Hochschule Karlsruhe
University of
Applied Sciences

Deskriptive Statistik
Kapitel 1

Prof. Dr. Andrea Wirth

Gliederung

1. **Einführung und Aufgaben**
   Data Science und Statistik**,** deskriptive Statistik, Wahrscheinlichkeitsrechnung, induktive Statistik, Vorgehensweisen und Beispiele
2. **Grundbegriffe, statistische Untersuchung und Darstellung der Daten**
   Datenquellen, Merkmale, Grundgesamtheit, Stichprobe, Messskalen, Vorgehensweise bei statistischen Untersuchungen Planung, Datenerhebung und –techniken, Datenaufbereitung, tabellarische und grafische Darstellung, Interpretation
3. **Eindimensionale Häufigkeitsverteilungen**
   Lageparameter, Streuungsparameter, Schiefe und Konzentration
4. **Mehrdimensionale Häufigkeitsverteilungen**
   Randhäufigkeiten, bedingte Verteilungen, Unabhängigkeit, Korrelation und Regression
5. **Bestandsanalyse, Zeitreihenanalyse und Prognoseverfahren**
   Verhältnis- und Indexzahlen, Zeitreihenanalyse
6. **Zusammenfassung**

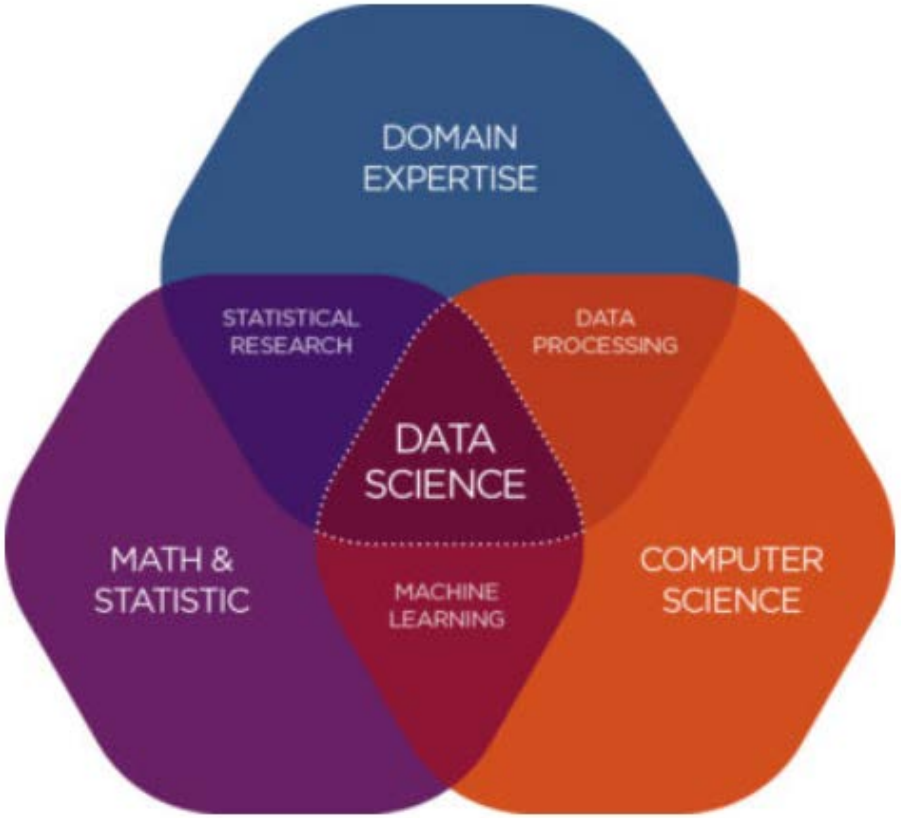
Literaturempfehlungen

- Günter Bamberg, Franz Baur, Michael Krapp: Statistik. de Gruyter GmbH, 18. Aufl. 2017
- Günther Bourier: Beschreibende Statistik, Praxisorientierte Einführung mit Aufgaben und Lösungen, Springer 13. Aufl. 2018
- Joel Grus: Einführung in Data Science, O´Reily, Deutsche Übersetzung: dpunkt.verlag GmbH 2. Aufl. 2020
- Reimar Hofmann: Vorlesung Statistik, Hochschule Karlsruhe, 2020
- Irene Rößler, Albrecht Ungerer: Statistik für Wirtschaftswissenschaftler, Springer, Gabler, BA kompakt, 6.Aufl. 2019
- Walter Krämer: Statistik für alle, Springer Verlag, 2015
- Walter Krämer: So lügt man mit Statistik, Campus Verlag, 2015

……..und einige weitere, jeweils in den Quellen genannt

Gliederung

1. **Einführung und Aufgaben**
   Data Science und Statistik**,** deskriptive Statistik, Wahrscheinlichkeitsrechnung, induktive Statistik, Vorgehensweisen und Beispiele



https://bigdataforbanking.com/de/ueber-data-science/

1. Data Science

Unsere Wirtschaft und Gesellschaft hängen in zunehmendem Maß von Daten und deren Interpretation ab. Durch die voranschreitende Digitalisierung werden immer mehr Daten zugänglich und darauf aufbauenden Analysen nötig und möglich. Parallel dazu haben sich Nutzungsmöglichkeiten und Einsatzfelder insbesondere durch Anwendungen wie maschinelles Lernen und künstlichen Intelligenz stark weiterentwickelt.

Ziel des Studiengangs ist es, Sie in die Lage zu versetzen,

- für komplexe Problemstellungen im Zusammenhang mit großen Datenmengen
- Lösungen zu konzipieren, zu bewerten und mit den jeweils aktuellen Technologien und Werkzeugen zu realisieren.

Die Absolvierenden des Studiengangs gestalten durch ihr Können und ihr Wissen die Art, wie wir zukünftig in einer von Daten abhängigen Gesellschaft gut zusammen-arbeiten und zusammenleben können.

1. Data Science – Der Aufstieg der Daten

»Daten! Daten! Daten!«, schrie er ungeduldig. »Ohne Lehm kann ich keine Ziegel herstellen.« – Arthur Conan Doyle

Der Aufstieg der Daten

Wir leben in einer Welt, die in Daten ertrinkt. Webseiten erfassen jeden Klick jedes Benutzers. Ihr Smartphone speichert Ihren Aufenthaltsort und Ihr Tempo jede einzelne Sekunde des Tages. »Quantified Selfer« tragen aufgemotzte Schrittmesser, die Herzfrequenz, Bewegungsgewohnheiten, Ernährung und Schlafzyklen registrieren. Intelligente Autos sammeln Informationen über Fahrgewohnheiten, intelligente Häuser sammeln Informationen über Lebensgewohnheiten, und intelligente Marketingleute sammeln Konsumgewohnheiten. Das Internet selbst stellt ein gewaltiges Netzwerk des Wissens dar, das (unter anderem) eine enorme Enzyklopädie mit Querverweisen darstellt – domänenspezifische Datenbanken über Filme, Musik, Sportergebnisse, Flippergeräte, Memes und Cocktails, außerdem viel zu viele Behördenstatistiken (einige davon sind sogar wahr!) von viel zu vielen Regierungen, bis Ihnen schwindelig wird.

Quelle: Joel Grus, Einführung in Data Science

1. Warum studieren Sie Data Science ?

Aufgabe:

Gründe für Ihr Data Science Studium.....

1. Was ist Data Science ?

Ein Witz sagt, dass ein Data Scientist jemand ist, der mehr über Statistik weiß als ein Informatiker und mehr über Informatik als ein Statistiker

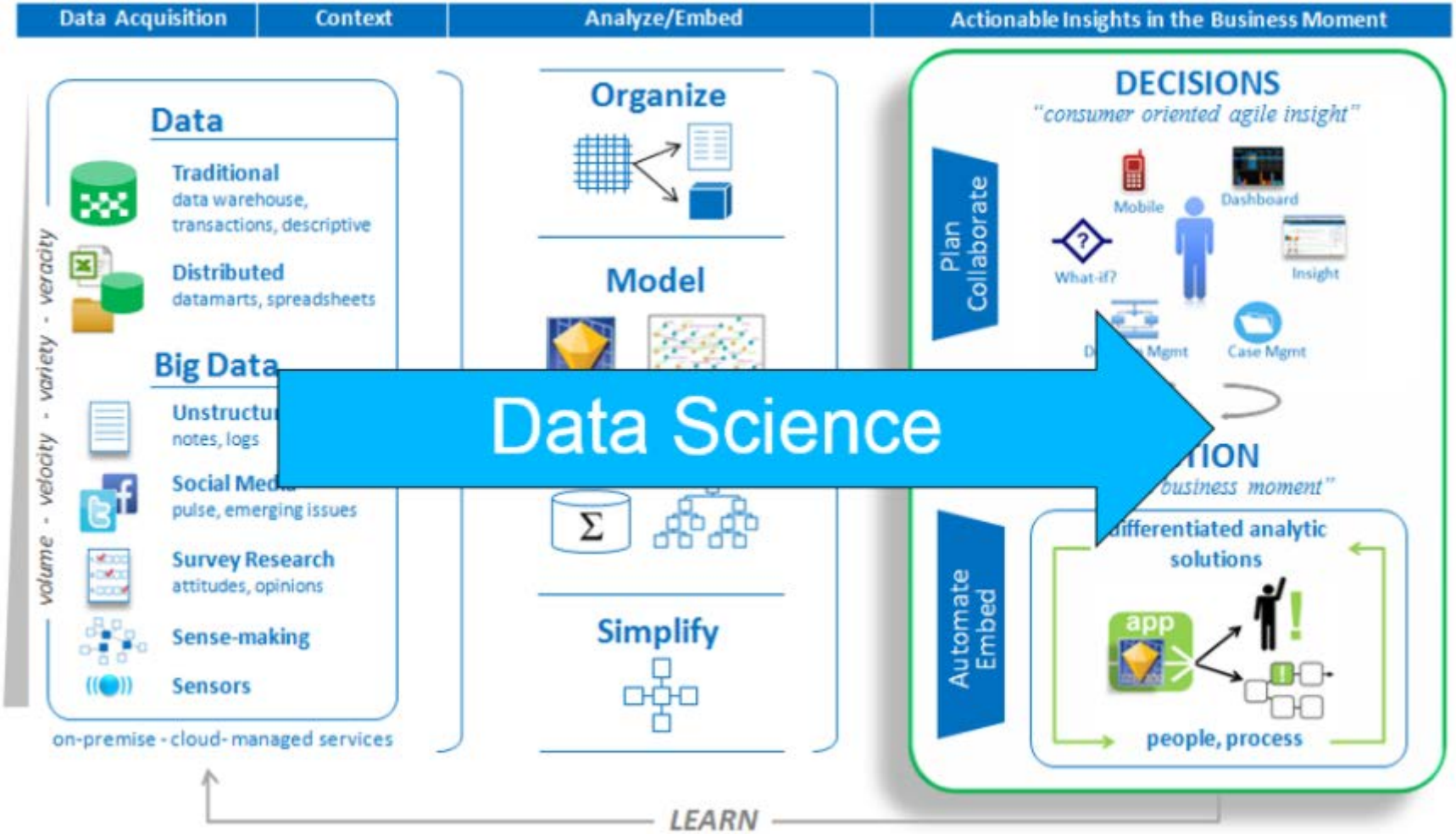
Ein Data Scientist ist jemand, der Erkenntnisse aus chaotischen Daten gewinnt.

Facebook fragt Sie nach Ihrer Heimatstadt und Ihrem gegenwärtigen Aufenthaltsort – vorgeblich, um es Ihren Freunden zu erleichtern, Sie zu finden und sich zu befreunden. Aber Facebook analysiert die Orte auch, um in der globalen Migration (*https://www.facebook.com/notes/facebook-data-science/coordinated-migration/10151930946453859*) und den Wohnorten von Footballfans (*https://www.facebook.com/notes/facebook-data-science/nfl-fans-on-facebook/10151298370823859*) Muster zu erkennen.

An der Wahlkampagne von Obama nahmen 2012 Dutzende von Data Scientists teil, die Daten durchwühlten und damit experimentierten, um Wähler mit besonderem Zuwendungsbedarf zu identifizieren, optimale auf Spender zugeschnittene Spendenaufrufe zu starten und Aufrufe zur Wahlbeteiligung auf die vielversprechendsten Gegenden zu fokussieren. Und im Jahr 2016 probierte die Trump-Kampagne eine große Zahl von Online-Ads aus (*https://www.wired.com/2016/11/facebook-won-trump-election-not-just-fake-news/*) und analysierte dann die Daten, um herauszufinden, welche funktionieren und welche nicht.

Quelle: Joel Grus, Einführung in Data Science

1. Was ist Data Science ?



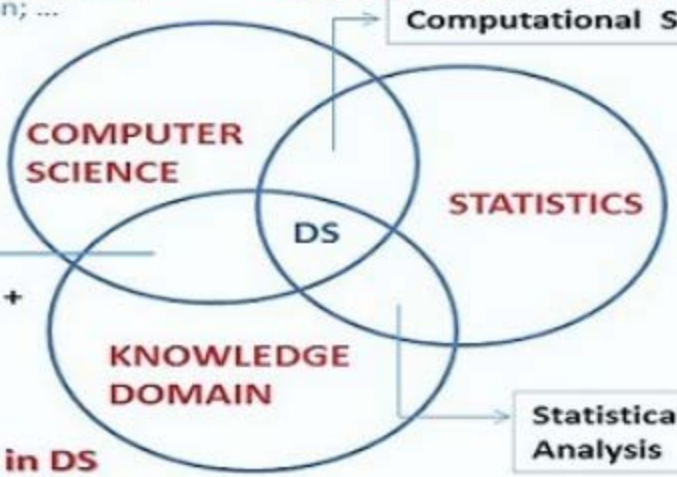
https://www.linkedin.com/pulse/definition-data-science-carlo-lauro/ Sept. 2020

1. Was ist Data Science ?

The Data Science curvilinear triangle
a DS definition by Carlo Lauro

Some CS tools:
Data extraction and preparation; Data Warehousing; High Performance Computing; R; Hadoop; Python; SAS; Rapid Miner; Tableau; Visualization ; Data Mining; A. I.; ANN; Machine Learning; Simulation; ...

Data Science (DS) is an interdisciplinary approach to meet the challenges of the Information Society, based on the methods of Computer Science and Statistics supplemented by Knowledge of the different domains. **Computer Science represents the language of the Data Science whereas Statistics is its Logic. The Knowledge of various domains of interest constitute the prerequisite of a Data Science.**



Computational Statistics

Some Stat tools:
Exploratory methods ; Density estimation; Regression; Time series; Causal Models and SEM; Bayesian models; Factorial analysis and PCA; Cluster analysis; Classification; SNA .....

Some Knowledge representation tools:
Logical rules; Hierarchical rules; Probability models; Graphs; Network; Metadata; Ontologies....

Computational Data Analysis

CDA -> Data = Algorithm + Accuracy

Statistical Data Analysis

SDA -> Data = Model + Error

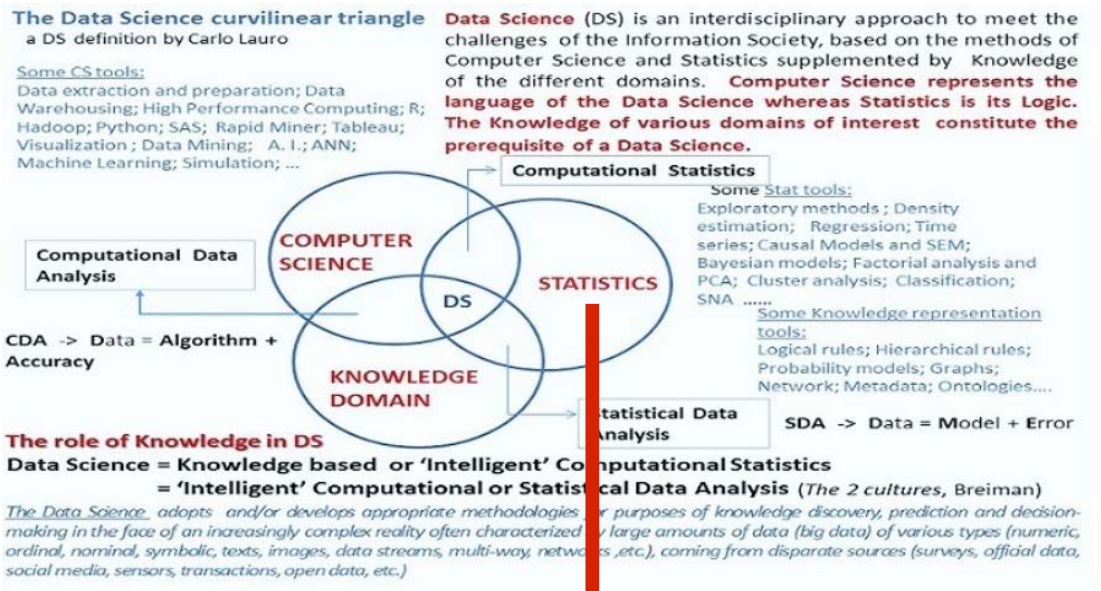
The role of Knowledge in DS

Data Science = Knowledge based or 'Intelligent' Computational Statistics
= 'Intelligent' Computational or Statistical Data Analysis (*The 2 cultures*, Breiman)

The Data Science adopts and/or develops appropriate methodologies for purposes of knowledge discovery, prediction and decision-making in the face of an increasingly complex reality often characterized by large amounts of data (big data) of various types (numeric, ordinal, nominal, symbolic, texts, images, data streams, multi-way, networks ,etc.), coming from disparate sources (surveys, official data, social media, sensors, transactions, open data, etc.)

http://www.anovaanalytics.com/platform/ Sept. 2020

1. Data Science und Statistik



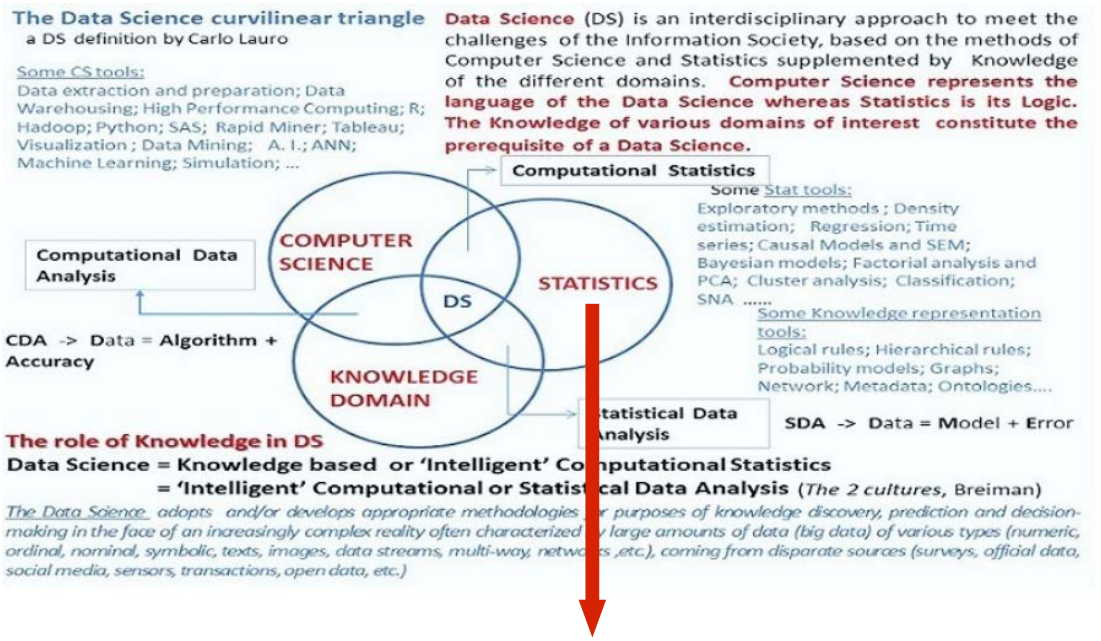
„Nichts lügt so sehr wie die Statistik“

„Traue keiner Statistik, die du nicht selbst gefälscht hast“.

Statistik als zentrale Voraussetzung für fundierte Datenanalyse und Erkenntnisse aus Daten

Diskutieren Sie die Definition der Statistik und deren Bereiche

1. Data Science und Statistik



Statistik: Entwicklung und Anwendung von Methoden zur Erhebung, Aufbereitung, Analyse und Interpretation von Daten.

Unterteilt in:

- ***Deskriptive Statistik***
- ***Wahrscheinlichkeitsrechnung***
- ***Induktive Statistik***

Quelle: Bourier: Beschreibende Statistik, S.1

1. Historischer Überblick - Statistik

- Seit 2600 v.Chr. fanden in Ägypten, seit 2300 v-Chr. in China oder zu 500 v.Chr. im persischen Großreich Bevölkerungszählungen statt. Seit 550 v.Chr. in Rom periodische Bevölkerungserhebungen; seit 800 n.Chr. unter Karl dem Großen „statistischen Zählungen“. Hier fand überwiegend die Ausrichtung auf messbare Eigenschaften /Zahlen statt. Seit Beginn der Neuzeit stellte man immer mehr Gebiete beschreibend statistisch dar.
- Ab 1650 Entwicklung der Wahrscheinlichkeitsrechnung (Korrespondenz Pascal und Fermat). 1713/1718 erste Lehrbücher der Wahrscheinlichkeitsrechnung von Bernoulli.
- Im 18./19 Jahrhundert Anwendung der Wahrscheinlichkeitsrechnung bei wirtschaftlichen und naturwissenschaftlichen Fragen. Ab Ende des 18. Jahrhundert Gründung staatlicher statistische Ämter in praktisch allen Ländern der Welt.
- Das Wort Statistik stammt von dem lateinischen Begriff „status“, der etwa Staat oder auch Zustand beschreibt ab. Im 18/19. Jahrhundert war die Statistik vorwiegend die Zustandsbeschreibung des Staates also das Sammeln von Daten über Heer, Gewerbe und Bevölkerung.
- Geschichtliche Entwicklung führte zu zwei Einsichten: Hinter empirisch erhobenen Daten können sich Gesetzmäßigkeiten oder Regelmäßigkeiten verbergen. In vielen Fällen lassen sich diese Regelmäßigkeiten nur erfassen, wenn man sich eines abstrakten mathematischen Wahrscheinlichkeitsbegriffs bedient.

Quelle: Hartung, Statistik 2002

1. Deskriptive Statistik

Die Methoden der ***deskriptiven Statistik*** (beschreibenden Statistik) zielen darauf ab:

- ***bekannte Daten*** zu beschreiben und darzustellen,
- diese durch ***aussagekräftige Kennzahlen und Diagramme*** zu charakterisieren
- und anschließend zu ***interpretieren***.

Unterscheidung:

- Im Extremfall wird genau eine Zahl zur Charakterisierung der Daten verwendet. So charakterisiert ein Lageparameter die Größenordnung aller Einzeldaten, ein Preisindex die Veränderung aller Preise eines umfangreichen Warenkorbs, ein Korrelationskoeffizient die Abhängigkeit zwischen den Ausprägungen zweier verschiedener Merkmale usw.
- Bei der Wahrscheinlichkeitsrechnung und der schließenden Statistik ist im Unterschied der Kenntnisstand über das interessierende Untersuchungsobjekt unvollständig.

Anwendungsfelder:

- Unternehmenssteuerung und kennzahlenorientierte Darstellungen
- Business Intelligence
- Alle Arten von Statistiken: https://de.statista.com/:
  Unternehmen, Branchen, Industrien, Märkte und Ländern

Quelle: Hofmann, Vorlesung Statistik 2020 und Bamberg: Statistik, 2017

1. Wahrscheinlichkeitsrechnung

Die ***Wahrscheinlichkeitsrechnung*** befasst sich damit:

- wie man Systeme, deren ***Verhalten nicht exakt vorhersehbar*** und deren Ausgang zufällig und nicht mit Sicherheit vorhersehbar ist,
- trotzdem ***quantitativ*** (in Zahlen/Funktionen) ***zu charakterisieren***,
- und das Ausmaß der Sicherheit auszudrücken und anschließend interpretieren kann.

Unterscheidung:

- Ableitung von Wahrscheinlichkeitsverteilungen aus Daten/Stichproben und zur Charakterisierung verwendet.
- Erwartungswerte beschreiben beispielsweise wie wahrscheinlich es ist, dass Max ein Würfelspiel gewinnt, eine Pumpe ausfällt oder dass ein potentieller Kunde auf den Werbelink Ihres Unternehmens klickt.
- Bei der Wahrscheinlichkeitsrechnung und der schließenden Statistik ist im Unterschied der Kenntnisstand über das interessierende Untersuchungsobjekt unvollständig.

Anwendungsfelder:

- Produktionssteuerung, Lebensdaueranalysen
- Business Intelligence als Prognose- und Vorhersagenfunktion
- Wichtige Voraussetzung für die Schließende Statistik

KA

Quelle: Hofmann, Vorlesung Statistik 2020 , Bourier: Beschreibende Statistik, Bamberg: Statistik,

1. Induktive Statistik

Die ***induktive*** (schließende) ***Statistik*** befasst sich damit:

- wie man bei ***unvollständigen Daten*** Kenntnisse über das gesamte Objekt erlangen kann,
- d.h. es werden Rückschlüsse von einer Teilgesamtheit (Stichprobe) auf die ***Eigenschaften der übergeordneten Gesamtheit*** gezogen.

Unterscheidung:

- Anstelle von Total- werden Teilerhebungen durchgeführt.
- Der Rückschluss ist mit einem Fehlerrisiko verbunden, das unter bestimmten Bedingungen mit Hilfe der Wahrscheinlichkeitsrechnung quantifiziert werden kann.
- Die klassischen Teilbereiche der induktiven Statistik, sind die Punkt-Schätzung, die Intervall-Schätzung und das Testen von Hypothesen.

Anwendungsfelder:

- Data Mining
- Alle Arten von Prognosen, wie Lastprognosen (Web-Server) oder Absatzprognose (Handel)
- Meinungsforschung
- Maschinelles Lernen

Quelle: Hofmann, Vorlesung Statistik 2020 , Bourier: Beschreibende Statistik, Bamberg: Statistik,

1. Statistik... und das gilt für alle Bereiche

Der Begriff ***Statistik ist mehrdeutig***:

„Nichts lügt so sehr wie die Statistik“

- ***Statistik als Zahlenergebnis***, d.h. als quantifizierte Information zu Massenerscheinungen in der Empirie in Form von Maßzahlen, Tabellen und Grafiken.

- ***Statistik als Methodenlehre***, d.h. die Gesamtheit von Methoden zur Gewinnung und Verarbeitung quantitativer empirischer Befunde.

„Traue keiner Statistik, die du nicht selbst gefälscht hast“.

Zweckbestimmung der Statistik:

- Die Statistik sollte niemals Selbstzweck sein, sondern stets einer konkreten Aufgabe dienen. Modellcharakter der Statistik: Ein Modell ist ein vereinfachendes Abbild der Wirklichkeit und entsteht durch die Beschränkung auf das Wesentliche.

- Erst ein Modell macht den zu untersuchenden Sachverhalt überschaubar und einer Analyse zugänglich. Der Statistiker muss sich stets um einen Kompromiss zwischen notwendiger Vereinfachung und verfälschender Vergröberung entscheiden.

Aufgaben

Unterscheiden Sie die Bereiche der Statistik !

Definieren Sie bitte eigene - erst mal unplausible - Beispiele und korrigieren Sie diese, so dass sie aussagekräftig(er) werden.

Achten Sie darauf, dass Sie mindestens ein Beispiel für die unterschiedlichen Bereiche der Statistik finden !

1. Beispiele

Beispiel: Ein Taxiunternehmen mit 15 Fahrzeugen desselben Typs verwendet unterschiedliche Reifenmarken. Ist der Mittelwert als Kennzahl bei folg. Fahrleistungen geeignet?

Reifen	Fahrleistungen der Fahrzeuge in tkm					Mittelwert
Marke A	26	29	24	31	20	26
Marke B	24	25	28	22	26	25
Marke C	31	30	32	29	33	31

Beispiel: 3 Sanierungsgebiete: zu entscheiden ist, welches zuerst saniert werden soll. Entscheidungskriterium sei das Alter der Häuser:

- Mittleres Alter / Durchschnittsalter der Häuser = 67 Jahre
- 85% der Häuser jünger als 51 Jahre, 15% älter
- 50% der Häuser jünger als 41 Jahre; 50% älter
- Ist der Mittelwert als Kennzahl geeignet ?

Anzahl der Häuser n_i	Alter der Häuser in Jahren a_i		Prozentsatz der Häuser die höchstens a_i Jahre alt sind
1	30	1x30=30	
9	40	9x40=360	
7	50	7x50=350	
2	100	2x100=200	
1	400	1x400=400	
20		1340	

1. Diskussionsbeispiele

SPIEGEL Panorama

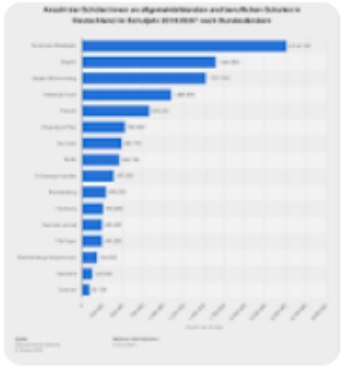
Medienbericht

Bundesweit 50.000 Schüler in Corona-Quarantäne

Pandemie im Klassenzimmer: In Deutschland sind laut einem Bericht aktuell 50.000 Schüler in vorsorglicher Corona-Quarantäne. Die Zahlen werden sich noch vervielfachen, glaubt der Lehrerverband.

26.09.2020, 10.09 Uhr

Insgesamt gab es im Schuljahr 2019/2020 in **Deutschland** nach vorläufigen Angaben ungefähr 10,91 Millionen **Schüler** an allgemeinbildenden und beruflichen Schulen. 12.03.2020



https://de.statista.com/statistik/daten/studie/981823/umfrage/anzahl-der-schueler-an-allgemeinbildenden-schulen/#:~:text=Insgesamt%20gab%20es%20im%20Schuljahr,an%20allgemeinbildenden%20und%20beruflichen%20Schulen.

entspricht 0,468 % - Ist das viel oder wenig?

1. Data Science – Umsetzung in Python

Ein motivierendes Szenario: DataSciencester

Herzlichen Glückwunsch! Sie wurden soeben als Leiter der Abteilung für Data Science bei DataSciencester angeheuert, *dem* sozialen Netzwerk für Data Scientists.

Finden von Schlüsselpersonen

Es ist Ihr erster Arbeitstag bei DataSciencester, und der Vizepräsident für Netzwerkarbeit steckt voller Fragen über Ihre Nutzer. Bisher hatte er niemanden, den er fragen konnte, und daher ist er begeistert, Sie dabeizuhaben.

Insbesondere möchte er herausbekommen, welche die »Schlüsselpersonen« unter den Data Scientists sind. Dazu stellt er Ihnen eine vollständige Kopie des Netzwerks von DataSciencester zur Verfügung. (Im wirklichen Leben übergibt man Ihnen die benötigten Daten eher selten.

```
users = [
    { "id": 0, "name": "Hero" },
    { "id": 1, "name": "Dunn" },
    { "id": 2, "name": "Sue" },
    { "id": 3, "name": "Chi" },
    { "id": 4, "name": "Thor" },
    { "id": 5, "name": "Clive" },
    { "id": 6, "name": "Hicks" },
    { "id": 7, "name": "Devin" },
    { "id": 8, "name": "Kate" },
    { "id": 9, "name": "Klein" }
]
```

Der Vizepräsident gibt Ihnen auch noch Daten über »Freundschaften« als eine Liste von `id`-Paaren:

```
friendship_pairs = [(0, 1), (0, 2), (1, 2), (1, 3), (2, 3), (3, 4),
                    (4, 5), (5, 6), (5, 7), (6, 8), (7, 8), (8, 9)]
```

Zum Beispiel zeigt das Tupel `(0, 1)` an, dass der Data Scientist mit der `id` 0 (Hero) und der Data Scientist mit der `id` 1 (Dunn) befreundet sind. Das komplette Netzwerk ist in Abbildung 1-1 dargestellt.

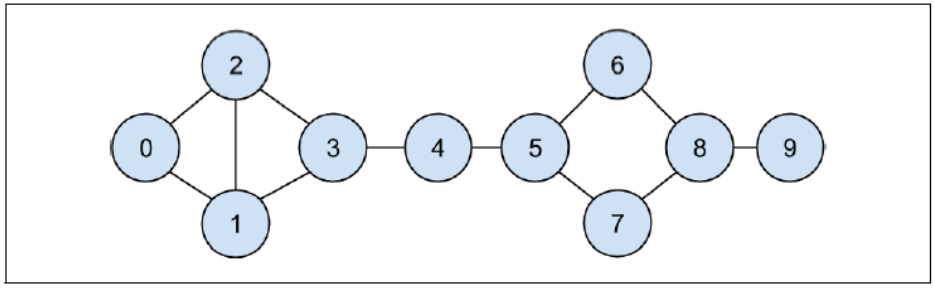


Abbildung 1-1: Das DataSciencester-Netzwerk

Quelle: Joel Grus: Einführung in Data Science

1. Vorgehensweisen in der Statistik

1. ***Planung***: exakte Formulierung des Untersuchungsziels und die Festlegung des Erhebungsprogramms.
2. ***Datenerhebung***: dient der Gewinnung des statistischen Datenmaterials.
3. ***Datenaufbereitung und -darstellung***: Verdichtung und Ordnung hin zu Tabellen und Schaubildern.
4. ***Datenanalyse und - interpretation***: Anwendung mathematischer / statistischer Methoden. Die erhaltenen Ergebnisse werden interpretiert und zusammengefasst.

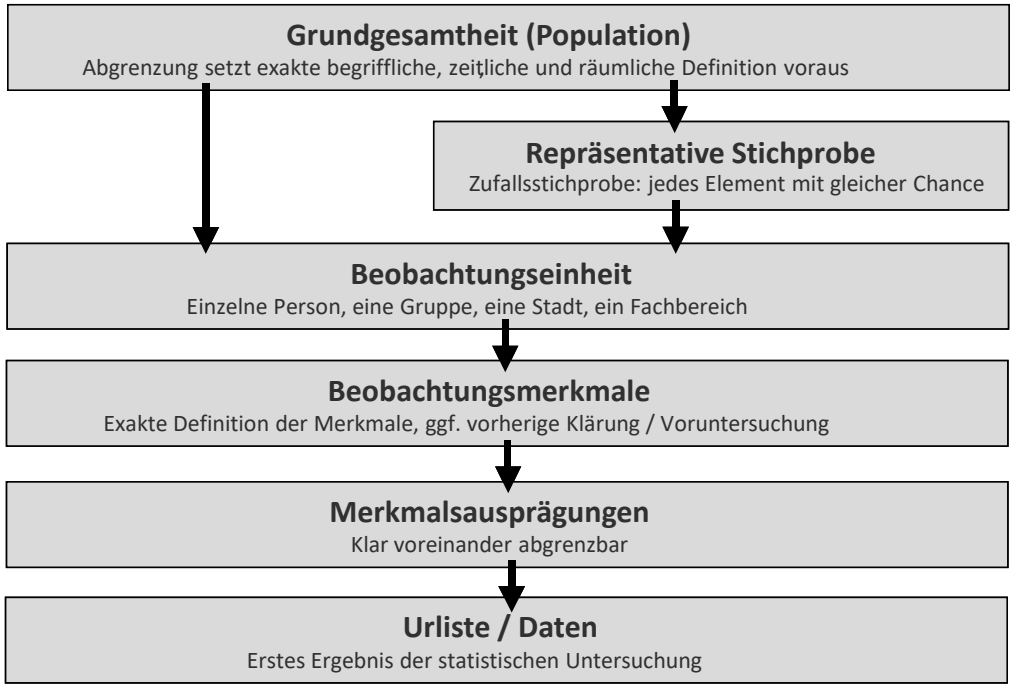
alternativ:

- Operationale Formulierung des Untersuchungsziels
- Festlegung der statistischen Gesamtheit
- Auswahl und Definition der Erhebungsmerkmale
- Wahl des Erhebungs- und Auswahlverfahrens
- Planung, Organisation, Durchführung und Kontrolle
- Aufbereitung und Erstellung von Dateien
- Auswertung
- Untersuchungszielbezogene Präsentation von ausgewählten Auswertungsergebnissen
- unter Umständen Dokumentation und Integration in ein Erhebungssystem

Quelle: Bourier: Beschreibende Statistik

1. Vorgehensweisen in der Statistik

datenorientiert:



Im nächsten Kapitel nach den Definitionen schauen wir uns diese Vorgehensweisen noch genauer an.

Als eine vorbereitende Aufgabe formulieren Sie bitte ein selbst gewähltes Beispiel für eine statistische Untersuchung.

Lernziele dieser Veranstaltung

Die Statistik stellt Informationen für deskriptive, analytische und operative Zwecke bereit. Die Adressaten dieser Informationen sind wirtschaftspolitische Entscheidungsträger, Produktionsunternehmen oder die Wissenschaft.

Absatz	Umsatzstatistik, zukünftige Markt- und Absatzpolitik
Produktion	Qualitätskontrolle, Ablauf- und Kostenplanung, Logistik
Beschaffung	Bestell- und Lagerstatistiken, Preise, Rabatte, Lieferzeiten
Finanzen	Einnahmen- und Ausgabenprognosen, Kreditwürdigkeit

- Kennenlernen zentraler Grundbegriffe der Statistik.
- Beherrschung wichtiger Lage- und Streuungskenngrößen sowie deren Interpretation.
- Fähigkeit zur Analyse mehrdimensionalen Datenmaterials mithilfe von Zusammenhangsmaßen und deskriptiver Regressionsanalyse.
- Erlangung elementaren Wissens hinsichtlich Bestimmung, Interpretation sowie Transformation von Preisindizes und Verhältniszahlen.
- Erwerb von Kenntnissen über den Umgang mit additiven Zeitreihenmodellen.

KA

Quelle: Bamberg: Statistik