Modul 2 – Skalen und Klassierung

Algorithmus zu Merkmalstypen

1.) *Ist die Menge überabzählbar*?
    a. **JA** → stetiges Merkmal ➔ Quantitatives/metrisches Merkmal
    b. **NEIN** → diskretes Merkmal ➔ quantitativ/metrisch oder qualitativ
2.) *Ist das Merkmal zeitlich*?
    a. **JA** → quantitativ/metrisch
    b. **NEIN** → qualitativ
3.) *Ist das Merkmal räumlich*?
    a. **JA** → qualitativ
    b. **NEIN** → quantitativ/metrisch

Algorithmus zur Bestimmung des Skalenniveau

1.) *Ist das Merkmal metrisch [quantitativ] (überabzählbar, z. B. reele Zahlen)*?
    a. **JA** → *Besitzt das Merkmal einen natürlichen Nullpunkt*?
        i. **JA** → *Sind die Ausprägungen absolut (ganze Zahlen)*?
            1. **JA** → <u>Absolutskala</u> [Ratioskala] (z.B. Zahl der Beschäftigten)
            2. **NEIN** → <u>Verhältnisskala</u> (z.B. Umsatz, Körpergröße, Einkommen, Temperatur in Kelvin)
        ii. **NEIN** → <u>Intervallskala</u> (z.B. IQ, Temperatur in Celsius, Längendifferenzen)
    b. **NEIN** → *Hat das Merkmal eine Rangfolge*?
        i. **JA** → <u>Ordinalskala</u> (z B. Konfektionsgrößen, Schulnoten, Windstärke)
        ii. **NEIN** → <u>Nominalskala</u> (z.B. Geschlecht, Familienstand, Steuerklasse, PLZ)

Skalierungen

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

Modul 3 – Häufigkeiten und Häufigkeitsverteilung

Absolute und relative Häufigkeiten

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

absolute Häufigkeit

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

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

relative Häufigkeit

Bezug zur Grundgesamtheit

Algorithmus zur Häufigkeitsverteilung

1.) Sortieren der Rohdaten
2.) Verdichten der Rohdaten (deren Merkmalsausprägung zählen)
3.) Darstellen der verdichteten Rohdaten → Häufigkeitstabelle
4.) (optional) Klassierung der Daten
    a. Einteilung in Klassen → klassierte Daten
    b. Verdichtung der klassierten Daten → klassierte Verteilung
    c. Darstellung der klassierten Daten → Häufigkeitstabelle mit klassierten Daten

Summenhäufigkeiten

Absolute Summenhäufigkeiten (absolute kumulierte Häufigkeiten)

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

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

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

$$\dots$$

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

$$\dots$$

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

Relative Summenhäufigkeiten (relative kumulierte Häufigkeiten)

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

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

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

$$\ldots$$

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

$$\ldots$$

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

Klassenbreite b_i

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

Mit x_{k-1} für obere Klassengrenze und x_k für die untere Klassengrenze

Klassenmitte m_i (arithmetisches Mittel der Klassengrenzen)

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

Mit x_{k-1} für obere Klassengrenze und x_k für die untere Klassengrenze

Kreuztabelle (zweidimensionale Häufigkeitsverteilung) $G \rightarrow M$

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

Farbe	Bedeutung
Grau	Absolute Häufigkeit der Merkmalsausprägungskombination
Rot	Relative Zeilenhäufigkeiten (bedingte relative Häufigkeiten)
Blau	Relative Spaltenhäufigkeiten (bedingte relative Häufigkeiten)
Grün	Relative Häufigkeiten der Merkmalsausprägungskombination

Eigenschaften empirische Verteilungsfunktion $F(x)$ (Treppenfunktion)

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

Diagramme erstellen

Algorithmus (außer für Kreisdiagramm und Histogramm)

1.) Koordinatenkreuz einzeichnen
2.) Achsen mit Merkmalsausprägung(en) aus Häufigkeitstabelle als Beschriftung
3.) Werte der Häufigkeiten der Merkmalsausprägungen aus Häufigkeitstabelle ablesen
4.) Säule, Stab, Balken in Koordinatenkreuz eintragen

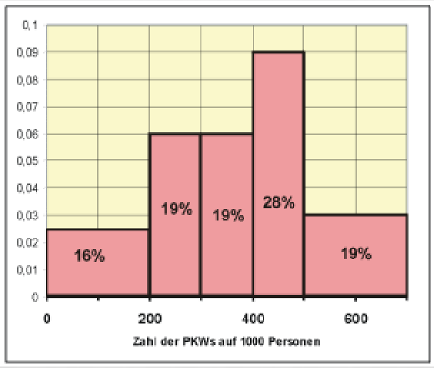
Algorithmus Kreisdiagramm

1.) Kreis zeichnen
2.) Prozentuale (relative Häufigkeit) Anteile des Kreises einzeichnen
3.) Merkmalsausprägung und relative Häufigkeit außen am Tortenstück eintragen
4.) Ggf. farbliche oder verschiedene Schraffur der Tortenstücke

Algorithmus Histogramm

1.) Häufigkeitstabelle um Spalte für Klassenbreite und Rechteckhöhe ergänzen
2.) Klassenbreite b_i berechnen
3.) Rechteckhöhe $r_i = h_i/b_i$ berechnen
4.) Von links nach rechts (oben nach unten in der Tabelle) die Klassen einzeichnen
   a. Breite des Rechtecks ist Breite der Klasse
   b. Höhe des Rechtecks entspricht Klassenhäufigkeit

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



Modul 4 – Lageparameter

Modus (Modalwert) $\bar{x}_D$

- ➔ Die am häufigsten auftretende Merkmalsausprägung (maximale Häufigkeit)

Modus bei klassierte Daten

- ➔ Mitte der Klasse mit den größten Häufigkeiten
- ➔ Klasse ist die Modalklasse

Median (Zentralwert) $\bar{x}_Z$

- ➔ Exakte Mitte ➔ 50 % der Werte links und 50 % der Werte rechts vom Median

Für ordinale und metrische Merkmale definiert als:

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

Mit n für die Anzahl an Merkmalsausprägungen und x für die Merkmalsausprägungen.

Algorithmus

1.) Daten in Reihenfolge sortieren
2.) Anzahl der Merkmalsausprägungen n ermitteln
3.) Mitte der Reihenfolge bestimmen
⇨ Median

HINWEIS: Bei ordinal skalierten Merkmalen (z.B. Schulnoten) existiert bei gerader Anzahl an Merkmalsausprägungen nur dann der Median, wenn die beiden infrage kommenden Merkmalsausprägungen gleich sind.

Median bei klassierten Daten

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

- x_k : obere Klassengrenze der Einfallsklasse
- x_{k-1} : untere Klassengrenze der Einfallsklasse
- F_{k-1} : relative Summenhäufigkeit der Klasse unterhalb (in Tabelle vor!) der Einfallsklasse
- f_k : relative Häufigkeit der Einfallsklasse

Die **Einfallsklasse** k ist jene Klasse, bei der die relative Summenhäufigkeit $F(x) = 50\ \%$ den Wert 0,5 bzw. 50 % erreicht bzw. überschreitet.

HINWEIS: Bei klassierten Daten ist Median nicht exakt bestimmbar → Näherungswert für Median

Arithmetisches Mittel (Mittelwert, Durchschnitt) $\bar{x}$

Sinnvoll für beliebige metrische Merkmale (Absolut-/Ratioskala, Intervallskala, Verhältnisskala).

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

- $x_i \coloneqq x_1, \ldots, x_n$: Merkmalsausprägungen
- n : Anzahl der Merkmalsausprägungen

Eigenschaften

- Summe der Abweichungen der Einzelwerte stehts 0 → $\sum(x_i - \bar{x}) = 0$
- Bekanntester Mittelwert
- Nur für quantitative Merkmale sinnvoll
- Empfindlich gegen Ausreißer (schiefe Verteilungen!)

Arithmetisches Mittel aus Häufigkeitstabellen

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

- n : Anzahl der Merkmalsausprägungen
- $x_1, \ldots, x_j$: Merkmalsausprägungen
- $h(x_1), \ldots, h(x_j)$: absolute Häufigkeit der Merkmalsausprägungen
- $f(x_1), \ldots, f(x_j)$: relative Häufigkeiten der Merkmalsausprägungen

Arithmetisches Mittel bei klassierten Daten

HINWEIS: Der arithmetische Mittelwert ist bei klassierten Daten ein Näherungswert, da Verteilung in den Klassen unbekannt. Es wird behelfsmäßig die Klassenmitte als Mittelwert für die Klasse angenommen und zur weiteren Berechnung genutzt.

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

- n : Anzahl der Merkmalsausprägungen (Klassenanzahl)
- $m_1, \ldots, m_k$: Klassenmitten (Mitte der jeweiligen Klasse)
- $h_1, \ldots, h_k$: absolute Klassenhäufigkeit
- $f_1, \ldots, f_k$: relative Klassenhäufigkeit

Median vs. Arithmetischen Mittelwert

- Median unempfindlich gegen Ausreißer
- Arth. Mittel empfindlich gegen Ausreißer
- ⇨ Median häufig der bessere Mittelwert, da Verteilung besser beschrieben!

Faustformeln für Neigung/Schiefe

Wenn gilt:

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

- ⇨ Rechtsschiefe (linkssteile) Häufigkeitsverteilung

Wenn gilt:

Modus > Median > arith. Mittel → $\bar{x}_D > \bar{x}_Z > \bar{x}$

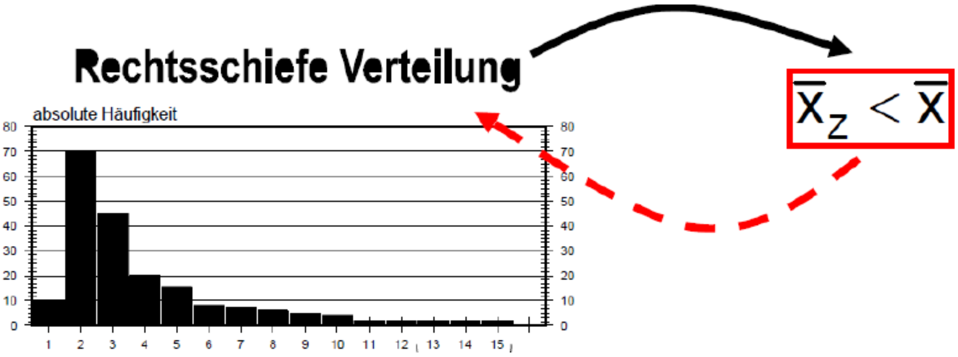
- ⇨ Linksschiefe (rechtssteile) Häufigkeitsverteilung

Wenn gilt:

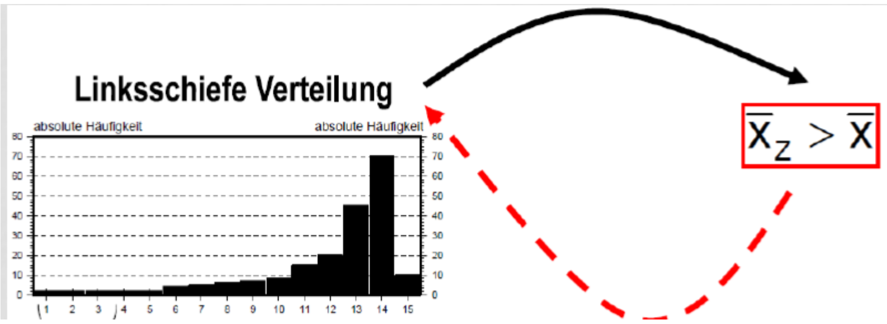
Modus ungefähr gleich Median ungefähr gleich arith. Mittel → $\bar{x}_D \approx \bar{x}_Z \approx \bar{x}$

- ⇨ Unimodale symmetrische Häufigkeitsverteilung

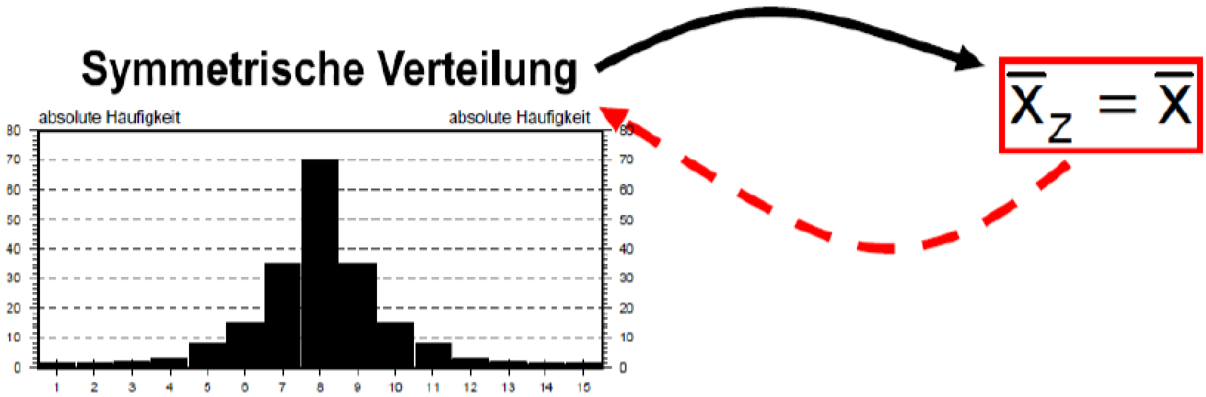
Rechtsschiefs (linkssteile) Verteilung



Linksschiefe (rechtssteile) Verteilung



Symmetrische Verteilung



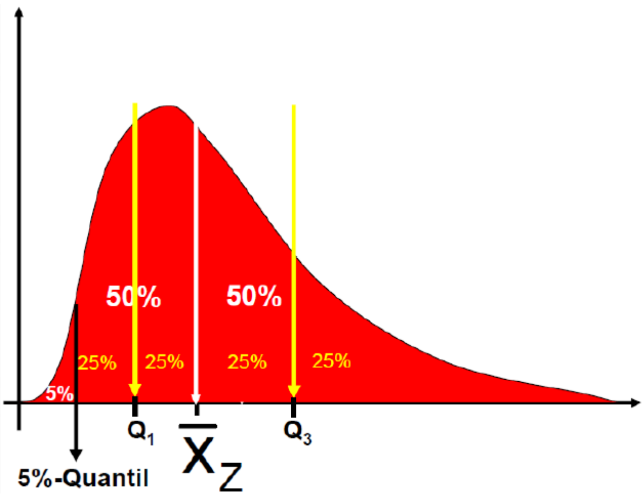
Quantile

- Ist Lagemaß
- Teilt Verteilungen in gleichgroße Abschnitte mit gleicher Häufigkeit

Spezielle Quantile

Benennung der Quantile $\tilde{x}_p$	Anzahl der Intervalle	
Terzile	3	1/3
Quartile	4	1/4
Quintile	5	1/5
Dezile	10	1/10
Vigintile	20	1/20
Perzentile (Zentile)	100	1/100

Quantilseinteilung einer Häufigkeit (mit oberen und unteren Quartil)



Modul 5 – Streuungsparameter

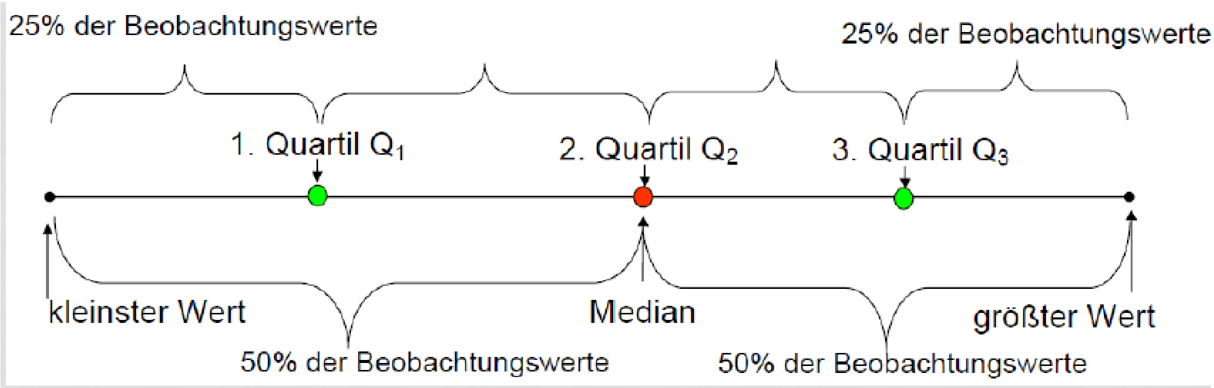
Streuungsparameter vs. Lageparameter

- Lageparameter geben keine Information über Streuung der Daten
- Arith. Mittel und Median verdecken oft große Ungleichheiten in Daten.

Eigenschaften von Streuungsparameter

- Hilft bei Ermittlung des Bezugspunkts (Lageparameter), um den Daten streuen
- Alle Beobachtungswerte werden berücksichtigt
- Streuung = 0 ➔ alle Werte sind gleich ➔ Streuungsparameter = 0
- Streuungsparameter proportional zur Streuung
  → je größer die Streuung, desto größer der Streuungsparameter
- Streuungsparameter unabhängig von Anzahl der Beobachtungswerte n

Quartil mit 5-Punkt-Zusammenfassung



- **Punkt 1**: *Minimum* → kleinster Wert
- **Punkt 2**: *1. Quartil* Q_1 → Mitte der unteren 50 % der Beobachtungswerte (25 %)
- **Punkt 3**: *2. Quartil* Q_2 = Median → Markiert die Mitte der Beobachtungswerte
- **Punkt 4**: 3. Quartil Q_3 → Mitte der oberen 50 % der Beobachtungswerte (75 %)
- **Punkt 5**: *Maximum* → größter Wert

⇨ Der Quartilsabstand wird als **Streuungsmaß** verwendet.

Berechnung des (Inter-)Quartilsabstands (interquartile range, IQR)

$$Q_A = IQR = Q_3 - Q_1 = 50\ \%$$

⇨ 50 % der Beobachtungswerte befinden sich zwischen Q_1 und Q_3.

Berechnung der Quartils

Q_1 *bzw. 25 % Quartil*

$$Q_{0,25} = \begin{cases} \frac{1}{2}\left(x_{(n \cdot 0,25)} + x_{(n \cdot 0,25+1)}\right) & \textbf{wenn } n \cdot 0,25 \textbf{ ganzzahlig} \\ x_{(\lceil n \cdot 0,25 \rceil)} & \textbf{wenn } n \cdot 0,25 \textbf{ nicht ganzzahlig} \end{cases}$$

Q_3 *bzw. 75 % Quartil*

$$Q_{0,75} = \begin{cases} \frac{1}{2}\left(x_{(n \cdot 0,75)} + x_{(n \cdot 0,75+1)}\right) & \textbf{wenn } n \cdot 0,75 \textbf{ ganzzahlig} \\ x_{(\lceil n \cdot 0,75 \rceil)} & \textbf{wenn } n \cdot 0,75 \textbf{ nicht ganzzahlig} \end{cases}$$

Spannweite (Variationsbreite) w

- Gibt Information über Ausdehnung der Werte
  → Maß für Breite des Streubereichs einer Häufigkeit
- Ordinale und metrische Merkmale

Berechnung über

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

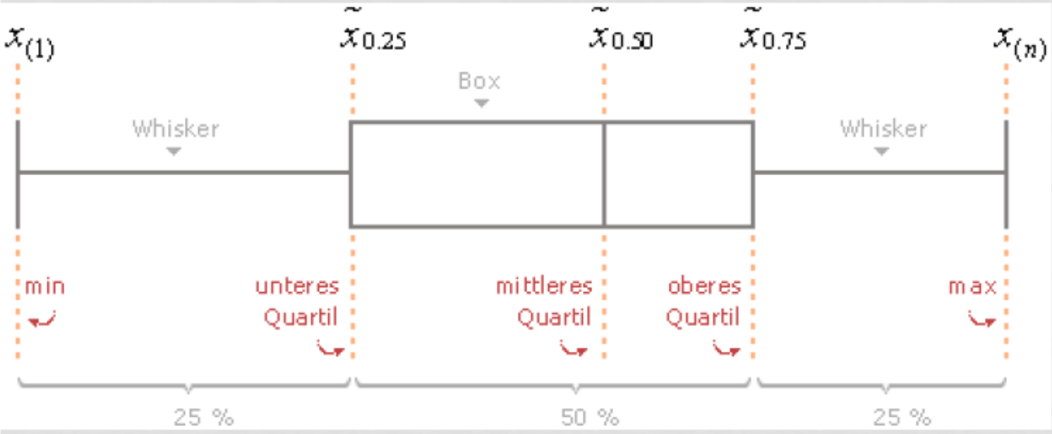
- x_{max} : Größte Merkmalsausprägung
- x_{min} : Kleinste Merkmalsausprägung

Quartilsabstand vs. Spannweite

- Ausreißer:
  - Quartilsabstand: unabhängig von Ausreißern
  - Spannweite: anfällig für Ausreißer
- Repräsentierte Merkmalsausprägungen
  - Quartilsabstand: Breite des mittleren Bereichs, in dem 50 % der Werte liegen
  - Spannweite: Breite des Gesamtbereichs

Boxplot (Box-and-Whisker-Plot)

- Ist grafische Darstellung der 5-Punkte-Zusammenfassung
- Liefert Informationen zu:
  - Lokalisation (Lage des Median)
  - Streuungsmaße:
    - Spannweite → Ausdehnung des gesamten Boxplots
    - Quartilsabstand → Ausdehnung der Box
  - Schiefe der Verteilung → Fläche der Boxhälften mit Median als Mitte



Algorithmus zum Boxplot

1.) Minimum und Maximum bestimmen
2.) Quartile $Q_1 = Q_{0,25} = Q_{25\,\%}, Q_2 = Q_{0,5} = Median\ und\ Q_3 = Q_{0,75} = Q_{75\,\%}$ berechnen (siehe oben)
3.) Werte in Koordinatensystem einzeichnen
4.) Box einzeichnen (s.o.)

Varianz s^2

- Ist das arithm. Mittel der Abweichungsquadrate
- Hängt eng mit Mittelwert und Standardabweichung zusammen

Eigenschaften

- Wichtiger Streuungsparameter
- Setzt metrische Merkmale voraus
- Grundwert für Standardabweichung und Variationskoeffizienten

Berechnung anhand der Merkmalsausprägungen

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

- n : Anzahl der Ausprägungsmerkmale
- x_i : Merkmalsausprägungen
- $\bar{x}$: arithmetischer Mittelwert

Berechnung der Varianz anhand von Häufigkeitstabellen

Formel (1):

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

Formel (2):

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

- n : Anzahl der Ausprägungsmerkmale
- x_i : Anzahl der Ausprägungsmerkmale
- $\bar{x}$: Anzahl der Ausprägungsmerkmale
- $h(x_i)$: Anzahl der Ausprägungsmerkmale
- $f(x_i)$: Anzahl der Ausprägungsmerkmale

Varianz aus Häufigkeitstabellen mit klassierten Daten

… mit absoluten Häufigkeiten

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

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

- n : absolute Summenhäufigkeit / absolute kumulierte Häufigkeit
- h_i : absolute Häufigkeit der i-ten Klasse
- k : Anzahl der Klassen
- m_i : Klassenmitten der i-ten Klasse

… mit relativen Häufigkeiten

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

- k : Anzahl der Klassen
- f_i : relative Häufigkeit der i-ten Klasse
- m_i : Klassenmitte der i-ten Klasse

Standardabweichung s

- Maß wie hoch Aussagekraft des Mittelwertes ist.
  - Kleine Standardabweichung
    → alle Beobachtungswerte nahe am Mittelwert gestreut → kleine Streuung
  - Große Standardabweichung
    → alle Beobachtungswerte weit um Mittelwert gestreut → große Streuung
- Bei normalverteilten Daten liegen 95 % der Beobachtungswerte im Intervall $[\bar{x} - 2s,\ \bar{x} + 2s]$
  - Arthm. Mittel +/- 2x Standardabweichung

Berechnung

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

- s^2 : Varianz
- n : Anzahl der Merkmalsausprägungen
- x_i : i-te Merkmalsausprägung
- $\bar{x}$: arithmetisches Mittel

Variationskoeffizient v (Abweichungskoeffizient)

- Relatives Streuungsmaß

Berechnung

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

- s : Standardabweichung
- $\bar{x}$: arithmetisches Mittel

Modul 6 – Korrelation und Regression

Methoden der Zusammenhangsanalysen bei nicht metrischen Merkmalen

- Rangkorrelationskoeffizient
    - Maß für Stärke des Zusammenhangs zwischen ordinalskalierte Merkmale
    - Spezialfall von Pearsons Korrelationskoefizent
    - Keine Annahmen über lineare Beziehung zwischen Variablen notwendig
    - Robust gegen Ausreißer
- Kontigenzkoeffizient
    - Maß für Stärke des Zusammenhangs nominaler oder ordinaler Merkmale
    - Auf beliebig große Kreuztabellen anwendbar
    - Für Kontigenzkoeffizient C gilt $0 \leq C \leq 1$
- Phi-Koeffizient (Vierfelder-Korrelationskoeffizient)
    - Maß für Stärke des Zusammenhangs zweier dichotomer Merkmale
    - Basiert auf Kontigenztabelle mit gemeinsamer Häufigkeitsverteilung der Merkmale

Streudiagramm (scatter plot)

- Grafische Darstellung der beobachteten Wertepaare der zwei Merkmale in kartesisches Koordinatensystem
- Aus Lage und Form der Punkte lassen sich Stärke und Richtung des Zusammenhangs ablesen
- Liefert erste Hinweise über mögliche Abhängigkeiten

Algorithmus

1.) (optional) Tabelle mit Wertepaare aufstellen
2.) Wertepaare in Koordinatensystem eintragen
3.) Richtung des Zusammenhangs ablesen (in welche Richtung zeigen die Werte von links nach rechts?)
4.) Stärke des Zusammenhangs ablesen (wie dicht liegen die Punkte zueinander?)
5.) (optional) Regressionsgerade einzeichnen
    a. $x_0 = 0, x_{max} = \max(x)$ [höchster Wert der x-Achse einsetzen]
    b. Punkte x_0 und x_{max} verbinden.

Korrelation

- Zahlenmäßiger statistischer Zusammenhang zwischen zwei Merkmalen
- Unterscheidung in positive und negative Korrelation
    - Positiv: gleichförmige Entwicklung der Merkmale
      → hohe Werte bei X ➔ hohe Werte bei Y
    - Negativ: gegenläufige entwicklung
      → hohe Werte bei X ➔ niedrige Werte bei Y
- Korrelation sagt nichts über kausalen Zusammenhang aus
- Korrelation sagt nichts über Kausalitätsrichtung aus
- ➔ Korrelation ➔ es folgt entweder Kausalität oder Scheinkorrelation

Kausaler Zusammenhang

- Ergibt sich aus Zusammenhang zwischen den Merkmalen X, Y
- Ursache-Wirkungs-Beziehung zwischen X, Y
  → Veränderungen an Merkmal Y können eindeutig auf Veränderung von X zurückgeführt werden.
- Identifikation über Abhängigkeitsanalyse
- Korrelation ist notwendige, aber nicht hinreichende Voraussetzung für kausalen Zusammenhang
- Quantifizierung erfolgt über Korrelationskoeffizenten.

Scheinkorrelation

- Korrelation (statistischer Zusammenhang) besteht, es liegt jedoch kein kausaler Zusammenhang vor!
- Entscheidung anhand des „gesunden Menschenverstands“

Korrelationskoeffizient r_{XY}

- Für Korrelationskoeffizent gilt $-1 \leq r_{XY} \leq +1$
- Statistische Kennzahl
- Gibt Auskunft über:
  - <u>Stärke</u> des linearen Zusammenhangs
  - <u>Richtung</u> des linearen Zusammenhangs
- Dimensionslose Größe
- ➔ Maß für den Grad des linearen Zusammenhangs

Schlussfolgerung anhand des Korrelationskoeffizentwertes

Wert	Folgerung	Erklärung
$\boldsymbol{r > 0}$	positiver Zusammenhang	Bei beiden Variablen treten hohe Werte gemeinsam auf.
$\boldsymbol{r < 0}$	Negativer Zusammenhang	Hohe werte bei der einen Variablen treten gemeinsam mit niedrigen Werten bei der anderen auf.
$\boldsymbol{r = -1}$	Extrem stark negativer Zusammenhang	Punktwolke liegt auf einer Geraden mit <u>negativer</u> Steigerung.
$\boldsymbol{r = +1}$	Extrem stark positiver Zusammenhang	Punktwolke liegt auf einer Geraden mit <u>positiver</u> Steigerung.
$\boldsymbol{r = 0}$	Kein linearer Zusammenhang	-

Korrelationsrechnung (nach Pearson)

Für zwei mindestens intervallskalierten Merkmalen X, Y mit jeweils positiver Standardabweichung s_X, s_Y und Kovarianz $COV(X, Y)$ ist Korrelationskoeffizent (nach Pearsonscher Maßkorrelationskoeffizient) definiert durch:

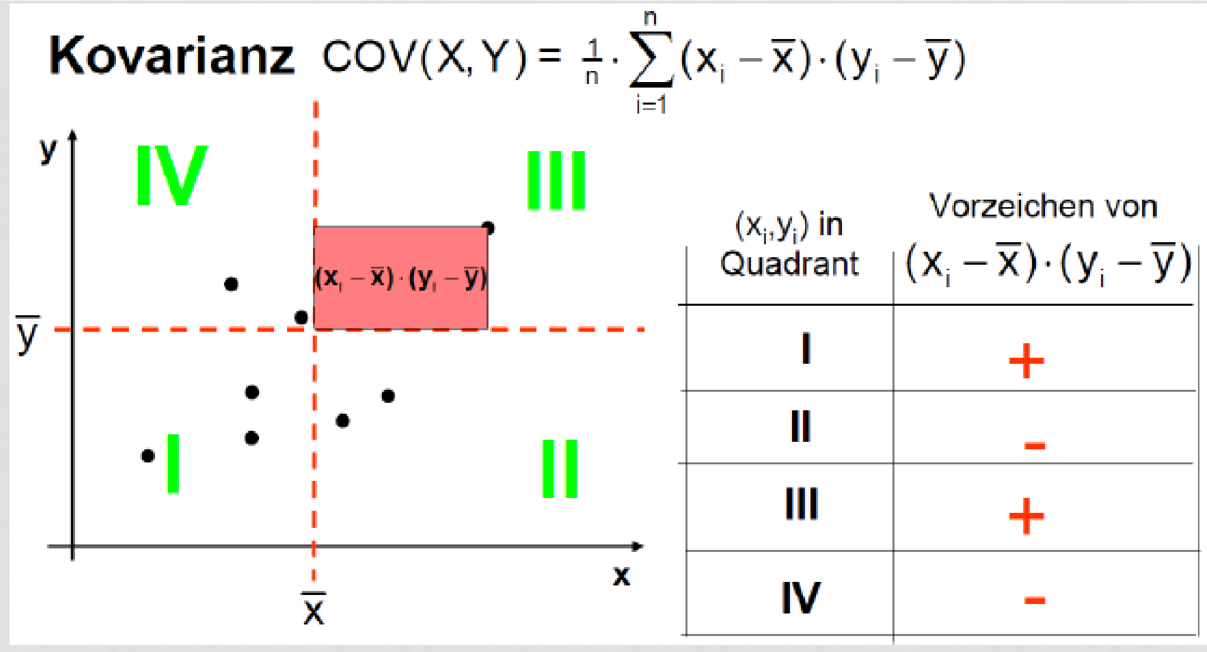
$$r_{XY} = \frac{COV(X,Y)}{s_X \cdot s_Y}$$

Es muss also gelten:

- $s_X > 0$
- $s_Y > 0$
- $COV(X, Y) > 0$

Kovarianz

- Informiert über gemeinsame Variabilität von X, Y.
- Positiver Zusammenhang → positive Kovarianz
- Negativer Zusammenhang → negative Kovarianz
- Kein linearer Zusammenhang → Kovarianz $COV(X, Y) \approx 0$



Standardisierung der Kovarianz

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

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

Beispiel

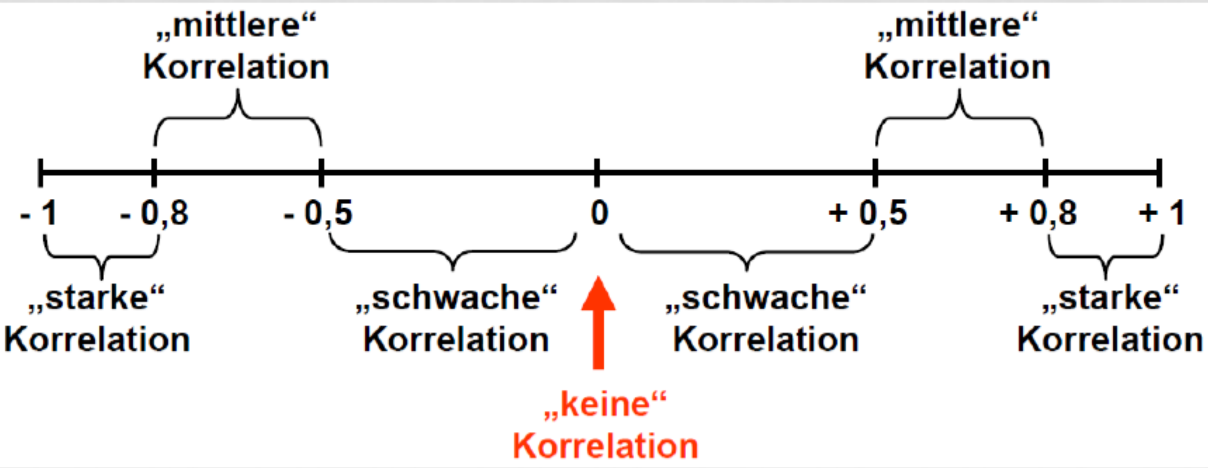
Filiale Nr. i	x_i Verkaufsfläche (1000qm)	y_i Filialumsatz (Mio €)	$x_i \cdot y_i$	x_i^2	y_i^2
1	3	30	90	9	900
2	2	10	20	4	100
3	6	40	240	36	1.600
4	5	20	100	25	400
Summe	16	100	450	74	3.000

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

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

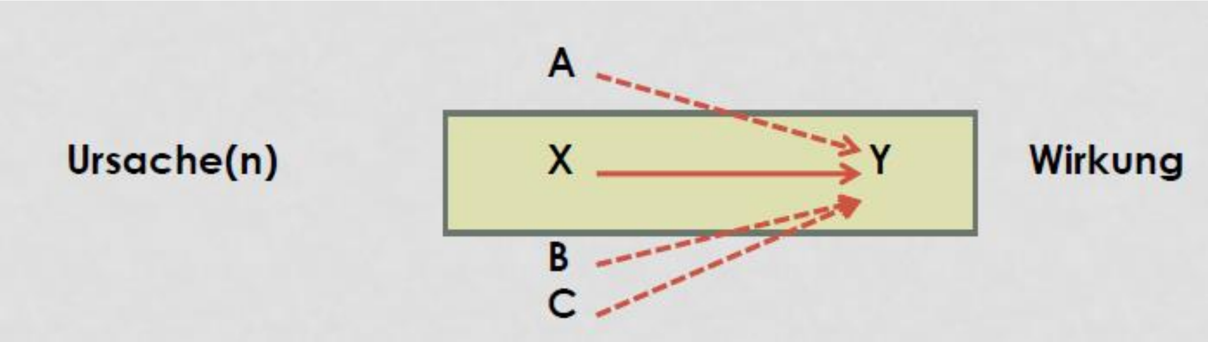
Interpretation des Korrelationskoeffizenten

Als **Orientierung** gilt:



Regressionsanalyse

Problem der Regressionsanalayse ist: „Es gibt meist nicht nur einen Einflussfaktor“ (Probleme sind selten monokausal …)



Unterscheidung in:

- *Einfache Regressionsanalyse*
  - Zwei metrische Merkmale X, Y wobei X die Einflussgröße ist und Y die Zielgröße.
  - Gerade wird mithilfe von zwei Parametern durch Punktwolke gelegt, sodass linearer Zusammenhang zwischen X, Y möglichst gut beschrieben wird.
- *Multiple Regressionsanalyse*
  - Verallgemeinerung der einfachen linearen Regression mit k Regressionen, welche abhängige Variable erklären sollen.
  - Mehrere metrische Größen $X_1, \dots, X_n$ als Einfussgrößen, Y als Zielgröße.

Regressionsfunktion $\hat{y}$

- Bestimmung erfolgt anhand der Methode der kleinsten Quadrate
- Regressionskoeffizienten a, b (Kurvenparameter) werden so bestimmt, dass Summe der quadratischen Abweichungen der Kurve von beobachteten Punkten minimal sind.

Grundformel

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

Methode der kleinsten Quadrate (OLS = ordinary least squares)

$$\boldsymbol{OLS(a,b)} = \sum_{i=1}^{n} \overbrace{(y_i - \hat{\boldsymbol{y}}_i)}^{\text{Residuen } u_i}{}^2 = \sum_{i=1}^{n} (y_i - (a + b * \boldsymbol{x}_i))^2 \rightarrow \boldsymbol{min}$$

Methode der kleinsten Quadrate (OLS = ordinary least squares)

Voraussetzungen

- X, Y sind quantitatives (metrisches) Merkmal
- Es existiert ein Zusammenhang $X \rightarrow Y$

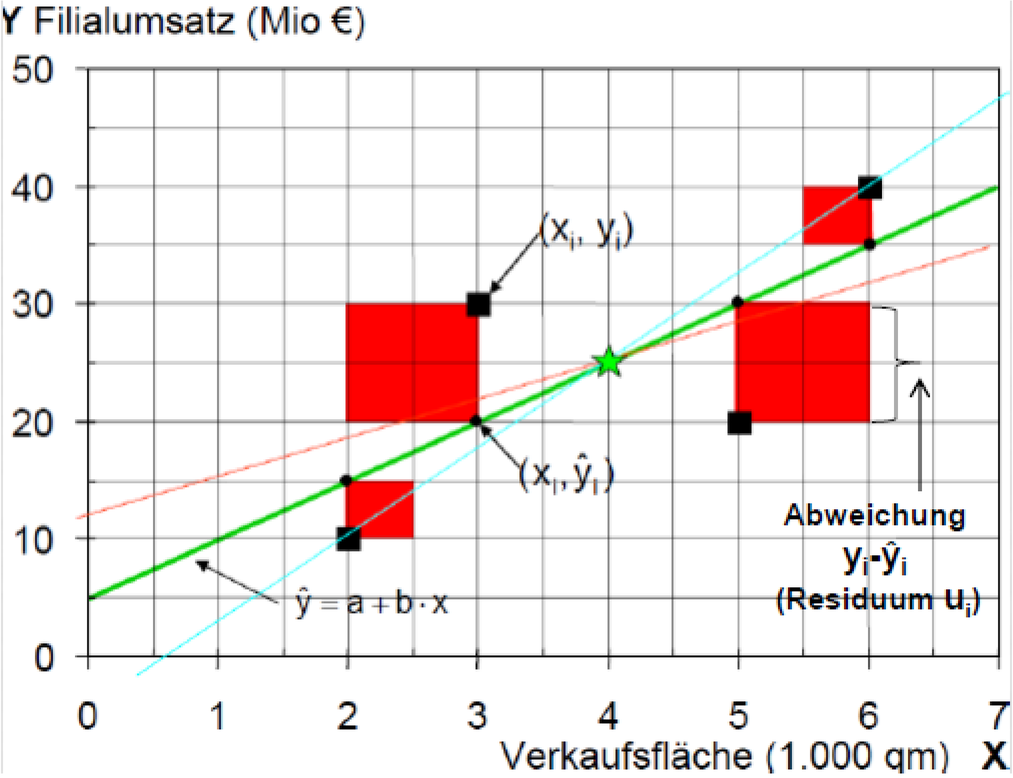
Algorithmus

1.) Sinnhaftigkeit der Abhängigkeitsanalyse überprüfen
   a. Falls nicht sinnvoll → keine Regressionsfunktion ermitteln
2.) Daten erheben für X, Y
3.) Wertepaare $(x_1, y_1), \ldots (x_n, y_n)$ für X, Y bilden
4.) Streudiagramm erstellen (qualitative Abhängigkeitsanalyse)
5.) (optional) Auswahl eines Funktionstyps (im Modul auf lineare Funktionen beschränkt!)
6.) Berechnung der Regressionsfunktion (nach Methode der kleinsten Quadrate)

Algorithmus Prognosewerte berechnen

1.) Ermittelte Regressionsfunktion $\hat{y} = a + bx$
2.) Wert des zu prognostizierenden Merkmals X in errechnete Regressionsfunktion einsetzen
   $\hat{y}(x) = a + bx$

Beispiel



Regressionsrechnung anhand von Berechnungstabelle mit Hilfsgrößen

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

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

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

Hilfsgrößen der Berechnungstabelle

- Merkmale und deren Werte sind gegeben

Vorgehen zur Berechnung der Hilfsgrößen

1.) Spalte für $x_i \cdot y_i$ anlegen und zeilenweise berechnen
2.) Spalte für x_i^2 anlegen und zeilenweise berechnen
3.) Spalte für y_i^2 anlegen und zeilenweise berechnen
4.) Spalte für Ergebnisse der Regressionsfunktion $\widehat{y}_i = a + bx_i$
5.) Spalte für Berechnung der Residualwerte $u_i = (y_i - \widehat{y}_i)$
6.) Summenzeile hinzufügen und je Spalte die Summe berechnen

Objekte	x_i	y_i	$x_i \cdot y_i$	x_i^2	y_i^2	$\widehat{y}_i$	u_i
1	...	...	...	...	...	...	...
2	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...
N	...	...	...	...	...	...	...
Summe	...	...	...	...	...	...	...

Interpretation der Ergebnisse

- Faktor b ist die Steigerung der Regressionsgeraden
- Variable a kann als Schwellwert verstanden werden, der den Schnittpunkt zur y-Achse markiert.

Anwendungsgebiete der Regressionsanalyse

- Vorhersagemodell erstellen
  - $\widehat{y}_i$ sind dann Prognosewerte
- Stärke des Zusammenhangs analysieren und quantifizieren
  - Zeigt an, welche x_i gar keinen Zusammenhang mit y haben oder auch jene x_i die redundanten Informationen über y enthalten.

Problemstellung bei der Regression: wie gut passen Modell und Realität zusammen?

- Empirische Werte (Werte der Tabelle) sind Werte aus der Realität
- Regressionsfunktion ist ein theoretisches Modell zu den empirischen Werten

Relevante Fragestellungen:

- Wie gut beschreibt das Modell (die Regressionsfunktion) die Realität?
- Wie gut wird die Realität durch das Modell wiedergegeben?
- Wie gut ist die Anpassungsgüte des modells?
- Wie gut beschreibt die Regressionsfunktion die Abhängigkeit?
  → Wie hoch ist die Erklärungskraft des Modells?
- Wie gut wird die Unterschiede zwischen den Merkmalen erklärt?
  → Wie viel Varianz wird durch das Modell nicht erklärt?

⇨ Es wird ein Gütemaß für die Schätzung der Parameter benötigt!

Gütemaße für die Regression

Abweichungszerlegung

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

- $\bar{y}$: arithmetisches Mittel für y_i's
- y_i : i-ter Wert des Merkmals Y
- $\hat{y}_i$: Regressionsfunktionswert/Prognosewert für y_i

Komponenten der Abweichungszerlegung

- $(y_i - \bar{y})$: Gesamtabweichung der Beobachtung y_i zum Mittelwert $\bar{y}$
  - Diesen Fehler würden wir machen, wenn wir mit dem Mittelwert die entsprechende Beobachtung vorhersagen würden
- $(y_i - \hat{y}_i)$: Unerklärte Abweichung / Residuum der Beobachtung y_i zur Regressionsgeraden (Residuum u_i)
  - Diesen Teil der Abweichung können wir auch durch Hinzunahme der unabhängigen Variablen x nicht vermeiden.
- $(\hat{y}_i - \bar{y})$: Erklärte Abweichung der Regressionsgerade zum Mittelwert $\bar{y}$
  - Diesen Fehler können wir durch Hinzunahme der unabhängigen Variablen x vermeiden.

Varianz der Regressionswerte

Varianzzerlegung

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

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

- n : Anzahl der Merkmalsausprägungen des Merkmals Y
- y_i : i-te Merkmalsausprägung des Merkmals Y
- $\bar{y}$: arithmetisches Mittel der Merkmalsausprägungen des Merkmals Y
- u_i : Residuum der i-ten Merkmalsausprägung y_i zur Regressionsgerade
- $\bar{u}$: arithmetisches Mittel der Residuen
- $\hat{y}_i$: i-te Regressionsfunktionsergebnisse / Prognosewerte
- $\bar{\hat{y}}$: arithmetisches Mittel der Regressionsfunktionsergebnisse / Prognosewerte

- s_y^2 : Varianz der Beobachtungswerte / Merkmalsausprägungen von Y
- s_u^2 : Varianz der Residuen
- $s_{\hat{y}}^2$: Varianz der Regressionsfunktionsergebnisse / Prognosewerte

Alternative Formeln zu den Varianzen

Varianz der Merkmalsausprägungen Y

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

- n : Anzahl der Merkmalsausprägungen (Beobachtungswerte) von Y
- y_i : i-te Merkmalsausprägung (Beobachtungswert) von Y
- $\bar{y}$: arithmetisches Mittel von Y

Varianz der Merkmalsausprägungen X

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

- n : Anzahl der Merkmalsausprägungen (Beobachtungswerte) von X
- x_i : i-te Merkmalsausprägung (Beobachtungswert) von X
- $\bar{x}$: arithmetisches Mittel von X

Varianz der Regressionsfunktionsergebnisse / Prognosewerte $\hat{Y}$

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

- n : Anzahl der Regressionsfunktionsergebnisse (Prognosewerte)
- $\hat{y_i}$: i-tes Regressionsfunktionsergebnis (i-ter Prognosewert)
- $\bar{\hat{y}}$: arithmetisches Mittel der Regressionsfunktionsergebnisse (Prognosewerte)

Varianz der Residuen U

$$s_U^2 = \left(\frac{1}{n} \cdot \sum_{i=1}^{n} u_i^2\right) - \bar{u}^2$$

- n : Anzahl der Residuen (Beobachtungsobjekte)
- u_i : i-tes Residuum
- $\bar{u}$: arithmetisches Mittel der Residuen

Varianz des unabhängigen Merkmals X

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

- $s_{\hat{y}}^2$: Varianz der Regressionsfunktionsergebnisse / Prognosewerte
- b^2 : Quadrierung des Regressionsfunktionsparameters b
- s_x^2 : Varianz des unabhängigen Merkmals X

Bestimmtheitsmaß der linearen Regression R^2

- Bestimmtheitsmaß bemisst die Erklärungskraft des Modells
- Ist Gütemaß der linearen Regression
- Gibt an, wie gut die unabhängige Variable Y geeignet ist, die Varianz der abhängigen Variablen X zu erklären.
- Nutzt Konzept der Varianzzerlegung → Varianz des abhängigen Merkmals kann in erklärte und unerklärte Varianz (Residualvarianz) zerlegt werden
- Anteil der Varianz der abhängigen Variablen, der sich durch die Varianz der unabhängigen Variablen erklären lässt.

Wertebereich

(unbrauchbares Modell) **0%** $\leq$ **R²** $\leq$ **100%** (perfekte Modellanpassung)

- Höhe des Wertes abhängig vom Anwendungsgebiet
  - Wenn menschliches Verhalten erklärt werden soll (z.B. Marketing) ist R^2 meist gering
  - In der Physik meist ein sehr höher Wert wegen exakt messbarer Größen
  - Auf Mikroebene gilt oft 10 % als gut.
  - Bei aggregierten Daten 40 % - 80 %

Formel (1)

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

- $\hat{y}_i$: i-ter Prognosewert (Regressionsfunktionsergebnis)
- $\bar{y}$: arithmetisches Mittel der Beobachtungswerte des Merkmals Y
- y_i : i-ter Beobachtungswert (Merkmalsausprägung) des Merkmals Y
- u_i : i-tes Residuum

Formel (2) – Verhältnis zwischen Streuung Prognosewerte und Gesamtstreuung von y_i

$$R^2 = \frac{{s_{\hat{y}}}^2}{{s_y}^2} \Leftrightarrow 1 - \frac{{s_u}^2}{{s_y}^2}$$

- $s_{\hat{y}}^2$: Varianz der Prognosewerte / Regressionsfunktionsergebnisse
- s_y^2 : Varianz der Beobachtungswerte des Merkmals Y
- s_u^2 : Varianz der Residuen

Formel (3) – nur bei einfacher linearer Regression (nur eine unabhängige Variable)

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

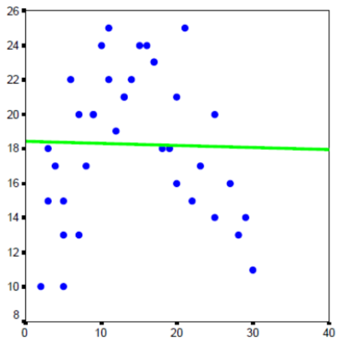
- r_{XY} : Korrelationskoeffizent nach Pearson

Probleme bei linearer Regression und Korrelation

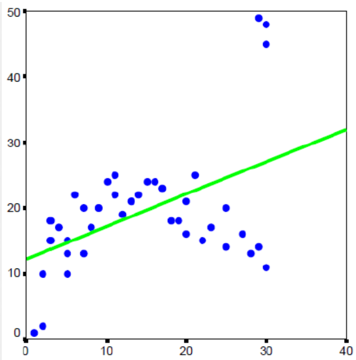
- Es werden nur lineare Zusammenhänge erfasst, nicht lineare Zusammenhänge (z. B. wenn Punktwolke eine Hyperbel anzeigt) können so nicht beschrieben werden.
- Ausreißer (Extremwerte) üben starken Einfluss aus
  - Wegen Quadrierung
  - Wegen Multiplikation

Beispiele

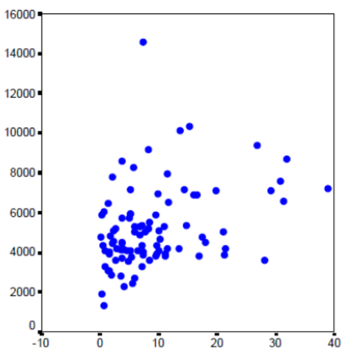
Nicht lineare Zusammenhänge



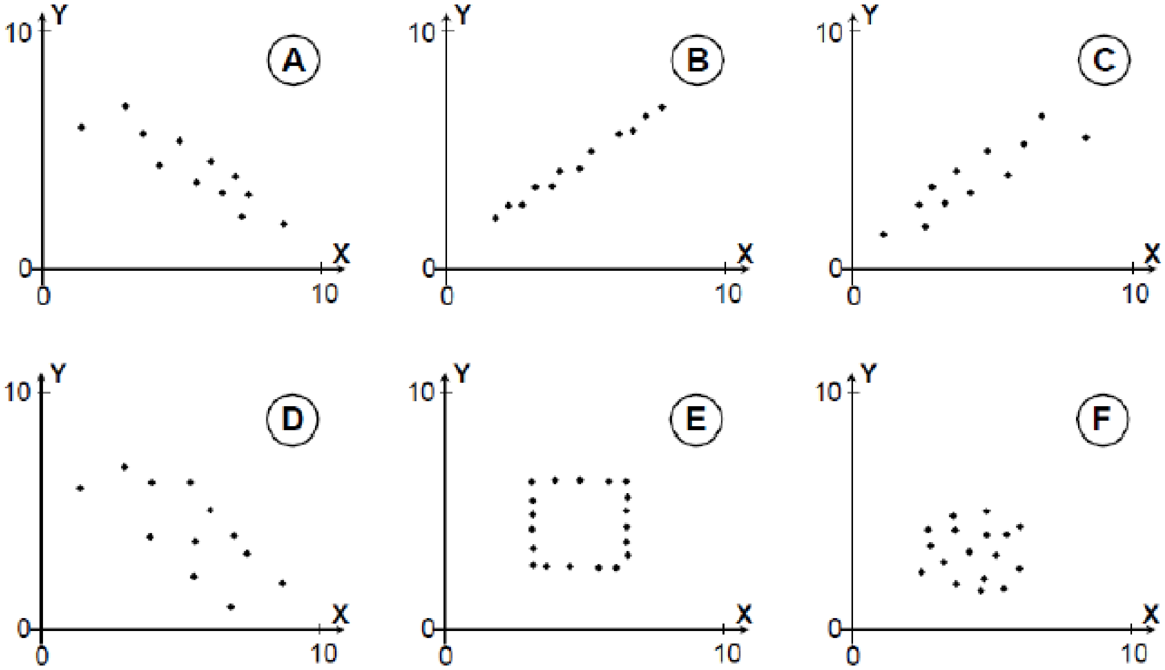
Auswirkungen der Extremwerte (Ausreißer) in der Punktwolke durch Quadrierung



Auswirkung der Extremwerte (Ausreißer) in der Punktwolke durch Multiplikation



Beispiele #2



<u>Lösung der Aufgabe 4</u>

$$-1 < r_A < r_D < r_F \approx 0 \approx r_E < r_C < r_B < +1$$

Modul 7 – Wahrscheinlichkeitsrechnung

Begriffe

Begriff	Symbol	Erläuterung
Experiment	-	Erhebung eines Merkmals an einem Merkmalsträger.
Elementarereignis	ω	Die verschiedenen möglichen Ergebnisse oder Realisationen eines Experiments. Eine einelementige Teilmenge des Ergebnisraums.
Ereignisraum (Ereignismenge)	Ω	Summe aller Elementarereignisse.
Stichprobenumfang	n	Anzahl der Wiederholungen des Experiments.
Ereignis	A	Tritt ein, wenn bei der Versuchsdurchführung ein $\omega \in A$ ergibt.
Mächtigkeit des Ereignisraums	$\|\Omega\| = n$	Anzahl der Ergebnisse des Ereignisraums.
Gegenereignis	-	Menge aller Ergebnisse, die nicht zum Ereignis gehören.
Sicheres Ereignis	Ω	Das sichere Ereignis bedeutet, dass das Ereignis immer eintritt.
Unmögliches Ereignis	$\emptyset$	Das unmögliche Ereignis tritt niemals ein.
Komplementärereignis	A^C	Gegenereignis, also das Ereignis ohne A.
Unvereinbar/disjunkt	$A \setminus B = \emptyset$	Die Ereignisse A, B sind miteinander unvereinbar und disjunkt, also haben keine Übereinstimmungen.
Wahrscheinlichkeit	$p(A)$, p_A, $Prob(A)$	Wahrscheinlichkeit wird einem Ereignis in der Wahrscheinlichkeitsrechnung zugeordnet.

Eigenschaften der Wahrscheinlichkeiten

- Wahrscheinlichkeiten in Verbindung mit den relativen Häufigkeiten, da das Ergebnis m für A bei n Wiederholungen des Zufallsexperiments per Quotient $^{m}/_{n}$ ausdrücken, was der relativen Häufigkeit $f(x)$ entspricht.
    - → empirisches Gesetz der großen Zahlen: $p(A) \coloneqq \lim_{n\to\infty} h_n(A)$
- Relative Häufigkeit $f(x)$ liegt im Bereich $0 \leq h(A) \leq 1$
  ➔ der Bereich jeder Wahrscheinlichkeit liegt im Bereich $0 \leq p(A) \leq 1$
  ➔ für die n-malige Durchführung eines Zufallsexperiments gilt für das Ereignis A: $0 \leq m \leq n$
- Wenn das Ereignis A immer im Zufallsexperiment eintritt, so wird dies als *„sicheres Ereignis"* bezeichnet und es gilt:
    - $h(A) = {}^{n}/_{n} = 1 \to p(A) = 1 = \Omega$
- Tritt das Ereignis A mit Sicherheit niemals ein, so wird dies als *„unmögliches Ereignis"* bezeichnet und es gilt:
    - $h(A) = {}^{0}/_{n} = 0 \to p(A) = 0$
- Das Gegenereignis A^C ist die Komplementärmenge zum Ereignis A, für diese gilt:
    - $A^C = 1 - A$
    - $p(A^C) = 1 - p(A)$ oder $1 = p(A) + p(A^C)$

Axiome von Kolmogorow

1.) Wahrscheinlichkeit für Eintreten eines Ereignisses A ist immer eine reele Zahl zwischen 0 und 1
   a. $0 \leq p(A) \leq 1$

2.) Das sichere Ereignis Ω hat die Wahrscheinlichkeit 1
   a. $p(A) = 1$ → A tritt mit Sicherheit ein
   b. $p(A) = 0$ → A tritt mit Sicherheit <u>nicht</u> ein
   c. $0 \leq p(A) \leq 1$ → Sonstiger Fall, wenn nicht unmögliches oder sicheres Ereignis

3.) Wahrscheinlichkeit einer Vereinigung abzählbar vieler disjunkter Ereignisse ist gleich der Summe der Wahrscheinlichkeiten der einzelnen Ereignisse
   a. → σ-Additivität
   b. → $P(A_1 \cup A_2 \cup \ldots \cup A_n) = \sum_{i=1}^{n} P(A_i)$

Grundsätzlicher Algorithmus

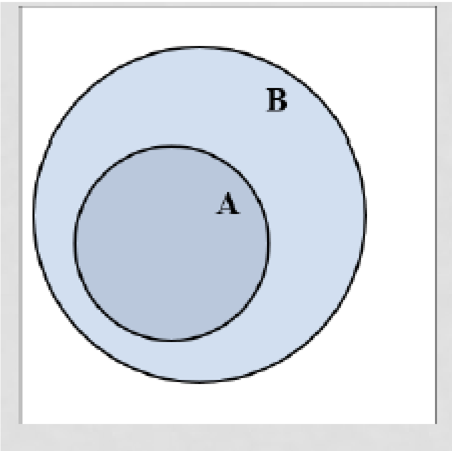
1.) Menge der möglichen Ereignisse Ω definieren
2.) Anzahl der möglichen Ereignisse $|\Omega|$ berechnen
3.) Die einzelnen Mengen der günstigen Ereignisse $A_1, \ldots, A_n$ definieren
4.) Menge des Ergebnisraums $E = P(A \cup B) = P(A) + P(B) - P(A \cap B)$ definieren
5.) Anzahl der günstigen Ereignisse des Ergebnisraums $|E|$ berechnen
6.) Wahrscheinlichkeit $p(A) = \frac{|E|}{|\Omega|}$ berechnen

Mengentheoretische Konzepte

Mengenoperationen:

∩	**Schnittmenge**
∪	**Vereinigung**
\	**Mengendifferenz**
c	**Komplementbildung**

(Echte) Teilmenge



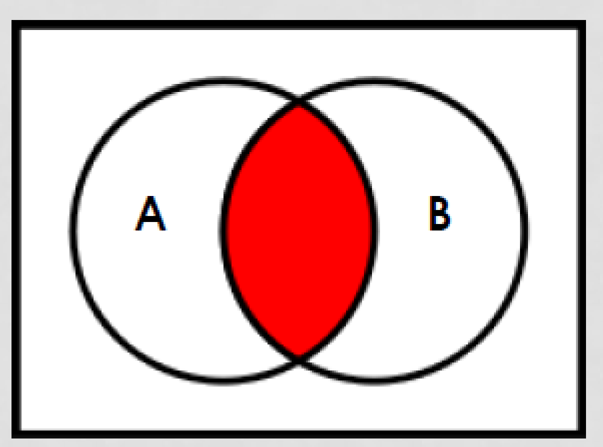
Eine Menge A heißt **Teilmenge** einer Menge B, wenn jedes Element von A auch Element von B ist.

<u>Formal</u>: $\mathbf{A \subseteq B \;:\Leftrightarrow \forall x\, (x \in A \rightarrow x \in B)}$

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

<u>Formal</u>: $\mathbf{A = B :\Leftrightarrow \forall x\, (x \in A \leftrightarrow x \in B)}$

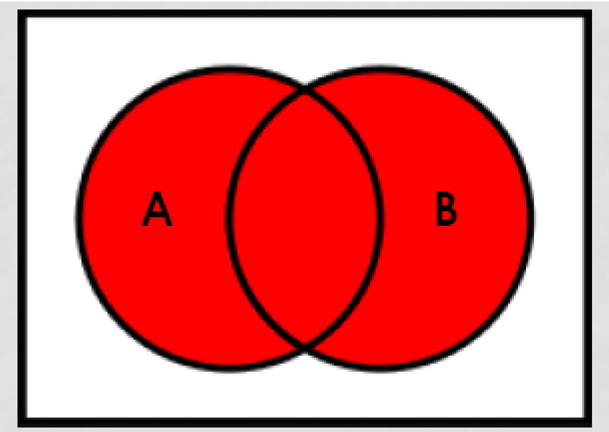
Schnittmenge



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

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

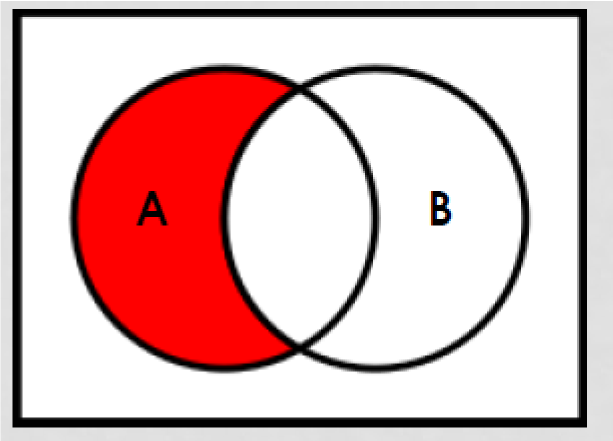
Vereinigungsmenge



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

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

Differenzmenge / Komplement



A ohne B

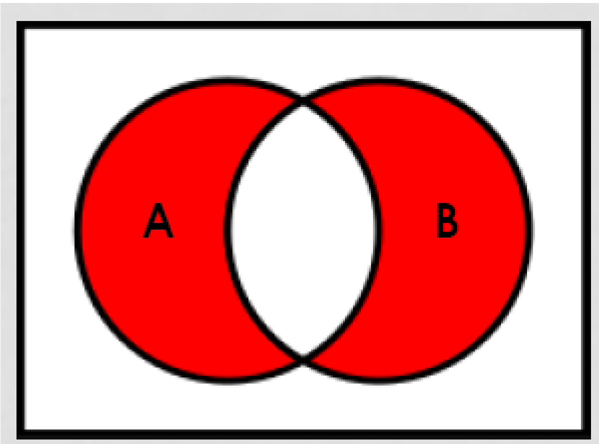
Die **Differenzmenge** (wird nur für <u>2 Mengen</u> definiert) von A und B ist die Menge der Elemente, die in A aber nicht in B enthalten sind.

<u>Formal</u>: $A \setminus B := \{ x \mid (x \in A) \land (x \notin B) \}$

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

<u>Formal</u>: $B^C := \{ x \mid x \notin B \}$

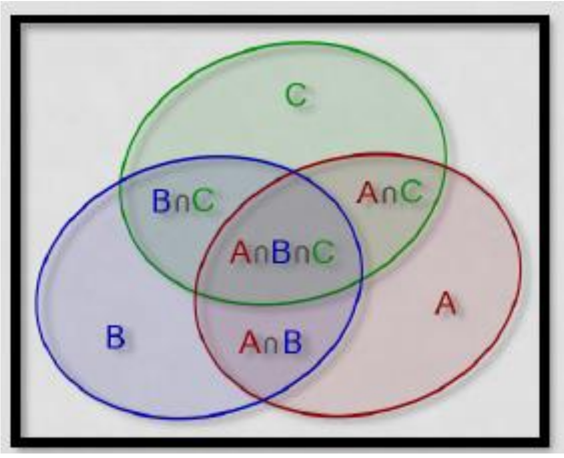
Symmetrische Differenz



<u>Formal</u>: $A \triangle B := \{ x \mid (x \in A \land x \notin B) \lor (x \in B \land x \notin A) \}$

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

Einschluss-Ausschluss-Verfahren (Prinzip von Inklusion und Exklusion, Prinzip der Einschließung und Ausschließung



Für <u>zwei</u> endliche disjunkte Mengen A und B gilt (**Summenregel**):

$$|\mathbf{A} \cup \mathbf{B}| = |\mathbf{A}| + |\mathbf{B}| - |\mathbf{A} \cap \mathbf{B}|$$

Noch allgemeiner gilt die **Summenregel** für <u>drei</u> Mengen:

$$|\mathbf{A} \cup \mathbf{B} \cup \mathbf{C}| = |\mathbf{A}| + |\mathbf{B}| + |\mathbf{C}| - |\mathbf{A} \cap \mathbf{B}| - |\mathbf{A} \cap \mathbf{C}| - |\mathbf{B} \cap \mathbf{C}| + |\mathbf{A} \cap \mathbf{B} \cap \mathbf{C}|$$

Gesetzmäßigkeiten der Mengenlehre

Für alle A , B , C ⊆ X gilt:

Antisymmetrie:	**$A \subseteq B$ und $B \subseteq A$**	→	**$A = B$**
Transitivität:	**$A \subseteq B$ und $B \subseteq C$**	→	**$A \subseteq C$**

Die Mengen-Operationen Schnitt ∩ und Vereinigung ∪ sind kommutativ, assoziativ und zueinander distributiv:

Assoziativgesetz:	$(A \cup B) \cup C = A \cup (B \cup C)$
	$(A \cap B) \cap C = A \cap (B \cap C)$
Kommutativgesetz:	$A \cup B = B \cup A$
	$A \cap B = B \cap A$
Distributivgesetz:	$A \cup (B \cap C) = (A \cup B) \cap (A \cup C)$
	$A \cap (B \cup C) = (A \cap B) \cup (A \cap C)$
De Morgansche Gesetze:	$(A \cup B)^C = A^C \cap B^C$
(Regeln von de Morgan)	$(A \cap B)^C = A^C \cup B^C$
Absorptionsgesetz:	$A \cup (A \cap B) = A$
	$A \cap (A \cup B) = A$

Für die Differenzmenge gilt:

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

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

Für die symmetrische Differenz gilt:

Assoziativgesetz:	$(A \triangle B) \triangle C = A \triangle (B \triangle C)$
Kommutativgesetz:	$A \triangle B = B \triangle A$
Distributivgesetz:	$(A \triangle B) \cap C = (A \cap C) \triangle (B \cap C)$

$A \triangle \emptyset = A$

$A \triangle A = \emptyset$

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

$A \triangle B = A^C \triangle B^C$

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

Für die Komplementmenge gilt:

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

$$\emptyset^C = \Omega \qquad \Omega^C = \emptyset$$

Doppeltes Komplement:

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

Idempotenzgesetze:

$$A \cap A = A \qquad A \cup A = A$$

Neutrale Elemente:

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

Dominanzgesetze:

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

Lösungsalgorithmus für Mengenoperationen

1.) Mengenbeschreibung in konkrete Mengen mit ihren Elementen transformieren
   a. Z.B.
      i. $A = \{x \in N | 3 \leq x \leq 8\} \rightarrow \{3, 4, 5, 6, 7, 8\}$
      ii. $B = \{x \in N | x\ ist\ Vielfaches\ von\ 5\} \rightarrow \{10, 15, 20, 25, \dots\}$

2.) Transformierte Mengen in Formel mit Mengenoperation einsetzen

3.) Mengenoperation auflösen und Ergebnismenge bilden
   a. Z.B.
      i. $A \cup B = \{1, 2, 3\} \cup \{3, 4, 5\} = \{1, 2, 3, 4, 5\}$
      ii. $A \cap B = \{1, 2, 3\} \cap \{3, 4, 5\} = \{3\}$

4.) Ergebnismenge notieren

Regeln der Wahrscheinlichkeitsrechnung

$$P(\Omega) = 1$$

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

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

1) $P(A \cup B) = P(A \setminus B) + P(A \cap B) + P(B \setminus A)$
   $P(A) = P(A \setminus B) + P(A \cap B)$
   $P(B) = P(A \cap B) + P(B \setminus A)$
   $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
2) sind A und B voneinander unabhängige Ereignisse:
   $P(A \cup B) = P(A \text{ oder } B) = P(A) + P(B)$
   $P(A \cap B) = P(A \text{ und } B) = P(A) * P(B)$

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

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

$1 -$

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

$P(B_1 \cap B_2 \cap B_3)]$

Beispiel:

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

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

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

Arten der Zufallsexperimente

- **Bernoulli-Experiment**:
  - Experiment hat nur zwei Versuchsausgänge
- **Laplace-Experiment**:
  - Alle Ereignisse haben die gleiche Wahrscheinlichkeit p

Formen und Berechnung von Wahrscheinlichkeiten

Allgemeine Berechnung (Bernoulli-Wahrscheinlichkeit)

Annahme: jedes Ziehungsergebnis hat die gleiche Wahrscheinlichkeit.

- Ω : die von k und n abhängige Menge aller möglichen Ziehungsergebnisse
- E : Teilmenge von Ω mit den günstigen Ziehungsereignissen

Die Wahrscheinlichkeit ist:

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

Laplace-Experiment

Alle Ereignisse haben die gleiche Eintrittswahrscheinlichkeit!

$$p = \frac{Zahl\ der\ günstigen\ Fälle}{Zahl\ der\ möglichen\ Fälle}$$

z.B. Würfelwurf, Münzwurf etc.

Gegenwahrscheinlichkeit

Sind für das gesuchte Ereignis A keine direkten Informationen gegeben, aber für die Gegenereignisse $B_1, B_2, \ldots, B_n$ kann die Wahrscheinlichkeit $p(A)$ anhand der Gegenereignisse über die Gegenwahrscheinlichkeit berechnet werden.

$$p(A) = 1 - p(B)$$

Mit $p(B) = p(B_1) + p(B_2) + \cdots + p(B_n)$.

Kombinatorik

Kombinatorik wird immer dann genutzt, wenn in einem Durchgang eines Experiments mehrere Wiederholungen stattfinden und bei dieser die Fragestellung ob mit oder ohne Reihenfolge als auch ob mit oder ohne Zurücklegen eine Rolle spielen.

Benennung

Kriterium ist die Reihenfolge:

- Mit Reihenfolge → Variation mit/ohne Wiederholung
- Ohne Reihenfolge → Kombination mit/ohne Wiederholung

Multiplikationsregel der Kombination

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

$$m_1 \cdot m_2 \cdot m_3 \cdot \ldots$$

Wichtigsten Bausteine von Kombinatorik-Formeln

1. **Fakultät**
   Für $n \in N_0$ gibt **n!** (n-Fakultät) die Anzahl der möglichen Permutationen (=Vertauschungen) von **n** verschiedenen Objekten an

$$n! = n \cdot (n-1) \cdot ... \cdot 1$$
$$0! = 1$$
$$1! = 1$$

2. **Binomialkoeffizient**
   $\binom{n}{k}$ → Tupel → Spachweise: **„n über k"**, oder „k aus n"
   Man kann auf $\binom{n}{k}$ Weisen **k** Elemente aus einer Menge mit **n** Elementen auswählen. Dies entspricht genau der Anzahl der Ziehungsergebnisse <u>ohne Zurücklegen und ohne Anordnung</u> (K_{oW}).

Grundmuster der Kombination

1. mit Reihenfolge, mit Zurücklegen

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

2. mit Reihenfolge, ohne Zurücklegen

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

3. ohne Reihenfolge, ohne Zurücklegen (Binomialkoeffizient)

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

4. ohne Reihenfolge, mit Zurücklegen

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

Mit

- k : Anzahl der zu ziehenden Elemente (Anzahl der Ziehungen)
- n : Anzahl der Elemente der Menge A, $A = \{1, \dots, n\}$
- $|\Omega|$: Anzahl der möglichen Ereignisse

Beispiele zur Identifikation der Grundmuster

Mit Reihenfolge, mit Zurücklegen

- Wie viele „Wörter“ (auch unsinnige) können aus 5 Buchstaben zustande kommen aus einem Alphabet vom Umfang 26?
  → $n = 26, k = 5, k < n$
  - 1.) $|\Omega| = V_{mW}$ berechnen
- Fünfmaliges Werfen eines Würfels (= oder Werfen von fünf Würfeln).
  → $n = 6, k = 5, k < n$
  - 1.) $|\Omega| = V_{mW}$ berechnen
- …

Mit Reihenfolge, ohne Zurücklegen

- Auf wie viele Arten können sich 5 Personen auf 5 freue (unterscheidbare) Plätze verteilen?
  → $n = 5, k = 5, n = k$
  - 1.) $|\Omega| = V_{oW}$ berechnen
- Es gibt eine genaue Sitzordnung für die Personen. Wie groß ist die Wahrscheinlichkeit, dass sich jede Person auf den ihr zugedachten Platz zufällig gesetzt hat?
  → Achtung! Laplace-Experiment: Die „Zahl der möglichen Fälle“ ist $|\Omega|$, die „Zahl der günstigen Fälle“ ist in diesem Beispiel $|E| = 1$.
  - 1.) $|\Omega| = V_{oW}$ berechnen
  - 2.) $|E|$ berechnen, in diesem Fall gilt $|E| = 1$, durch „Es gibt eine genaue …“
  - 3.) $p = \frac{|E|}{|\Omega|} = \frac{1}{|\Omega|}$ berechnen
- Wie ist die Wahrscheinlichkeit, beim viermaligen Werfen eines Würfels lauter verschiedene Augenzahlen zu erzielen?
  → $n = 6, k = 4, k < n$
  - 1.) $|\Omega| = V_{oW}$ ermitteln
  - 2.) Menge günstige Ereignisse ermitteln $|E|$
  - 3.) Wahrscheinlichkeit berechnen $p = \frac{|E|}{|\Omega|}$
- …

Ohne Reihenfolge, ohne Zurücklegen

- Gleichzeitiges Ziehen von 2 Kugeln aus der Urne U_6 (= Urne mit 6 Kugeln) ohne zurücklegen. Wie viele Paare sind möglich, wenn man die Kugeln durchnummeriert?
  - Aus „wie viele Paare“ → Ereignisraum $|\Omega| = \{ \{1,2\}, \{1,3\}, \dots \{5,6\} \}$ definieren (alle möglichen Kombinationen!).
  - 1.) $|\Omega| = K_{oW}$ berechnen
- Einer Urne mit 6 roten und 4 grünen Kugeln werden gleichzeitig 5 Kugeln entnommen. Wie ist die Wahrscheinlichkeit, dass genau 2 der Kugeln rot sind?
  → $n = 10, n_1 = 6, n_2 = 4, k = 5, k < n$
  - 1.) $|\Omega| = K_{oW}$ berechnen
  - 2.) $|E|$ berechnen
  - 3.) $p = \frac{|E|}{|\Omega|}$ berechnen
- …

Ohne Reihenfolge, mit Zurücklegen

- 10 Sportlerinnen nehmen an 3 Wettbewerben teil, bei denen es jeweils genau eine Siegerin gibt. Auf wie viele Arten können die Preise verteilt werden?
  → $n = 10, k = 3$
  - 1.) $|\Omega| = K_{mW}$ berechnen

Algorithmus für Kombinatorikaufgaben

Anordnung der Elemente einer Menge mit Wiederholung (Permutation)

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

Ist $\mathbf{n = n_1 + n_2 + \ldots + n_k}$, dann ist die Anzahl der Permutationen

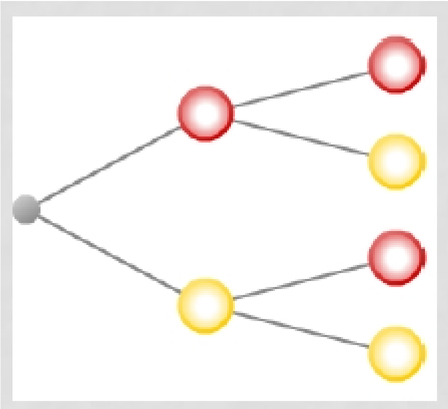
$$P_{mW} = \frac{n!}{n_1!\, n_2! \ldots n_k!}$$

Beispiele

- Der Name **RAFAELLA** ist eine Anordnung der Buchstabenmenge $\{R, A, F, E, L\}$ mit der Besonderheit, dass der Buchstabe L zweimal und der Buchstabe A dreimal auftritt.
  - 1.) Anzahl der Vorkommen der einzelnen Buchstaben zählen
    - $n_1 = 1x\ R, n_2 = 3x\ A, n_3 = 1x\ F, n_4 = 1x\ E, n_5 = 2x\ L$
  - 2.) Es gibt also $P_{mW} = \frac{8!}{1!3!1!1!2!} = \frac{8 \cdot 7 \cdot 6 \cdot 5 \cdot 4 \cdot 3 \cdot 2}{3 \cdot 2 \cdot 2} = 3.360$
  - Permutation → Es können also 3.360 unterschiedliche „Wörter“ aus dieser Buchstabenmenge zusammengestellt werden.

Baumdiagramm (auch Wahrscheinlichkeitsgraph, W-Graph)

- **Grafische Darstellung** und Lösung von Wahrscheinlichkeitsproblemen
    - Gerichteter Graph in Baumstruktur → Baumdiagramm
- Einzutragende Ereignisse müssen disjunkt sein.
- **Knoten** sind die möglichen Ausgänge $A, \bar{A}, B, \bar{B}, \dots$
- **Kanten/Linien** enthalten die Angaben zu den Wahrscheinlichkeiten $p(A), p(\bar{A}), p(B), p(\bar{B}), \dots$



Pfadregeln für das Baumdiagramm

1.) **Multiplikationssatz**:
   a. Die Wahrscheinlichkeit für das Eintreten eines Pfades ist das **Produkt** aller der längs dieses Pfades verzeichneten Wahrscheinlichkeiten.
   b. $p_{AB} = p(A) \cdot p(B), p_{A\bar{B}} = p(A) \cdot p(\bar{B}), \dots$

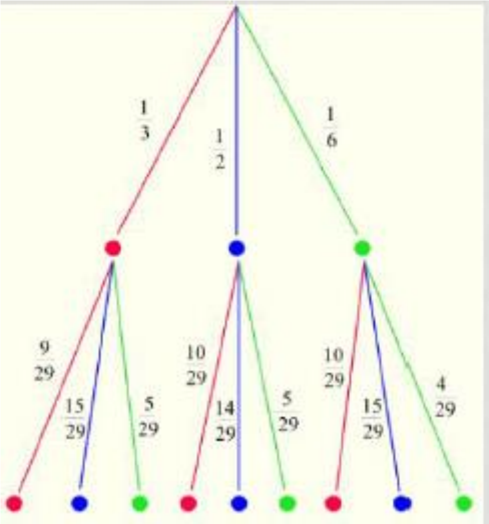
2.) **Satz von der totalen Wahrscheinlichkeit**:
   a. Die Wahrscheinlichkeit eines Ereignisses A ist gleich der **Summe** der Wahrscheinlichkeiten aller Pfade die zu einem Zustand führen, bei dem das Ereignis A eintritt.
   b. $p = p(A) \cdot p(B) + p(A) \cdot p(\bar{B})$

Wahrscheinlichkeitsmengen

Schlagwort	Mengenformel
Beide, alle, zufällig, Und, …	$p(A \cap B)$
Wenn … dann, wenn …, …	$p(A \vert B)$ oder $p(B \vert A)$
Höchstens, …	$p = 1 - p(A \cup B)$ (?)
Wenigstens, mindestens, Oder, …	$p(A \cup B)$

P (A ∪ B) = P (A oder B) = P (A) + P (B)
P (A ∩ B) = P (A und B) = P (A) * P (B)

Beispiel Baumdiagramm



$$p(\text{rote Kugel}) = \frac{1}{3}$$
$$p(\text{blaue Kugel}) = \frac{1}{2}$$
$$p(\text{grüne Kugel}) = \frac{1}{6}$$

Die Summe dieser Wahrscheinlichkeiten ist 1.

Multiplikationsregel

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

$$p(\text{erst rot}, \text{dann blau}) = \frac{1}{3} \cdot \frac{15}{29} = \frac{5}{29}$$
$$p(\text{erst blau}, \text{dann rot}) = \frac{1}{2} \cdot \frac{10}{29} = \frac{5}{29}$$

Additionsregel (Satz der totalen Wahrscheinlichkeit)

Die Pfade sind disjunkte Ereignisse, daher werden die Wahrscheinlichkeiten addiert:

$$p(A) = p(\text{erst rot}, \text{dann blau}) + p(\text{erst blau}, \text{dann rot}) = \frac{5}{29} + \frac{5}{29} = \frac{10}{29}$$

Bedingte Wahrscheinlichkeit

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

Voraussetzung: $P(B) > 0$

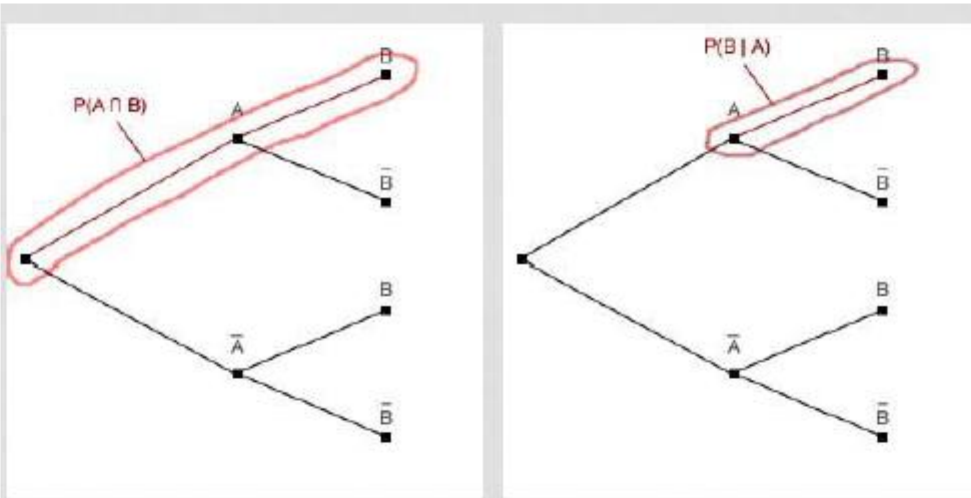
$$P(A|B) := \frac{P(A \cap B)}{P(B)}$$

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

Unterschied zwischen $P(A \cap B)$ und $P(B|A)$

- $P(A \cap B)$ und $P(B|A)$ sind verschiedene Wahrscheinlichkeiten!

➔ $P(A \cap B)$: Erst tritt Ereignis A ein, dann tritt noch Ereignis B ein.
➔ $P(B|A)$: Ereignis A ist bereits eingetreten, es ist nur noch die Eintrittswahrscheinlichkeit des Ereignisses B wichtig.



Totale Wahrscheinlichkeit

- Hilfsmittel, um mit Hilfe bekannter Wahrscheinlichkeiten weitere unbekannte Wahrscheinlichkeiten zu ermitteln.

Formeln:

Für **zwei Ereignisse** A und B:

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

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

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

Satz von Bayes

- Ergibt sich direkt aus dem Satz der totalen Wahrscheinlichkeit!
- Lösungsansatz zum Berechnen von bedingten Wahrscheinlichkeiten
- Lösung u.a. mittels Baumdiagramm

Formeln:

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

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

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

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

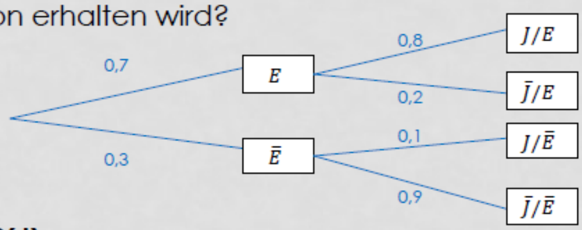
Beispiel für totale Wahrscheinlichkeit und Satz von Bayes

Beispiel:

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

E: Ende des Studiums

J: Job



Tipp:

s. auch Folie 48

Gefragt ist hier nach **P(J)**.

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

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

Algorithmus Modellierung

1.) Wonach ist gefragt?
    a. Anzahl der möglichen Ereignisse (wie viele, …)
    b. Wahrscheinlichkeit
2.) Form des Zufallsexperiments ermitteln
    a. Bernoulli-Experiment, Laplace-Experiment, Gegenwahrscheinlichkeit, Kombinatorik, bedingte Wahrscheinlichkeit, …
3.) Ereignisraum Ω definieren
    a. Alle möglichen Ereignisse!
4.) Anzahl der möglichen Ereignisse $|\Omega|$ berechnen
    a. Wenn nach der Anzahl der möglichen Ereignisse (z.B. „wie viele“, …) gefragt wird, endet an dieser Stelle die Aufgabe.
5.) Ereignis A definieren (Was sind die günstigen/erwünschten/gefragten Ereignisse?)
6.) Anzahl $|E|$ der günstigen Ereignisse von A berechnen.
7.) Wahrscheinlichkeit $p(A)$ berechnen
    a. Ggf. mittels Gegenwahrscheinlichkeit $p(A) = 1 - p(B)$
    b. Bei Bedingten Wahrscheinlichkeiten (hängt eine Wahrscheinlichkeit vom Eintreten eines Ereignisses ab?)
        i. Ggf. Über Satz der Totalen Wahrscheinlichkeit
        ii. Ggf. Vier-Felder-Tafel
        iii. Baumdiagramm