Statistik

1. Datenerfassung - Begriffe

Grundgesamtheit	Die Gesamtheit der Individuen oder Objekte, die Gegenstand einer statistischen Untersuchung sind.
Zufallsprinzip	Sicherstellung, dass jedes Element der Grundgesamtheit bei der Erhebung einer Stichprobe ausgewählt werden kann.
Stichprobe	Ausgewählter Teil der Grundgesamtheit
Rohdaten	Die in einer Stichprobe erfassten Daten
Urliste	Liste, in der die Rohdaten der Stichprobe eingetragen werden
Verdichtete Daten	Verarbeitete Rohdaten
Strichliste	Verdeutlichung der Merkmalsausprägung
Merkmalsklasse	Zusammenfassung von unterschiedlichen Merkmalen
Merkmalsausprägung	Spezifische Daten der Merkmalsklasse
Nominalskala	Erfassung eines **qualitativen** Merkmals **ohne** eindeutige Rangfolge
Ordinalskala	Merkmale **mit** eindeutiger Rangfolge
Metrische Skala	Erfassung eines **quantitativen** Merkmals **mit** eindeutiger Rangfolge

2. Häufigkeiten

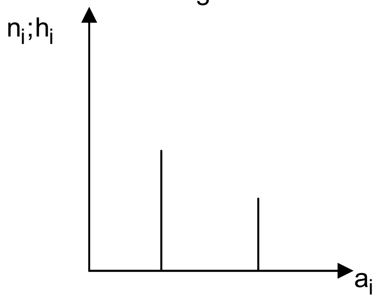
Absolute Häufigkeit n_i	Die Anzahl, mit der eine Merkmalsausprägung a_i vorkommt.	
Gesamtheit n	Summe der absoluten Häufigkeiten $n = \sum_{i=1}^{k} n_i = n_1 + n_2 + \ldots + n_k$	
Relative Häufigkeit $h(a_i)$	$h(a_i) = \frac{n_i}{n}$;	n_i : einzelne Beobachtungswerte n: Gesamtheit der Beobachtungswerte
Summe der relativen Häufigkeiten	Für k Merkmalsausprägungen gilt: $\sum_{i=1}^{k} h(a_i) = 1$	

3. Diagramme

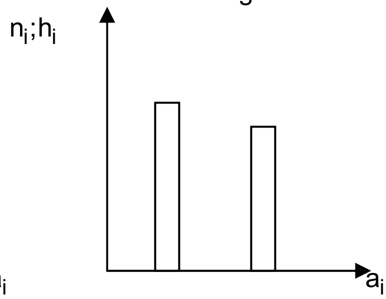
Stabdiagramm

Säulendiagramm

Stabdiagramm

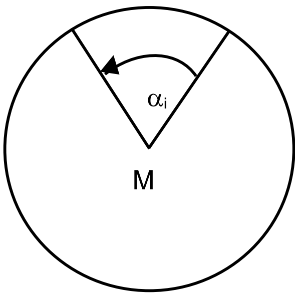


Säulendiagramm



n_i : absolute Häufigkeit
h_i : relative Häufigkeit
a_i : Merkmalsausprägung

Kreisdiagramm:



$\alpha_i = h(a_i) \cdot 360°$

M: Mittelpunkt
α_i: Mittelpunktswinkel
$h(a_i)$: relative Häufigkeit

4. Lagemaße

Arithmetisches Mittel $\bar{x}$

Das arithmetische Mittel $\bar{x}$ ist der Durchschnittswert aller Beobachtungswerte.

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

$$\bar{x} = \frac{1}{n} \cdot (x_1 + x_2 + \ldots + x_n)$$

n: Anzahl der Beobachtungswerte
x_i : Beobachtungswerte
i: Laufvariable von i = 1 bis zur Anzahl n der Beobachtungswerte

Arithmetisches Mittel $\bar{x}$ über relative Häufigkeit

$$\bar{x} = \sum_{i=1}^{k} a_i \cdot h(a_i)$$

$$\bar{x} = a_1 \cdot h(a_1) + a_2 \cdot h(a_2) + \ldots + a_k \cdot h(a_k)$$

a_i : Merkmalsausprägungen
$h(a_i)$: relative Häufigkeiten
k: Anzahl der Merkmalsausprägungen

Zentralwert z **Median x_{med}**	Der Zentralwert z (auch Median x_{med} genannt) ist derjenige Wert, der die geordneten Beobachtungswerte x_i in zwei Hälften teilt. Der Zentralwert steht in der Mitte der Rangwertliste. Fall 1: n ungeradzahlig $z = x_{med} = x_{\frac{n+1}{2}}$ Fall 2: n geradzahlig: $z = x_{med} = \frac{1}{2} \cdot \left(x_{\frac{n}{2}} + x_{\frac{n+2}{2}} \right) = \frac{1}{2} \cdot \left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right)$
Spannweite w	Die Differenz zwischen dem größten Beobachtungswert x_{max} und dem kleinsten Beobachtungswert x_{min} einer Reihe von Beobachtungswerten x_i wird als Spannweite w bezeichnet. $w = x_{max} - x_{min}$
Modalwert x_{mod}	Der Modalwert x_{mod} ist der am häufigsten vorkommende Beobachtungswert.
Verteilung der Lagemaße	linksschiefe Verteilung: $x_{mod} > x_{med} > \overline{x}$ rechtsschiefe Verteilung: $\overline{x} > x_{med} > x_{mod}$ symmetrische Verteilung: $\overline{x} \approx x_{med} \approx x_{mod}$
5. Streumaße	
Quartil $Q_{1;2;3}$ **Quartilsabstand Q**	Werden alle Messwerte einer Stichprobe der Größe nach geordnet und in vier gleiche Bereiche eingeteilt, so werden die drei Grenzen zwischen diesen vier Bereichen als Quartile bezeichnet. **1. Quartil:** $Q_1 = x_{\frac{n+1}{4}}$ **2. Quartil:** $Q_2 = x_{\frac{n+1}{2}} = x_{med}$ **3. Quartil:** $Q_3 = x_{\frac{3n+3}{4}}$ Der Quartilsabstand ist die Breite, die die beiden mittleren Bereiche einnehmen. **Quartilsabstand: $Q = Q_3 - Q_1$**
Mittlere Abweichung d_{med} vom Median	$d_{med} = \frac{1}{n} \cdot \sum_{i=1}^{n} \lvert x_i - x_{med} \rvert$ n: Anzahl der Beobachtungswerte x_i: Beobachtungswerte x_{med}: Median

Mittlere Abweichung e vom arithmetischen Mittel

$$e = \frac{1}{n} \cdot \sum_{i=1}^{n} \left| x_i - \overline{x} \right|$$

$\overline{x}$: arithmetisches Mittel

Varianz v

Die Varianz v ist ein Maß für die Streuung der Messgrößen

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

Standardabweichung s

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

Varianz v für absolute Häufigkeiten

$$v = \frac{1}{n} \cdot \sum_{i=1}^{k} \left(x_i - \overline{x} \right)^2 \cdot n_i$$

k: Anzahl der Klassen
n_i: absolute Häufigkeiten

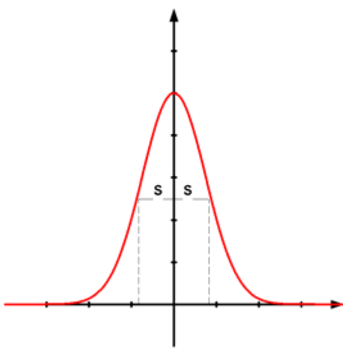
Varianz v für relative Häufigkeiten

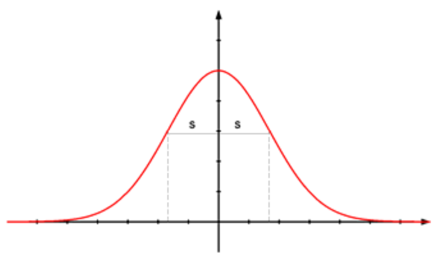
$$v = \sum_{i=1}^{k} \left(x_i - \overline{x} \right)^2 \cdot h_i$$

h_i: relative Häufigkeit

Normalverteilung

Verteilungen, die die Form einer Glockenkurve aufweisen, nennt man Normalverteilungen oder auch Gauß-Verteilungen.
Die Normal- oder Gauß-Verteilung ist ein wichtiger Typ kontinuierlicher Wahrscheinlichkeitsverteilungen.





Die Standardabweichung s beschreibt die Breite der Normalverteilung. Berücksichtigt man die tabellierten Werte der Verteilungsfunktion, gilt näherungsweise folgende Aussage:
Rund 68% aller Werte liegen im Intervall $\left[\overline{x} - s;\ \ \overline{x} + s\right]$,
rund 95% aller Werte liegen im Intervall $\left[\overline{x} - 2s;\ \ \overline{x} + 2s\right]$,
mehr als 99% aller Werte liegen im Intervall $\left[\overline{x} - 3s;\ \ \overline{x} + 3s\right]$.

Stochastik

Zufallsexperiment

- Alle möglichen Ergebnisse sind vorab bekannt
- Einzelne Experimentergebnisse sind zufällig
- Beliebige Wiederholbarkeit unter gleichen Startbedingungen

Häufigkeit

Relative Häufigkeit:
Tritt ein Ereignis E bei einer Versuchsreihe mit n Versuchen genau n_i-mal auf, so wird der Quotient $\frac{n_i}{n}$ als **relative Häufigkeit** des Ereignisses E bezeichnet.

Absolute Häufigkeit:
n_i heißt die **absolute Häufigkeit** des Ereignisses E.

Wahrscheinlichkeit für das Eintreffen eines Ereignisses P(E)

$\mathbf{P(E) = \frac{g}{m}}$ — g: Anzahl der **günstigen** Ereignisse

m: Anzahl der **möglichen** Ereignisse

$0 \le P(E) \le 1$ — M: Ereignisraum

$M=\{ m_1; m_2; \ldots; m_n \}$; — m_i: Elementarereignisse

$\mathbf{\sum_{i=1}^{n} P(m_i) = 1}$ — $P(m_i)$: Wahrscheinlichkeit des Eintreffens eines Elementarereignisses

Gegenereignis $\bar{E}$

$P(\bar{E}) = 1 - P(E)$

$P(E) + P(\bar{E}) = 1$

LAPLACE-Experiment

Sind bei einem Zufallsexperiment alle Ergebnisse des Ereignisraums **gleichwahrscheinlich**, so wird dies als LAPLACE-Experiment bezeichnet.

$\mathbf{P(m_1) = P(m_2) = \ldots = P(m_n)}$;

$P(m_i) = \frac{1}{n}$

Wahrscheinlichkeit des Eintreffens eines Ereignisses

Eintreffen des Ereignisses	
gewiss	$\mathbf{P(E) = 1}$
wahrscheinlich	$\mathbf{1 > P(E) > 0{,}5}$
zweifelhaft	$\mathbf{P(E) = 0{,}5}$
unwahrscheinlich	$\mathbf{0{,}5 > P(E) > 0}$
unmöglich	$\mathbf{P(E) = 0}$

Pfadregel

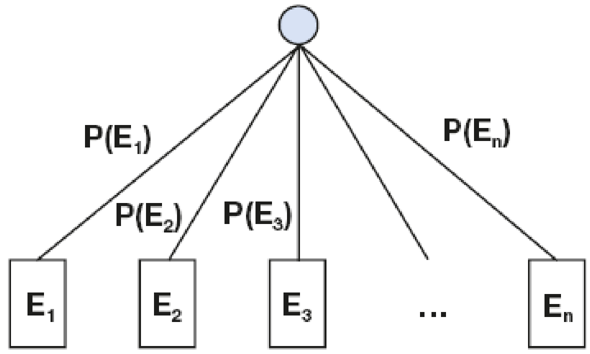
Wahrscheinlichkeit, dass das Ereignis $\mathbf{E_1}$ **oder** das Ereignis $\mathbf{E_2}$ eintritt:

$$\mathbf{P(E_1 \vee E_2) = P(E_1) + P(E_2)}$$

$$P(\bar{E}) = 1 - P(E)$$

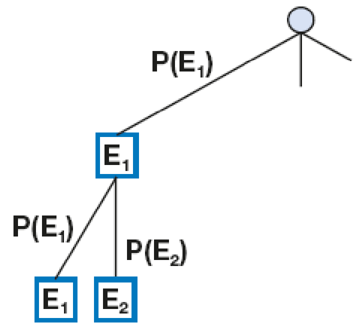
Baumdiagramme

einstufiges Baumdiagramm



Pfadadditionsregel: $P(E_1;...;E_n) = P(E_1) + P(E_2) + ... + P(E_n) = 1$

zweistufiges Baumdiagramm



Pfadmultiplikationsregel $P(E_1;...;E_1) = P(E_1) \cdot ... \cdot P(E_1)$

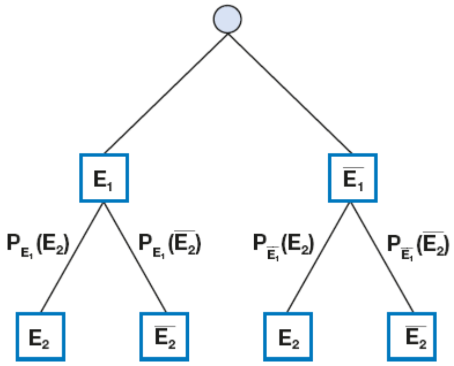
$$P(E_1;E_1) = P(E_1) \cdot P(E_1)$$
$$P(E_1;E_2) = P(E_1) \cdot P(E_2)$$

Bedingte Wahrscheinlichkeit

Wahrscheinlichkeit für ein Ereignis nach dem Eintreffen eines vorhergegangenen Ereignisses

$\mathbf{P_{E_m}(E_n)}$: Wahrscheinlichkeit für das n-te Ereignis, nachdem das m-te Ereignis eingetreten ist.

Baumdiagramm
1. und 2. Stufe



Vierfeldertafel

	E_2	$\overline{E_2}$
E_1	$E_1 \cap E_2$	$E_1 \cap \overline{E_2}$
$\overline{E_1}$	$\overline{E_1} \cap E_2$	$\overline{E_1} \cap \overline{E_2}$

Satz von Bayes

$$P_{E_1}(E_2) = \frac{P(E_1 \cap E_2)}{P(E_1)} \quad \text{mit } P(E_1) \neq 0$$

Inverses Baumdiagramm

Die Ereignisse der ersten und zweiten Stufe gegenüber dem ursprünglichen Baumdiagramm sind vertauscht.

Hypothesentest

Überprüfung des Wahrheitsgehalts einer Hypothese

- Nullhypothese H_0: für wahr angesehene Behauptung

- Alternativhypothese H_A: alternative Theorie zur Nullhypothese

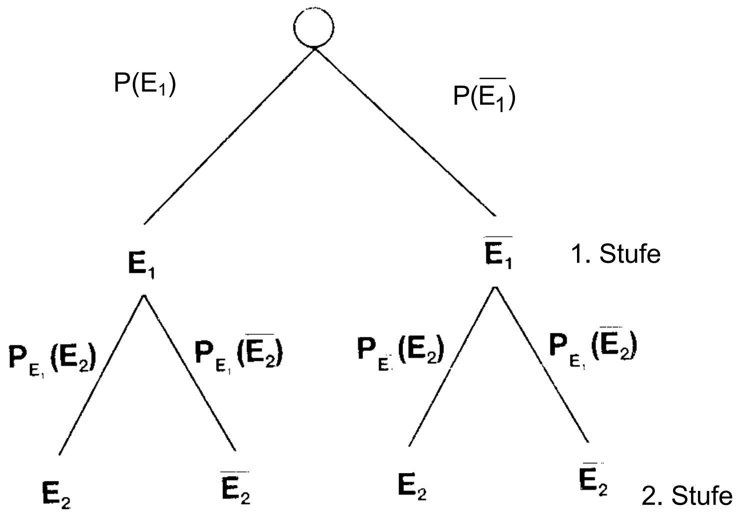
Bei jedem Hypothesentest wird ein Zufallsexperiment in Form einer Stichprobe der Länge **n** durchgeführt. Die **Testgröße Z** gibt die Zahl der Treffer im Stichprobenergebnis an.

Die Grenze zwischen **Annahmebereich A** und **Ablehnungsbereich $\overline{A}$** heißt **kritischer Wert c.**

$A = \{1; \ldots; c\}$;
$\overline{A} = \{c + 1; \ldots; n\}$
Der **Fehler 1. Art** ist der Fehler, eine wahre Hypothese abzulehnen.
Der **Fehler 2. Art** ist der Fehler, eine falsche Hypothese anzunehmen.

Entscheidung beim Hypothesentest		Wahrheit: H_0	Wahrheit: H_A
Ergebnis der Stichprobe	H_0	richtig	Fehler 2.Art
	H_A	Fehler 1.Art	richtig

Binomialverteilung



...
k-te Stufe

Die Wahrscheinlichkeitsberechnung ist abhängig von der Gesamtzahl **n** der Einzelexperimente, der Wahrscheinlichkeit **p** des Eintreffens für ein Einzelexperiment und der Anzahl **k** der Treffer.

$$B(n;p;k) = \binom{n}{k} \cdot p^k \cdot (1-p)^{n-k}$$

für $k \neq 0$: $\binom{n}{k} = \frac{n}{1} \cdot \frac{n-1}{2} \cdot \frac{n-2}{3} \cdot \ldots \cdot \frac{n-k+1}{k}$

für $k = 0$: $\binom{n}{0} = 1$