# WIRTSCHAFTSSTATISTIK

## MODUL 1: EINFÜHRUNG

WS 2020/21

DR. E. MERINS

# LERNFORM

- **Online-Studiengang** → selbständiges Lernen soll geübt werden
- **Hochschule ≠ Schule** → der Stoff wird schneller vermittelt
- **Vor- und Nachbereitung unbedingt notwendig**
- **kontinuierliche Arbeit unbedingt notwendig** → nicht erst kurz vor der Klausur mit dem Lernen beginnen
- **kontinuierliches Üben zielführend (bzgl. Klausurerfolg)**

# UNTERSTÜTZUNG BEIM LERNEN …

- **Ca. jede vierte Woche ein Webseminar 90 Minuten (bitte Termine beachten) → Vorstellung neuer Themen mit Rechenbeispielen**
  - Zum Ablauf: 1. Präsentation vom Betreuer, 2. Fragen per Chat / Mail von Teilnehmern
  - Dokumentation und Aufzeichnungen vom Webseminar werden online auf **moodle.oncampus.de** abgelegt
  - Übungsaufgaben werden online auf **moodle.oncampus.de** abgelegt
  - online zeitlich versetzt: Lösungen zu den Übungsaufgaben
- **Nach Bedarf weitere Webseminare** (Termine werden separat mitgeteilt) → Beantwortung vorher zugesandten Fragen
- **Betreuung per E-Mail**
- **Online-Studienmodul mit Lernmaterial zum Selbststudium**
- **Tipp:** Formelsammlung → kann privat erstellt werden, hilft beim Üben

# ZUR GESCHICHTE DER STATISTIK

- Die „praktische Statistik“ ist 4.000 – 5.000 Jahre alt
- Der Ursprung ist nicht die Mathematik. Die Mathematik kam erst vor rund 300 Jahren dazu über die Wahrscheinlichkeitsrechnung
- Ausgangspunkt der Statistik: (Staats-)Verwaltung/Management von großen Projekten
- Das Wort Statistik stammt von lateinisch *statisticum* „den Staat betreffend“ und italienisch *statista* Staatsmann oder Politiker, was wiederum aus dem griechischen στατίζω (einordnen) kommt
- Die deutsche Statistik, eingeführt von Gottfried Achenwall 1749, bezeichnete ursprünglich die „Lehre von den Daten über den Staat“. Im 19. Jahrhundert hatte der Schotte John Sinclair das Wort erstmals in seiner heutigen Bedeutung des allgemeinen Sammelns und Auswertens von Daten benutzt.

# WO BRAUCHT MAN STATISTIK?

**In welchen Gebieten (Wissenschaften) braucht man Statistik?**

- **In den empirischen Wissenschaften (auch Realwissenschaften bzw. Erfahrungswissenschaften genannt)**
  - Naturwissenschaften
  - Sozialwissenschaften
  - Biologie / Medizin (Biometrie)
  - Ingenieurswissenschaften (Technometrie)
  - Verhaltenswissenschaften (Psychometrie)

  → man will neue Erkenntnisse gewinnen über einen Ausschnitt der Realität. Dazu werden empirische Untersuchungen durchgeführt; hierbei fallen Daten an, die mit statistischen Methoden ausgewertet werden
- **In allen Bereichen, in denen große Datenmengen anfallen, aus denen man Erkenntnisse gewinnen will**

# WAS IST EINE „STATISTIK“?

- **Was ist eine „Statistik“?**

  → eine systematische Zusammenstellung von Zahlen und Daten

- **Wozu?**

  → zur Beschreibung bestimmter Zustände, Entwicklungen und Phänomene

- **Ziel:**

  → Gewinnung von Information aus unübersichtlichen und/oder unstrukturierten und/oder großen Datenmengen

**Statistik ist die Lehre von Verfahren und Methoden zur Gewinnung, Erfassung, Analyse, Charakterisierung, Abbildung, Nachbildung und Beurteilung von beobachtbaren Daten über die Wirklichkeit (Empirie).**

# GEGENSTAND DER STATISTIK

- **Datengewinnung**

  Es gibt verschiedene Möglichkeiten, wie man Daten erhalten kann. Für die Wirtschaftsstatistik werden neben amtlichen Erhebungen vor allem Berichte, Umfragen und betriebliche Quellen verwendet

  - **Datenerhebung = jede systematische Datengewinnung** → 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 einer Gesamtheit
    **Teilerhebung** → $n < N$

# GEGENSTAND DER STATISTIK

- **Datenanalysen**

  Anwendung statistischer Verfahren zum Zweck der Erkenntnisgewinn

- **Datencharakterisierung**

  **Beschreibung**, **Visualisierung**, **Kennzahlen**: die grafische und tabellarische Darstellung von Daten sowie die Berechnung von zusammenfassenden, den empirischen Sachverhalt beschreibenden Kennzahlen, wird als Datencharakterisierung bezeichnet.

  Sie ist Gegenstand der deskriptiven Statistik

# GEGENSTAND DER STATISTIK

- **Datenbeurteilung**

  Die Beurteilung von Daten erfolgt durch:

  - Schlüsse auf der Basis unvollständiger Daten, z. B. Schlüsse von der **Stichprobe** auf ihre **Grundgesamtheit**
  - Allgemeiner: auf der Basis unsicherer Daten, unter Anwendung der **Wahrscheinlichkeitsrechnung**. Dies ist Gegenstand der induktiven (schließenden) Statistik.

# GEGENSTAND DER STATISTIK

- **Datenaufbereitung**

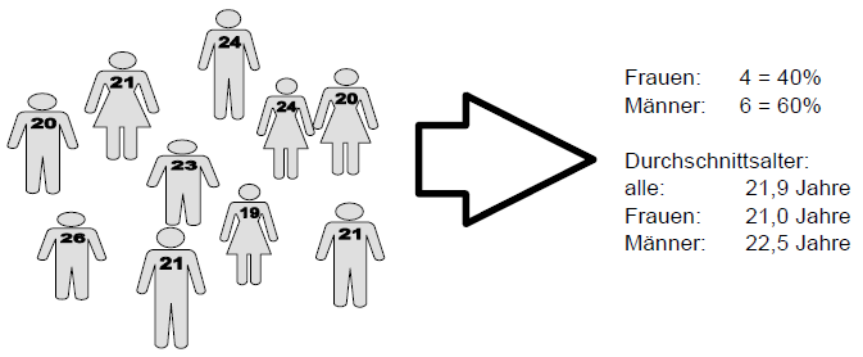
  **Ordnung, Zusammenfassung und Darstellung** des erhobenen statistischen Datenmaterials in Datendateien, Tabellen und/oder geeigneten Grafiken.

- **Datenmissbrauch**

  Man sieht statistischen Ergebnissen nicht an, ob sie manipuliert wurden. Der Missbrauch von Daten ist kein Problem der Statistik, sondern eines der Personen, die mit Daten umgehen

# BEREICHE DER STATISTIK

- **deskriptive oder beschreibende Statistik**



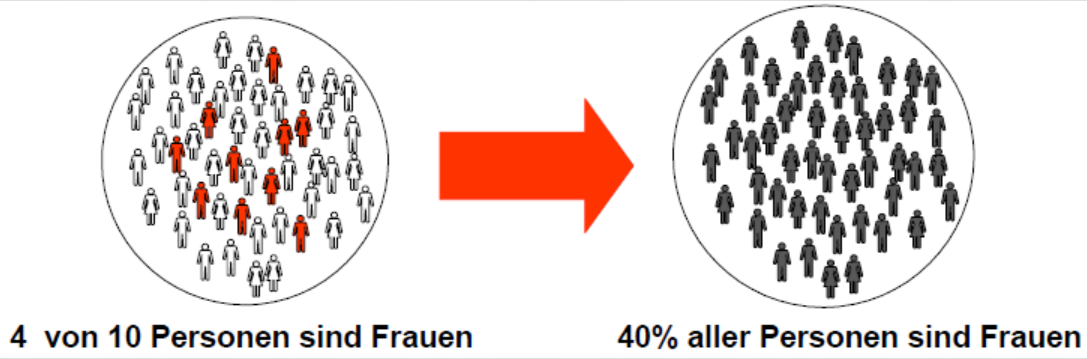

Die deskriptive Statistik (lat.: descriptio - Beschreibung) dient der Betrachtung der Daten an sich. Die gewonnene Daten werden verdichtet bzw. so dargestellt, dass das Wesentliche deutlich hervortritt.

Für eine übersichtliche Darstellung muss das, oft sehr umfangreiche, Material auf geeignete Art und Weise zusammengefasst werden.
Dazu werden insbesondere die drei Darstellungsformen benutzt:

- **Tabellen**
- **grafische Darstellungen**
- **charakteristische Maßzahlen**

# BEREICHE DER STATISTIK

- **induktive oder schließende Statistik → Der Schluss vom Teil aufs Ganze**




**Probleme der Stichprobe:**

**Stichprobenfehler**

**"Repräsentativität"**

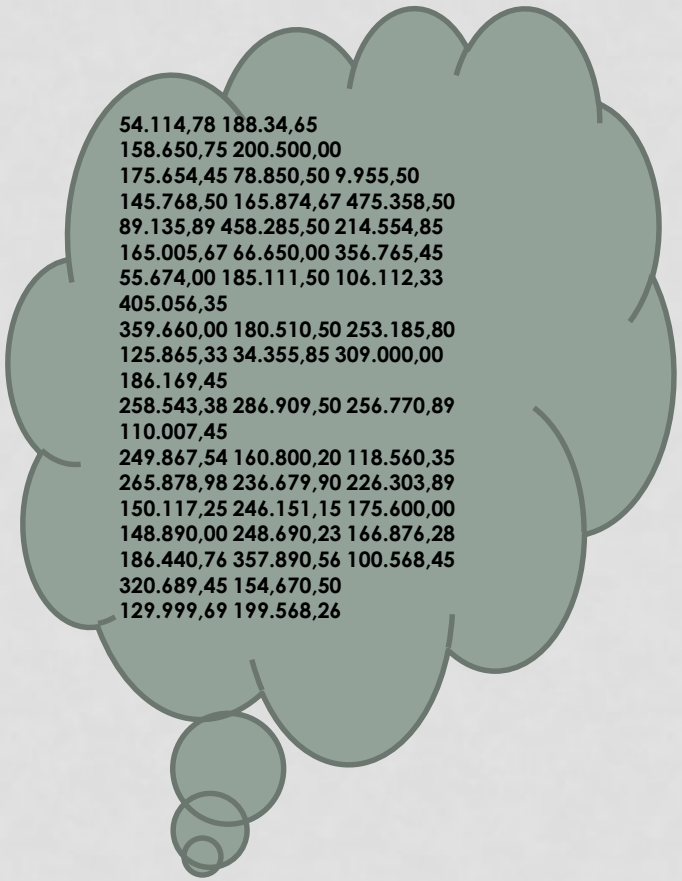
Die induktive Statistik (lat.: inductio - Hineinführen) dient dazu, aus den erhobenen Fakten Schlüsse auf die Ursachenkomplexe zu ziehen, die zu diesen Daten geführt haben. Die induktive Statistik basiert auf der Wahrscheinlichkeitstheorie.

Die Einteilung in deskriptive und induktive Statistik wurde verwendet, um die unterschiedliche Zielsetzung der in diesen beiden Bereichen verwendeten Methoden herauszustellen („Beschreiben" im Gegensatz zu „geplant Analysieren").

Weitere Synonyme für induktive Statistik: analytische oder inferentielle Statistik.

# STATISTIK IN TABELLEN UND GRAFIKEN

## INPUT




## OUTPUT

**Umsätze der Meyer AG über die Großhändler in NRW im Jahr 2008**

| Umsatzklasse in Tsd. € | Anzahl Großhändler (absolute Häufigkeit) | Anteil Großhändler von Gesamt in % (relative Häufigkeit) |
|---|---|---|
| 0 bis unter 100 | 7 | 14% |
| 100 bis unter 200 | 23 | 46% |
| 200 bis unter 300 | 12 | 24% |
| 300 bis unter 400 | 5 | 10% |
| 400 bis unter 500 | 3 | 6% |
| Summe | 50 | 100% |

(Quelle: Umsatzstatistiken der Vertriebsabteilung, 2008)
**Tabelle 1**

# STATISTIK IN TABELLEN UND GRAFIKEN

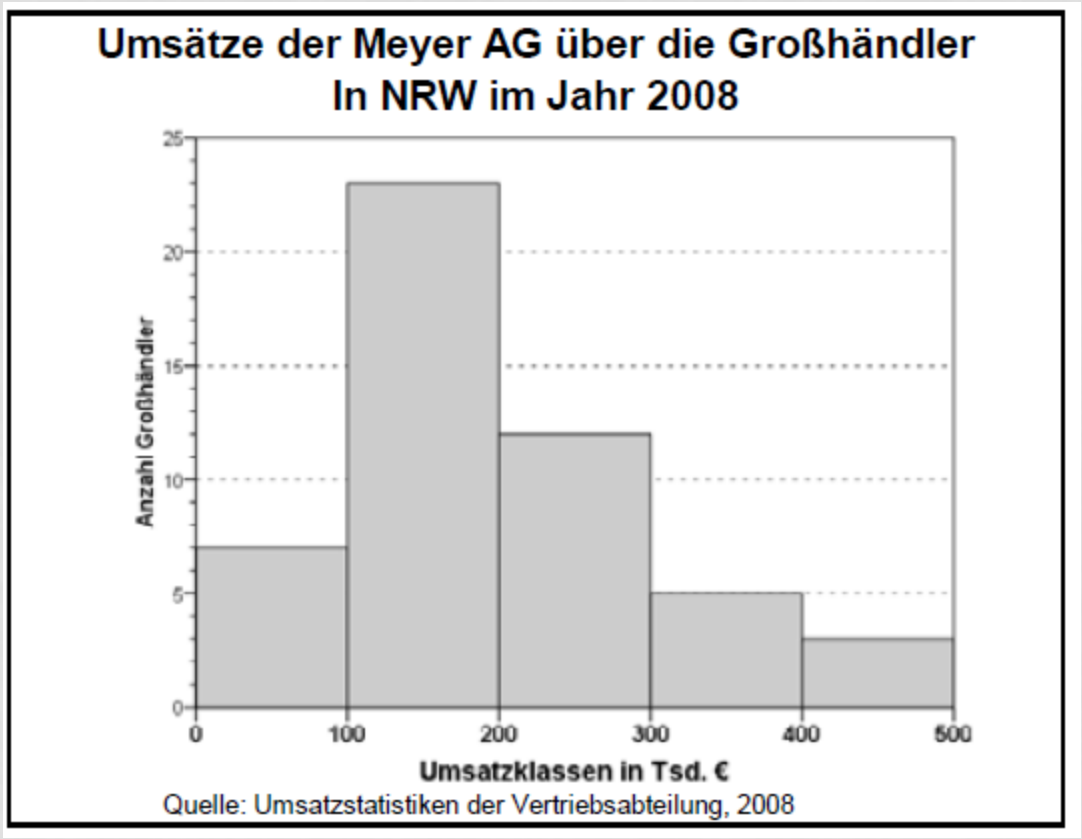



**Abbildung 1**

# TABELLEN VS. GRAFIKEN

**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.

# FACHTERMINOLOGIE

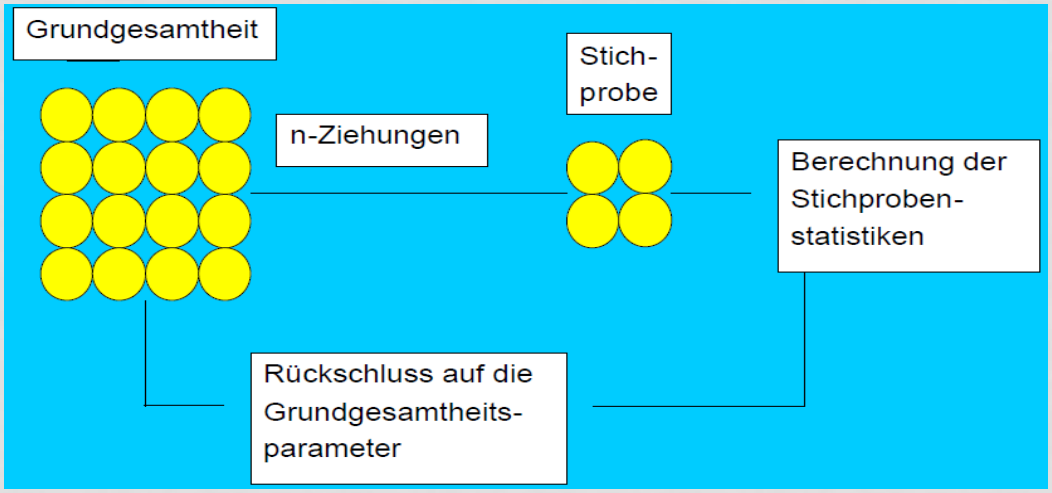
**Fachterminologie**

| | |
|---|---|
| **Untersuchungseinheiten = statistische Einheiten (Träger der Information)** | **einzelne Großhändler** |
| **Grundgesamtheit = statistische Masse** | **Alle Großhändler der Meyer AG in NRW im Jahr 2008** |
| **Umfang einer Gesamtheit = Anzahl ihrer Einheiten (Elemente)** | **Anzahl der Großhändler** |
| **Merkmal = Variable** | **Umsatz im Jahr 2008** |
| **Information der Tabelle** | **(klassierte) Häufigkeitsverteilung mit absoluten und relativen (Klassen-)Häufigkeiten** |

# 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.

# 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 der ausgewählten Klassen an dem Test teil.

# WILLKÜRLICHE UND BEWUSSTE AUSWAHLEN

- **Willkürliche Auswahlen (Auswahlen aufs Geratewohl)**

  unkontrollierte Aufnahme eines Elementes der Grundgesamtheit in die Stichprobe

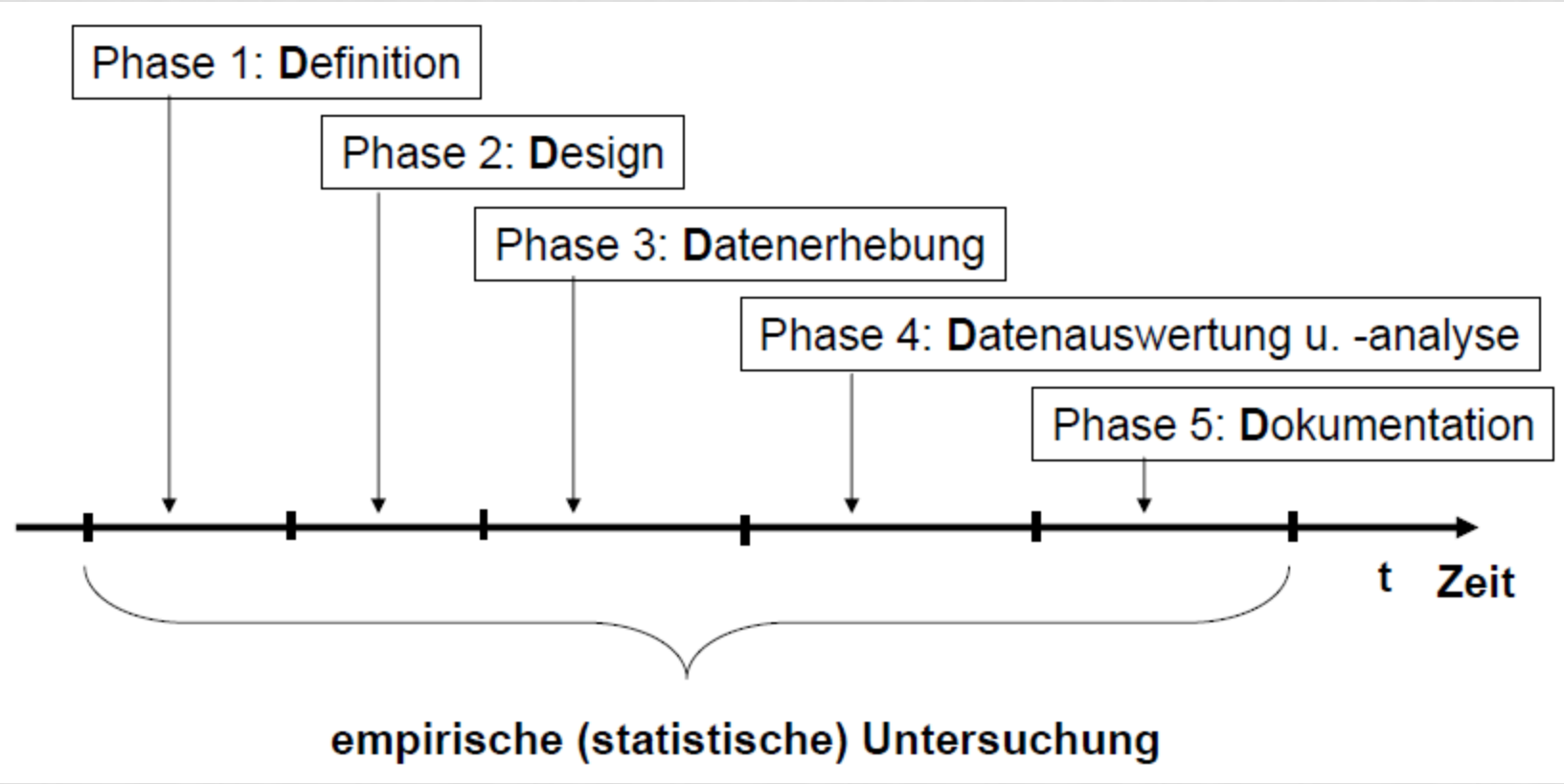
- **Bewusste Auswahlen (Auswahlen nach Gutdünken)**

  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)**

# ABLAUF EINER EMPIRISCHEN UNTERSUCHUNG

**Die 5 D´s**



# 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!

# DIE 5 D´S

- **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

# DIE 5 D´S

- **Design-Entscheidungen (Phase 2)**
  - **„Konstruktion“ von Messinstrumenten (Pretest der z.B. Fragebögen)**

    **Ziel:** das Risiko des Misserfolgs zu reduzieren und vorab Gründe für ein eventuelles Versagen zu finden. Außerdem können nach den Pretests eventuell noch Verbesserungen vorgenommen werden.
  - **Auswertungsdesign**
    - Methodischer Zugang zur Auswertung
  - **Entscheidungsspielraum wird eingeschränkt bzw. beeinflusst durch:**
    - Budget
    - Zeit
    - Thema/Aufgabenstellung
  - **Interdependenzen (gegenseitige Abhängigkeit und Beeinflussung) zwischen den Design-Entscheidungen**

# DIE 5 D´S

- **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

# DIE 5 D´S

- **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**

# DATENANALYSE MIT STATISTIK-SOFTWARE

**Zur Datenanalyse verwendet man in der Praxis unterschiedliche Statistik-Software.**

**Marktführer sind:**

- **EXCEL (Tabellenkalkulation, einfache statistische Methoden, Grafiken, etc.)**

  Nicht für anspruchsvolle statistische Aufgaben geeignet!

# DATENANALYSE MIT STATISTIK-SOFTWARE

- **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.

# DATENANALYSE MIT STATISTIK-SOFTWARE

- **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.

# DATENANALYSE MIT STATISTIK-SOFTWARE

- **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.

# DATENANALYSE MIT STATISTIK-SOFTWARE

- **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.

# DATENANALYSE MIT STATISTIK-SOFTWARE

- **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.
  - Mathematica und Matlab (numerische Problemstellungen)
  - Etc.

# STATISTIK-SOFTWARE IM VERGLEICH

| | Stärken | Schwächen |
|---|---|---|
| **R** | •Sehr großer Funktionsumfang (weit über 2000 Pakete)<br>•Sehr gut automatisier- und integrierbar (z.B. LaTeX, ODBC, MS …)<br>•Sehr guter Community-Support sowie kostenpflichtiger Support über Drittanbieter<br>•Umfangreiche Hilfe-Ressourcen frei verfügbar (Manuals, Tutorials<br>•Alle gängigen Plattformen werden unterstützt (Windows, Linux...)<br>•Zukunftssicher durch große, aktive Entwickler-Community | •Einarbeitung in die R-Syntax kann eine Einstiegshürde darstellen<br>•Stabilität/Qualität wenig genutzter Pakete z.T. nicht auf dem hohen Niveau<br>•Bei Verwendung sehr großer Datensätze wird leistungsfähige Hardware benötigt |
| **SAS** | •Schnelle Integration neuer statistischer Verfahren, sehr stabile und zuverlässige Routinen<br>•Sehr gute Dokumentation und professioneller Support<br>•Vielzahl von (kostenpflichtigen) Modulen und Schnittstellen, eigene Business Intelligence Software<br>•Gut geeignet für Umgang mit großen Datensätzen<br>•Umfangreiches hauseigenes Schulungsangebot | •Verschiedene, teils komplizierte (aber mächtige) Programmsprachen<br>•Lizenzmodell verbunden mit hohen Kosten |
| **SPSS** | •Leicht erlernbar, Bedienung jedoch nicht immer intuitiv<br>•Erweiterbar über kommerzielle Module<br>•Umfangreiche Literatur vorhanden | •Versionen für Windows und MacOS<br>•kurzes Update-Zyklus (1 Jahr)<br>•schwierig automatisier- und integrierbar |
| **STATA** | •Großer Funktionsumfang - nahezu jede statistische Methode<br>•Einfacher Einstieg durch GUI<br>•Automatisierbar & mit alten Versionen kompatibel<br>•Guter Support durch die STATA-Community, umfangr. Literatur<br>•Lauffähig unter Windows, Mac, Unix<br>•Im Vgl. zur kommerz. Konkurrenz vergleichsweise preiswert<br>•Investitionssicherheit durch 3-jährigen Release-Zyklus | •Eher träge bei der Einarbeitung neuer Methoden (Versionsupdates)<br>•Integration von und in andere Software ist umständlich<br>•Beschränkung auf einen gleichzeitig geöffneten Datensatz |