also $\Omega = 6^2$.

    Es ist recht logisch, dass die Wahrscheinlichkeit $W(A \mid B)$ nicht dieselbe wäre, wenn B nicht eingetroffen, sondern B eingetroffen wäre.
    A = {(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)} Die Roten kommen auch in $\bar{B}$ vor.
    $\bar{B}$ = im ersten Wurf kein 1 ≙ 36 – 6 = 30

    $$W(A \mid \bar{B}) = \frac{W(A \cap \bar{B})}{W(\bar{B})} = \frac{\frac{3}{36}}{\frac{30}{36}} = \frac{1}{10}$$

    Deshalb handelt es sich um abhängige Ereignisse.
  - Beispiel mit unabhängigen Ereignissen
    Für Ereignis C = („1“ im zweiten Wurf) beträgt die Wahrscheinlichkeit $W(C) = \frac{1}{6}$
    Dieses Ereignis ist unabhängig davon, ob 1 im ersten Wurf eingetreten ist oder nicht. Das lässt sich wie folgt beweisen:
  - 1 ist im ersten Wurf eingetreten:
    C = {(1,1),(2,1),(3,1),(4,1),(5,1),(6,1)}
    B = im ersten Wurf ein 1 {(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)}

    $$W(C \mid B) = \frac{W(C \cap B)}{W(B)} = \frac{\frac{1}{36}}{\frac{6}{36}} = \frac{1}{6}$$

    1 ist nicht im ersten Wurf eingetreten:
    C = {(2,1),(3,1),(4,1),(5,1),(6,1)} (1 weniger da (1,1) nicht vorkommen kann)
    $\bar{B}$ = im ersten Wurf kein 1 ≙ 36 – 6

    $$W(C \mid \bar{B}) = \frac{W(C \cap \bar{B})}{W(\bar{B})} = \frac{\frac{5}{36}}{\frac{30}{36}} = \frac{1}{6}$$

    Es spielt also keine Rolle für das Ereignis C ob B eintrifft oder nicht.
- **Unabhängigkeit von Ereignissen: Satz / Rechenregel**
  - **Zwei Ereignisse sind voneinander <u>unabhängig</u> wenn gilt:**
    $W(A) = W(A \mid B)$ oder
    $W(A) = W(A \mid \bar{B})$ oder
    $W(A \mid B) = W(A \mid \bar{B})$

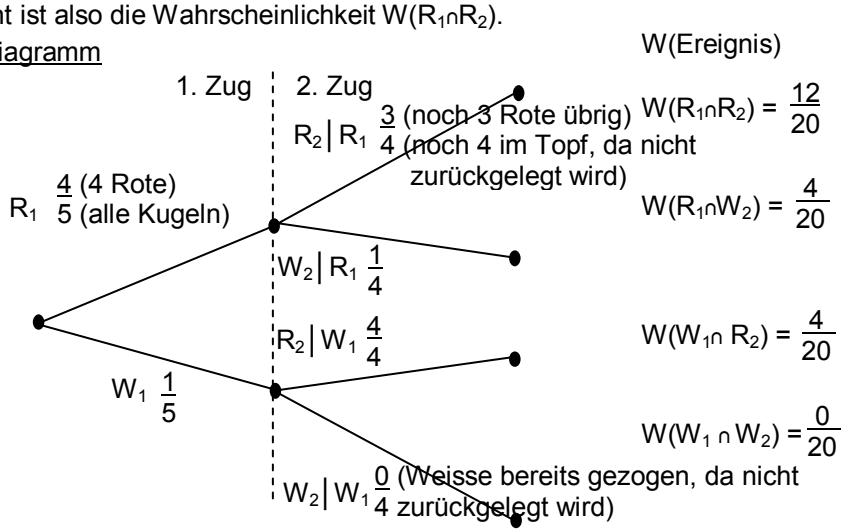
y	S	$\bar{S}$	Summe
M	60	20	80
$\bar{M}$	5	15	20
Summe	65	35	100

M Mathematik bestanden
$\bar{M}$ Mathematik nicht bestanden
S Statistik bestanden
$\bar{S}$ Statistik nicht bestanden

- **Beispiele**
  - Klausur
    Sind die beiden Ereignisse S (Bestehen der Statistikprüfung) und M (Bestehen der Mathematikprüfung) unabhängig?
    - W(S) = 65 %
    - $W(S \mid M) = \frac{W(S \cap M)}{W(M)} = \frac{0.6}{0.8} = 75\ \%$
    - Die Ereignisse sind abhängig voneinander.
  - Qualitätskontrolle von Tellern
    U Teller mit Unebenheiten, W(U) = 20 %
    D Teller mit Dekorfehlern, W(D) = 12 %
    $U \cap D$ Teller mit Unebenheiten und Dekorfehlern, $W(U \cap D) = 8\ \%$
    - Sind die Ereignisse D und U unabhängig?
    - $W(D \mid U) = \frac{W(D \cap U)}{W(U)} = \frac{0.08}{0.20} = 40\ \%$
    - W(D) = 12 %
    - Die Ereignisse sind abhängig voneinander.

o **4. Multiplikationssatz**

- **Definition**
  - Ausgehend von zwei oder mehr Ereignissen stellt sich die Frage, wie gross die Wahrscheinlichkeit ist, dass alle Ereignisse eintreten.
  - Gefragt ist also nach dem **Eintritt des Durchschnitts** dieser Ereignisse.
- **Einführungsbeispiel**
  - In einem Topf befinden sich fünf Kugeln: 4 Rote und 1 Weisse.
  - Wie gross ist die Wahrscheinlichkeit, dass bei zwei Ziehungen ohne Zurücklegen zwei rote Kugeln gezogen werden?
  - Die Ereignisse lauten: R_1 = {rot im 1. Zug}
    R_2 = {rot im 2. Zug}
    $R_1 \cap R_2$ = {rot im 1. Zug und rot im zweiten Zug)
    Gesucht ist also die Wahrscheinlichkeit $W(R_1 \cap R_2)$.
  - Baumdiagramm



  - Die Endwahrscheinlichkeiten ergeben sich durch multiplizieren der auf dem Weg liegenden Wahrscheinlichkeiten:

$$W(R_1 \cap R_2) = \frac{4}{5} \cdot \frac{3}{4} = \frac{12}{20} = 60\ \%$$

$$W(R_1 \cap W_2) = \frac{4}{5} \cdot \frac{1}{4} = \frac{4}{20} = 20\ \%$$

$$W(W_1 \cap R_2) = \frac{1}{5} \cdot \frac{4}{4} = \frac{4}{20} = 20\ \%$$

$$W(W_1 \cap W_2) = \frac{1}{5} \cdot \frac{0}{4} = \frac{0}{20} = 0\ \%$$

- **Multiplikationssatz: Satz / Rechenregel**
  - Die Endwahrscheinlichkeiten ergeben sich durch multiplizieren der auf dem Weg des Baumdiagramms liegenden Wahrscheinlichkeiten. Dies ergibt sich durch einfaches Umstellen der Formel aus der bedingten Wahrscheinlichkeit.

- **Die Wahrscheinlichkeit, dass <u>zwei Ereignisse A und B gemeinsam eintreten</u> beträgt:**

  $\mathbf{W(A \cap B) = W(A) \cdot W(B \mid A)}$ oder $\mathbf{W(A \cap B) = W(B) \cdot W(A \mid B)}$

  Bei n Ereignissen gilt entsprechend:

  $\mathbf{W(A_1 \cap A_2 \cap A_3 \cap \ldots \cap A_n) = W(A_1) \cdot W(A_2 \mid A_1) \cdot W(A_3 \mid A_1 \cap A_2) \cdot \ldots \cdot W(A_n \mid A_1 \cap A_2 \cap \ldots \cap A_{n-1})}$

- **Multiplikationssatz für unabhängige Ereignisse:**

  $\mathbf{W(A \cap B) = W(A) \cdot W(B)}$

  Bei n Ereignissen gilt entsprechend:

  $\mathbf{W(A_1 \cap A_2 \cap A_3 \cap \ldots \cap A_n) = W(A_1) \cdot W(A_2) \cdot W(A_3) \cdot \ldots \cdot W(A_n)}$

- **Beispiele**
  - <u>Wareneingangskontrolle Glühbirnen</u>
    Aus einer Lieferung von 50 Glühbirnen werden 4 ohne Zurücklegen entnommen und geprüft. Die Lieferung wird angenommen, wenn alle 4 Glühbirnen brennen. Wie gross ist die Wahrscheinlichkeit, dass eine Lieferung mit 10 % Ausschuss (5 defekte Glühbirnen) angenommen wird?
    - Die Wahrscheinlichkeit eine brennende Glühbirne zu ziehen beträgt beim ersten Ziehen: $\frac{45}{50}$

      Angenommen, die erste Glühbirne brennt, wie gross ist die Wahrscheinlichkeit bei der zweiten Ziehung wiederum eine zu ziehen die brennt (immer noch unter der Annahme, dass fünf defekt sind): $\frac{44}{49}$
    - Für A_i = {Glühbirne Nr. i ist in Ordnung) gilt:

      $W(A_1 \cap A_2 \cap A_3 \cap A_4) = W(A_1) \cdot W(A_2 \mid A_1) \cdot W(A_3 \mid A_1 \cap A_2) \cdot W(A_4 \mid A_1 \cap A_2 \cap A_3)$

      $= \frac{45}{50} \cdot \frac{44}{49} \cdot \frac{43}{48} \cdot \frac{42}{47} = \underline{64.696\ \%}$

- **5. Wahrscheinlichkeit des Komplementärereignisses**
  - **Definition**
    - **Das Komplementärereignis $\bar{A}$ ist das Ereignis, das genau dann eintrifft, wenn das Ereignis A nicht eintrifft.**
    - Die Aufgabe besteht darin, die Wahrscheinlichkeit zu ermitteln, dass das Ereignis A nicht eintritt.
  - **Einführungsbeispiel**
    - Ein Unternehmen führt die drei Produkte A, B und C ein. Die Produkte sind voneinander unabhängig, d.h. sie konkurrieren nicht miteinander und ergänzen sich nicht. Die Marketingabteilung schätzt die Wahrscheinlichkeit einer erfolgreichen Einführung auf 0.9, 0.8 bzw. 0.95.
    - Wie gross ist die Wahrscheinlichkeit, dass mindestens ein Produkt am Markt erfolgreich sein wird?
    - Es gelte:
      - A = {A ist erfolgsreich}
      - B = {B ist erfolgsreich}
      - C = {C ist erfolgsreich}
      - E = {mindestens ein Produkt ist erfolgreich}

      Gesucht ist also W(E).
    - Wir können diese Wahrscheinlichkeit auf <u>verschiedenen Wegen</u> ermitteln:
      - (1) Wir können überlegen, dass die gesuchte Wahrscheinlichkeit der Wahrscheinlichkeit der Vereinigung der drei Ereignisse A, B und C entspricht. Die Wahrscheinlichkeit, dass mindestens eines von drei Ereignissen A und B eintritt beträgt <u>$W(E) = W(A \cup B \cup C)$.</u>

        Additionssatz:

        $$\begin{aligned} W(A \cup B \cup C) &= W(A) + W(B) + W(C) \\ &\quad - W(A \cap B) - W(A \cap C) - W(B \cap C) \\ &\quad + W(A \cap B \cap C) \\ &= 0.9 + 0.8 + 0.95 \\ &\quad - (0.9 \cdot 0.8) - (0.9 \cdot 0.95) - (0.8 \cdot 0.95) \\ &\quad + (0.9 \cdot 0.8 \cdot 0.95) \\ &= 0.999 = \underline{99.9\ \%} \end{aligned}$$

- **Beachte: Multiplikationssatz für** $A \cap B \cap C$
  Da die Ereignisse unabhängig sind, kann mit dem Multiplikationssatz für unabhängige Ereignisse gerechnet werden.

- (2) Man kann auch überlegen, welches die günstigen Fälle sind, dass mindestens ein Produkt erfolgreich ist [$(A \cap B \cap C)$, $(A \cap B \cap \overline{C})$, $(A \cap \overline{B} \cap C)$, $(\overline{A} \cap B \cap C)$, $(A \cap \overline{B} \cap \overline{C})$, $(\overline{A} \cap B \cap \overline{C})$, $(\overline{A} \cap \overline{B} \cap C)$] und dann die Wahrscheinlichkeiten dieser Fälle bestimmen und die so ermittelten Wahrscheinlichkeiten aufsummieren.
  Multiplikationssatz:
  - Da die Ereignisse unabhängig sind, kann mit dem Multiplikationssatz für unabhängige Ereignisse gerechnet werden.
  - $W(A \cap B \cap \overline{C}) = W(A) \cdot W(B) \cdot W(\overline{C}) = 0.684$
    $W(A \cap \overline{B} \cap \overline{C}) = W(A) \cdot W(\overline{B}) \cdot W(\overline{C}) = 0.036$
    $W(A \cap \overline{B} \cap C) = W(A) \cdot W(\overline{B}) \cdot W(C) = 0.171$
    $W(\overline{A} \cap B \cap \overline{C}) = W(\overline{A}) \cdot W(B) \cdot W(\overline{C}) = 0.076$
    $W(\overline{A} \cap \overline{B} \cap \overline{C}) = W(\overline{A}) \cdot W(\overline{B}) \cdot W(\overline{C}) = 0.009$
    $W(\overline{A} \cap \overline{B} \cap C) = W(\overline{A}) \cdot W(\overline{B}) \cdot W(C) = 0.019$
    $W(A \cap B \cap \overline{C}) = W(A) \cdot W(B) \cdot W(\overline{C}) = 0.004$
    Summe $= 0.999 = \underline{99.9\ \%}$

- (3) Als dritte Möglichkeit können wir die Wahrscheinlichkeit des Komplementärereignisses (Welche Fälle sind denkbar, dass kein Produkt erfolgreich ist) bestimmen und die gesuchte Wahrscheinlichkeit ist dann 1 minus die Wahrscheinlichkeit des Komplementärereignisses.
  - Nur der Fall $(\overline{A} \cap \overline{B} \cap \overline{C})$ führt zum Ergebnis, dass die Forderung, mindestens ein Produkt sei erfolgreich, verletzt wird. Das bedeutet, dass man die Wahrscheinlichkeit dieses Falles bestimmen sollte, die sogenannte **Komplementärwahrscheinlichkeit** zu W(E). Diese Methode ist besser, da das Resultat sehr einfach zu bestimmen ist.
  - $W(E) = 1 - (\overline{A} \cap \overline{B} \cap \overline{C})$
    $= 1 - (0.1 \cdot 0.2 \cdot 0.05) = 1 - 0.001 = 0.999 = \underline{99.99\ \%}$

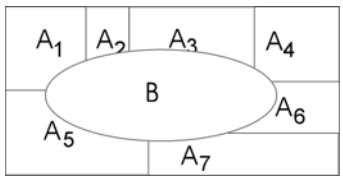
- **Wahrscheinlichkeit des Komplementärereignisses: Satz / Rechenregel**
  - **$W(A) = 1 - W(\overline{A})$** oder **$W(A) = 1 - W(\overline{A})$**

- **Beispiele**
  - Werfen mit zwei Würfeln
    Wie gross ist die Wahrscheinlichkeit, bei zehnmaligem Werfen mit zwei Würfeln mindestens einen Sechserpasch zu werfen.
    - A_i = {Sechserpasch im Wurf i}
    - Wie gross ist also die Wahrscheinlichkeit in zehn Würfen keinen Sechserpasch zu werden?
      $W(\overline{A_1} \cap \overline{A_2} \cap \overline{A_3} \cap \overline{A_4} \cap \overline{A_5} \cap \overline{A_6} \cap \overline{A_7} \cap \overline{A_8} \cap \overline{A_9} \cap \overline{A_{10}}) =$
      $1 - W(\overline{A_1}) \cdot W(\overline{A_2}) \cdot W(\overline{A_3}) \cdot W(\overline{A_4}) \cdot W(\overline{A_5}) \cdot W(\overline{A_6}) \cdot W(\overline{A_7}) \cdot W(\overline{A_8}) \cdot W(\overline{A_9}) \cdot W(\overline{A_{10}})$

      $A = (6,6)$
      Ω (1 Wurf mit zwei Würfel) $= 6^2 = 36$

      $$= \frac{35}{36} \cdot \frac{35}{36} \cdot \frac{35}{36} \cdot \frac{35}{36} \cdot \frac{35}{36} \cdot \frac{35}{36} \cdot \frac{35}{36} \cdot \frac{35}{36} \cdot \frac{35}{36} \cdot \frac{35}{36}$$
      $$= \left(\frac{35}{36}\right)^{10} = 1 - 0.7545 = \underline{24.55\ \%}$$
  - Arzneimittelforschung
    Ein Unternehmen schätzt die Wahrscheinlichkeit, dass von 10 entwickelten Medikamenten kein einziges eine Marktzulassung erreicht, auf 5 %.
    - Geben Sie das komplementäre Ereignis und dessen Wahrscheinlichkeit an:
    - $\overline{Z}_i$ = {Medikament i ist ohne Zulassung}
    - $W(\overline{Z}_i) = 0.05 = \underline{5\ \%}$
    - $W(Z_i) = 1 - 0.05 = 0.95 = \underline{95\ \%}$

- **6. Die totale Wahrscheinlichkeit**
  - **Definition**
    - Gegeben sind die **Ereignisse A_i** mit (i=1, …, n), die ein **vollständiges Ereignissystem** bilden (Alle A_i bilden als paarweise **disjunkte** Ereignisse Ω). Gegeben sei weiter das **Ereignis B**, das aus Elementarereignissen der Ereignisse A_i zusammengesetzt ist.



    - Die Wahrscheinlichkeiten $W(A_i)$ und $W(B \mid A_i)$ für i=1, …, n seien bekannt.
    - Gesucht ist die Wahrscheinlichkeit für das Ereignis B.
    - Fasst man B als Vereinigung aller Durchschnitte der A_i mit B auf $(A_1 \cap B) \cup (A_2 \cap B) \cup \ldots (A_n \cap B)$, dann kann die Wahrscheinlichkeit für B als Total der (=Summe) der Durchschnittswahrscheinlichkeiten berechnet werden: $\bigcup_{i=1}^{n} (A_i \cap B)$
      Deshalb nennt man sie totale Wahrscheinlichkeit.
  - **Einführungsbeispiel**
    - Drei Maschinen 1, 2, und 3 produzieren den gleichen Artikel. Die Maschinen haben einen Produktionsanteil von 60 %, 10 % und 30 %. Die Ausschussquote beträgt 5 %, 2 % und 4 %. Wie gross ist die Wahrscheinlichkeit, dass ein zufällig entnommener Artikel Ausschuss ist.
    - A_i = {Artikel stammt von Maschine i}, i = 1, 2, 3
      B = {Artikel ist Ausschuss)
      $B \mid A_i$ = {Artikel ist Ausschuss wenn er von Maschine A_i stammt}
    - $W(A_1) = 0.60$ $W(B \mid A_1) = 0.05$
      $W(A_2) = 0.10$ $W(B \mid A_2) = 0.02$
      $W(A_3) = 0.30$ $W(B \mid A_3) = 0.04$
    - Mit Hilfe des allgemeinen Multiplikationssatzes können wir die Wahrscheinlichkeiten der Durchschnitte berechnen:
      $W(A_1 \cap B) = W(A_1) \cdot W(B \mid A_1) = 0.60 \cdot 0.05 = 0.030$ (Die Wahrscheinlichkeit, dass ein Ausschuss bei Maschine 1 eintritt beträgt also 3 %.)
      $W(A_2 \cap B) = W(A_2) \cdot W(B \mid A_2) = 0.10 \cdot 0.02 = 0.002$
      $W(A_3 \cap B) = W(A_3) \cdot W(B \mid A_3) = 0.30 \cdot 0.04 = 0.012$
    - Die Wahrscheinlichkeit, dass ein zufällig entnommener Artikel Ausschuss ist, ergibt sich aus der Addition der drei berechneten Wahrscheinlichkeiten:
      $W(B) = W(A_1 \cap B) + W(A_2 \cap B) + W(A_3 \cap B) = 0.044 = \underline{\underline{4.4\ \%}}$
  - **Die totale Wahrscheinlichkeit: Satz / Rechenregel**
    Bilden die Ereignisse A_1, A_2 … A_n ein **vollständiges Ereignissystem** und ist B ein beliebiges Ereignis, dann gilt: $W(B) = \sum_{i=1}^{n} W(A_i) \cdot W(B \mid A_i)$
    $W(B \mid A_i)$ = Wahrscheinlichkeit für das Ereignis B unter der Bedingung des Ereignisses A_i
  - **Beispiele**
    - Klausur
      Aus der Tabelle ist ersichtlich, dass insgesamt 80 % die Mathematikprüfung bestanden haben.
      Wir treffen die Annahme, dass hier nur die folgenden Informationen gegeben sind:
      $W(S)$ = {Bestehen der Statistikprüfung} = 0.65
      $W(\overline{S})$ = Nicht-bestehen der Statistikprüfung) = 0.35
      $W(M \mid S)$ = {Wahrscheinlichkeit die Mathematikprüfung zu bestehen, wenn die Statistikprüfung bestanden ist} = 60/65
      $W(M \mid \overline{S})$ = {Wahrscheinlichkeit die Mathematikprüfung zu bestehen, wenn die Statistikprüfung nicht bestanden ist} = 20/35

y	S	$\overline{S}$	Summe
M	60	20	80
$\overline{M}$	5	15	20
Summe	65	35	100

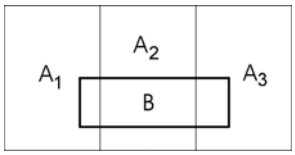
M Mathematik bestanden
$\overline{M}$ Mathematik nicht bestanden
S Statistik bestanden
$\overline{S}$ Statistik nicht bestanden

      - Wie gross ist die Wahrscheinlichkeit W(M), dass die Mathematikprüfung bestanden wird?
      - $W(M) = 0.65 \cdot 60/65 + 0.35 \cdot 20/35 = 0.8 = \underline{\underline{80\ \%}}$
    - Glücksspiel
      Der Spielgegner darf, für sie nicht sichtbar, 10 grüne Kugeln (G) und 10 rote Kugeln (R) beliebig au zwei Urnen A und B verteilen. Sie bestimmen die Urne aus der dann eine Kugel gezogen wird. Ist die Kugel grün, erhalten Sie 10 GE, ist sie rot, zahlen Sie 5 GE. Gehen Sie auf dieses Spiel ein?
      - Der Spielgegener legt eine rote Kugel in Urne A und die restlichen Kugeln in Urne B. Folgende Wahrscheinlichkeiten ergeben sich:
        W(Urne A) = 0.5 W(Urne B) = 0.5

$W(G \mid A) = 0$ $W(G \mid B) = 10/19$
$W(R \mid A) = 1$ $W(R \mid B) = 9/19$

- Wie viel beträgt die Wahrscheinlichkeit für die Entnahme einer grünen Kugel?
  - $W(G) = 0.5 \cdot 0 + 0.5 \cdot 10/19 = 0.2632$
- Wie viel beträgt die Wahrscheinlichkeit für die Entnahme einer roten Kugel?
  - $W(R) = 0.5 \cdot 1 + 0.5 \cdot 9/19 = 0.7368$
- Das heisst auf 19 Spiele wird 5 Mal die grüne (26.32 % von 19 Spielen) und 14 Mal die rote Kugel (73.68 % von 19 Spielen) gezogen.
  - Sie bezahlen 14 Mal 5 GE 70 GE
    Sie erhalten 5 Mal 10 GE 50 GE
    Im Schnitt verlieren Sie auf 19 Spiele 20 GE

- **7. Der Satz von Bayes**
  - **Definition**



    - Gegeben sind die **Ereignisse A_i** mit (i=1, …, n), die ein **vollständiges Ereignissystem** bilden (Alle A_i bilden als paarweise **disjunkte** Ereignisse Ω). Gegeben sei weiter das **Ereignis B**, das aus Elementarereignissen der Ereignisse A_i zusammengesetzt ist.
    - Die Wahrscheinlichkeiten $W(A_i)$ und $W(B \mid A_i)$ für i=1, …, n seien bekannt.
    - Gesucht ist nun nicht die Wahrscheinlichkeit für das Ereignis B, sondern $W(A_i \mid B)$: Wahrscheinlichkeit für das Ereignis A_i unter der Bedingung des Ereignisses B.
  - **Einführungsbeispiel**
    - Die totale Wahrscheinlichkeit, dass ein zufällig entnommener Artikel auf Maschine i produziert wird, beträgt:
      $W(A_1) = 0.60$
      $W(A_2) = 0.10$
      $W(A_3) = 0.30$
    - Die Wahrscheinlichkeit, dass ein zufällig entnommener Artikel Ausschuss ist, wenn er auf Maschine i produziert wurde, beträgt:
      $W(B \mid A_1) = 0.05$
      $W(B \mid A_2) = 0.02$
      $W(B \mid A_3) = 0.04$
    - Gesucht ist die Wahrscheinlichkeit, dass ein Artikel auf Maschine i produziert wurde, wenn von diesem Artikel bekannt ist, dass er Ausschuss (B) ist, also $W(A_i \mid B)$.
    - Von der totalen Wahrscheinlichkeit wissen wir:
      $W(A_1 \cap B) = W(A_1) \cdot W(B \mid A_1) = 0.60 \cdot 0.05 = 0.030$
      $W(A_2 \cap B) = W(A_2) \cdot W(B \mid A_2) = 0.10 \cdot 0.02 = 0.002$
      $W(A_3 \cap B) = W(A_3) \cdot W(B \mid A_3) = 0.30 \cdot 0.04 = 0.012$
      und
      $W(B) = W(A_1 \cap B) + W(A_2 \cap B) + W(A_3 \cap B) = 0.044 = \underline{4.4\ \%}$
    - Damit können wir $W(A_i \mid B)$ errechnen:
      Sind die Ereignisse D und U unabhängig?
      - $W(A_1 \mid B) = \frac{W(A_1 \cap B)}{W(B)} = \frac{0.030}{0.044} = \underline{68.18\ \%}$

        Die Wahrscheinlichkeit, dass ein zufällig ausgewählter Artikel auf Maschine 1 Produziert wurde beträgt 60 %. Falls der zufällig ausgewählte Artikel Ausschuss ist, beträgt die Wahrscheinlichkeit dass er auf Maschine 1 produziert wurde 68.18 %.
      - $W(A_2 \mid B) = \frac{W(A_2 \cap B)}{W(B)} = \frac{0.002}{0.044} = \underline{4.55\ \%}$
      - $W(A_3 \mid B) = \frac{W(A_3 \cap B)}{W(B)} = \frac{0.012}{0.044} = \underline{27.27\ \%}$

Die totale Wahrscheinlichkeit: Satz / Rechenregel
Bilden die Ereignisse $A_1, A_2 \dots A_n$ ein vollständiges Ereignissystem und ist B ein beliebiges Ereignis, dann gilt für das Ereignis $A_i \mid B$:

$$W(A_i \mid B) = \frac{W(A_i) \cdot W(B \mid A_i)}{\sum_{i=1}^{n} W(A_i) \cdot W(B \mid A_i)}$$

(Allgemeiner Multiplikationssatz für $W(A_i \cap B)$)
(Satz der totalen Wahrscheinlichkeit für W(B))

$W(A_i \mid B)$ = Wahrscheinlichkeit für das Ereignis A_i unter der Bedingung des Ereignisses B

- **Beispiele**
  - Qualitätsprüfung
    Eine automatische Messanlage prüft die Bruchfestigkeit von Rohrgestellen für Kopfstützen. Ein Rohrgestellt, das den Anforderungen nicht genügt, wird mit einer Wahrscheinlichkeit von 99.9 % als Fehlerhaft eingestuft. Ein Rohrgestellt, dass den Anforderungen genügt, wird mit einer Wahrscheinlichkeit von 2 % fälschlicherweise als fehlerhaft eingestuft. Der Anteil der fehlerhaften Rohrgestellte in der gesamten Produktion beträgt 3 %.
    - RG = {Rohr in Ordnung} W(RG) = 0.97
      RF = {Rohr ist fehlerhaft} W(RF) = 0.03
      F = {Rohr als „fehlerhaft“ eingestuft} $W(F \mid RG) = 0.02$ $W(F \mid RF) = 0.999$
      G = {Rohr als „in Ordnung“ eingestuft} $W(G \mid RF) = 0.001$ $W(G \mid RG) = 0.98$
    - Wie gross ist die Wahrscheinlichkeit, dass ein Rohrgestell fehlerhaft ist, wenn obwohl es als „in Ordnung“ eingestuft wurde.

      $$W(RF \mid G) = \frac{W(RF) \cdot W(G \mid RF)}{W(RF) \cdot W(G \mid RF) + W(RG) \cdot W(G \mid RG)}$$

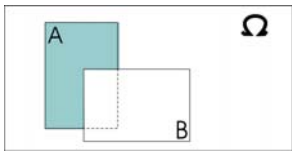
      $$W(RF \mid G) = \frac{0.03 \cdot 0.001}{0.03 \cdot 0.001 + 0.97 \cdot 0.98} = \underline{0.0031558\ \%}$$

  - Kundenzufriedenheit
    Eine Brauerei bewirtschaftet drei Biergärten A, B und C. Wiederholt wurde geklagt, dass die Bedienung sehr unfreundlich war. Im Biergarten A fühlten sich 10 %, in Biergarten B 40 % und in Biergarten C 70 % der Gäste unfreundlich bedient. Die Gäste verteilen sich im Verhältnis 60 %, 30 % und 10 % auf die drei Biergärten.
    - Wie gross ist die Wahrscheinlichkeit, dass ein Gast aus dem Biergarten A (bzw. B bzw. C) kommt, wenn er unzufrieden ist.
    - A = {Gäste im Biergarten A} W(A) = 0.60
      B = {Gäste im Biergarten B} W(A) = 0.30
      C = {Gäste im Biergarten C} W(A) = 0.10
      $U \mid A$ = {Unzufriedene Gäste in Biergarten A} $W(U \mid A) = 0.10$
      $U \mid B$ = {Unzufriedene Gäste in Biergarten B} $W(U \mid B) = 0.40$
      $U \mid C$ = {Unzufriedene Gäste in Biergarten C} $W(U \mid C) = 0.70$
    - $$W(A \mid U) = \frac{W(A) \cdot W(U \mid A)}{W(A) \cdot W(U \mid A) + W(B) \cdot W(U \mid B) + W(C) \cdot W(U \mid C)}$$

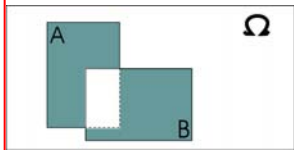
      $$W(A \mid U) = \frac{0.60 \cdot 0.10}{0.60 \cdot 0.10 + 0.30 \cdot 0.40 + 0.10 \cdot 0.70}$$

      $$W(A \mid U) = \underline{25\ \%}$$

- **8. Weitere Rechenregeln**
  - **Logische Differenz**
    - **A \ B umfasst also alle Elementarereignisse von A, die nicht Elementarereignisse von B sind.**
    - **Satz/Rechenregel**
      $W(A \setminus B) = W(A) - W(A \cap B)$



  - **Symmetrische Differenz**
    - **Als symmetrische Differenz bezeichnen wir:**
      $A \triangle B = A \setminus B \cup B \setminus A$
    - **Die symmetrische Differenz der Ereignisse A und B ist das Ereignis, das genau aus den Elementarereignissen besteht, die entweder nur zu Ereignis A oder nur zu Ereignis B gehören.**
    - **Satz/Rechenregel**
      $W(A \triangle B) = W(A) - W(A \cap B) + W(B) - W(A \cap B)$
      $= W(A) + W(B) - 2 \cdot W(A \cap B)$



- **Vollständiges Ereignissystem**
  - **Als vollständiges Ereignissystem wird jede Zerlegung des Ereignisraumes Ω in paarweise disjunkte Ereignisse bezeichnet.**
  - **Ein vollständiges Ereignissystem ist eine Zusammenstellung von Ereignissen derart, dass jedes Elementarereignis des Ereignisraumes Ω in genau einem der Ereignisse enthalten ist.**
  - **Satz/Rechenregel**

    $$\sum_{i=1}^{n} W(A_i) = 1$$