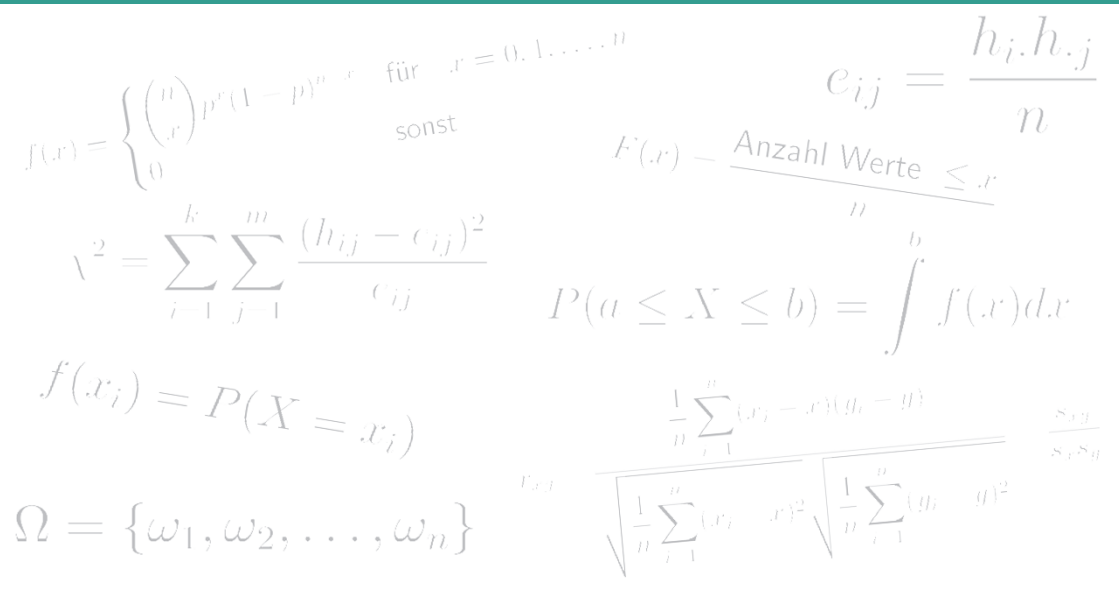
Für Statistiker/-innen… und solche, die es werden wollen oder müssen



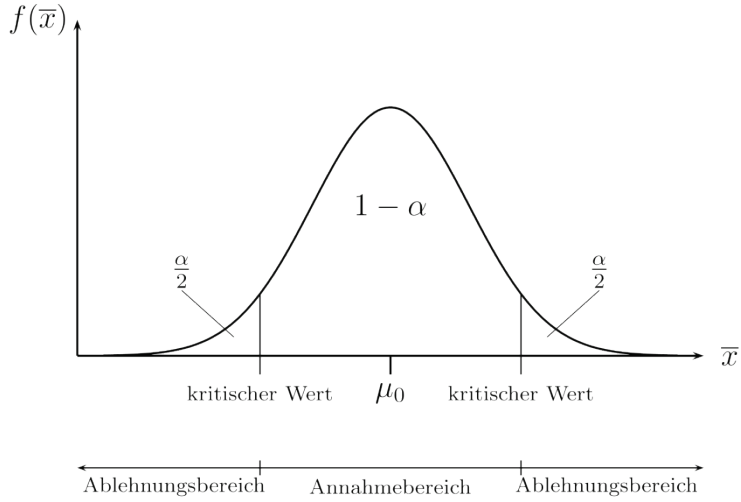
Prof. Dr. Bianca Krol
Prof. Dr. Karsten Lübke

» Wörterbuch Statistik

die wichtigsten Begriffe mit Formeln

FOM Hochschule
ifes
Institut für Empirie & Statistik
der FOM Hochschule
für Oekonomie & Management

Ablehnungsbereich Fällt das Ergebnis eines Hypothesentests in den Ablehnungsbereich, so wird die Nullhypothese verworfen.



Absolute Häufigkeit Absolute Häufigkeit der j-ten Gruppe eines Merkmals oder absolute Häufigkeit der j-ten Merkmalsausprägung einer Häufigkeitsverteilung: Die absolute Häufigkeit gibt als Absolutzahl an, wie oft ein Merkmal bzw. eine Merkmalsausprägung vorkommt.

$$h_j$$

Achsenabschnitt (Regression) Der Achsenabschnitt a einer linearen Regression gibt den Wert für y an der Stelle x=0 an. Der optimale Schätzer für a ist der Mittelwert von Y minus b mal den Mittelwert von X.

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

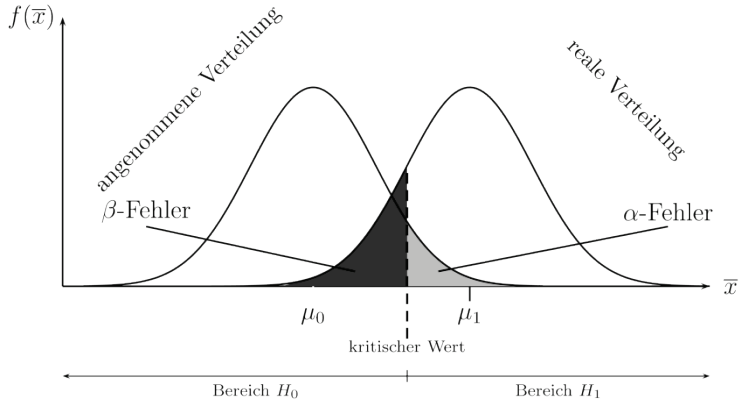
Achsenabschnitt (Zeitreihe) Der Achsenabschnitt a eines linearen Trends gibt den Wert von y zum Zeitpunkt t=0 an.

$$\hat{a} = \bar{y} - \frac{n+1}{2}\hat{b}$$

Additives Zeitreihenmodell Im Additiven Zeitreihenmodell setzt sich der Wert von y zum Zeitpunkt t zusammen aus: Linearer Trend, bestehend aus Achsenabschnitt und Steigung + evtl. Saisoneffekt (zyklisch) + evtl. Rest (Fehler, weitere Effekte wie z.B. Konjunkturzyklus).

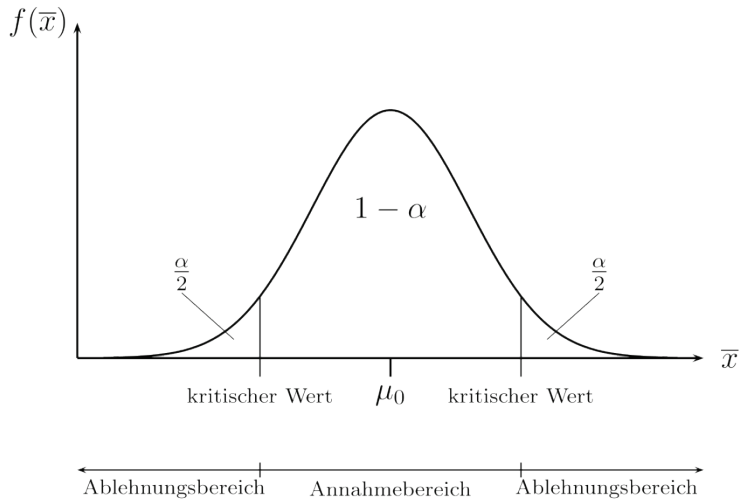
$$y_t = g_t + z_t + r_t$$

Alpha-Fehler Der a-Fehler (auch Fehler 1. Art) wird dann begangen, wenn die Nullhypothese H0 im Rahmen eines Hypothesentests abgelehnt wird, obwohl sie eigentlich richtig ist.



Alternativhypothese Die Alternativhypothese (H1) ist das (logische) Gegenteil der Nullhypothese (H0). Beide Hypothesen werden im Rahmen eines Hypothesentests aufgestellt.

Annahmebereich Fällt das Ergebnis eines Hypothesentests in den Annahmebereich, so wird die Nullhypothese nicht abgelehnt.



Anteilswert Liegt eine dichotome Grundgesamtheit vor, so gibt der Anteilswert an, welcher Anteil der Merkmalsträger in der Grundgesamtheit die interessierende Merkmalsausprägung aufweist.

arithmetisches Mittel Das arithmetische Mittel ist ein Lagemaß für metrische Daten. Das arithmetische Mittel wird oft als Durchschnitt bezeichnet und berechnet sich, indem die Summe der Merkmalswerte aller Daten durch die Gesamtzahl der Daten dividiert wird.

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

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{m} h_i x_i = \sum_{i=1}^{m} f_i x_i$$

Balkendiagramm Das Balkendiagramm zeigt die Häufigkeitsverteilung der Merkmalsausprägungen. Die Höhe der Balken ent-

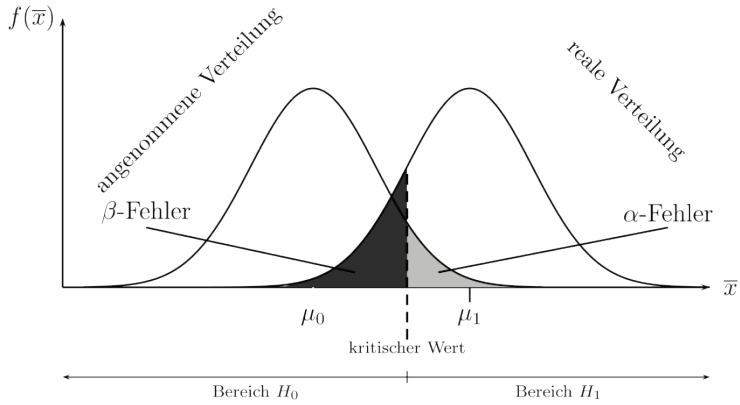
spricht dabei der absoluten oder relativen Häufigkeit der jeweiligen Merkmalsausprägung.

Beobachtung Der Merkmalsträger inklusive der jeweiligen Merkmalsausprägung ist die Beobachtung, die man bei der Datenerhebung macht. Die Beobachtung ist dabei üblicherweise Teil einer Stichprobe. Die Gesamtanzahl der Beobachtungen ist n und ergibt den Umfang der Stichprobe.

Bestimmtheitsmaß Das Bestimmtheitsmaß ist das Quadrat der Korrelation zwischen X und Y. Das Bestimmtheitsmaß nimmt Werte zwischen 0 und 1 an. Je höher es ist, desto besser ist die Modellanpassung einer linearen Regression.

$$R^2 = 1 - \frac{s_{\hat{e}}^2}{s_y^2} = r_{xy}^2$$

Beta-Fehler Der b-Fehler (auch Fehler 2. Art) wird dann begangen, wenn die Nullhypothese H0 im Rahmen eines Hypo-thesentests beibehalten wird, obwohl sie eigentlich falsch ist.



Beziehungszahlen Beziehungszahlen setzen sachlich verschiedene Maßzahlen - bei einem sinnvollen Zusammenhang - ins Verhältnis (z.B. Einkaufstage pro Kunde = Einkaufstage / aktive Kunden).

Binomialverteilung Die Binomialverteilung ist eine diskrete Wahrscheinlichkeitsverteilung, deren Grundlage ein Zufallsexperiment ist, das lediglich zwei sich gegenseitig ausschließende Ergebnisse zulässt (Bernoulli-Experiment). Wenn dieses Zufallsexperiment n-mal unabhängig voneinander wiederholt wird, die Stichprobe von n Elementen also mit Zurücklegen gezogen wird, resultiert die Binomialverteilung. Die Wahrscheinlichkeitsfunktion lautet:

$$f(x) = \begin{cases} \binom{n}{x} p^x (1-p)^{n-x} & \text{für} \quad x = 0, 1, \ldots, n \\ 0 & \text{sonst} \end{cases}$$

Business Intelligence Unter Business Intelligence (Abk. BI) werden Aktivitäten zusammengefasst, die der (zeitnahen) Zusammenführung und der Transformation der operativen Daten und der anschließenden zweckgebundenen Wissensgenerierung in Form von Reports und Expertisen für das Management dienen, um letztlich betrieblichen Mehrwert zu schaffen.

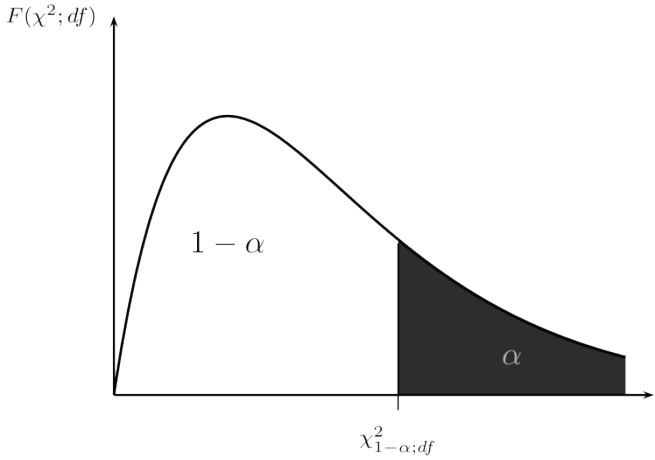
Chi-Quadrat-Koeffizient (Kontingenzkoeffizient nach Pearson) Zusammenhangsmaß innerhalb einer Kreuztabelle. Wird gebildet als Summe über alle Zellen der Kreuztabelle der quadratischen Abweichung der beobachteten Häufigkeiten von den erwarteten Häufigkeiten bei Unabhängigkeit dividiert durch die erwarteten Häufigkeiten bei Unabhängigkeit.

$$\chi^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{(h_{ij} - e_{ij})^2}{e_{ij}}$$

Chi-Quadrat-Test Überbegriff für eine Klasse von statistischen Tests, die auf der Chi-Quadrat-Verteilung aufsetzen. Man unterscheidet zwischen Chi-Quadrat-Unabhängigkeitstest, -Homogenitätstest und -Anpassungstest. Im Rahmen der Tests wird die Abweichung zwischen der Verteilung der vorliegenden Stichprobe und einer theoretisch zu erwartenden Verteilung untersucht.

Chi-Quadrat-Verteilung Verteilung einer Zufallsvariablen. Liegen Zi unabhängige, standardnormalverteilte Zufallsvariablen vor,

dann ist die Summe der quadrierten Zufallsvariablen Zi chi-quadrat-verteilt mit n Freiheitsgraden. Die Chi-Quadrat-Verteilung wird für verschiedene Chi-Quadrat-Tests sowie für die Berechnung von Konfidenzintervallen für Varianzen benötigt.



Clusteranalyse Multivariates Verfahren, mit dem Beobachtungen anhand von (häufig metrischen) Merkmalen gruppiert werden.

Deskriptive Statistik Im Rahmen der deskriptiven Statistik werden erhobene Daten durch möglichst wenige, möglichst aussagefähige Kennzahlen und graphische Darstellungen beschrieben.

Dichtefunktion Für stetige Zufallsvariablen können keine Wahrscheinlichkeiten für das Auftreten einzelner Werte angegeben werden. Hier lassen sich lediglich Wahrscheinlichkeiten für bestimmte Intervalle einer Funktion f(x) angegeben. Diese Funktion wird Dichtefunktion genannt. Aus der Dichtefunktion lässt sich die Verteilungsfunktion einer stetigen Zufallsvariablen erstellen.

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

Diskretes Merkmal Wenn die Anzahl der Merkmalsausprägungen endlich oder zumindest abzählbar unendlich ist, dann liegt ein diskrete Merkmal vor. So sind nominalskalierte Merkmale diskret, aber auch ordinale Merkmale und metrische Merkmale, bei denen etwas gezählt wird (z.B. Anzahl Kinder, Kunden, produzierte Autos), sind diskrete Merkmale.

Diskriminanzanalyse Multivariates Verfahren, mit dem der Einfluss von (häufig metrischen) Merkmalen auf ein nominales Zielmerkmal untersucht wird

Elementarereignis s. Ereignis

$$\omega_i$$

Empirische Verteilungsfunktion Die empirische Verteilungsfunktion F(x) an einer Stelle x ist die Anzahl der Beobachtungen kleiner gleich x dividiert durch die Anzahl Beobachtungen.

$$F(x) = \frac{\text{Anzahl Werte } \leq x}{n}$$

Ereignis Jeder mögliche Ausgang eines Zufallsexperimentes wird Ereignis genannt. Wenn das Zufallsexperiment z.B. das Werfen eines Würfels ist, dann ist die Zahl, die geworfen wird, das Ereignis, welches eingetreten ist. Ereignisse, die sich gegenseitig ausschließen und sich nicht weiter zerlegen lassen heißen Elementarereignisse oder Ergebnis eines Zufallsexperiments. Das Ereignis "ungerade Augenzahl" (also die Zahlen 1,3,5) ist ebenfalls ein mögliches Ereignis des Zufallsexperimentes. Da es aus anderen Ereignissen zusammengesetzt ist, ist es kein Elementarereignis.

$$e_{ij} = \frac{h_{i.} h_{.j}}{n}$$

Erwartungswert Der Erwartungswert einer Zufallsvariablen ist der Wert, der im Durchschnitt als Ausgang eines Zufallsexperimentes erwartet wird. Er kennzeichnet somit das Zentrum der Zufallsver-

teilung. Er entspricht damit dem arithmetischen Mittel in der deskriptiven Statistik. Allerdings bezieht er sich nicht auf gegebene Daten, sondern auf den theoretisch erwarteten Ausgang des Zufallsexperiments.

$$E(X) \quad \text{oder auch} \quad \mu$$

Faktorenanalyse Multivariates Verfahren, um auf der Basis von beobachteten Daten auf wenige zugrundeliegende latente Variablen zu schließen.

Gauß-Markov-Modell Mit Hilfe des Gauss-Markov-Modell lassen sich Regressionsgleichungen mit mehreren unabhängigen Variablen erwartungstreu schätzen.

Glatte Komponente (Zeitreihe) s. Linearer Trend (Zeitreihe)

Gleichverteilung Im Falle einer diskreten Zufallsvariablen liegt eine Gleichverteilung vor, wenn alle möglichen Ausprägungen xi mit derselben Wahrscheinlichkeit P(xi) vorkommen. Beim Werfen eines Würfels beträgt die Wahrscheinlichkeit für alle sechs möglichen Ausgänge des Experimentes 1/6. Im Falle einer stetigen Zufallsvariablen liegt eine Gleichverteilung vor, wenn die Dichtefunktion innerhalb eines endlichen Intervalls konstant und außerhalb des Intervalls Null ist.

$$f(x) = \begin{cases} \frac{1}{n} & \text{für} \quad x_1, x_2, \ldots, x_n \\ 0 & \text{sonst} \end{cases}$$

$$f(x) = \begin{cases} \frac{1}{b-a} & \text{für} \quad a \leq x \leq b \\ 0 & \text{sonst} \end{cases}$$

Gleitender Durchschnitt Beim gleitenden Durchschnitt wird der lokale Trend zum Zeitpunkt t durch den arithmetischen Mittelwert eines Fensters um den Zeitpunkt berechnet. Das Fenster besteht aus den q Beobachtungen vor t und den q Beobachtungen nach t und der Beobachtung t selbst.

$$\hat{g}_t = \frac{1}{2q+1} \sum_{j=-q}^{q} y_{t+j} = \frac{1}{2q+1}(y_{t-q} + \cdots + y_t + \cdots + y_{t+q})$$

Gliederungszahlen Gliederungszahlen sind Anteilswerte, d.h. die Berichtsgröße ist Teil der Basisgröße (z.B. Aktivitätsquote = aktive Kunden / alle Kunden).

Grundgesamtheit Die Grundgesamtheit ist die Menge an Objekten, deren Merkmale untersucht werden, z.B. alle Kunden, Bevölkerung, EU-Länder.

Gruppierung Bei der Gruppierung werden Merkmalsausprägungen zusammengefasst, um die Häufigkeitsverteilung zu ermitteln. Gründe dafür sind z.B. viele verschiedene Merkmalsausprägungen, oder dass Gruppen manchmal leichter zu erheben sind als exakte Werte (z.B. Einkommen zwischen 1000 und 2000 Euro). Die Gruppen (auch Klassen) fassen benachbarte Merkmalsausprägungen zusammen (Intervalle). Die Häufigkeit einer Gruppe ist die Anzahl der Beobachtungen mit Merkmalsausprägungen in dem entsprechenden Intervall.

Häufigkeitsverteilung Die Häufigkeitsverteilung gibt für jede Merkmalsausprägung an, wie oft diese auftritt.

Hauptkomponentenanalyse Multivariates Verfahren in dem hochdimensionale Daten in den Richtungen der größten Streuung linear zusammengefasst werden.

Hypergeometrische Verteilung Wenn eine dichotome Grundgesamtheit vorliegt, aus der eine zufällige Stichprobe von n Elementen ohne Zurücklegen gezogen wird, resultiert als diskrete Wahrscheinlichkeitsverteilung die hypergeometrische Verteilung.

$$f(x) = \begin{cases} \dfrac{\binom{M}{x}\binom{N-M}{n-x}}{\binom{N}{n}} & \text{für} \quad \max\{0, n+M-N\} \le x \le \min\{n, M\} \\ 0 & \text{sonst} \end{cases}$$

Hypothese Eine Hypothese ist eine Vermutung, die man hinsichtlich eines bestimmten Sachverhaltes in einer Grundgesamtheit hat. Das können z.B. Vermutungen über Zusammenhänge oder

Unterschiede zwischen Merkmalen einer Grundgesamtheit sein. Die Hypothese wird im Rahmen eines Hypothesentests überprüft.

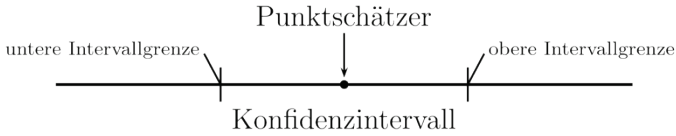
Hypothesentest Mittels eines Hypothesentests wird eine zuvor aufgestellte Hypothese hinsichtlich ihrer Gültigkeit in der Grundgesamtheit überprüft. Dazu wird eine Nullhypothese H0 aufgestellt, die i.d.R. ausdrückt, dass der vermutete Zusammenhang oder Unterschied nicht besteht. Die Alternativhypothese H1 ist das genaue Gegenteil zur in H0 aufgestellten Vermutung. Es wird eine Teststatistik berechnet, die angibt, ob der in den vorliegenden Stichprobendaten zu beobachtende Zusammenhang oder Unterschied die Nullhypothese unterstützt. Das Ergebnis des Hypothesentests ist entweder die Beibehaltung (die Teststatistik überschreitet den kritischen Wert nicht) oder die Ablehnung (die Teststatistik überschreitet den kritischen Wert) der Nullhypothese. Da Hypothesentests immer auf der Grundlage einer Zufallsstichprobe durchgeführt werden, beinhaltet jeder Hypothesentest das Risiko einer Fehlentscheidung. Es gibt zwei Arten von Fehlern, die man beim Durchführen von Hypothesentests begehen kann: den Fehler 1. Art (a-Fehler) und den Fehler 2. Art (b-Fehler).

Indexzahlen Indexzahlen ermöglichen Aussagen über einen zeitlichen oder räumlichen Vergleich einer ganzen Gruppe verschiedener, aber ähnlicher Merkmale.

Indizes s. Indexzahlen

Induktive Statistik Im Rahmen der induktiven Statistik werden die Aussagen einer Stichprobe verallgemeinert und mit Hilfe der vorhandenen Daten werden statistische Schlüsse ermöglicht.

Intervallschätzung Um den aus der Stichprobe berechneten Punktschätzer wird ein Intervall konstruiert, das den zu schätzenden Parameter mit einer vorgegebenen Wahrscheinlichkeit (üblicherweise 90%, 95% oder 99%) überdeckt. Dieses Intervall wird Konfidenzintervall genannt.



Intervallskala s. metrische Skala

Irrtumswahrscheinlichkeit Die Irrtumswahrscheinlichkeit (auch Signifikanzniveau) a gibt die Wahrscheinlichkeit an, mit der man sich bei der Festlegung eines Konfidenzintervalls irrt und der unbekannte Parameter außerhalb des Konfidenzintervalls liegt. Die Irrtumswahrscheinlichkeit ist die Gegenwahrscheinlichkeit zum Konfidenzniveau.

$$\alpha$$

Klassierung s. Gruppierung

Klassifikationsanalyse s. Clusteranalyse und Diskriminanzanalyse

Konfidenz (A→B) Anteil der gemeinsamen Transaktionen von A und B an allen Transaktionen von A. Bedingte Wahrscheinlichkeit von B gegeben A.

$$Konfidenz(A \rightarrow B) = P(B|A)$$

Konfidenzintervall Ergebnis einer Intervallschätzung. Das Konfidenzintervall wird (meist symmetrisch) um einen Punktschätzer gelegt. Die Intervallgrenzen sind abhängig von der Irrtumswahrscheinlichkeit a, der Streuung der Daten sowie dem Stichprobenumfang.

$$P(\text{untere Intervallgrenze} \leq Parameter \leq \text{obere Intervallgrenze}) = 1 - \alpha$$

Konfidenzniveau Das Konfidenzniveau gibt im Rahmen einer Intervallschätzung an, mit welcher Wahrscheinlichkeit der unbekannte Parameter (z.B. der Erwartungeswert in einer Grundgesamtheit) vom Konfidenzintervall überdeckt wird. Das Konfidenzniveau beträgt üblicherweise 90%, 95% oder 99% und wird durch den Term 1-a ausgedrückt. Die Gegenwahrscheinlichkeit a wird als Irrtumswahrscheinlichkeit bezeichnet.

$$1 - \alpha$$

Kontingenzkoeffizient C Der Kontingenzkoeffizient misst den Zusammenhang zweier nominaler Merkmale. C nimmt Werte zwischen 0 und 1 an, wobei die Extremwerte nie erreicht werden. Der Kontingenzkoeffizient ist umso größer, je größer die Abhängigkeit zwischen den Merkmalen ist. Daumenregel: C<0,2 : geringer Zusammenhang, C>0,6 :starker Zusammenhang. Berechnet wird C wie folgt: Wurzel aus dem Pearsonschen Chi-Quadrat dividiert durch Chi-Quadrat plus der Anzahl Beobachtungen.

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

Korrelationsanalyse Die Korrelationsanalyse untersucht den Grad des linearen Zusammenhangs zwischen zwei metrischen Merkmalen. Dabei werden an jedem Merkmalsträger zwei Merkmale erhoben.

Korrelationskoeffizient Der Korrelationskoeffizient nach Bravis-Pearson normiert die Kovarianz mit Hilfe der einzelnen Standardabweichungen. Der Korrelationskoeffizient nimmt Werte zwischen -1 und 1 an. Eine Korrelation größer als Null besagt, dass hohe Werte von x tendenziell mit hohen Werten bei y einhergehen, eine Korrelation kleiner als Null besagt, dass hohe Werte bei x eher mit niedrigen Werte bei y auftreten – und jeweils umgekehrt. „Hoch“ und „niedrig“ werden dabei relativ zum Mittelwert definiert. Je größer der absolute Korrelationskoeffizient desto größer der (lineare) Zusammenhang zwischen den Merkmalen. Der Korrelationskoeffizient ist symmetrisch, d.h. die
Korrelation zwischen den Merkmalen X und Y ist dieselbe wie zwischen Y und X.

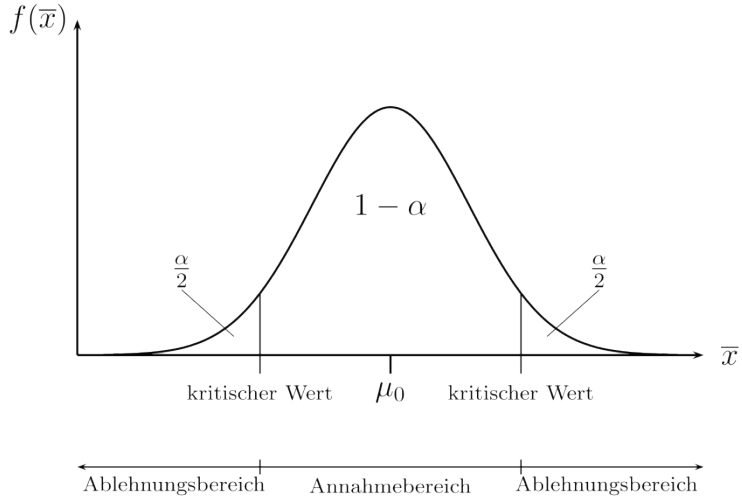
$$r_{xy} = \frac{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}} = \frac{s_{xy}}{s_x s_y}$$

Kovarianz Die Kovarianz misst die durchschnittliche Übereinstimmung in der Streuung von zwei Merkmalen. Sie ist Ausdruck der Stärke des linearen Zusammenhangs.

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

Kreuztabelle Die Kreuztabelle stellt die gemeinsame Verteilung zweier oder mehrerer nominaler Merkmale dar.

kritischer Wert Der kritische Wert bezeichnet bei einem Hypothesentest die Grenze zwischen dem Annahme- und Ablehnungsbereich für die Nullhypothese.



Lagemaß Lagemaße geben den Schwerpunkt innerhalb der Verteilung der Daten an. Zu den wichtigsten Lagemaßen gehören Modus, Median und arithmetisches Mittel.

Lift (A→B) Steigerung der relativen Häufigkeit von B durch den Kauf von A. Quotient aus der bedingten Wahrscheinlichkeit von B gegeben A und der (unbedingten) Wahrscheinlichkeit von B.

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

Lineare Regression Lineares Modell innerhalb der Regressionsanalyse. Im einfachsten Fall hängt ein metrisches Merkmal Y linear über Achsenabschnitt a und Steigung b von einem unabhängigen Merkmal X ab. Da diese Erklärung in der Regel fehlerbehaftet ist, wird im Modell zusätzlich ein Residuum e verwendet.

$$y = a + bx + e$$

Linearer Trend (Zeitreihe) Der lineare Trend einer Zeitreihe setzt sich analog zur linearen Regression zusammen aus Achsenabschnitt (Zeitreihe) und Steigung (Zeitreihe).

$$g_t = a + bt$$

Median Der Median (auch Zentralwert) ist das Element, das bei einer vom kleinsten zum größten Wert geordneten Beobachtungsreihe genau in der Mitte liegt.

$$x_{med} = \begin{cases} x_{(\frac{n+1}{2})} & \text{, falls n ungerade} \\ \frac{1}{2}\left(x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}\right) & \text{, falls n gerade} \end{cases}$$

Mengenindex Laspeyres Der Quotient aus der Summe der Mengen der Berichtsperiode multipliziert mit den Preisen der Basisperiode geteilt durch die Summe der Mengen der Basisperiode multipliziert mit den Preisen der Basisperiode. Dieser Quotient wird mit 100 multipliziert.

$$Q_{0t}^{L} = \frac{\sum_i q_{ti} p_{0i}}{\sum_i q_{0i} p_{0i}} 100$$

Mengenindex Paasche Der Quotient aus der Summe der Mengen der Berichtsperiode multipliziert mit den Preisen der Berichtsperiode geteilt durch die Summe der Mengen der Basisperiode multipliziert mit den Preisen der Berichtsperiode. Dieser Quotient wird mit 100 multipliziert.

$$Q_{0t}^{P} = \frac{\sum_i q_{ti} p_{ti}}{\sum_i q_{0i} p_{ti}} 100$$

Merkmal Das Merkmal ist die Eigenschaft, die im Rahmen einer statistischen Untersuchung analysiert wird, z.B. Umsatz, Kosten, Anzahl Mitarbeiter, Alter, Benotung, Geschlecht.

Merkmalsausprägung Die Merkmalsausprägung sind die jeweiligen Werte, die das Merkmal annehmen kann, z.B. 50,00€, 45 Jahre, weiblich, Schulnote 4.

Merkmalsträger Die Merkmalsträger sind die Objekte, die untersucht werden, z.B. Kunden, Filialen, Produkte, Zeitungen.

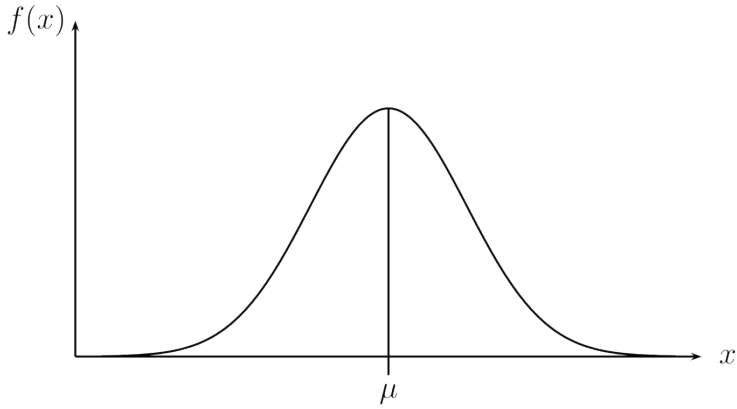
Messzahlen Messzahlen entstehen, wenn sachlich gleiche, aber zeitlich oder örtlich verschiedene Merkmalswerte aufeinander bezogen werden (z.B. Veränderung Umsatz = Umsatz 2010 / Umsatz 2009).

Metrische Skala Wenn sowohl die Rangordnung als auch die Abstände zwischen Merkmalsausprägungen bestimmbar sind, dann liegen metrisch skalierte Daten vor: z. B. Umsätze (100T€ sind 40T€ mehr als 60T€), Einkommen (4T€ sind doppelt so viel wie 2T€), Preise, Temperatur.

Modus Der Modus (auch Modalwert) ist derjenige Wert, der in den vorliegenden Daten am häufigsten vorkommt.

Nominalskala Unterschiedliche Merkmalsausprägungen ohne Rangordnung nennt man nominalskaliert: z.B. Farbe (gelb, grün, blau...), Familienstand (ledig, verheiratet ...).

Normalverteilung Die Normalverteilung ist die wohl wichtigste Verteilung in der Statistik. Es handelt sich um eine stetige Verteilung, die lediglich von den Parametern m (Erwartungswert) und s (Standardabweichung) abhängt. Die Dichtefunktion ist eine glockenförmige Funktion, deren Breite durch s festgelegt wird. Sie ist symmetrisch zum Erwartungswert m und das Maximum der Funktion liegt an der Stelle x=m. Überführt man eine normalverteilte Variable in eine standardnormalverteilte Variable, dann findet man die Verteilungsfunktion tabelliert und kann somit sehr einfach die Wahrscheinlichkeiten ablesen.



Nullhypothese Die Nullhypothese (H0) ist eine der beiden benötigten Hypothesen, die einem Hypothesentest zugrunde liegen. (Die andere wird Alternativhypothese genannt). Die Nullhypothese ist die Hypothese, die im Rahmen des Hypothesentests überprüft wird. Man unterscheidet zwischen ein- und zweiseitigen Nullhypothesen. Mit einer zweiseitigen Nullhypothese wird überprüft, ob ein Parameter einen bestimmten Wert besitzt. Mit einer einseitigen Nullhypothese wird überprüft, ob ein Parameter einen bestimmten Wert nicht unter- oder überschreitet.

Ordinalskala Liegt eine sinnvolle Rangordnung zwischen verschiedenen Merkmalsausprägungen vor, so spricht man von

ordinalskalierten Daten: z.B. Schulnote (1, 2, 3,...), Güteklasse bei Lebensmitteln.

p-Quantil Das p-Quantil einer Stichprobe ist diejenige Zahl, x(p) so dass n·p Beobachtungen kleiner oder gleich dem p-Quantil sind und der Rest der Beobachtungen ist größer oder gleich x(p). Der Median ist das 0,5-Quantil. Weitere wichtige Quantile sind die sogenannten Quartile, welche die Beobachtungen im Verhältnis 0,25:0,75 bzw. 0,75:0,25 teilen.

Preisindex Laspeyres Der Preisindex nach Laspeyres zeigt wie viel der Warenkorb der Basisperiode in der Berichtsperiode kostet, d.h. der Quotient der hypothetischen Gesamtausgaben der Berichtsperiode geteilt durch die tatsächlichen Gesamtausgaben der Basisperiode. Dieser Quotient wird mit 100 multipliziert.

$$P_{0t}^{L} = \frac{\sum_i q_{0i} p_{ti}}{\sum_i q_{0i} p_{0i}} 100$$

Preisindex Paasche Der Preisindex nach Paasche ist der Quotient der tatsächlichen Gesamtausgaben in der Berichtsperiode geteilt durch die fiktiven Ausgaben der Basisperiode für den Warenkorb der Berichtsperiode. Dieser Quotient wird mit 100 multipliziert.

$$P_{0t}^{P} = \frac{\sum_i q_{ti} p_{ti}}{\sum_i q_{ti} p_{0i}} 100$$

Primärstatistik Speziell für die statistische Analyse erhobenes Datenmaterial wird als Primärstatistik bezeichnet. Die Erhebung erfolgt z.B. durch Beobachtung, Befragung, Experiment.

Prüfgröße s. Testgröße

Punktprognose (Regression) Die Prognose von y für einen (neuen) Wert x ist die Summe aus dem Schätzer für den Achsenabschnitt a und x mal dem Schätzer für die Steigung b. Der Wert für das unabhängige Merkmal x ist dabei entweder gesteuert (z.B. Marketingetat) oder ist aus anderen Gründen bekannt (z.B. Nachfrage).

p-Wert Anstatt bei einem Hypothesentest den kritischen Wert mit der Testgröße zu vergleichen, um eine Entscheidung über die Annahme oder die Ablehnung der Nullhypothese treffen zu können, kann auch der p-Wert betrachtet werden. Dieser wird von vielen Statistik-Programmen automatisch ausgegeben. Der p-Wert ist die Wahrscheinlichkeit, bei einer zutreffenden Nullhypothese die aus der Stichprobe berechnete Testgröße (oder einen unter H0 noch unwahrscheinlicheren Wert) zu erhalten. Die Nullhypothese wird abgelehnt, wenn der p-Wert kleiner ist als das zuvor festgelegte Signifikanzniveau a. Je kleiner der p-Wert ist, desto eher wird die Nullhypothese verworfen.

Quartile s. p-Quantil

Randhäufigkeit Die Randhäufigkeit ist die Zeilen- oder Spaltensumme einer Kreuztabelle.

$$h_{i\cdot} = \sum_{j=1}^{m} h_{ij} \text{ bzw. } h_{\cdot j} = \sum_{i=1}^{k} h_{ij}$$

Rang Der Rang gibt die Position einer Merkmalsausprägung in der aufsteigend sortierten Liste an.

$$r_{sp} = \frac{\frac{1}{n}\sum_{i=1}^{n}(R_{xi} - \bar{R}_x)(R_{yi} - \bar{R}_y)}{\sqrt{\frac{1}{n}\sum_{i=1}^{n}(R_{xi} - \bar{R}_x)^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}(R_{yi} - \bar{R}_y)^2}}$$

Regressionsanalyse Multivariates Verfahren, mit dem der Einfluss von metrischen Merkmalen auf ein metrisches Zielmerkmal untersucht wird.

Relative Häufigkeit Relative Häufigkeit der j-ten Gruppe eines Merkmals oder relative Häufigkeit der j-ten Merkmalsausprägung einer Häufigkeitsverteilung: Die relative Häufigkeit gibt als Prozentzahl an, wie oft ein Merkmal bzw. eine Merkmalsausprägung vorkommt.

$$f_j = \frac{h_j}{n}$$

Relative Summenhäufigkeit Die aufsummierte relative Häufigkeit ist die relative Summenhäufigkeit.

$$F_j = f_1 + f_2 + \ldots + f_j$$

Residuum (Regression) Das Residuum ist der Fehlerterm einer linearen Regression oder der Prognosefehler bei der Anwendung einer linearen Regression.

$$\hat{e}_i = y_i - \hat{y}_i$$

Saisonbereinigung Korrektur eines additiven Zeitreihenmodells um den Saisoneffekt.

$$\hat{y}_t^{\,s} = y_t - \hat{z}_t$$

Saisoneffekt Saisonale Komponenten können auf mehreren Ebenen zu finden sein, z.B. quartalsweise oder monatlich. Dabei werden folgende Annahmen verwendet: Die saisonale Komponente ist konstant über die Zeit, d.h. zum Beispiel der Quartalseffekt ist über die Jahre gleich und die saisonalen Komponenten heben sich über die Zeit auf (z.B. innerhalb eines Jahres). Jeder Zyklus bestehe dabei aus k Perioden, so dass insgesamt m=n/k Zyklen vorliegen.

$$\hat{z}_t = \hat{z}_{t+k} = \hat{z}_{t+2k} = \cdots = \frac{1}{m} \sum_{j=0}^{m-1} (y_{t+jk} - \hat{g}_{t+jk})$$

Schätzer Mittels eines Schätzers (auch Schätzfunktion) wird aus den Stichprobendaten eine (angenäherter) Wert für den unbekannten Parameter (z.B. Erwartungswert) einer Grundgesamtheit bestimmt. Man unterscheidet Punkt- und Intervallschätzungen.

Schätzfehler Der Schätzfehler im Rahmen einer Intervallschätzung entspricht der halben Breite des (symmetrischen) Konfidenzintervalls. Er berechnet sich als Produkt aus dem kritischen Wert und Standardfehler.

$$e = z_{1-\frac{\alpha}{2}} \cdot \frac{s}{\sqrt{n}}$$

Schätzfunktion s. Schätzer

Sekundärstatistik Statistische Daten und Auswertungen von zu anderen Zwecken erhobene Daten (z.B. aus operativen Prozessen wie Kassensysteme, Produktionsdaten) bezeichnet man als Sekundärstatistik.

Signifikanzniveau s. Irrtumswahrscheinlichkeit

Signifikanztest s. Hypothesentest

Skala Die Skala (auch Skalierung) der Daten gibt das Mess- oder Informationsniveau der Merkmalsausprägungen an: nominal, ordinal, metrisch.

Standardabweichung Die Standardabweichung ergibt sich aus der Wurzel der Varianz. Dadurch liegt ein Streuungsmaß in derselben Dimension wie die Ausgangsdaten vor.

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

Standardfehler Als Standardfehler wird die Streuung der Stichprobenkennwerte um den unbekannten Parameter der Grundgesamtheit bezeichnet. Wird z.B. der Erwartungswert einer Grundgesamtheit gesucht, dann liefern verschiedene Stichproben aus dieser Grundgesamtheit verschiedene Mittelwerte, die als Schätzer verwendet werden können. Die Streuung zwischen diesen Mittelwerten wird durch den Standardfehler ausgedrückt.

$$\hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Standardnormalverteilung Die Normalverteilung mit dem Erwartungswert m=0 und der Standardabweichung s=1 wird Standardnormalverteilung genannt. Jede normalverteilte Variable X

lässt sich durch Transformation in die standardnormalverteilte Variable Z überführen (z-Transformation).

$$Z = \frac{X - \mu}{\sigma}$$

Steigung (Regression) Die Steigung b einer linearen Regression gibt den Wert der Veränderung von y an, wenn x um eins erhöht wird. Der optimale Schätzer für b ist die Kovarianz von X und Y geteilt durch die Varianz von X.

$$\hat{b} = \frac{s_{xy}}{s_x^2}$$

Steigung (Zeitreihe) Die Steigung b eines linearen Trends gibt den Wert der Änderung von y an, wenn sich t um eins erhöht. Dabei werden aber noch keine Saisoneffekte etc. berücksichtigt.

$$\hat{b} = \frac{s_{ty}}{s_t^2} = \frac{\frac{1}{n}\sum_{t=1}^{n} t(y_t - \bar{y})}{\frac{n^2-1}{12}}$$

Stetiges Merkmal Wenn (zumindest theoretisch) zwischen zwei Merkmalsausprägungen unendlich viele verschiedene Zwischenwerte liegen (reelle Zahlen, z.B. Umsatz, Größe, Temperatur, Alter), dann liegt ein stetiges Merkmal vor.

Streudiagramm In einem Streudiagramm werden die zwei metrischen Merkmale jedes Merkmalsträgers zusammen abgetragen. Das Streudiagramm ermöglicht einen Eindruck des Zusammenhangs zweier metrischer Merkmale (Korrelationsanalyse, Regressionsanalyse).

Streuungsmaß Während die Lagemaße die Mitte des Datensatzes beschreiben, machen die Streuungsmaße deutlich, wie weit die einzelnen Werte von dieser Mitte entfernt sind. Die Varianz und die daraus abgeleitete Standardabweichung sind die prominentesten Streuungsmaße in der Statistik.

Supervised Learning s. Diskriminanzanalyse

Support (A→B) Anteil der gemeinsamen Transaktionen von A und B an allen Transaktionen. Wahrscheinlichkeit von A und B.

$$Support(A \rightarrow B) = P(A \text{ und } B)$$

Teilerhebung Wenn lediglich eine Stichprobe der Grundgesamtheit untersucht wird, wird eine Teilerhebung durchgeführt.

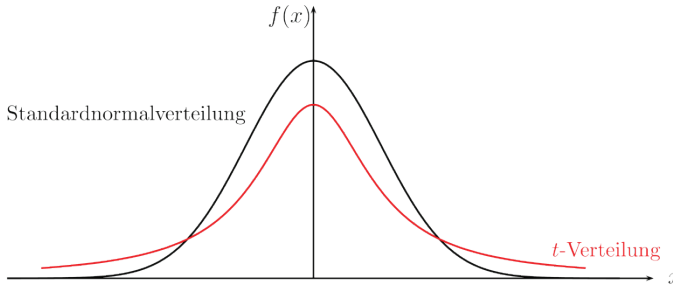
Test s. Hypothesentest

Testgröße Die Testgröße ist eine Stichprobenfunktion, die im Rahmen eines Hypothesentests verwendet wird, um eine Entscheidung bezüglich der Gültigkeit der Nullhypothese zu treffen. Bei Normalverteilung sieht die Testgröße für den Erwartungswert wie folgt aus:

$$T = \frac{\bar{X} - \mu}{\hat{\sigma}_{\bar{x}}}$$

Teststatistik s. Testgröße

t-Verteilung Die t-Verteilung (auch Student-Verteilung) ist eine stetige Verteilung, die der Standardnormalverteilung ähnelt. Sie liegt symmetrisch um den Nullpunkt, verläuft aber etwas flacher und nähert sich mit zunehmendem Stichprobenumfang der Standardnormalverteilung an. Die t-Verteilung kommt vor allem bei Schätz- und Testverfahren zum Erwartungswert einer Grundgesamtheit zum Einsatz, wenn deren Varianz nicht bekannt ist.



Umsatzindex Der Umsatzindex ist der Quotient der realen Ausgaben der Berichtsperiode geteilt durch die realen Ausgaben der Basisperiode. Dieser Quotient wird mit 100 multipliziert.

$$U_{0t} = \frac{\sum_i q_{ti} p_{ti}}{\sum_i q_{0i} p_{0i}} 100$$

Unabhängigkeit Zwei Merkmale sind voneinander unabhängig, wenn die Kenntnis des einen Merkmals keine Informationen über die Verteilung des anderen Merkmals liefert. Wenn zwei Merkmale unabhängig sind, dann ist die Wahrscheinlichkeit von A und B ist die Wahrscheinlichkeit von A mal der Wahrscheinlichkeit von B.

$$P(A \text{ und } B) = P(A) \cdot P(B)$$

Unsupervised Learning s. Clusteranalyse

Varianz Die Varianz misst die Streuung von metrischen Daten. Dazu werden zunächst die Abweichungen der Beobachtungswerte zum arithmetischen Mittel quadriert. Diese Abstandsquadrate werden dann addiert und schließlich durch die Gesamtzahl der Daten dividiert. Durch die Quadrierung wird auch die Dimension der Daten quadriert. Daher berechnet man üblicherweise im Anschluss die Standardabweichung. Wenn die Grundgesamtheit bekannt ist, spricht

man von der deskriptiven Varianz, die berechnet werden kann. Ist die Grundgesamtheit unbekannt, berechnet man die empirische Varianz, die sich auf eine Stichprobe bezieht. Die empirische Varianz liefert eine asymptotisch erwartungstreue Schätzung für eine unbekannte Varianz aus der nicht zu erfassenden Grundgesamtheit.

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

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

$$s^2 = \sum_{i=1}^{m}(x_i - \bar{x})^2 \cdot f_i = \sum_{i=1}^{m}f_i x_i^2 - \bar{x}^2$$

Varianzanalyse Multivariates Verfahren, mit dem der Einfluss von nominalen Merkmalen auf ein metrisches Zielmerkmal untersucht wird.

Varianz-Kovarianz-Matrix Die Matrix aller Kovarianzen zwischen unabhängigen Variablen und ihren Varianzen im Rahmen einer linearen Regression wird als Varianz-Kovarianz-Matrix bezeichnet. Sie ergibt sich aus der Multiplikation der Elemente der inversen quadratischen Koeffizientenmatrix mit der Varianz der Residuen.

Variationskoeffizient Der Variationskoeffizient berechnet sich, indem die Standardabweichung durch das arithmetische Mittel dividiert wird. Dadurch erhält man eine dimensionslose Größe, die prozentual zu interpretieren ist und sich insbes. für den Vergleich von Daten mit unterschiedlichen Dimensionen anbietet.

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

Verbraucherpreisindex Der Verbraucherpreisindex ist ein Preisindex nach Laspeyres. Der Warenkorb besteht aus über 700 Güterarten und monatlich werden dafür über 300.000 Einzelpreise erfasst.

Verbundkaufanalyse s. Assoziationsanalyse

Verhältniszahl Eine Verhältniszahl ist der Quotient zweier Maßzahlen. Die Maßzahl im Zähler wird Berichtsgröße, die im Nenner Basisgröße genannt.

Verteilungsfunktion Analog zur (relativen) Summenhäufigkeit in der deskriptiven Statistik lassen sich auch die Wahrscheinlichkeiten für das Auftreten bestimmter Werte von diskreten Zufallsvariablen summieren. Dadurch erhält man die Verteilungsfunktion einer diskreten Zufallsvariablen. Sie gibt die Wahrscheinlichkeit dafür an, dass ein bestimmter Wert x nicht überschritten wird. Im Falle stetiger Zufallsvariablen ergibt sich die Verteilungsfunktion nicht durch Summation, sondern durch Integration. Damit entspricht die Verteilungsfunktion stetiger Zufallsvariablen dem Flächeninhalt unter dem Graphen der Dichtefunktion von -¥ bis zum Wert x. Die Verteilung gibt somit - wie im stetigen Fall auch - die Wahrscheinlichkeit dafür an, dass ein bestimmter Wert x nicht überschritten wird.

$$F(x) = P(X \leq x) = \sum_{x_i \leq x} f(x_i)$$

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)dt$$

Vertrauensintervall s. Konfidenzintervall

Vertrauensniveau s. Konfidenzniveau

Vertrauenswahrscheinlichkeit s. Konfidenzniveau

Vollerhebung Wenn die Grundgesamtheit untersucht wird, dann spricht man von einer Vollerhebung.

Wahrscheinlichkeit Der Ausgang eines Zufallsexperimentes ist ungewiss. Den Ereignissen können aber gewisse Wahrscheinlichkeiten zugeordnet werden. Diese quantifizieren die Chancen des Eintretens der verschiedenen Ereignisse. Wird z.B. ein Würfel geworfen, so ist die Wahrscheinlichkeit für das Auftreten der Augenzahlen 1 bis 6 gleichwahrscheinlich und beträgt jeweils 1/6. Es gibt verschiedene Definitionsansätze für den Begriff Wahrscheinlichkeit. Wenn alle Elementarereignisse die gleiche Wahrscheinlichkeit aufweisen, kann die klassische Definition nach Laplace verwendet werden. Diese wird wie folgt ausgedrückt:

$$P(A) = \frac{A}{\Omega} = \frac{\text{Anzahl der für A günstigen Ereignisse}}{\text{Anzahl aller möglichen Ereignisse}}$$

Wahrscheinlichkeitsdichte s. Dichtefunktion

Wahrscheinlichkeitsfunktion Die Wahrscheinlichkeitsfunktion f(x) weist jeder Ausprägung xi einer diskreten Zufallsvariablen eine Wahrscheinlichkeit P(xi) für deren Auftreten zu. So ist die Wahrscheinlichkeit für das Werfen einer 4 bei einem Würfelwurf 1/6. Aus der Wahrscheinlichkeitsfunktion lässt sich die Verteilungsfunktion einer diskreten Zufallsvariablen erstellen.

$$f(x_i) = P(X = x_i)$$

Wahrscheinlichkeitsverteilung s. Wahrscheinlichkeitsfunktion

Zeitreihenanalyse Bei der Zeitreihenanalyse wird die Entwicklung eines Merkmals y über die Zeit t untersucht. Häufig werden für t die Werte t=1,…,n verwendet (bei äquidistanten Zeitpunkten).

zentraler Grenzwertsatz Zieht man n unabhängige Stichprobenwerte xi einer Zufallsvariablen X, dann konvergiert die Verteilung des Stichprobenmittelwertes mit zunehmendem n gegen die Normalverteilung. Da dies unabhängig von der ursprünglichen Verteilung der Zufallsvariablen X gilt, kommt der Normalverteilung in der Statistik eine überragende Bedeutung zu.

Zufällige Stichprobe Eine Zufallsstichprobe resultiert daraus, dass jeder Merkmalsträger die gleiche Wahrscheinlichkeit hat, erhoben zu werden. Daher gibt es keine verzerrenden Effekte.

Zufallsexperiment Ein Vorgang, der zumindest gedanklich beliebig oft wiederholt werden kann und mindestens zwei mögliche Ausgänge hat, die im Voraus nicht bestimmbar sind, wird Zufallsexperiment (auch Zufallsvorgang) genannt. Als Beispiele werden oft das Werfen eines Würfels oder die Ziehung der Lottozahlen angeführt.

Zufallsvariable Eine Zufallsvariable ist definiert als numerisches Ergebnis eines Zufallsexperiments. Wie bei den statistischen Merkmalen unterscheidet man zwischen diskreten und stetigen Zufallsvariablen. Bei diskreten Zufallsvariablen wird mittels einer Wahrscheinlichkeitsfunktion f(x) jedem Wert xi eine Wahrscheinlichkeit P(xi) für das Auftreten dieses Wertes zugewiesen. Für stetige Zufallsvariablen können keine Wahrscheinlichkeiten für das Auftreten einzelner Werte angegeben werden. Hier lassen sich lediglich Wahrscheinlichkeiten für bestimmte Intervalle einer Funktion f(x) angegeben. Diese Funktion wird Dichtefunktion genannt.Sowohl aus der Wahrscheinlichkeitsfunktion als auch aus der Dichtefunktion lässt sich die Verteilungsfunktion einer Zufallsvariablen erstellen.

Zufallsvorgang s. Zufallsexperiment

„Wir ertrinken in Informationen, aber wir hungern nach Wissen … „
John Naisbitt

Statistik hilft uns u. a.
- Stärken und Schwächen zu erkennen und zu analysieren
- Potenziale und Entwicklungen einzuschätzen
- Zusammenhänge zu erkennen und zu nutzen
- besser zu sein als der Wettbewerb

Vor diesem Hintergrund möchten wir Ihnen mit dieser Sammlung die wichtigsten Begriffe und Formeln der Statistik an die Hand geben. Wir wünschen Ihnen bei Ihrem Forschungsvorhaben viel Erfolg und hoffen, dass Sie dieses Werk ein wenig dabei unterstützen kann.

Prof. Dr. Bianca Krol & Prof. Dr. Karsten Lübke
ifes Institut für Empirie & Statistik
www.fom-ifes.de

FOM Hochschule | ifes
Institut für Empirie & Statistik
der FOM Hochschule
für Oekonomie & Management