Statistik Modul 1: Einführung in die Statistik

Statistik 4-5k Jahre alt. Mathematik kam erst vor 300 Jahren dazu,
Ausgangspunkt: (Statts-) Verwaltung/Management von gr. Projekten
Stat. = lat. Statisticum „den Staat befreffend“ ital. Statista = Staatsmann

In welchen Gebieten benötigt man die Statistik?

In den Empirischen Wissenschaften: (Real- bzw. Erfahrungswissenschaft.)
- Natur-, Sozial-, Ingeniers-, Verhaltenswissensch., Biologie / Medizin

Ziel: Gewinn neuer Erkenntnisse durch Ausschnitt der Realität. Dazu durchführung empirischer Untersuchungen, Auswertung der Daten

Was ist eine Statistik?

- systematische Zusammenstellung von Zahlen und Daten
- **Wozu?** – Beschrei. Bestim. Zustände, Entwicklungen / Phänomene
- **Ziel:** Gewinnung Inform. Aus unübersichtl. / unstrukt. Datenmengen

Statistik: Lehre von Verfahren und Methoden zur Gewinnung, Erfassung, Analyse, Charakterisierung, Abbildung, Nachbildung und Beurteilung von beobachtbaren Daten über die Wirklichkeit.

Gegenstand der Statistik: Datengewinnung / -erhebung / Quellen:
Amtliche Erhebungen, Berichte, Umfragen, betr. Quellen

Datenerhebung: Vorgang zur Ermittlung und zur Erfassung von Ausprägungen eines statistischen Merkmals
Primärerhebung: Erhebung neuer Daten nach Vorgaben
Sekundärerhebung: aus bereits vorhandenem Datenmaterial
Vollerhebung: Untersuchung aller statistischen Einheiten e. Gesamtheit
Teilerhebung: $n < N$
Datenanalysen: Anwend. stat. Verfahren zum Zweck d. Erkenntnisgew.
Datencharakteriierung: graf. / tabeller. Darstellung von Daten sowie berechnung von zusamm.f., den emp. Sachverhalt beschr. Kennzahl

Vor- und Nachteil einer „Statistik“ in tabellarischer Darstellung und einer „Statistik“ in graphischer Darstellung:

- Tabellarische Darstellung:

Vorteil: liefert detailliertere Informationen, man kennt die genauen Werte das ist insbesondere bei Planungsaufgaben wichtig.
♣ Nachteil: Tabellen sind schwerer zu lesen, man braucht Zeit, um die Information zu verarbeiten. Tabellen sind „langweilig“.
♣ Graphische Darstellung:
♣ Vorteil: Man kann sich sehr schnell ein Bild von den quantitativen Verhältnissen machen, man erkennt sehr schnell die wesentlichen Informationen (wenn das Diagramm gut gestaltet ist ...).
♣ Nachteil: Nur mit Mühe lassen sich genaue Werte ablesen.

STICHPROBE

Grundgesamtheit ◊ die Menge aller möglichen Erhebungseinheiten
Stichprobe ◊ eine n-elementige Teilmenge der Grundgesamtheit mit N Elementen (Merkmalsträgern)
Ein **Auswahlverfahren** ist die Art und Weise, wie die Elemente der Stichprobe möglichst zweckmäßig ausgewählt werden.

ZUFALLSSTICHPROBE

Einfache Zufallsstichproben: jede mögliche Stichprobe und auch jedes Element besitzen dieselbe Chance ausgewählt zu werden. Dies ist dann eine echte Zufallsstichprobe (meist unrealistisch), der Idealfall einer Stichprobe. Sie ist ein genaues Abbild der Grundgesamtheit, so dass der Schluss von der Stichprobe auf die Grundgesamtheit gewährleistet ist.
Geschichtete Zufallsstichproben: Die Elemente der Grundgesamtheit werden so in Gruppen (Schichten, strata) eingeteilt, dass jedes Element der Grundgesamtheit zu einer – und nur zu einer– Schicht gehört. Danach werden einfache Zufallsstichproben aus jeder Schicht gezogen.

DIE 5 D´S

♣ **Definition (Phase 1)**
♣ Definition des Informationsbedarfs, der Hypothesen, der Begriffe, der Untersuchungseinheiten, über die man Information haben will. Nur durch eindeutige und verständliche Formulierung der Zielsetzung kann gewährleistet werden, dass wirklich das erforscht wird, was erforscht werden soll!
♣ **Design-Entscheidungen (Phase 2)**
♣ **Abgrenzung der Grundgesamtheit, evtl. Stichprobenumfang**
♣ **Erhebungsart:**
- Querschnitt- oder Längsschnittuntersuchung
- Primärerhebung oder Sekundärerhebung
- Vollerhebung oder Teilerhebung

Erhebungstechnik:
- Befragung (persönlich, telefonisch, schriftlich oder online)
- Beobachtung (offen oder verdeckt), - Dokumentenanalyse

♣ **Datenerhebung (Phase 3)**
♣ entsprechend der getroffenen Entscheidungen
♣ Datenauswertung und –analyse (Phase 4)
♣ Vorbereitung der „maschinellen“ Datenauswertung / –analyse (mit Software)
♣ Dateiaufbau festlegen (Variablendefinition und –codierung), Datenimport
♣ Datenbereinigung
♣ Datenqualitätssicherung (Kontrolle auf Vollständigkeit und Plausibilität)
♣ Datenaufbereitung (Sortierung der Daten, Klassenbildung, ...)
♣ Datenauswertung und Datenanalyse (univariate und multivariate Datenanalysen mit Anwendung geeigneter statistischer Methoden)
♣ Einsatz von Statistik-Software

DATENANALYSE MIT STATISTIK-SOFTWARE

Zur Datenanalyse verwendet man in der Praxis unterschiedliche Statistik-Software.Marktführer sind:
♣ **EXCEL (Tabellenkalk, einfache statistische Methoden, Grafiken,etc.)**
Nicht für anspruchsvolle statistische Aufgaben geeignet!
♣ **R (die mächtige Open Source-Lösung, kostenfrei) www.r-project.org**
Eine populäre Open Source-Statistik-Umgebung, die durch Pakete nahezu beliebig erweiterbar ist und sich zunehmender Beliebtheit erfreut. Mit RStudio existiert eine komfortable Entwicklungsumgebung, die lokal oder in einer Client- Server-Installation über den Webbrowser genutzt werden kann. R-Applikationen lassen sich über Shiny auch direkt interaktiv im Web nutzen. R kann insbesondere Viel-Nutzern, die die Bereitschaft mitbringen, sich intensiver mit Statistik auseinanderzusetzen, uneingeschränkt empfohlen werden.
♣ **SAS (kommerzielle Statistik-Software, der Mercedes unter den Statistik-Programmen)**
SAS ist ein mächtiges und sehr stabiles Tool, welches insbesondere in größeren
Organisationen eingesetzt wird und sich im Pharma-Bereich zum Quasi-Standard für viele Analysen entwickelt hat. Die Software besteht aus unterschiedlichen Modulen, die z.T. völlig verschiedene Bedienkonzepte verfolgen. Entsprechend aufwändig ist die Einarbeitung. Im Vergleich zur kommerziellen Konkurrenz gehört SAS (auch aufgrund der Ausrichtung auf größere Unternehmen/Organisationen) zu den teuersten Lösungen.
Eine professionelle Statistiksoftware, welche insbesondere in der Biometrie, der klinischen Forschung und im Banken-Sektor Anwendung findet.

„OPERATIONALISIERUNG“ EINES BEGRIFFS

"Operationalisierung" eines Begriffs ist die Angabe derjenigen Vorgehensweisen und Forschungsoperationen, mit deren Hilfe zu entscheiden ist, ob und in welchem Ausmaß der mit dem Begriff bezeichnete Sachverhalt in der Realität vorliegt, was bedeutet, dass man beobachtbare Kriterien dafür anzugeben hat, wann ein Sachverhalt vorliegt bzw. je nach Skalenniveau auch in welcher Ausprägung er auftritt. Etwas weniger abstrakt: „Operationalisierung“ definiert, wie man den Begriff konkret misst. Die Operationalisierung ist besonders wichtig bei Begriffen ohne direkten empirischen Bezug (so genannte „latente Variable“), z.B. Kundenzufriedenheit, Teamfähigkeit, Intelligenz, Werbewirkung u.a.Der Ausdruck Operationalisierung bezeichnet im weitesten Sinne die Entwicklung eines Forschungsdesigns für eine konkrete Fragestellung, während es im engeren Sinne um die Formulierung von Messvorschriften geht, d.h., um die Bestimmung von Indikatoren, mit deren Hilfe ein Konstrukt gemessen werden kann.
◊ die Festlegung der Vorgehensweise (Operation) bei der Definition der Untersuchungsvariablen in einer Untersuchung.
Beispiel: Intelligenz kann operational durch die Anzahl der Lösungen von Intelligenzaufgaben in
einem konkreten Intelligenztest definiert werden. 25
„OPERATIONALISIERUNG“ EINES BEGRIFFS

Informationsbedarf ◊ empirische (statistische) Untersuchung
Bei einer empirischen Untersuchung messen wir Merkmale bei ausgewählten Untersuchungseinheiten mit einem Messinstrument auf einer Skala. **Ergebnis: Messwerte = Merkmals- = Beobachtungswerte**
Wir messen bei Kind und seiner Mutter das Merkmal Körpergröße mit einem cm-Maß auf einer cm-Skala. Messergebnisse:Kind: 121 cm, Mutter: 168 cm.

SKALENNIVEAU

Nach der Art des Merkmals richtet sich, auf welche Weise die Beobachtungswerte bei der statistischen Untersuchung gemessen werden können (Messung = Eindeutige Zuordnung einer Beobachtung zu einem Punkt auf einer Messskala) Vom **Skalenniveau** hängt auch ab, welche Rechenoperationen mit den Beobachtungswerten
und welche statistischen Auswertungsmethoden zulässig sind.
Man unterscheidet folgende **Skalenniveaus:**
I. Nicht metrische Skalen ◊ Anwendung bei qualitat. Merkmalen. Keine Rechenoperationen mit den Merkmalsausprägungen zulässig:
• Nominalskala
• Ordinalskala
II. Metrische Skalen (Kardinalskalen) ◊Anwendung bei quantitativen Merkmalen. Skala hat Nullpunkt und Maßeinheit. Rechenoperationen sind zulässig:• Intervallskala, • Verhältnisskala (Ratioskala), • Absolutskala

KLASSIERUNG BEI QUALITATIVEN MERKMALEN

Beispiel:
Merkmal: Beruf, Merkmalsausprägung:
◊ Berufsgruppe: Handwerker = Klasse von z.B.
♣ Maurer, ♣ Dachdecker, ♣ Schreiner, ♣ Fliesenleger
Zielkonflikt: Übersichtlichkeit versus Informationsverlust

ENTSCHEIDUNGEN BEI KLASSIERUNG

♣ Anzahl der Klassen
♣ Klassenbreite(n) ◊ alle gleich oder unterschiedlich
♣ Klassengrenzen (Klassen definieren) ◊ untere Klassengrenzen, obere Klassengrenzen
♣ untere/obere offene Randklasse? ◊ „bis unter 50 kg“ bzw. „120 kg und

Formen als Dokumentation der Daten:

Einzelwerte (Einzelbeobachtungen) ◊ ungeordnete Reihe (Urliste, Rohdaten, Primärdaten) ◊ INPUT-Blase auf der Folie 2
◊ Die Urliste ist im Bereich Statistik direkte Ergebnis e. Datenerheb
Vorteile:
Die Urliste enthält alle Beobachtungswerte und damit: keine Auslassungen, keine
Übertragungsfehler und keine verlorene Information
Nachteile:
Urlisten können in der Praxis tausende oder Millionen von Datensätze enthalten, die für sich genommen unübersichtlich und nicht auswertbar sind; außerdem können bei einer unkorrigierten Urliste noch offensichtliche Fehler, wie Zahlendreher oder unplausible Daten enthalten sein

SUMMENHÄUFIGKEITEN

◊ sinnvoll nur für Rangmerkmale und metrische Merkmale
absolute Summenhäufigkeiten relative Summenhäufigkeiten
(absolute kumulierte Häufigkeit) (relative kumulierte Häufigkeit)
$H(x_1) = h(x_1)$ $F(x_1) = f(x_1)$
$H(x_2) = h(x_1)+h(x_2)$ $F(x_2) = f(x_1)+f(x_2)$
$H(x_3) = h(x_1)+h(x_2)+h(x_3)$ $F(x_3) = f(x_1)+f(x_2)+f(x_3)$
... ...
$H(x_j) = h(x_1)+h(x_2)+ \dots +h(x_j)$ $F(x_j) = f(x_1)+f(x_2)+ \dots +f(x_j)$
... ...
$H(x_i) = h(x_1)+h(x_2)+ \dots +h(x_i) = n$ $F(x_i) = f(x_1)+f(x_2)+ \dots +f(x_i) = 1$ (100%)

HÄUFIGKEITSVERTEILUNGEN

Die Daten einer Urliste müssen in der Praxis also aufbereitet werden, um ihren Zweck zu erfüllen. Das geschieht meist durch das Bilden von Häufigkeitsverteilungen:
Schritt 1: Sortieren der Daten ◊ geordnete Reihe nach irgendeiner Ordnung, z. B. alpha-betische Ordnung der Merkmalsträger oder Größenordnung der Merkmalsausprägung
Schritt 2: Verdichten der sortierten Daten auf Merkmalsausprägungen und zählen wie oft
diese vorkommen ◊ geordnete Menge von Wertepaaren (Merkmalsausprägung und
zugehörige Häufigkeit) heißt Häufigkeitsverteilung
Schritt 3: Darstellen tabellarisch von nach Merkmalsausprägungen sortierten
Häufigkeitsverteilungen ◊ die Häufigkeitstabelle

Für klassierte Daten:

Schritt 1: Einteilung der Werte in Klassen ◊ klassierte Daten (Sortierung nicht nötig)
Schritt 2: Verdichten der klassierten Daten ◊ Häufigkeitsverteilung für klassierte Daten
(klassierte Verteilung)
Schritt 3: Darstellen der klassierten Daten ◊ Häufigkeitstabelle für klassierte Daten

EINLEITUNG

Problem der Lageparameter:

Die Lageparameter schweigen sich aus über die Streuung der Daten. Das arithmetische Mittel (der Durchschnitt) und auch der Median verdecken oft eine große Ungleichheit.
◊ die Berechnung des Durchschnitts ist nicht immer sinnvoll
◊ der Durchschnitt kann offensichtlich nicht immer alles beschreiben

STREUUNGSPARAMETER

Forderungen an eine „gute“ Kennzahl zur Messung der Streuung:
♣ Bezugspunkt, um den die Werte streuen (◊ Lageparameter)
♣ alle Beobachtungswerte werden berücksichtigt
♣ Streuung = 0 (alle Werte sind gleich) ◊ Streuungsparameter = 0
♣ je größer die Streuung, umso größer der Streuungsparameter
♣ der Streuungsparameter ist unabhängig von der Anzahl der Beobachtungswerte n

QUARTILSABSTAND

Zusammenfassung:
Der Median teilt einen nach Größe sortierten Datensatz in der Mitte
◊ links und rechts vom Median liegen gleich viele Beobachtungswerte.
Unterteilt man die linke und die rechte Hälfte nach gleicher Vorschrift, wie man den Median bestimmt, so erhält man 4 gleich große Bereiche, die durch drei Quartils aufgeteilt werden.
25% aller geordneten Beobachtungswerte sind kleiner als das 1. Quartil.
50% aller geordneten Beobachtungswerte sind kleiner als das 2. Quartil.
75% aller geordneten Beobachtungswerte sind kleiner als das 3. Quartil.
Zwischen dem 1. und 3. Quartil liegen 50% aller Beobachtungswerte.
Dieser Bereich wird auch Quartilsabstand genannt.

VARIANZ

In der beschreibenden Statistik nennt man das arithmetische Mittel der Abweichungsquadrate die Varianz.
Eigenschaften:
♣ wichtiger Streuungsparameter
♣ Voraussetzung: metrisches Merkmal
♣ Ausgangswert für weitere folgende Streuungsparameter:
• Standardabweichung
• Variationskoeffizient
◊ Mittelwert und Varianz bzw. Standardabweichung hängen eng zusammen.

STREUDIAGRAMM

Streudiagramm (oder Streuungsdiagramm)
Ein Streudiagramm (engl. scatter plot) ist die graphische Darstellung von beobachteten Wertepaaren zweier Merkmale. Diese Wertepaare werden in ein kartesisches Koordinatensystem eingetragen, wodurch sich eine Punktwolke ergibt. Bei beiden Boxplots stimmt der eingetragene Median fast mit der Koordinatenachse überein, es gibt also jeweils etwa gleich viele positive und negative Werte, gemeinsam nehmen die Variablen aber fast nur Werte im I. und im III. Quadranten ein. Aus Lage und Form der dargestellten Punktwolke lassen sich die Stärke und die Richtung des Zusammenhangs der Merkmale ablesen. Das Streudiagramm liefert erste Hinweise über eine mögliche Abhängigkeit zwischen Merkmalen.

ZUSAMMENHANGSANALYSE

Zusammenhangsanalyse (Interdependenzanalyse)
Es wird eine Wechselwirkung der Variablen untereinander untersucht. Ein Zusammenhangsmaß, auch Assoziationsmaß genannt, gibt in der Statistik die Stärke und ggf. die Richtung eines Zusammenhangs
Zusammenhangsanalyse zwischen zwei metrischen Merkmalen X und Y
♣ Zusammenhangsanalyse ◊ Korrelationsanalyse
♣ Zusammenhangsmaß ◊ Korrelationskoeffizient $-1 \leq r_{xy} \leq +1$
♣ Grafische Darstellung ◊ Streudiagramm
♣ Korrelationsanalyse (oder Maßkorrelationsanalyse) ◊ wird geprüft, ob zwei Variablen X und Y linear zusammenhängen und wie stark dieser Zusammenhang ist
♣ Korrelationsanalyse mit dem Spezialfall der Rangkorrelationsanalyse ◊ Zusammenhang zweier ordinalskalierter Merkmale mit Hilfe von Rangzahlen. Der
Rangkorrelationskoeffizient nach SPEARMAN hat eine besondere praktische
Bedeutung wegen seiner einfachen Berechnung
♣ Kontingenzanalyse (oder Assoziationsanalyse, lat.: contingentia - Zufälligkeit)◊ Zusammenhangsanalyse auf der Basis einer Kontingenztabelle (=Häufigkeitstabelle, s. Modul 03). Je größer der Unterschied zwischen den Häufigkeiten in den Tabellenfeldern ist, umso stärker ist der Zusammenhang bzw. die Abhängigkeitzwischen den Merkmalen Hinweis: Typen und Arten des Zusammenhanges sind in den Kurs-Materialien „Abschnitt IV.
Multivariate Daten Teil 10: ZHA-Zusammenhänge“ gut beschrieben 5

KORRELATIONSKOEFFIZIENT

Korrelationskoeffizient r_{xy} $-1 \leq r_{xy} \leq +1$
Pos. Zsmhang $r > 0$: hohe Werte in der einen Variablen treten tendenziell gemeinsam mit hohen Werten in der anderen Variablen auf.
Neg. Zsmhang $r < 0$: hohe Werte in der einen Variablen treten tendenziell gemeinsam mit niedrigen Werten in der anderen Variablen auf.
Korrelationskoeffizient $r = -1$: es liegt ein extrem starker neg. lin. Zsmhang vor ◊ die Punktewolke liegt auf einer Geraden mit negativer Steigung.
Korrelationskoeffizient $r = +1$: es liegt ein extrem starker positiver linearer Zusammenhang vor ◊ die Punktewolke liegt auf einer Geraden mit positiver Steigung.
Korrelationskoeffizient $r = 0$: es liegt kein linearer Zusammenhang vor.
Voraussetzungen:
♣ X und Y quantitative (metrische) Merkmale
♣ X ◊ Y (es existiert ein Zusammenhang)
Vorbereitende Arbeiten:
♣ Überprüfung, ob Abhängigkeitsanalyse sinnvoll ist
♣ Erhebung von Daten für X und Y ◊ $(x_1,y_1), \dots, (x_n,y_n)$
1. Schritt: Visualisierung im Streudiagramm
(qualitative Abhängigkeitsanalyse)
2. Schritt: Auswahl eines Funktionstyps
(hier: Beschränkung auf lineare Funktionen)
3. Schritt: Berechnung der Regressionsfunktion
(nach Methode der kleinsten Quadrate)

QUARTILSABSTAND VS. SPANNWEITE

Vergleich zwischen Quartilsabstand und Spannweite:
Quartilsabstand: Von Ausreißern unabhängig
Gibt die Breite des mittleren Gibt die Gesamtbreite an, in dem
Bereichs an, in dem ca. 50% aller Werte liegen
Spannweite Vom kleinsten und größten Wert abhängig
Gibt die Gesamtbreite an, in dem alle Werte liegen

BOXPLOT

Aus einem Boxplot lassen sich Informationen über die:
♣ Lokalisation (Lage des Median)
♣ Streuungsmaße:
• Spannweite ◊ Ausdehnung eines Boxplots (Differenz w = xmax – xmin)
• Quartilsabstand ◊ Ausdehnung der Box (Differenz IQR = Q3 – Q1)
♣ Schiefe (Vergleich der beiden Hälften der Box oder der Längen der Whisker) eines Datensatzes sowie über den evtl. vorliegenden Ausreißer gewinnen. Eine der Definitionen der Whisker besteht darin, die Länge der Whisker auf maximal das 1,5-Fache des Interquartilsabstands (1,5×IQR) zu beschränken. Der Whisker endet nicht genau nach dieser Länge, sondern bei dem Wert aus den Daten, der noch innerhalb dieser Grenze liegt. Die Länge der Whisker wird also durch die Datenwerte und nicht allein durch den IQR bestimmt. Dies ist auch der Grund, warum die Whisker nicht auf beiden Seiten gleich lang sein müssen. Gibt es keine Werte außerhalb der Grenze von 1,5×IQR, wird die Länge des Whiskers durch den maximalen und minimalen Wert festgelegt. Andernfalls werden die Werte außerhalb der Whisker separat in das Diagramm eingetragen.

ABHÄNGIGKEITSANALYSE

Abhängigkeitsanalyse (Dependenzanalyse)
Es wird zwischen unabhängigen und abhängigen Merkmalen unterschieden. Es geht um einen gerichteten Zusammenhang. Man hat vorab eine sachlogisch begründete Vorstellung über den Zusammenhang zwischen den Merkmalen, d.h. man weiß oder vermutet, welche der Merkmale auf andere Merkmale einwirken (können).
Abhängigkeitsanalyse zwischen zwei metrischen Merkmalen X und Y
♣ Abhängigkeitsanalyse ◊ Regressionsanalyse
♣ Abhängigkeitsmaß ◊ Regressionsfunktion ŷ = a + b*x
♣ Grafische Darstellung ◊ Streudiagramm
+ Regressionsgerade

MULTIVARIATE ANALYSEMETHODEN

X, Y, Z Merkmale
Beispiel: Zusammenhangsanalyse (Interdependenzanalyse)
Mitarbeiterzufriedenheit X
Kundenzufriedenheit Y Z Motivation der Mitarbeiter
Abhängigkeitsanalyse (Dependenzanalyse)
Verkaufsfläche X
Anzahl Personal Y Z Filialumsatz 8

ZUSAMMENHANGSANALYSE BEI NICHT METRISCHEN MERKMALEN

Rangkorrelationskoeffizient
♣ ein Maß für die Stärke des Zusammenhangs zweier ordinalskalierter Merkmale
♣ der Spearmansche Rangkorrelationskoeffizient nutzt Ränge statt der Beobachtungswerte ◊ ein Spezialfall von Pearsons Korrelationskoeffizient, bei dem die Daten in Ränge konvertiert werden, bevor der Korrelationskoeffizient berechnet wird
♣ benötigt keine Annahme, dass die Beziehung zwischen den Variablen linear ist
♣ robust gegenüber Ausreißern Kontingenzkoeffizient
♣ ein Maß für die Stärke des Zusammenhangs zweier (oder mehrerer) nominaler oder ordinaler Merkmale. Er basiert auf dem Vergleich von tatsächlich ermittelten Häufigkeiten zweier Merkmale mit den Häufigkeiten, die man bei Unabhängigkeit dieser Merkmale erwartet hätte
♣ kann bei beliebig großen Kreuztabellen angewendet werden
♣ der Kontingenzkoeffizient C liegt zwischen 0 und +1, d.h., $0 \leq C \leq 1$.
Phi-Koeffizient (auch Vierfelder-Korrelationskoeffizient)
♣ ein Maß für die Stärke des Zusammenhangs zweier dichotomer Merkmale
♣ basiert auf einer Kontingenztafel, die die gemeinsame Häufigkeitsverteilung der Merkmale
enthält

Das wohl berühmteste Beispiel für eine Scheinkorrelation:

Der Storch bringt die Babys!
Der Wissenschaftler Robert Matthews fand 2001 eine Korrelation nicht unerheblicher Höhe von 0,62 zwischen der Geburtenrate eines Landes und der Anzahl dort lebenden Störche.
Wie kommt diese Korrelation zustande?
Bei Matthews basiert diese hohe Korrelation zu einem großen Teil auf der Größe des Landes: in größeren Ländern leben mehr Störche. Und dort werden mehr Kinder geboren als in kleineren Ländern. Auch eine andere Erklärung wäre denkbar ◊ **„Urbanität vs.Ländlichkeit“.** In der Stadt leben weniger Störche als auf dem Land. Gleichzeitig ist auf dem Land aufgrund soziokultureller Unterschiede die Geburtenrate höher als in der Stadt. Daraus ergibt sich,dass in Gegenden, in denen viele Störche leben, auch die Geburtenrate höher ist. **Auf jeden Fall:** Ein kausaler Zusammenhang liegt nicht vor, der Storch bringt keine Kinder!

EMPIRISCHE VERTEILUNGSFUNKTION

Eigenschaften:
♣ Die empirische Verteilungsfunktion F(x) ist (relative) Summenhäufigkeitskurve,
relative Summenfunktion
♣ Die empirische Verteilungsfunktion F(x) gibt für jede beliebige reelle Zahl x den Anteil der Merkmalsträger an, für die das Merkmal X einen Wert x_i annimmt, der kleiner oder gleich x ist
♣ Wertebereich: $0 \leq F(x) \leq 1$
♣ F(x) ist monoton nichtfallend (steigt oder ist konstant)
♣ F(x) ist eine Treppenfunktion mit Sprungstellen bei $x_1, x_2, ..., x_i$
♣ Die Größe der Sprünge beträgt $f_i = F(x_i) - F(x_{i-1})$

GRAFISCHE DARSTELLUNG DER HÄUFIGKEITSVERTEILUNG

♣ Ziel:
• ein anschauliches Bild der Daten
• das Wesentliche der Verteilung aufzuzeigen
♣ Wahlentscheidung:
• Form der grafischen Darstellung
• Achsenmaßstab
• Evtl. Ausschnitt darstellen
◊ Manipulationen sind denkbar (optische Täuschung!)
♣ Die am weitesten verbreiteten grafischen Darstellungsformen:
• Säulendiagramm
• Stabdiagramm
• Balkendiagramm
• Kreisdiagramm
• Histogramm (bei klassierten Daten)

Histogramm

♣ grafische flächenproportionale Darstellung der Häufigkeiten von klassierten Daten
♣ Im Unterschied zum Säulendiagramm muss bei einem Histogramm die x-Achse immer eine Skala
sein, deren Werte geordnet sind und gleiche Abstände haben
♣ direkt nebeneinanderliegende Rechtecke (keine Abstände dazwischen) der Breite der jeweiligen Klasse
♣ Absolute oder relative Häufigkeiten der Klassen werden durch die Flächen der Rechtecke
dargestellt: Fläche = Breite x Höhe
• Die Breite der Rechtecke entspricht der Breite der Klasse
• Die Höhe der Rechtecke entspricht den Klassenhäufigkeiten
• Die Fläche eines Rechtecks = $c \cdot f(x_j)$, wobei $f(x_j)$ die relative Klassenhäufigkeit der Klasse j und c ein Proportionalitätsfaktor ist. Ist c gleich dem Stichprobenumfang (c = n), so ist die Fläche eines jeden Rechtecks gleich der absoluten Klassenhäufigkeit. Das Histogramm wird absolut genannt wenn Summe der Flächeninhalte aller Rechtecke = n. Verwendet das Histogramm die relativen Klassenhäufigkeiten (c = 1), wird das Histogramm relativ oder normiert genannt (Summe der Flächeninhalte aller Rechtecke ist 1).

LAGEPARAMETER

◊ Lageparameter beschreiben die “Lage” der Elemente der Grundgesamtheit bzw. der Stichprobe in Bezug auf die Messskala
◊ noch Lokationsmaße genannt

ALLGEMEINE LAGEPARAMETER

Allgemeine Mittelwerte:
♣ Modus x_D, ♣ Median x_Z, ♣ arithmetisches Mittel x
♣ Quantil x_p, Spezielle Mittelwerte:
♣ geometrisches Mittel x_G
♣ harmonisches Mittel x_H
♣ Modus (oder Modalwert) x_D
Der Modus oder Modalwert ist die am häufigsten auftretende Merkmals-ausprägung (maximale Häufigkeit). Er wird hauptsächlich für nominaleMerkmale verwendet, ist aber auch für alle anderen (diskreten) Merkmalstypen sinnvoll.
Bei klassierten Daten ist der Modalwert die Mitte der Klasse mit den größten Häufigkeiten. Diese Klasse nennt man die Modalklasse.
Bemerkung:
Gibt es mehrere Merkmalsausprägungen mit der gleichen maximalen Häufigkeit, so
existieren mehrere Modalwerte ◊ Multimodale Verteilungen (bimodale Verteilung:
zwei Modalwerte; trimodale Verteilung: drei Modalwerte; usw.)
♣ Median (oder Zentralwert) x_Z
Mindestens 50% der Werte liegen links und mindestens 50% rechts des Medians (den Median selbst ggf. mit eingerechnet).
Median ist ein sehr robustes Lokationsmaß. Robuste statistische Kenngrößen sind wenig anfällig gegen Datenausreißer. Man muss die Hälfte der Daten gegen +∞ oder −∞ verschieben, um den Median selbst gegen ±∞ wandern zu lassen.

Median (oder Zentralwert) x_Z
Bemerkungen:
Falls das betrachtete Merkmal nur ordinal skaliert ist (z.B. Zeugnisnoten), so ist bei geradem n zu beachten, dass der Median nur dann existiert, wenn beide infrage kommenden Merkmalsausprägungen gleich sind.
Beispiel:
bei den Zeugnisnoten 1 2 3 4 5 6 existiert kein Median, denn 3,5 als Zeugnisnote ist nicht üblich.
Aber: 1 2 3 3 4 5 hat den Median 3.

MEDIAN BEI KLASSIERTEN DATEN

♣ Median (oder Zentralwert) x_Z
Für metrische Daten in Klassen, kann die exakte Merkmalsausprägung des Medians nicht bestimmt werden ◊ Näherungswerte für Median
$x_Z := x_{k-1} + x_k - x_{k-1} * 0,5 - F_{k-1}$
f_k wobei k = Einfallsklasse (Klasse mit F(x)=50%)

ARITHMETISCHES MITTEL

♣ Arithmetisches Mittel x
Eigenschaften:
♣ Die Summe der Abweichungen der Einzelwerte vom arithmetischen
Mittel ist stets gleich null $\sum x_i - x = 0$
♣ bekanntester Mittelwert
♣ nur für quantitative Merkmale sinnvoll
♣ empfindlich gegen Ausreißer (Vorsicht bei schiefen Verteilungen!)

Datenbeurteilung: Erfolgt durch: Schlüsse auf Basis unvollst. Daten, z.B. Schlüsse von der Stichprobe auf Grundgesamtheit
- Allgemeiner: auf Basis unsicherer Daten, unter Anwendung der Wahrsch lichkeitsrechnung. Dies ist Ggst. Der induktiven (schließenden) Statistik
Datenaufbereitung: Ordnung, Zusammenfassung und Darstellung des erhobenen statist. Datenmaterials in Datendateien, Tabellen / Grafik.
Datenmissbrauch: Statistischen Ergebnissen nicht klar ob manipuliert. Missbrauch von Daten kein Problem d. Statistik, sondern Schuld Person
Deskreptive / beschreibende Statistik: dient der Betrachtung der Daten an sich. Gewonnene Daten werden verdichtet / so dargestellt, dass das Wesentliche deutlich hervortritt. Für übersichtliche Darstellung muss das oft sehr umfangreiche, Material auf geeignete Art und Weise zusammengefasst werden. **Darstellungsformen**: Tabellen, graf. Darstellungen und charakteristische Maßzahlen.
induktive / Schließende Statistik: Schluß vom Teil aufs Ganze
Probleme der Stichprobe: Stichprobenfehler / „Repräsentativität“
Ind. Statistik: dient dazu, aus den erhobenen Fakten Schlüsse auf die Ursachenkomplexe zu ziehen, die zu diesen Daten geführt haben.
Die ind. Statistik bastiert auf Wahrsch.l.keitstheorie. Die Einleitung in deskreptive und indukte Statistik wurde verwendet, um die Unterschiedliche Zielsetzung der in diesen beiden Bereichen verwendeten Methoden herauszustellen. Synonyme: analytische / inferentielle Statistik.

KLUMPENSTICHPROBE

Klumpenstichprobe
eine einfache Zufallsauswahl, bei der die Auswahlregeln nicht auf die Elemente der Grundgesamtheit, sondern auf zusammengefasste Elemente (Klumpen, Cluster) angewendet werden und dann jeweils die Daten aller Elemente des ausgewählten Clusters erhoben werden. Ein Nachteil dieses Verfahrens: es kann kein Stichprobenumfang n vorgegeben werden.
Beispiel: Es soll ein Leistungstest an deutschen Schulkindern durchgeführt werden. Im ersten Schritt werden ‘Gemeinden‘ als Klumpen ausgewählt. Als ‚Liste‘ kann das Telefonvorwahlverzeichnis benutzt werden. Darin sind ca. 8.000 Gemeinden zu finden, aus denen eine Stichprobe gezogen werden kann. Einige der Gemeinden werden über keine Schulen verfügen. Eine Liste der Schulen ist ebenfalls als ‚Liste‘ (über das verantwortliche Schulamt) vorhanden. Aus den zur Verfügung stehenden Schulen wird dann eine Stichprobe gezogen, anschließend aus den dort existierenden Klassen. Schließlich nehmen Kinder ausgewählten Klassen an dem Test teil.

WILLKÜRLICHE UND BEWUSSTE AUSWAHLEN

Willkürliche Auswahlen (Auswahlen aufs Geratewohl)
unkontrollierte Aufnahme eines Elementes der Grundgesamtheit in die Stichprobe. **Bewusste Auswahlen (Auswahlen nach Gutdünken)**
nach einem Auswahlplan (anhand von Listen und festgelegten Regeln) und diesem Plan zugrunde liegenden angebbaren Kriterien. Es gibt viele verschiedene Arten bewusster Auswahlen: **Auswahl extremer Fälle**
♣ Auswahl typischer Fälle, ♣ Konzentrationsprinzip, ♣ Schneeball-Verfahren , Quotaverfahren (bestimmte Merkmale in der Stichprobe sollen exakt in derselben Häufigkeit (in %) vorkommen wie in der Grundgesamtheit)

• Dokumentation (Phase 5)
Dokumentation in Tabellen und Schaubildern und Interpretation der Ergebnisse Beispiel für die Gliederung einer Ergebnisstudie:
♣ Problemstellung
♣ Vorgehensweise, Beschreibung und Begründung aller Design-Entscheidungen
♣ Hauptteil: Ergebnisse der empirischen Untersuchung
♣ Folgerungen, Empfehlungen, Wertungen
♣ Anhang: Fragebogen, Literatur-, Abbildungs- und Tabellenverzeichnis
Mögliche Reaktionen auf die Ergebnisse der empirischen Untersuchung
„Na klar!“ ◊ Vermutungen bestätigt, „Aha!!!“ ◊ Ergebnisse überraschen

KORRELATION

♣ Korrelation ◊ zahlenmäßiger statistischer Zusammenhang zwischen zweiMerkmalen X und Y.
Eine positive Korrelation liegt vor, wenn die beiden Merkmale sich gleichförmig
entwickeln ◊ bei höheren Werten von X auch Y hohe Werte hat.
Eine negative Korrelation liegt vor, wenn X und Y sich gegenläufig entwickeln ◊ bei
höheren Werten von X liegen niedrigere Werte von Y vor.
♣ Ein kausaler Zusammenhang zwischen X und Y liegt vor, wenn es zwischen X und Y eine Ursache-Wirkungs-Beziehung gibt, d.h., wenn eine Veränderung desabhängigen Merkmals Y eindeutig auf eine Veränderung von X zurückzuführen ist.
♣ Eine Korrelation sagt nichts über einen kausalen Zusammenhang aus und auch nichts über eine Kausalitätsrichtung.

♣ SPSS (Statistik für Dummies)
SPSS gilt als besonders einfach zu bedienen, da die Software in den jüngeren Versionen stark in Richtung eines Tools entwickelt wurde, welches Auswertungen weitgehend automatisiert durchführt, ohne dass dem Benutzer besondere Methodenkenntnisse abverlangt werden. Die Stabilität hat gelitten. Während SPSS einige speziellere Module (z.B. für das Direktmarketing) mitbringt, ist das Spektrum gut unterstützter Methoden insgesamt geringer als z.B. bei R oder SAS.
Insbesondere in den Sozialwissenschaften und der Psychologie war SPSS auch im universitären Bereich fest verankert. Der ursprünglich eigenständige Anbieter wurde mittlerweile von IBM übernommen.
♣ STATA (Mehr als nur Panel-Analysen)
Obwohl STATA eine ausgereifte, sehr stabile und leistungsstarke Software ist, ist die Verbreitung - gerade in Unternehmen - gering. Dabei ist STATA für Anwender, die Wert auf ein breites Methodenspektrum, Stabilität, ein ausgereiftes Bedienkonzept inkl. Skriptsprache und einen fairen Preis legen, der teureren kommerziellen Konkurrenz überlegen.
STATA ist eine kommerzielle Statistiksoftware und wird insbesondere in der Ökonometrie angewendet.
♣ **Weitere Programme**
Daneben existieren etliche Programme, die sich auf bestimmte Methoden spezialisiert haben. Einige dieser Programme seien in dieser unvollständigen Übersicht zumindest kurz erwähnt:
• Eviews (Ökonometrie, Zeitreihenanalyse), • SPSS Amos (Modellierung und Schätzung von Strukturgleichungsmodellen)
• WinBUGS und OpenBUGS (speziell für Bayes'sche Statistik). Mit RBugs und R2OpenBUGS existieren Pakete, die die Funktionalität in R integrieren.

ANWENDUNG DER REGRESSIONSANALYSE

Regressionsverfahren haben viele praktische Anwendungen. Die meisten Anwendungen fallen in eine der folgenden beiden Kategorien:

♣ zum Erstellen eines Vorhersagemodells

♣ um die Stärke des Zusammenhangs zu quantifizieren: so können diejenigen xj ermittelt werden, die gar keinen Zusammenhang mit y haben oder diejenigen Teilmengen xi, ... , xj, die redundante Information über y enthalten.

MODELL VS. REALITÄT

Beispiel:

Verkaufsflächen ◊ Filialumsatz

unterschiedlich groß unterschiedlich hoch

WARUM?

♣ Wie gut erklären die Unterschiede bei den Verkaufsflächen die Unterschiede bei den Filialumsätzen?

◊ Wie viel Varianz wird durch das Modell nicht erklärt?

♣ Wie gut erklärt die Regressionsfunktion die Abhängigkeit zwischen Verkaufsfläche und Filialumsatz?

◊ Wie hoch ist die Erklärungskraft des Modells?

BESCHREIBENDE STATISTIK UND WAHRSCHEINLICHKEITSTHEORIE

Die beschreibende Statistik kommt ohne den Begriff Wahrscheinlichkeit aus Beschreibende Statistik Wahrscheinlichkeitstheorie

Relative Häufigkeit Wahrscheinlichkeit

Häufigkeitsverteilung Wahrscheinlichkeitsverteilung

Stichprobe Zufallsvariablen

Mittelwert Erwartungswert

Standardabweichung Streuung

Varianz Varianz

Median Median

Quantile Quantile

AXIOME VON KOLMOGOROW

Die axiomatische Begründung der Wahrscheinlichkeitstheorie wurde in den 1930er Jahren von Andrei Kolmogorow entwickelt.

Ein Wahrscheinlichkeitsmaß (kurz W-Maß) muss demnach folgende drei Axiome erfüllen:

1. Die Wahrscheinlichkeit für das Eintreten eines Ereignisses A ist immer eine reelle Zahl zwischen 0 und 1: $0 \le p(A) \le 1$
2. Das sichere Ereignis Ω hat die Wahrscheinlichkeit 1:

$p(A) = 1$ ◊ A tritt mit Sicherheit ein

$p(A) = 0$ ◊ A tritt mit Sicherheit nicht ein

$0 < p(A) < 1$ ◊ die Werte dazwischen drücken Grade an Sicherheit aus. Je größer die Wahrscheinlichkeit p(A), umso „eher" ist anzunehmen, dass das Ereignis A eintritt.

3. Die Wahrscheinlichkeit einer Vereinigung abzählbar vieler disjunkter Ereignisse ist gleich der Summe der Wahrscheinlichkeiten der einzelnen Ereignisse ◊ σ-Additivität (‚Sigma'-Additivität):

MENGENTHEORETISCHE KONZEPTE

Ereignisse sind Teilmengen des Ereignisraums.

◊ Ereignisse können ihre Beziehungen in Begriffen der Mengenlehre ausdrücken

◊ Ereignisse können wie Mengen miteinander verknüpft werden

Mengenoperationen:

⋂ Schnittmenge

⋃ Vereinigung

\ Mengendifferenz

c Komplementbildung

"Wie hoch muss mein R^2 sein?"

♣ Die übliche Größenordnung des R^2 variiert, je nach dem um welches Anwendungsgebiet es sich handelt. In Bereichen wie dem klassischen Marketing, indenen es hauptsächlich darum geht, menschliches Verhalten zu erklären bzw. vorherzusagen, sind meist geringe R^2 (deutlich kleiner 50%) zu erwarten. In anderen Bereichen wie bspw. der Physik sind höhere R^2 die Regel. Dies ist wenig überraschend, da auf das menschliche Verhalten zahlreiche und häufig nicht direkt messbare Einflüsse wirken. In der Physik hingegen werden oft Zusammenhänge zwischen wenigen exakt messbaren Größen untersucht. Dies geschieht zusätzlich meist unter experimentellen Bedingungen, unter denen sich Störeinflüsse minimieren lassen.

♣ Während auf der Mikro-Ebene in vielen Fällen bereits ein R^2 von 10% als gut gelten kann, erwarten viele bei stärker aggregierten Daten ein R^2 von 40% bis 80% oder sogar mehr. Ein Modell mit geringem R^2 - selbst bei stärker aggregierten Daten – ist nicht nutzlos, da die Alternative dazu oft gar kein Modell darstellt, was einem R^2 von 0 entspricht. Im übertragenen Sinne bedeutet das, dass eine systematische Prognose auf Basis eines Modells mit beschränktem R^2 oft schon besser ist als eine unsystematische Planung, die ausschließlich auf Bauchgefühl setzt. Generell ist die Aussagekraft von Modellen mit geringem R^2 nicht zwangsläufig schlecht.

WAHRSCHEINLICHKEITSTHEORIE

Die Verbindung zur Wahrscheinlichkeitstheorie wird auch über den Zufallsaspekt einer Stichprobe hergestellt. Historisch ist die Wahrscheinlichkeitsrechnung eng mit dem Glücksspiel verbunden.

Ein (Zufalls-) Experiment ist ein beliebig oft (unter identischen Bedingungen) wiederholbarer Vorgang, dessen Ergebnis „vom Zufall abhängt", d.h. nicht exakt vorhergesagt werden kann.

Die verschiedenen möglichen Ergebnisse oder Realisationen des Experiments heißen Elementarereignisse ω (‚Klein-Omega'). Sie bilden zusammen den Ereignisraum Ω (‚Groß-Omega'). Experiment ◊ die Erhebung eines Merkmals an einem Merkmalsträger Elementarereignisse ◊ die Merkmalsausprägungen Stichprobe vom Umfang n ◊ die n-malige Wiederholung des Experiments Annahme: die Ausgangssituation bei der n-fachen Wiederholung des Experiments ist immer dieselbe. In der Praxis ist dies jedoch unrealistisch. So sind z.B. bei einem Test zur Wirkung eines Medikaments an 20 Versuchspersonen die Bedingungen (Alter, frühere Krankheiten etc.) bei jeder der 20 Versuchswiederholungen (hier also Versuchspersonen) andere.

A ist eine (echte) Teilmenge von B

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

Formal: $A \subseteq B :\Leftrightarrow \forall x (x \in A \rightarrow x \in B)$

Zwei Mengen heißen gleich, wenn sie dieselben Elemente enthalten.

Formal: $A = B :\Leftrightarrow \forall x (x \in A \leftrightarrow x \in B)$

Schnittmenge von A und B

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

Vereinigungsmenge von A und B

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

A ohne B

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

Formal: $A \setminus B := \{ x \mid (x \in A) \wedge (x \notin B) \}$

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

Formal: $BC := \{ x \mid x \notin B \}$

Monty-Hall-Problem oder Ziegenproblem

USA-Spielshow „Let's Make a Deal" ◊ Deutsche Variante „Geh aufs Ganze!"

Angenommen Sie hätten die Wahl zwischen drei Toren. Hinter einem der Tore ist ein Auto, hinter den anderen sind Ziegen. Sie wählen ein Tor, z.B. Tor Nr. 1, und der Moderator, der weiß, was hinter jedem Tor ist, öffnet ein anderes Tor, z.B. Nr. 3, hinter dem eine Ziege steht. Er fragt Sie nun:

„Möchten Sie auf das Tor Nr. 2 wechseln?"

Ist es von Vorteil, die Wahl des Tores zu ändern?

Antwort (ohne Berücksichtigung einer bestimmten Motivation des Moderators):

Ja, Sie sollten wechseln!

Das zuerst gewählte Tor hat die Gewinnchance von 1/3, aber das zweite Tor hat eine Gewinnchance von 2/3.

Hier ist ein Weg, sich das Geschehen vorzustellen: angenommen, es gäbe 1 Million Tore und Sie wählen Tor Nr. 1. Dann öffnet der Moderator, der das eine Tor mit dem Preis immer vermeidet, alle Tore bis auf das Tor Nummer 777.777. Sie würden doch sofort zu diesem Tor wechseln, oder?

ZUFALLSEXPERIMENTE

Beispiele für Zufallsexperimente:

1. Bernoulli-Experiment: Werfen einer Münze: Ω = {Kopf, Wappen} oder Ω = {0, 1}
2. Würfeln: Ω = {1, 2, 3, 4, 5, 6}
3. Lotto 6 aus 49: $\Omega = \{ \omega \mid \omega = \{j1, ..., j6\} , j1, ..., j6 \in \{1, 2, 3, ..., 48, 49\} \}$

Da Mengen nur verschiedene Elemente enthalten, gilt $|\{j1, ..., j6\}| = 6$.

4. Anzahl der Anrufe in einer Telefonvermittlung pro Tag: $\Omega = N0 := N \cup 0$
5. $\Omega = \{ \omega \mid \omega =$ Matrikelnummer eines Studenten im WS 2019/20 $\}$.
6. Verlauf der Körpertemperatur eines Lebewesens: $\{ \omega = (id, f) \mid id \in N, f \in {}^\circ C(R+) \}$.

Ergebnis des Experiments ist die Identifikationsnummer id des Lebewesens und eine (beschränkte) stetige Funktion auf der nichtnegativen reellen Achse. f(0) ist die Körpertemperatur bei der Geburt. Nach dem Tod (T > 0) des Lebewesens könnte man die Umgebungstemperatur zur Fortsetzung der Funktion f heranziehen.

GESETZMÄßIGKEITEN

Für alle $A , B , C \subseteq X$ 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: $(A \cup B)C = AC \cap BC$

(Regeln von de Morgan) $(A \cap B)C = AC \cup BC$

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: $(A \cap B) \setminus C = (A \setminus C) \cap (B \setminus C)$

$(A \cup B) \setminus C = (A \setminus C) \cup (B \setminus C)$

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

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

$A \setminus B = A \cap BC$

WAHRSCHEINLICHKEIT UND RELATIVE HÄUFIGKEIT

Was bedeutet Wahrscheinlichkeit?

Zufallsexperiment:

Würfeln mit einem Zufallsgenerator: für drei Werte von n wird die Anzahl des Auftretens von Augenzahl 6 in diesen n Versuchsdurchführungen und die dazugehörige relative Häufigkeit ermittelt. Dabei wurde jeder dieser n Versuche 5 mal durchgeführt. Die relativen Häufigkeiten werden mit wachsendem n einander immer ähnlicher. Die Wahrscheinlichkeit eines Ereignisses ist die für eine gegen unendlich strebende Anzahl n von Durchführungen des betreffenden Zufallsexperiments vorausgesagte relative Häufigkeit seines Eintretens.

◊ mathematische Idealisierung, da n in der Wirklichkeit nicht "gegen unendlich strebt"

WAHRSCHEINLICHKEITSEIGENSCHAFTEN

Eigenschaften der Wahrscheinlichkeit:

1. Die relative Häufigkeit jedes Ereignisses A liegt im Bereich $0 \le h(A) \le 1$, und daher gilt dies auch für jede Wahrscheinlichkeit.

(Beweis: tritt das Ereignis bei n-maliger Durchführung des Zufallsexperiments m mal ein, so gilt $0 \le m \le n$, woraus die Behauptung folgt). 2. Tritt ein Ereignis A mit Sicherheit ein, so tritt es bei n-maliger Durchführung des Zufallsexperiments immer, also n mal, ein. Seine relative Häufigkeit ist dann

$h(A) = n/n = 1$ ◊ $p(A) = 1$

3. Tritt ein Ereignis A mit Sicherheit nicht ein, so tritt es bei n-maliger Durchführung des Zufallsexperiments nie, also 0 mal, ein. Seine relative Häufigkeit ist dann $h(A) = 0/n = 0$ ◊ $p(A) = 0$

Ein Ereignis (event) ist eine Teilmenge A des Ereignisraums Ω.

$A \subset \Omega$

A tritt ein, falls sich bei Versuchsdurchführung ein $\omega \in A$ ergibt.

Die einelementigen Teilmengen des Ereignisraums (oder die Elemente von Ω) heißen Elementarereignisse (singleton) {ω}.

Ein Gegenereignis ist die Menge aller Ergebnisse, die nicht zum Ereignis gehören.

Ω heißt sicheres Ereignis ◊ tritt also immer ein

Ø heißt unmögliches Ereignis ◊ kann nie eintreten

Ac heißt Komplementärereignis ◊ Gegenereignis, ohne A

Teilmengen A und B heißen unvereinbar oder disjunkt, falls A \ B = Ø

WAHRSCHEINLICHKEIT

Um exakte Voraussagen über die Begrenzung unserer Möglichkeiten zu treffen, brauchen wir einen Maß für die Sicherheit (oder Unsicherheit). Einsolches Maß ist die Wahrscheinlichkeit p (engl.: probability).

Die Wahrscheinlichkeitsrechnung ordnet jedem Ereignis A eines Zufallsexperiments eine Wahrscheinlichkeit p(A) (oder pA oder Prob(A)) für sein Eintreten zu. Beispiel:

Beim Münzwerfen gibt es nur zwei Elementarereignisse, die gleichmöglich sind.

p(‚Kopf') = ½

p(‚Zahl') = ½

p(‚Kopf oder Zahl') = 2/2 = 1 ◊ entweder ‚Kopf' oder ‚Zahl' tritt beim ein

p(‚Kopf und Zahl') = 0/2 = 0 ◊ ‚Kopf' und ‚Zahl' können n gleichz. eintreten

WAHRSCHEINLICHKEIT UND RELATIVE HÄUFIGKEIT

Was bedeutet Wahrscheinlichkeit? Im normalen Sprachgebrauch wird die Wahrscheinlichkeit eines Ereignisses auch häufig als Prozentzahl angegeben, indem sie mit 100 multipliziert und dafür mit dem Zusatz 'Prozent' versehen wird. Spätestens an dieser Stelle fällt die Parallelität zu den relativen Häufigkeiten eines Merkmals auf. In der späteren Anwendung werden unter anderem die Wahrscheinlichkeiten eines Ereignisses oder einer Merkmalsausprägung durch die entsprechenden relativen Häufigkeiten geschätzt, die über die n-fache Wiederholung des Experiments gewonnen werden. Allgemein lassen sich natürlich auf diese Weise die Wahrscheinlichkeiten beliebiger Ereignisse näherungsweise bestimmen, wenn sie sich zum Beispiel nicht elementar logisch oder physikalisch herleiten lassen.

BESTIMMTHEITSMAß

Das Bestimmtheitsmaß R2 (Erklärungskraft des Modells) ist ein Gütemaß der linearen Regression.

Das R^2 gibt an, wie gut die unabhängige Variable Y geeignet ist, die Varianz der abhängigen Variable X zu erklären.

(unbrauchbares Modell) $0\% \leq R^2 \leq 100\%$ (perfekte Modellanpassung)

Das R^2 nutzt das Konzept der Varianzzerlegung und besagt, dass sich die Varianz des abhängigen Merkmals in erklärte Varianz und nicht erklärte Varianz (Residualvarianz) zerlegen lässt.

Bestimmtheitsmaß R2 ◊ Anteil der Varianz der abhängigen Variable, der sich durch die Varianz der unabhängigen Variable erklären lässt.

BESTIMMTHEITSMAß

Es folgt: R^2 ist das Verhältnis aus der Streuung der Prognosewerte und der Gesamtstreuung der y-Werte (s. Folie 38): Achtung!

◊ Bei einer einfachen linearen Regression (nur eine unabhängige Variable) entspricht das Bestimmtheitsmaß dem Quadrat des Korrelationskoeffizienten nach Pearson rXY $\boldsymbol{R2} = (rXY)\mathbf{2}$

Beispiel:

X = Verkaufsfläche, Y = Filialumsatz

rXY = 0,707 ◊ R2 = $0{,}707^2$ = 0,50 = 50 %

Bedeutung / Interpretation:

50 % der Varianz der Filialumsätze lassen sich durch die Varianz der Verkaufsflächen erklären. Die

anderen 50 % lassen sich nur durch andere Einflussfaktoren erklären.

WAHRSCHEINLICHKEIT UND RELATIVE HÄUFIGKEIT

Was bedeutet Wahrscheinlichkeit?

Würfel:Das Maß für die Sicherheit, die höchste Augenzahl 6 zu würfeln, könnte so formuliert werden: "Ungefähr bei jedem sechsten Würfel-Versuch wird die Augenzahl 6 auftreten". Das bedeutet: "Unter 6 Würfel-Versuchen wird ungefähr 1 mal die Augenzahl 6 auftreten".

Ganz sicher können wir natürlich nicht sein, dass bei nur 6 Versuchen die gewünschte Augenzahl genau 1 mal eintritt, also würfeln wir öfter: "Unter 6000 Würfel-Versuchen wird ungefähr 1000 mal die Augenzahl 6 auftreten". Das klingt schon plausibler. Gehen wir noch einen Schritt weiter: "Unter einer sehr großen Zahl n von Würfel- Versuchen wird ungefähr n/6 mal die Augenzahl 6 auftreten".

WAHRSCHEINLICHKEIT UND RELATIVE HÄUFIGKEIT

Was bedeutet Wahrscheinlichkeit?

Genauer: Wenn wir ein Zufallsexperiment in identischer Weise n mal durchführen und dabei genau m mal das Ereignis A eintritt, so nennen wir den Quotienten m/n die relative Häufigkeit h(A), mit der das Ereignis A eingetreten ist. Die relative Häufigkeit wird nicht bei jeder Reihe von n Versuchsdurchführungen gleich sein. Wenn aber n sehr groß ist, so ergibt sich jedes Mal ungefähr die gleiche relative Häufigkeit und wenn wir gedanklich n gegen unendlich wachsen lassen, so sollte die relative Häufigkeit einen fixen, nur vom Zufallsexperiment und dem betrachteten Ereignis A abhängigen Wert annehmen. Diesen Wert nennen wir die Wahrscheinlichkeit des Ereignisses.

$\boldsymbol{P\ A} := \mathbf{lim}$

$\boldsymbol{n \to \infty\ hn(A)}$

◊ empirisches Gesetz der großen Zahlen

Das letzte Beispiel zeigt, dass auch Funktionen als Ergebnisse eines Zufallsexperiments auftreten können.

Man interessiert sich also dafür, ob bei Durchführung des Zufallsexperiments bestimmte Ereignisse eintreten.

Zum Beispiel, ob:

1. beim Wurf einer Münze A = {Kopf} gefallen ist
2. beim Würfeln eine 5 oder 6, d. h. B = {5, 6} herauskam
3. im Lotto 6 aus 49 "sechs Richtige" angekreuzt wurden
4. mehr als 1000 Anrufe pro Tag in der Telefonvermittlung, D = { n | n > 1000}, auftraten
5. K = {ω | Matrikelnummer ω beginnt mit einer 7 }
6. die Körpertemperatur eines Lebewesens nie den Wert 40 °C überschritt. In jedem Beispiel handelt es sich bei Ereignissen um Teilmengen von Ω.

EREIGNISRAUM

Ereignisraum Ω ◊ auch Ergebnismenge oder Merkmalraum genannt

Ω ǂ 0 ◊ eine nichtleere Menge Ω ist die Menge aller möglichen Ergebnisse eines mathematischen Zufallsexperiments, die sog. Ergebnismenge oder Merkmalraum oder Ereignisraum. Man spricht auch vom Stichprobenraum (sample space). Die Anzahl der Ergebnisse der Menge Ω nennt man Mächtigkeit |Ω|= n. Ω kann endlich, abzählbar oder sogar überabzählbar unendlich sein. Ω heißt diskret, falls es höchstens abzählbar unendlich viele Elemente hat.