Statistik

1. Einführung

Begriff Statistik

- geordnete Zusammenstellung von Informationen
  Beispiele:
  - Arbeitslosenstatistik
  - Unfallstatistik
- Methoden zur Untersuchung von Massenerscheinungen

beschreibende Statistik (deskriptive Statistik, deduktive Statistik)

- Zusammenfassung
- Ordnung
- grafische Darstellung
- tabellarische Darstellung

⇓

Berechnung Kennzahlen

Interpretation

Untersuchung von Zusammenhängen

Erkennung von Mustern / Tendenzen

schließende Statistik (beurteilende Statistik, induktive Statistik)

- mit Hilfe von Wahrscheinlichkeitsrechnungen wird auf Basis einer Stichprobe auf die Gesamtheit geschlossen
  z.B. Qualitätskontrolle
  Wahlhochrechnungen

2. Häufigkeitsverteilungen

Rohdaten:

- gesammelte Daten
- zahlenmäßig nicht sortiert

Beispiel: Körpergrößen von 100 Studenten entnommen aus alphabetischer Liste

Zahlenfolgen:

- Anordnung numerischer Rohdaten sortiert nach Größe in auf- oder absteigender Reihenfolge

Spannweite:

- Differenz zwischen größter und kleinster Zahl

Beispiel:

Größter Student	199 cm
Kleinster Student	155 cm
⇒ Spannweite	44 cm

Klassen und Kategorien:

- Zusammenfassung von Rohdaten (in Klassen)
- Bestimmung Anzahl der Werte je Klasse
  ↳ Häufigkeit der Klasse

Häufigkeitsverteilung bzw. -tabelle:

- tabellarische Zusammenfassung + Häufigkeit

Beispiel:
Körpergröße von 100 Studenten

Körpergröße in cm	Anzahl Studenten
151 – 160	5
161 – 170	18
171 – 180	42
181 – 190	27
191 – 200	8
Gesamt	100

↳ 1. Klasse: 151 – 160 cm Häufigkeit=5

Klassenintervall:

- Symbol, das die Klasse definiert
  z. B. 151 – 160

Klassengrenzen:

- die jeweiligen Zahlen der Klassenintervalle
  z. B. 151, 160

untere Klassengrenze:

- die kleinere Zahl
  z. B. 151

obere Klassengrenze:

- die größere Zahl
  z. B. 160

offenes Klassenintervall:

- Klassenintervall hat entweder keine obere oder untere Klassengrenze
  z. B. 65 und älter

wahre Klassengrenzen:

- Grenzwerte unter Berücksichtigung der Meß- bzw. Erfassungsgenauigkeit
  z. B. Enthält das Klassenintervall 151 – 160 theoretisch alle Meßwerte zwischen 150,5 bis 160,5

untere wahre Klassengrenze:

z. B. 150,5

obere wahre Klassengrenze:

z. B. 160,5

Klassenbreite (-größe, -länge):

- Differenz zwischen der wahren oberen und der wahren unteren Klassengrenze
  z. B. 160,5 – 150,5 = 10

Klassenkennzahl (-mitte):

- Mitte des Klassenintervalls
- Addition der (wahren) unteren und oberen Klassengrenze und Division der Summe durch 2
  z. B. (151 + 160) / 2 = 155,5

Allgemeine Regeln zur Bildung von Häufigkeitsverteilungen

- Bestimmung des größten und kleinsten Wertes der Rohdaten
- Ermittlung der Spannweite
- Einteilung der Spannweite in eine sinnvolle Anzahl gleich großer Klassenintervalle
  falls nicht möglich:
  - verschiedene Größen
  - offene Intervalle

  Anzahl der Klassenintervalle zwischen 3 bis 10 (20)
- Bestimmung der Anzahl der Beobachtungen, die in jedes Klassenintervall fallen, d. h. Klassenhäufigkeit

Histogramm oder Häufigkeitshistogramm:

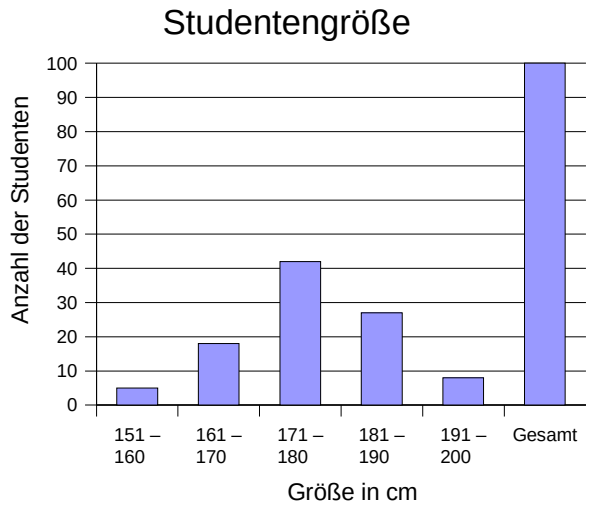
- grafische Darstellung der Häufigkeitsverteilung

Darstellungsregeln:

Anreihung von Rechtecken, die wie folgt bestimmt werden:

a) Grundseite auf horizontaler Achse (x-Achse)
   Mittelpunkt bei der Klassenkennzahl
   Breite entspricht der Klassengröße (-breite)
b) Flächen (der Rechtecke) sind proportional zur Klassenhäufigkeit ⇒
   i. Klassenintervalle gleich groß
      ⇒ Höhe ist proportional zur Klassenhäufigkeit
   ii. Klassenintervalle nicht gleich groß
      ⇒ Anpassung der Höhe

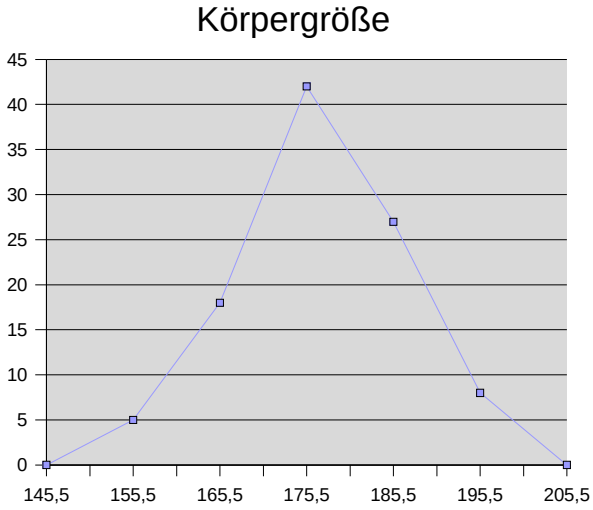
Beispiel: Körpergröße



Häufigkeitspolygon

- Polygonzug, bei dem die Klassenhäufigkeit über die Klassenkennzahl aufgetragen ist.

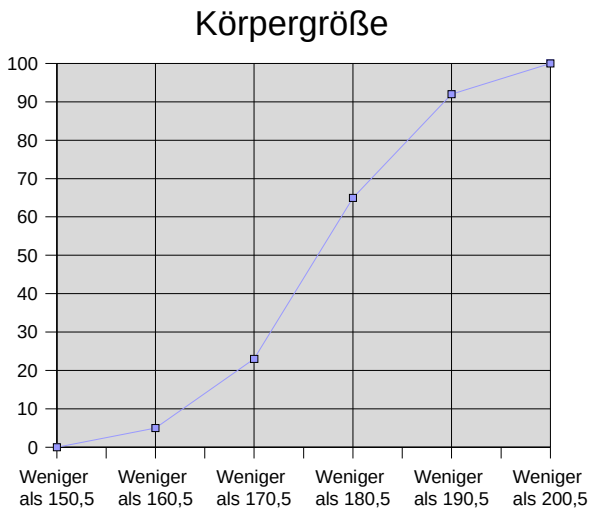
Körpergröße in cm	Anzahl Studenten
145,5	0
155,5	5
165,5	18
175,5	42
185,5	27
195,5	8
205,5	0



Kumulatives Häufigkeitspolygon

Die Gesamthäufigkeit aller Werte unter der oberen Klassengrenze eines gegebenen Klassenintervalls nennt man kumulative Häufigkeit.

Körpergröße [cm]	*Anzahl Studenten*
Weniger als 150,5	0
Weniger als 160,5	5
Weniger als 170,5	23
Weniger als 180,5	65
Weniger als 190,5	92
Weniger als 200,5	100



– grafische Darstellung der „weniger als“-Werte gegen die obere wahre Klassengrenze

„oder mehr“-Kumulativverteilung

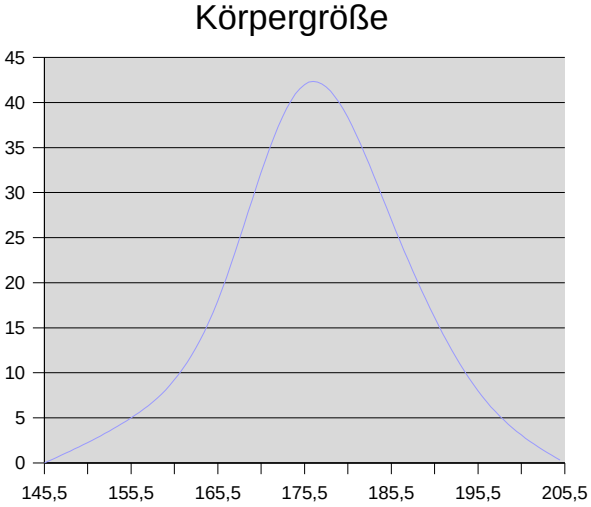
– enthält alle Werte, die größer als oder gleich der unteren Klassengrenze eines jeden Klassenintervalls sind.

Relative Häufigkeitsverteilung

$$\text{Relative Häufigkeit einer Klasse} = \frac{\text{Häufigkeit der Klasse}}{\text{Gesamtsumme (aller Klassen)}}$$

in der Regel in %

Geglättete Häufigkeitspolygone



sehr viele Werte

→ kleine Klassenintervalle

→ Häufigkeitspolygon besteht aus vielen Liniensegmenten

⇓

Häufigkeitsverteilung

3. Zentralmaße

Die gebräuchlichsten sind:

- arithmetisches Mittel
- der Zentralwert, Median
- der Modus
- das geometrische Mittel
- das harmonische Mittel

Ausgangssituation für die folgenden Definitionen:

Gegeben ist eine Liste von n reellen Zahlen $x_1, x_2, \ldots x_n$

Arithmetisches Mittel

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

Beispiel: Mittlerer Umsatz der 35 größten deutschen Unternehmen

$$\overline{x} = \frac{162{,}3 + 85{,}6 + \ldots + 8{,}5}{35} = 31{,}3 \text{ Mrd } €$$

Falls

a) Werte in einer Liste mehrfach vorkommen
   d.h. es tritt in einer Liste $x_1, x_2, \ldots x_n$ der Wert a_i (i = 1, 2 ... k, k ∈ n) mit der absoluten Häufigkeit H_i bzw. mit der relativen Häufigkeit h_i auf,

b) eine Häufigkeitsverteilung vorliegt
   d.h. es ist die Klassenkennzahl (a_i) einer Klasse und die absolute Häufigkeit H_i bzw. die relative Häufigkeit h_i bekannt,

so kann für

a) das arithmetische Mittel genau

und für

b) das arithmetische Mittel näherungsweise

nach folgender Formel berechnet werden:

$$\overline{x} \overset{a)\ b)}{=} (\approx) \frac{H_1 a_1 + H_2 a_2 + \ldots H_k a_k}{H_1 + H_2 + \ldots H_k} \overset{a}{=} (\overset{b}{\approx}) \frac{\sum_{i=1}^{n} H_i a_i}{\sum_{i=1}^{k} H_i} \overset{a}{=} (\overset{b}{\approx}) \frac{\sum_{i=1}^{k} h_i a_i}{\sum_{i=1}^{k} h_i = 1} = \sum_{i=1}^{k} h_i a_i$$

Beispiel: Übungsaufgabe 2 (Häufigkeitsverteilung)
Bestimmung mittleres Einkommen

$$\overline{x} \approx \frac{2 \cdot 22500 + 9 \cdot 27500 + \ldots + 1 \cdot 95000}{162} \approx 41327\ € \approx 41300\ €$$

Eigenschaften des arithmetischen Mittels

Definition:
$$\overline{x} = \frac{x_1 + x_2 \ldots + x_n}{n}$$
$$\Rightarrow n\,\overline{x} = x_1 + x_2 + \ldots + x_n$$

Für das arithmetische Mittel gilt:

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

d.h. die Abweichungen vom arithmetischen Mittel heben sich gegenseitig auf.

Beweis:

$$\sum_{i=1}^{n}(x_i-\overline{x})=x_1-\overline{x}+x_2-\overline{x}+\ldots+x_n\overline{x}=x_1+x_2+\ldots+x_n-n\,\overline{x}=0$$

Minimaleigenschaft des arithmetischen Mittels

Von keinem anderen Wert ist die Summe der Quadratabstände der x_i geringer als vom arithmetischen Mittel, d.h.

$\sum_{i=1}^{n}(x_i-x)^2$ hat an der Stelle $x<\overline{x}$ ein Minimum

Beweis:

$$f(x)=\sum_{n=1}^{n}(x_i-x)^2=\sum_{i=1}^{n}(x_i^2-2x_ix+x^2)$$

zu zeigen ist $f'(\overline{x})=0$ und $f''(\overline{x})>0$

$$f'(x)=-2x_1+2x-2x_2+2x-\ldots-2x_n+2x=-2\sum_{i=1}^{n}(x_i-x)$$

$$f'(x)=-2\sum_{n=1}^{n}(x_i-\overline{x})=0$$

$$f'(x)'=2+2+2+\ldots+2=2n>0$$

$$f''(\overline{x})=2n>0$$

Gesucht: Durchschnittsprozentsatz;

Steigerung des Umsatzes in 3 Jahren auf:

$$1{,}2\cdot 1{,}025\cdot 1{,}075=1{,}32225=132{,}25\%=\frac{4056519}{3067891}=1{,}32225$$

$$(1+i)\cdot(1+i)\cdot(1+i)=1{,}2\cdot 1{,}025\cdot 1{,}075$$

$$(1+i)^3=1{,}2\cdot 1{,}025\cdot 1{,}075$$

$$1+i=\sqrt[3]{1{,}2\cdot 1{,}025\cdot 1{,}075}\approx 1{,}098$$

$$i=0{,}098=9{,}5\%$$

Geometrisches Mittel

$$\widehat{x}=(1+i)=\sqrt[n]{x_1\cdot x_2\cdot\ldots\cdot x_n}$$

Das geometrische Mittel ist dann zu verwenden, wenn sich die Endgröße aufgrund von Multiplikationen ergibt.

Beispiele:

- Zinsen
- Kosten- u. Preisindizes
- Aktienindizes

- Bevölkerungswachstum

Harmonisches Mittel

Beispiel Durchschnittsgeschwindigkeit:

Gesamtfahrzeit

$$A_{ges} = \frac{240}{\tilde{V}} = \frac{120km}{80\frac{km}{h}} + \frac{120}{120\frac{km}{h}}$$

$$\frac{2}{\tilde{V}} = \frac{1}{80} + \frac{1}{120}$$

$$\tilde{V} = \frac{2}{\frac{1}{80} + \frac{1}{120}} = 96\frac{km}{h}$$

$$\tilde{V} = \frac{2}{\frac{1}{V_1} + \frac{1}{V_2}} \triangleq \text{ harmonisches Mittel}$$

Harmonisches Mittel

$$\tilde{x} = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \ldots + \frac{1}{x_n}}$$

Es eignet sich für Größen, die indirekt proportional zu anderen Größen sind.

Anwendungsgebiete:

Geschwindigkeit: $(\sim \frac{1}{Zeit})$

Stromstärke: $(\sim \frac{1}{R})$

Dichte: $(\sim \frac{1}{Volumen})$

Median (auch Zentralwert)

Mittelwerte haben die Eigenschaft, daß sich extreme Einzelwerte, sog. Ausreißer, sehr stark auswirken.

Beispiel Unternehmensstatistik

$$\overline{x_{\text{mit DaimlerChrysler}}} = 31,3 Mrd\ €$$

$$\overline{x_{\text{ohne DaimlerChrysler}}} = 27,5 Mrd\ €$$

Völlig unempfindlich gegenüber Ausreißern ist der Median.

Definition: Ordnet man eine Liste von Variablenwerten nach der Größe, so heißt

- bei einer ungeraden Anzahl von Werten der in der Mitte stehende Wert
- bei einer geraden Anzahl das arithmetische Mittel der beiden mittleren Werte

Median.

Beispiel: Unternehmensstatistik

m = 21,9 Mrd €

Der Median teilt die Liste in zwei gleich große Teile. In dem einem Teil finden sich die unterdurchschnittlichen Werte, im dem anderen die überdurchschnittlichen.

Minimaleigenschaft des Medians

Es sei $x_1, x_2, ..., x_n$ eine Liste reeller Zahlen mit dem Median m, dann hat

$\sum_{i=1}^{n} (x_i - x)$ an der Stelle x = m ein Minimum.

Quatile, Dezile und Percentile

Ausgangssituation: Eine nach Größe geordnete Reihe von Daten.
Median m: teilt die Reihe in 2 gleiche Teile.
Quatile Q_1, Q_2, Q_3: teilen die Reihe in 4 gleiche Teile.
Dezile $D_1 ... D_9$: teilen die Reihe in 10 gleiche Teile.
Percentile $P_1 ... P_{99}$: teilen die Reihe in 100 gleiche Teile.

Beispiel: Unternehmensstatistik
$Q_1 = Q_u = 13{,}1$ Mrd €
$Q_2 = m = 21{,}9$ Mrd €
$Q_3 = Q_o = 35{,}9$ Mrd €

$Q_2 = m = D_5 = P_{50}$
$Q_1 = P_{25}$
$Q_3 = P_{75}$

Modus

Der Modus einer Liste ist der Wert, der am häufigsten vorkommt.
Beispiele:
Die Reihe 2, 2, 5, 7, 9, 9, 9, 10, 10, 11, 12 und 18 hat den Modus 9.
Die Reihe 3, 5, 8, 10, 12, 15 und 16 hat keinen Modus.
Die Reihe 2, 3, 4, 4, 4, 5, 5, 7, 7, 7 und 9 hat zwei Modi (4, 7).

Skalenarten

Mittelwerte: metrische Daten
Median: metrische Daten, ordinalskalierte Daten (Rangfolge, z.B. Dienstgrade)
nominalskalierte Daten (z.B. Geschlecht): bestenfalls Modus als Zentralmaß

4. Streuungsmaße und Schiefe

Spannweite

$R = SP = x_{max} - x_{min}$
- sehr einfaches Maß
- in manchen Fällen aussagekräftig
- extrem ausreißerempfindlich

z.B. Unternehmensstatistik
SP(Umsatz) = 162,3 – 8,5 = 153,8 Mrd €

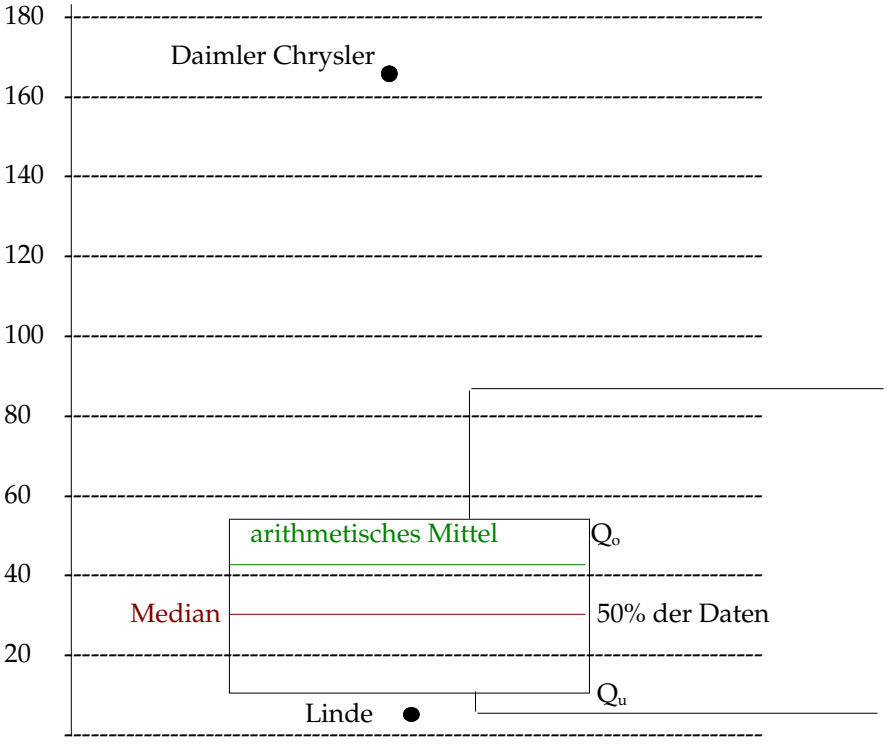
Interquantil-Spannweite

$ISP = Q_3 - Q_1 =$ Q_3, Q_o: 3. Quartil, oberes Quartil

$Q_o - Q_u$ Q_1, Q_u: 1. Quartil, unteres Quartil

Die Interquartil-Spannweite i_{SP} umfasst die Hälfte von Daten einer Liste.
Beispiel: Unternehmensstatistik
$i_{SP} = 35,9 - 13,1 = 22,8$ Mrd €

Boxplot (Box-Whisker-Plot, Kastenschaubild)



Mittlere absolute Abweichung

$$s^x = \frac{\sum_{i=1}^{n} |x_i - x_Z|}{n}$$

s^x: mittlere absolute Abweichung
n: Anzahl der Werte
x_i: Einzelwerte
x_Z: Zentralmaß (in Regel: Median = Zentralwert)

Beispiel Unternehmensstatistik:
Median = 21,9 Mrd €

$$s^* = \frac{1}{35}(|162,3 - 21,9| + \ldots + |8,5 - 21,9|) = \underline{17,9 \text{ Mrd } €}$$

Mittlere quadratische Abweichung oder empirische Varianz s^2

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

n_i: Anzahl der Meßwerte
x_i: Einzelwert
$\bar{x}$: arithmetisches Mittel

Beispiel Unternehmensstatistik

$$s^2 = \frac{1}{35}[(162,3 - 31,3)^2 + \ldots (8,5 - 31,3)^2] = 874,6 \text{ (Mrd €)}^2$$

empirische Standardabweichung

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

Beispiel Unternehmensstatistik

$$s = \sqrt{s^2} = \sqrt{874{,}6} = 29{,}6 \text{ Mrd } €$$

Minimaleigenschaft Median

➔ Streumaß: Mittlere absolute Abweichung

Minimaleigenschaft arithmetisches Mittel

➔ Streumaß: Mittlere quadratische Abweichung = empirische Varianz, empirische Standardabweichung

Empirische Varianz und empirische Standardabweichung sind ausreißerempfindlich und wenig robuste Steuerungsmaße.

s-Regeln (normal verteilter Daten)

Was heißt normalverteilt?

Daten liegen symmetrisch und konzentriert um das arithmetische Mittel.

Bei solchen Verteilungen wird man feststellen, daß

- ca. 2/3 aller Daten höchstens um s von $\overline{x}$
- ca. 95% aller Daten höchstens um 2s von $\overline{x}$
- ca. 99% aller Daten höchstens um 3s von $\overline{x}$

abweichen.

Verschiebungssatz

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

Beweis:

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

$$= \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\overline{x}\cdot\frac{1}{n}\underbrace{\sum_{i=1}^{n} x_i}_{n\overline{x}} + \frac{1}{n}n\,\overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - 2\,\overline{x^2} + \overline{x^2} = \frac{1}{n}\sum_{i=1}^{n} x_i^2 - \overline{x^2}$$

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

Berechnung der empirischen Standardabweichung aus Häufigkeitstabellen

Gegeben: a_i (i = 1, 2, ... k, k < n), H_i, k_i

$$s = \sqrt{\frac{1}{n}\sum_{n=1}^{k} H_i (a_i - \overline{x})^2}$$

$$s = \sqrt{\sum_{i=1}^{k} k_i (a_i - \overline{x})^2}$$

analoge Vorgehensweise für die mittlere absolute Abweichung.

$$s^* = \frac{1}{n}\sum_{i=1}^{k} H_i |a_i - xz|$$

$$s^* = \sum_{i=1}^{k} k_i |a_i - xz|$$

Vergleich von mittleren Abweichungen

Betrachtung von Streuungen im Verhältnis zum zugehörigen Zentralmaß (Normierung).

Variationskoeffizienten: $v = \frac{s}{\overline{x}}$, dimensionslos, meist %

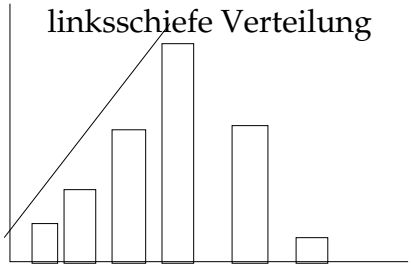
Variabilitätskoeffizient: $v = \frac{s^*}{Z}$, dimensionslos, meist %

Beispiele:

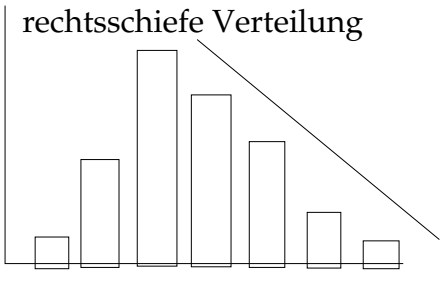
Zuckerpakete	*Pfeffertütchen*
Füllgewicht 1 kg	Füllgewicht 50g
S = 50 g	S = 2,5g
	⇓
	V = 5%

Schiefe

linksschiefe Verteilung



rechtsschiefe Verteilung



$$sch = \frac{1}{n}\sum_{i=1}^{n} (x_i - \overline{x})^3$$

sch < 0 ⇒ Verteilung linksschief

sch > 0 ⇒ Verteilung rechtsschief

sch = 0 ⇒ Verteilung symmetrisch

5. Zusammenhänge zwischen Merkmalen

Bisher: Eindimensionale beschreibende Statistik
Zusammenhänge zwischen Merkmalen gehen verloren
z.B. zwischen Beschäftigungszahl und Umsatz

Zweidimensionale beschreibende Statistik

Methoden zur Untersuchung des Zusammenhangs zwischen zwei Merkmalen.

Methoden sind z.B.

- Regressionsanalyse
- Korrelationsanalyse

für metrische Daten

5.1. Streudiagramm (und Trend)

- grafische Darstellung zweier metrisch skalierter Merkmale in einem kartesischen Koordinatensystem.
- Zusammengehörige Ausprägungen beider Merkmale werden als Zahlenpaare aufgefaßt.
- Zahlenpaare werden als Punkt im Koordinatensystem gezeichnet.

Bsp.: Unternehmensstatistik

Hervorheben der Trends durch Umrahmung der „Punktwolke“ und einzeichnen einer Trendgerade.

Ermittlung der Gleichung der Trendgerade (näherungsweise):

Umsatz ≈ 0,4 Mrd €/Tsd · Beschäftigungszahl – 10 Mrd €

z.B. entspricht eine Beschäftigtenzahl einem Umsatz von
y_{100}=0,4 · 100 – 10 = 30 Mrd €

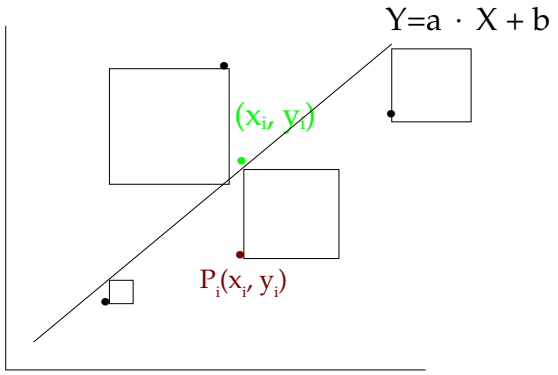
5.2. Lineare Regression und Korrelation

Methode der kleinsten Quadrate

Gegeben: n Punkte $P(x_i, y_i)$; i = 1, 2, ... n

Gesucht: Gerade, so daß die Summe der Quadratabstände der Punkte von der Geraden in vertikaler (oder auch horizontaler) Richtung möglichst klein ist.

vertikale Quadratabstände



Geradengleichung: $y = aX + b$

Abstand $P_i(x_i, y_i)$ in vertikaler Richtung zur Geraden:

$Y_i - y_i = aX_i + b - y_i = ax_i + b - y_i$

Summe der Quadratabstände aller n Punkte

$$\sum_{i=1}^{n} (a \cdot x_i + b - y_i)^2 ; i = 1,2,\ldots n$$

$$F(a,b) = \sum_{i=1}^{n} (ax_i + b - y_i)^2 ; x_i, y_i \text{ bekannte Punktkoordinaten} ; a, b \text{ Variablen}$$

Gesucht ist das Minimum von F(a, b):

$$\frac{\delta F}{\delta a} = 2 \sum_{i=1}^{n} (ax_i + b - y_i) \cdot x_i = 0$$

$$\sum_{i=1}^{n} (ax_i^2 + bx_i - x_i \cdot y_i) = 0 \qquad ①$$

$$\frac{\delta F}{\delta a} = 2 \sum_{i=1}^{n} (ax_i + b - y_i) \cdot 1 = 0$$

$$\sum_{i=1}^{n} (ax_i + b - y_i) = 0 \qquad ②$$

Umformen

① $$a \sum_{i=1}^{n} x_i^2 + b \sum_{i=1}^{n} x_i = \sum_{x=1}^{n} x_i \, y_i$$

② $$a \sum_{i=1}^{n} x_i + n \cdot b = \sum_{i=1}^{n} y_i$$

$$b = \frac{\sum_{i=1}^{n} y_i}{n} - \frac{a \cdot \sum_{i=1}^{n} x_i}{n} \quad \text{in ①}$$

$$a \sum_{i=1}^{n} x_i^2 + \frac{\sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i - a(\sum_{i=1}^{n} x_i)^2}{n} = \sum_{i=1}^{n} x_i \, y_i \quad | \, n$$

$$a \cdot (n \sum_{i=1}^{n} x_i^2 - (\sum_{i=1}^{n} x_i)^2) = n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i$$

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

Beispiel: Unternehmensstatistik

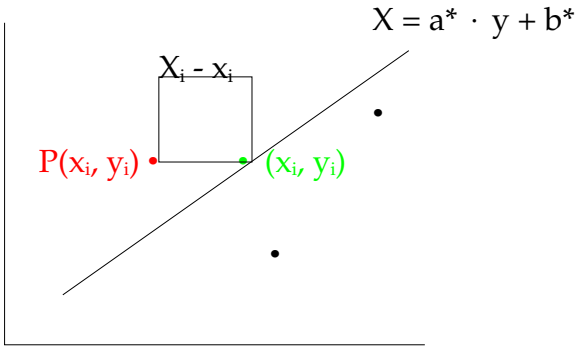
$$a = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2} = \frac{15 \cdot 107785 - 1381{,}5 \cdot 387{,}2}{15 \cdot 341877{,}8 - 1381{,}5^2} = 0{,}336$$

$$b = \frac{\sum_{i=1}^{n} y_i}{n} - \frac{a \cdot \sum_{i=1}^{n} x_i}{n} = \frac{387{,}2}{15} - 0{,}336\frac{1381{,}5}{15} = -5{,}13$$

Die Gleichung lautet somit

Umsatz [Mrd €] = 0,336 · Beschäftigtenzahl [Tsd] – 5,13 [Mrd €]

horizontale Quadratabstände



analoge Vorgehensweise

$X_i - x_i = a^* \cdot y_i + b^* - x_i = a^* \cdot y_i + b^* - x_i$

$$\sum_{i=1}^{n} (a^* \cdot y_i + b^* - x_i)^2 ; i = 1,2,\ldots n$$

$$F^*(a^*, b^*) = \sum_{i=1}^{n} (a^x \cdot y_i + b^* - x_i)^2$$

⇩ Vorgehensweise wie vorher

$$a^* = \frac{n \cdot \sum_{i=1}^{n} x_i \cdot y_i - \sum_{n=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i}{n \cdot \sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2}$$

$$b^* = \frac{\sum_{i=1}^{n} x_i}{n} - a^* \frac{\sum_{i=1}^{n} y_i}{n}$$

Beispiel: Unternehmensstatistik

$$a^* = \frac{15 \cdot 107785{,}0 - 1381{,}5 \cdot 387{,}2}{15 \cdot 36626{,}9 - 387{,}2^2} = 2{,}708$$

$$b^* = \frac{1381{,}5}{15} - 2{,}708 \cdot \frac{387{,}2}{15} = 22{,}20$$

$$\text{Beschäftigenzahl}\,|\,\text{TSO}\,| = 2{,}708\,\frac{\text{TSD}}{Mrd\ €}\cdot\ \text{Umsatz}\,|\,Mrd\ €\,| + 22{,}20\,|\,\text{TSO}\,|$$

$$\text{Umsatz}\,|\,Mrd\ €\,| = \frac{1}{2{,}708}\cdot\ \text{Beschäftigtenzahl} - \frac{22{,}20}{2{,}708} = 0{,}369\cdot\ \text{Beschäftigtenzahl} - 8{,}20$$

Lineare Regression

⇒ Unterschied zwischen beiden Geraden

Man bezeichnet:

1. Regressionsgerade: Minimierung der Quadratabstände in vertikaler Richtung
2. Regressionsgerade: Minimierung der Quadratabstände in horizontaler Richtung

Die Unterschiede zwischen den beiden Regressionsgeraden sind umso stärker, je stärker die einzelnen Punkte streuen.

- Der lineare Zusammenhang zwischen den beiden Merkmalen ist umso stärker,
- Ein linearer Trend ist umso stärker

je weniger die Steigungen der beiden Regressionsgeraden abweichen.

Korrelationsanalyse

- Quantifizierung des linearen Zusammenhangs

Die lineare Korrelation untersucht, um wieviel geringer die Steigung der ersten (also der flacheren) Regressionsgeraden im Verhältnis zur zweiten (also der steileren) Regressionsgeraden ist.

1. Regressionsgerade $Y = aX + b \Rightarrow$ Steigung $k_1 = a$
2. Regressionsgerade

$$X = a^* \cdot y + b^*$$

$$Y = \frac{1}{a^*}\cdot X - \frac{b^*}{a^*} \Rightarrow Steigung\ k_2 = \frac{1}{a^*}$$

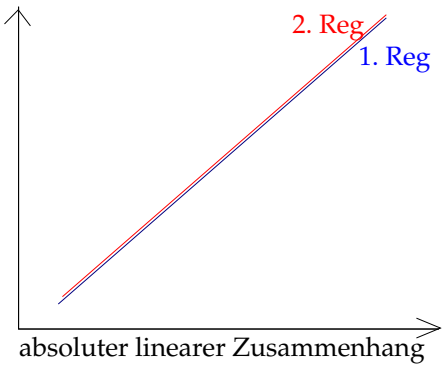
$$r^2 = \frac{h_1}{h_2} = \frac{a}{\frac{1}{a^*}} = a \cdot a^*$$ (Bestimmtheitsmaß)

$$|r| = \sqrt{a \cdot a^*}$$

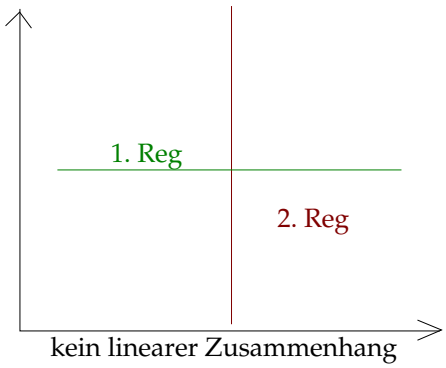
r: Pearsonscher Korrelationskoeffizient

Extremfälle:

$|r| = 1$ $\qquad$ $|r| = 0$



absoluter linearer Zusammenhang



kein linearer Zusammenhang

Beispiel Unternehmensstatistik

$a = 0{,}336 ; a^* = 2{,}708$

$|r| = \sqrt{0{,}336 \cdot 2{,}708} = 0{,}95$ $\Rightarrow$ deutlicher linearer Zusammenhang

Spearmanscher Korrelationskoeffizient / Rangkorrelationskoeffizient

Die Rangkorrelationskoeffizienten basieren auf Anordnung der Meßwerte innerhalb der Meßreihe.

Ausgangsdatenreihe: $(x_1, y_1) \ldots (x_n, y_n)$

Reihe der x-Werte: $x_1, \ldots x_n$

Ordnung der Reihe in aufsteigender Reihenfolge: $x_{(1)}, \ldots x_{(n)}; x_{(1)} \leq x_{(2)} \leq x_{(3)} \ldots x_{(n)}$

Position jedes Wertes in der Reihe wird als sein Rang $R(x_i)$ bezeichnet, analog $R(y_i)$

Beispiel: Unternehmensstatistik

$R(x_4) = R(198{,}7) = 13 = x_{(13)}$

$R(y_4) = R(31{,}6) = 12 = y_{(12)}$

In Analogie zum Pearsonschen Korrelationkoeffizienten definiert man deshalb den Spearmanschen Rangkorrelationskoeffizienten.

$$r_s = 1 - \frac{6}{n \cdot (n^2 - 1)} \cdot \sum_{i=1}^{n} (R(x_i) - R(y_i))^2$$

5.3. Nichtlineare Regression

Vorgehen:

- Anwendung der Methode der kleinsten Quadrate
- Bestimmung des Funktionstyps
  z.B.
  $y = b \cdot x^a$
  $y = b \cdot a^x$
  $y = ax^3 + bx^2 + cx + d$
  $y = \frac{x}{a + bx}$
- Bestimmung der Abweichungsquadrate
  $$F(a, b, \ldots) = \sum_{i=1}^{n} (Y_{(x_i)} - y_i)^2$$
- Bestimmung der partiellen Ableitungen
  $$\frac{\delta F(a, b, \ldots)}{\delta a}, \frac{\delta F(a, b, \ldots)}{\delta b}$$
- Nullsetzen der partiellen Ableitungen
- Lösung des Gleichungssystems und Berechnung der Koeffizienten a, b, ...

5.4. Analyse von Zeitreihen

- Analyse von Zeitreihen entspricht der Untersuchung des Zusammenhangs zwischen Merkmalen
- Veränderung der Ausprägung eines Merkmales in Abhängigkeit von der Zeit
- wesentliches Ziel: Prognose von Entwicklungen

Saisonale Schwankungen

⇒ erschwerte Prognose

Prognose auf Basis gleichender Mittelwerte

Grundannahmen:

- Es gibt einen globalen Trend
- saisonale Schwankungen sind einigermaßen regelhafte Abweichungen vom globalen Trend

Vorgehen:

- Modellierung des globalen Trends
- Modellierung der Saisonkomponenten (= rel. Abweichung vom globalen Trend)
  $y_t = T_t \cdot s_t$
  y_t: Wert der Zeitreihe zum Zeitpunkt bzw. Zeitraum t
  T_t: Trendwert zum Zeitpunkt bzw. Zeitraum t
  s_t: Saisonskomponente zum Zeitpunkt bzw. Zeitraum

Modellierung des globalen Trends

- „Glättung" der Zeitreihe für den globalen Trend
- Berechnung der Regressionsfunktion für „geglättete Zeitreihe"

Methode der gleitenden Durchschnitte

Unterscheidung:

Gleitender Durchschnitt ungerader Ordnung k

$$y^* := \frac{y_{i-\frac{(k-1)}{2}} + \ldots y_i \ldots y_{i+\frac{(k-1)}{2}}}{k} = \frac{1}{k} \cdot \sum_{i=i-\frac{(k-1)}{2}}^{i+\frac{(k-1)}{2}} y_i$$

Gleitender Durchschnitt gerader Ordnung

$$y_i^* = \frac{\frac{1}{2}y_{i-\frac{k}{2}} + y_{i-\frac{k}{2}+1} + \ldots + y_i + \ldots + y_{i+\frac{k}{2}-1} + \frac{1}{2}y_{i+\frac{k}{2}}}{k}$$

$$= \frac{1}{k}\left(\frac{1}{2}y_{i-\frac{k}{2}} + \sum_{i-\frac{k}{2}+1}^{i+\frac{k}{2}-1} y_i + \frac{1}{2}y_{i+\frac{k}{2}}\right)$$

Anwendung Quartals-Durchschnitt (Ordnung 4)

Beispiel Energieverbrauch einer Stadt

III Q 1996

t = 1: 142,6

t = 2: 96,4

t = 3: 88,9 $\qquad y_3^* = \frac{1}{4}(0{,}5 \cdot 142{,}6 + 96{,}3 + 88{,}9 + 136{,}4 + 0{,}5 \cdot 137{,}8) = 115{,}45$

t = 4: 136,4

t = 5: 137,8

Kommentierung der Ergebnisse bzw. Methode

- Ausreißer werden abgeschwächt
- saisonale Schwankungen werden ausgeglichen
- plausible Darstellung des globalen Trends
- sehr gute Korrelation, deutlicher linearer Trend

⇓

1. Regressionsgerade liefert eine brauchbare Beschreibung für den globalen Trend.

Modellierung der Saisonkomponenten

Berechnung der relativen Saisonabweichungen

$$s_t = \frac{y_t}{T_t}$$

Berechnung des arithmetischen Mittels für die jeweilige Saison

z.B. $s_I = \frac{(1{,}27 + 1{,}20 + 1{,}20 + 1{,}23 + 1{,}20)}{5} = 1{,}22$

und Interpretation als durchschnittliche Abweichung der Saison vom Trend und Verwendung für die Prognose.

z.B. Prognose für I/01 (t = 21)

$$T_t = 0{,}62 \cdot t + 112{,}1$$

$$T_{21} = 0{,}62 \cdot 21 + 112{,}1 = 125{,}1$$

$$y_{21} = T_{21} \cdot s_I = 125{,}1 \cdot 1{,}22 = 152{,}3$$

6. Grundlagen der Wahrscheinlichkeitsrechnung

Die Wahrscheinlichkeitstheorie beschäftigt sich mit der Berechnung der Auftretenswahrscheinlichkeit P (probability) zufälliger Ereignisse, bzw. Zufallsexperimenten.

Zufallsexperimente:

- beliegib oft unter gleichartigen Bedingungen wiederholbar
- Ausgang nicht mit Sicherheit vorherzusagen

z.B. Würfeln, Qualitätsprüfung, ...

Um Zufallsexperimente mathematisch behandeln zu können, müssen ihre Merkmale bzw. Ergebnisse beschrieben werden.

z.B. Ergebnisse eines Würfelwurfs

- {1}, {2}
- {1} oder {2}
- {gerade}
- ⋮

Qualitätsprüfung {intakt}, {defekt}

Ein Elementarereignis E (ω) ist nicht mehr in andere Ereignisse zerlegbar.

Elementarereignisse schließen sich gegenseitig aus.

z.B. Würfeln: {1}, {2}, ... {6}

Die Menge aller möglichen Elementarereignisse bildet den Ergebnissraum S oder den

Ereignissraum S (Ω)

Ereignissraum beim einmaligen Würfeln

S = {1, 2, ..., 6}

Beispiel: zweimaliges Werfen einer Münze

K für Kopf, Z für Zahl

Elementarereignisse: {K, K}, {K, Z}, {Z, K}, {Z, Z}

Ereignissraum: S = {KK, KZ, ZK, ZZ}

Als Komplementärereigniss $\overline{A}$ bezeichnet man alle Ereignisse der Menge aller Elementarereignisse eines Ereignissraums S, die nicht im betrachteten Ereignis A enthalten sind, bezeichnet

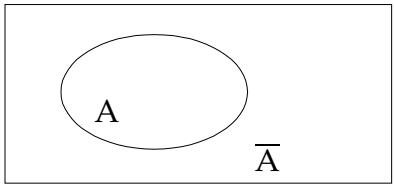
<u>Beispiel:</u> (Würfelwurf)

Ereignis A: „Werfen einer geraden Augenzahl“

A = {2, 4, 6}

Komplementärereignis

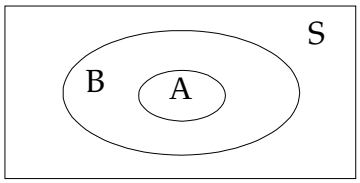
$\overline{A} = \{1, 3, 5\}$



Mengenoperationen

<u>Implikation und Äquivalente</u>

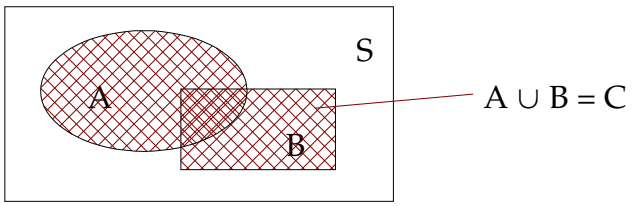
A zieht B nach sich (A impliziert B genau dann, wenn A ⊂ B



A und B sind gleichwertig (äquivalent), genau dann A ⊂ B und B ⊂ A

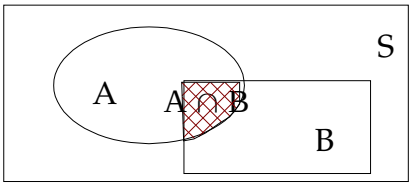
$$A \equiv B$$

<u>Vereinigung von Ereignissen</u>



$$A_1 \cup A_2 \cup A_3 \ldots A_n = \bigcup_{i=1}^{n} A_i$$

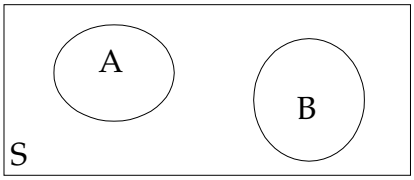
<u>Durchschnitt von Ereignissen</u>



$$A_1 \cup A_2 \cup A_3 \ldots A_n = \bigcap_{i=1}^{n} A_i$$

<u>Ausschließende Ereignisse</u> (disjunkte, unvereinbare Ereignisse)

Zwei Ereignisse heißen disjunkt (sich ausschließend), wenn ihr gleichzeitiges Eintreten unmöglich ist und damit der Durchschnitt die leere Menge enthält. $A \cap B = \emptyset$



Ein Ereignis ist stets zu seinem Komplementärereignis ausschließend (disjunkt).

Ausschließende Ereignisse sind jedoch nicht notwendig komplementär.

Beispiel: einmaliges Werfen eines Würfels

A = {1, 3, 5}, B = {2, 4, 6}

$\Rightarrow A \cap B = \emptyset$

$B \equiv \overline{A}$

$A \equiv \overline{B}$

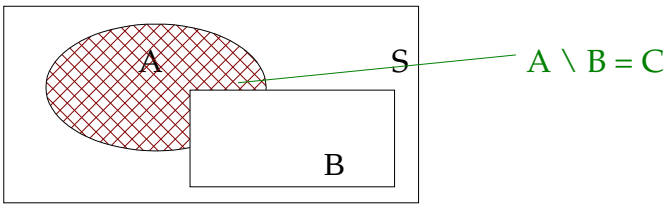
Die Ereignisse A und B sind komplementär und sich ausschließend.

$C = \{1,3\}; D = \{2,4\}$

$\Rightarrow C \cap D = \emptyset$

Die Ereignisse C und D schließen sich aus, sind aber nicht komplementär.

Logische Differenz von Ereignissen



Das Ereignis C ist die logische Differenz der Ereignisse A und B, wenn C das Ereignis charakterisiert, bei dem A, aber nicht B eintritt.

Beispiel: einmaliges Werfen Würfel

A = {1, 2, 3}

B = {3, 4}

$\Rightarrow A \setminus B = C = \{1, 2\}$

$B \setminus A = \{4\}$

Definition der Wahrscheinlichkeit

Die Wahrscheinlichkeit (P) ist Maß zur Quantifizierung der Sicherheit bzw. Unsicherheit über das Eintreten von Ereignissen im Rahmen eines Zufallsexperimentes.

Klassische Definition

$$P(A) = \frac{\textit{Anzahl der Elementarereignisse in A}}{\textit{Anzahl der Elementarereignisse in S}}$$

Beispiel: Würfeln

Ereignis: A = „gerade Augenzahl“

Elementarereignisse: {1}, {2}, {3}, {4}, {5}, {6}

in A enthaltene Elementarereignisse: {2}, {4}, {6}

$$P(A) = \frac{3}{6} = 0,5$$

Häufigkeitsinterpretation, statistischer Wahrscheinlichkeitsbegriff

Hier betrachtet man die Wahrscheinlichkeit P(A) als den Wert, gegen den die relative Häufigkeit des Ereignisses A bei

- unendlich vielen unabhängigen Wiederholungen des Experiments
- unter identischen Bedingungen

konvergiert.

Sei $H_n(A)$ die absolute Häufigkeit des Auftretens von A bei n-maliger Wiederholung des Zufallsexperimentes. Dann gilt für die relative Häufigkeit von A

$$h_n(A) = \frac{H_n(A)}{n}$$

Die Wahrscheinlichkeit von A ist nach dem statischen Wahrscheinlichkeitsbegriff definiert als

$$P(A) = \lim_{n \to \infty} h_n(A)$$

Rechnen mit Wahrscheinlichkeiten

Angaben zu den Beispielen siehe Blätter

Additionssatz der Wahrscheinlichkeiten

$$P(A \cup B) = P(A) + P(B)$$

$$P(A) = \frac{4}{52} = \frac{1}{13} \quad | \quad P(B) = \frac{1}{13}$$

$$P(A \cup B) = P(A) + P(B) = \frac{1}{13} + \frac{1}{13} = \frac{2}{13}$$

nicht ausschließende Ereignisse

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

$$P(A \cup B) = \frac{4}{52} + \frac{13}{52} - \frac{1}{52}$$

Verallgemeinerung des Additionssatzes auf 3 Ereignisse

$$P(A \cup B \cup B) = P(A) + P(B) + P(C) - P(A \cap B) - P(A \cap C) - P(B \cap C) + P(A \cap B \cap C)$$

Komplementäres Ereignis

$$A \cup \overline{A} = S$$

$$P(S) = 1 \Rightarrow P(A) = 1 - P(\overline{A})$$

$$P(A) = 1 - P(\overline{A}) = 1 - \frac{1}{36} = \frac{35}{36}$$

Bedingte Wahrscheinlichkeit

$$P(A/B)=\frac{P(A\cap B)}{P(B)};\quad P(B)>0$$

$$P(A/B)=\frac{\frac{1}{6}}{\frac{3}{6}}=\frac{1}{3}$$

in analoger Weise gilt: (B bei schon eingetroffenem A)

$$P(B/A)=\frac{P(A\cap B)}{P(A)}\quad P(A)>0$$

$$P(B/A)=\frac{\frac{1}{6}}{\frac{1}{6}}=1$$

Bedingte Wahrscheinlichkeit: Beispiel

$B=\{VWL\text{-}Student\}\quad P(B)=0,2$

$A=\{weiblich\}\quad P(A\cap B)=0,12$

$$P(A/B)=\frac{P(A\cap B)}{P(B)}=\frac{0,12}{0,2}=0,6$$

unabhängige Ereignisse

$$P(A/B)=P(A)$$

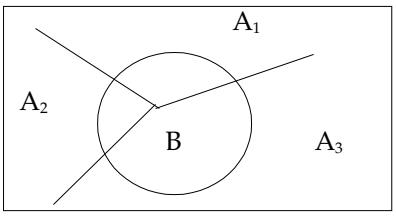
Multiplikationsgesetze

$$P(A\cap B)=P(A)\cdot P(B/A)=P(B)\cdot P(A/B)$$

$$P(A\cap B)=\frac{5}{10}\cdot\frac{4}{9}=\frac{20}{90}=\frac{2}{9}$$

$$P(A\cap B)=P(A)P(B)=\frac{5}{10}\cdot\frac{5}{10}=\frac{25}{100}=\frac{1}{4}$$

Satz von der totalen Wahrscheinlichkeit



$$\bigcup_{i=1}^{n} A_i = S$$

mit $A_i \cap A_j = \emptyset$

für $i \neq j$

$$P(B) = P(A_i \cap B) + P(A_2 \cap B) \ldots P(A_n \cap B)$$

$$P(A_i \cap B) = P(A_i) \cdot P(B/A_i)$$

$$P(B) = \sum_{i=1}^{n} P(A_i) \cdot P(B/A_i)$$

$x: \{\text{defektes Teil}\}$

$$P(x) = P(A) \cdot P(x/A) + P(B) \cdot P(x/B) + P(C) \cdot P(x/C) =$$

$$= 0,6 \cdot 0,02 + 0,3 \cdot 0,04 + 0,1 \cdot 0,06 = 0,03$$

Theorem von Bayes

1. $P(B) > 0$
2. $\bigcup_{i=1}^{n} A_i = S$
3. $A_i \cap A_j = \emptyset$ für $i \neq j$

$$P(A_j/B) = \frac{P(A_j \cap B)}{P(B)} = \frac{P(A_j) \cdot P(B/A_j)}{\sum_{i=1}^{n} P(A_i) \cdot P(B/A_i)}$$

$x: \text{defektes Teil}$

$$P(A/x) = \frac{P(A) \cdot P(x/A)}{P(x)} = \frac{0,6 \cdot 0,2}{0,03} = 0,4$$