

# 1. EINFÜHRUNG UND GRUNDLEGENDES

mit dem Wort **STATISTIK** im Deutschen zwei Bedeutungen verbunden:

- eine wissenschaftliche Disziplin  
(engl. „statistics“)
- das Ergebnis dieser wissenschaftlichen Betätigung  
(„statistic“)  
z.B.: Statistik d. Ehescheidungen, d. Erzeugerpreise, etc.

## ARBEITSDEFINITION:

Statistik befasst sich mit dem **SAMMELN**, der **PRÄSENTATION** und der **ANALYSE** von Daten  
(üblicherweise bezogen auf Populationen)

### **SAMMELN:**

Wie komme ich zu der Information, die ich benötige ?

### **PRÄSENTATION:**

Wie kann ich Information kommunizieren ?

(wesentlicher Aspekt: Reduktion von Daten,  
Großteil der Information soll erhalten bleiben –  
wenige Kenngrößen, einfache Grafiken)

### **ANALYSE:**

Welche Schlüsse kann ich ziehen ?

Wie kann ich aufgrund von Stichprobeninformation allgemeine Aussagen treffen ?

## **Beispiel 1:**

### **National Patent Development Corporation (NPDC)**

#### Problem:

- NPDC betreut neu patentierte Produkte
- neues Produkt CARIDEX:  
Paste, die kariöse Stellen an Zähnen auflöst - Bohren kann vermieden werden
- Voranalyse ergab, dass 10.000 Dentisten CARIDEX im ersten Jahr nach Markteinführung verwenden würden
- Gerät zum Auftragen der Paste würde zum Selbstkostenpreis von 200\$ weitergegeben
- CARIDEX kostet je behandeltem Zahn 0.50\$, soll um 2.50\$ verkauft werden
- Fixkosten für NPDC: 4 Mill \$ im Jahr

#### Fragestellung:

- kann CARIDEX schon im 1. Jahr profitabel sein ?
- wieviele Zähne werden im ersten Jahr mit CARIDEX behandelt ?

#### Studie:

Befragung von 400 Dentisten, wieviel Zähne sie in einer durchschnittlichen Woche mit CARIDEX behandeln würden

Daten (Ausschnitt):     7 4 4 5 4 4 8 4 6 4 3 4 ...

## Wie kann benötigte Information aus den Daten gewonnen werden ?

- Wieviel Behandlungen erwarten die 400 Dentisten durchschnittlich ?
- Erwarten die 400 Dentisten ungefähr die gleiche Anzahl von Behandlungen / Woche oder differieren die einzelnen Dentisten stark ?

### Zusammenfassung der Daten

#### **Deskriptive (beschreibende) Statistik**

Methoden, um Daten übersichtlich und informativ zu organisieren, zusammenzufassen und zu präsentieren

Graphische Methoden  
Numerische Methoden

- nicht nur die 400 befragten Dentisten interessant - wie viele Behandlungen sind von den 10.000 amerikanischen Dentisten zu erwarten

### Schlussfolgerungen aus Stichprobendaten

#### **Inferenzstatistik (schließende Statistik)**

Methoden, wie man ausgehend von Stichprobendaten Schlußfolgerungen auf Charakteristika einer Population ziehen kann

- Schätzen
- Testen
- Modellieren

## **Beispiel 2: Host Selling Commercials – Announcer Commercials**

### Problem:

- Announcer Commercials: übliche Werbespots
- Host Selling Commercials: wenn ein Show Master oder eine TV Figur Werbung für ein Produkt macht
- Studie an Kindern sollte untersuchen, ob Show Master Werbung effektiver ist

### Fragestellung:

Ist Werbung mit Show Master effektiver als normale Werbung ?

## Studie: (Experiment)

2 Gruppen von jeweils 121 Kindern (6-10 Jahre)

sehen Sendung mit Werbepausen

Produkt: Canary Crunch (cereal - Frühstückskost)

Gruppe SW: Werbung durch Show Master

Gruppe NW: normale Werbung

Nach Sendung:

- Kinder werden getestet, was sich gemerkt haben (max. 10 Punkte)
- Kinder dürfen aus 4 verschiedenen cereals eines aussuchen (Gratispackung):

FL ... Fruit Loops

BB ... Boo Berries

KH ... Kangaroo Hops

CC ... Canary Crunch (beworben)

## Daten (Ausschnitt):

Gruppe	Gedächtnis- Test Punkte	Gewählte Gratispackung
SW	6	FL
SW	9	CC
SW	7	KH
...	...	...
NW	8	BB
NW	6	FL
NW	10	CC
...	...	...

Wie kann benötigte Information aus den Daten gewonnen werden ?

- Merken sich Kinder mehr Details eines Produkts, wenn es von Show Master beworben wird ?
- Würden sie dieses Produkt dann eher auswählen ?
- Unterscheiden sich die beiden Gruppen ?

## **Beispiel 3:**

### **Duxbury Press (DP)**

#### Problem:

- DP ist Verlag, der akademische Bücher verkauft
- Akademischer Buchmarkt unterscheidet sich von anderen Produktmärkten:
  - Käufer meist Studenten, die sich an Literaturempfehlung des Professors halten
  - Vertreter versuchen Professoren zu überzeugen, ihr Buch zu verwenden
  - Freiexemplare an Professoren, um ihnen Entscheidung zu erleichtern und nahezulegen
- Curt Hindrichs, einer der Herausgeber von DP, untersucht Verkaufszahlen von Statistikbüchern
- er bemerkt, daß viele Freiexemplare verteilt werden, und fragt sich, ob es zu viele sind, oder ob noch mehr Freiexemplare günstiger wären

### Fragestellung:

Curt Hindrichs will wissen, ob eine direkte Beziehung zwischen der Anzahl verschenkter Exemplare und der Anzahl verkaufter Bücher, bzw. der Verkaufserträge besteht?

### Daten von 78 Vertretern (Ausschnitt):

Vertreter (Code)	Bruttoverkauf in Dollar	Anzahl Freiexemplare
1305	2086,00	106
1307	63093,00	337
1327	41017,00	182
1329	7621,00	192
1330	28725,00	161
...	...	...

Wie kann benötigte Information aus den Daten gewonnen werden ?

- graphische Darstellung der Daten  
(*Scatterplot – Streuungsdiagramm*)
- Berechnung eines Zusammenhangsmaßes  
(*Korrelationskoeffizient*)
- wenn möglich, Erstellung einer Prognose:  
wieviel Ertrag bei wieviel verkauften Exemplaren  
(*Regressionsanalyse*)



# GRUNDSÄTZLICHE BEGRIFFE

statistische Erhebung dient üblicherweise dazu, Informationen über eine bestimmte, wohldefinierte Menge von Beobachtungseinheiten (Personen, Objekten, etc) zu gewinnen.

- **Beobachtungseinheiten:**

auch Fälle (engl. cases), Merkmalsträger:  
Personen, Objekte, Tiere, etc. , die ein bestimmtes Merkmal haben, das interessiert

- **(statistische) Population:**

Gesamtmenge aller dieser Beobachtungseinheiten  
z. B.: alle Österreicher, die wahlberechtigt sind, die Studenten einer bestimmten Universität, PKWs einer bestimmten Hubraumklasse

- **Variablen:** (Merkmale)

Charakteristika von Beobachtungseinheiten  
z. B.: Präferenz für bestimmte Partei, Alter, Benzinverbrauch auf 100km

- **Vollerhebung:**

wenn Daten von allen Elementen der Population erhoben werden

- **Stichprobenerhebung:**

wenn eine Stichprobe (nur ein Teil der Population) gewählt wird und Daten nur von dieser Teilmenge der Population gesammelt werden



# Kriterien für gute Messung:

## OBJEKTIVITÄT

das zu ermittelnde Merkmal wird eindeutig festgestellt,  
das Messergebnis hängt nicht von der Person, die misst, ab.

z.B.: Körpergröße, Deutsch-Schularbeit

bei Problemen:

- formulieren von Beurteilungsrichtlinien
- mehrere Personen, die messen

## VALIDITÄT (Gültigkeit)

ein Messinstrument misst tatsächlich das, was es messen soll  
(z.B.: Intelligenztest)

## RELIABILITÄT (Zuverlässigkeit)

wenn Ergebnis der Messung exakt festgestellt wird, d.h. wenn  
bei mehrmaliger Messung (zumindest approximativ) das gleiche  
herauskommt

→ meist gibt es leichte Variation

reliabel heißt: kleine Fehler (Abweichungen vom wahren Wert)  
in beiden Richtungen, gleich häufig

verzerrt („biased“) heißt: Fehler systematisch in eine Richtung

# VERSCHIEDENE EINTEILUNGEN VON VARIABLEN:

## QUALITATIV

Ergebnis der Messung erfolgt durch Klassifikation (Einteilung) in Kategorien

„Es gibt nur das Eine oder das Andere“

Beispiele:  
Geschlecht  
Zivilstand  
Beruf

## QUANTITATIV

Ergebnis kommt durch eine Art Zählen zustande.

„Es gibt mehr oder weniger“

Beispiele:  
Anzahl der Geschwister  
Alter  
Körpergröße

## DISKRET

Messen (im Prinzip) nur mit ganzen Zahlen

Zahlen können nur auf bestimmten Punkten der Zahlengerade liegen

Beispiele.  
Konfektionsgröße  
Anzahl der Familienmitglieder

## STETIG

Messen (im Prinzip) mit reellen Zahlen

Zahlen können jeden beliebigen Wert auf Zahlengerade annehmen

Beispiele:  
Körpergröße  
Dauer eines Telefonats

# SKALENNIVEAUS (Einteilung nach Stevens 1959)

## (1) NOMINALSKALA

man kann einzelne Kategorien **zahlenmäßig nicht vergleichen**, Kategorien können in beliebiger Reihenfolge angeordnet werden

**Bsp.:** Farbe von Blumen, Geschlecht, Hunderasse

## (2) ORDINALSKALA (RANGSKALA)

Kategorien können in eine **sinnvolle Reihenfolge** gebracht werden. Zahlen dürfen nicht addiert oder subtrahiert werden. (etwas ist „größer“ als etwas anderes, aber man kann nicht angeben um wieviel)

**Bsp.:** Rangreihung bei Talentewettbewerb, militärischer Rang, Präferenzen bei Speisen, ...

## (3) INTERVALLSKALA

impliziert Messungen auf einer **Skala mit gleichgroßen Einheiten**, d. h. es gibt numerisch gleichgroße Distanzen

**Bsp.:** man benötigt gleichviel Energie, um einen Liter Wasser von  $10^{\circ}$  auf  $20^{\circ}$  oder von  $60^{\circ}$  auf  $70^{\circ}$  zu erwärmen, d. h., Temperaturmessung (in Celsius) erfolgt auf einer Intervallskala

man kann keine Verhältnisse bilden (etwas ist x-Mal größer), weil es keinen absoluten Nullpunkt gibt

## (4) RATIO(NAL)SKALA

wie Intervallskala mit absolutem Nullpunkt

**Bsp:** Temperatur in Kelvingraden, Körpergröße, etc.

## (5) ABSOLUTSKALA

wie (4), aber mit natürlichen Einheiten - ( bei (3) und (4) nicht )

**Bsp:** Zählungen, Wahrscheinlichkeiten

## Einteilung von Daten in dieser Vorlesung:

für statistische Behandlung von Daten ist relevant, welche Eigenschaften Daten haben (d.h. qualitativ – quantitativ, bzw. diskret – stetig, Skalenniveau)

man muss die Art der Daten beachten, bevor man entscheiden kann, welche statistischen Operationen und Methoden man anwenden kann

### **KATEGORIAL**

- qualitativ
- diskret
- Nominal- oder Ordinalskala

### **METRISCH**

- quantitativ
- diskret, stetig
- Intervall-, Ratio- oder Absolutskala

### **ORDINAL**