

# Statistik I

## Einheit 1: Grundlagen der Datenanalyse

12.04.2023 | Prof. Dr. phil. Stephan Goerigk

# Kontakt

Prof. Dr. phil. Stephan Goerigk

Psychologische Methodenlehre

Infanteriestraße 11a · 80797 München ·

[stephan.goerigk@charlotte-fresenius-uni.de](mailto:stephan.goerigk@charlotte-fresenius-uni.de)

Zoom Sprechstunde (bitte per Email anmelden):

Meeting-ID: 284 567 8838

Kenncode: 807174

Publikationen

[Commitment to Research Transparency](#)



# Übersicht Lehrveranstaltung

Termine:

- 14 Termine
- Mittwoch 10:50 - 12:20

Begleitendes Seminar:

- Donnerstag 13:05-13:50 (SR08, Pavillon 2)
- Dozentin: Sara Vragolic ([sara.vragolic@charlotte-fresenius-uni.de](mailto:sara.vragolic@charlotte-fresenius-uni.de))

Materialien:

- werden auf **Studynet** bereitgestellt

Prüfungsleistung:

- Klausur 90 min
- 1/3 geschlossene Fragen (z.B. MC) & 2/3 offene Fragen und Rechnungen

# Termine

Einheit	Datum	Thema
1	12.04.2023	Grundlagen der Datenanalyse
2	19.04.2023	Skalenniveaus und statistische Kennwerte (1)
3	26.04.2023	Statistische Kennwerte (2)
4	03.05.2023	Visualisierung
5	10.05.2023	Wahrscheinlichkeitstheorie und Verteilungen
6	17.05.2023	Stichprobe, Grundgesamtheit und Stichprobenfehler
7	31.05.2023	Hypothesen und Hypothesentests
8	07.06.2023	t-Test (1)
9	14.06.2023	t-Test (2)
10	21.06.2023	Chi^2 Test und Mann-Whitney U-Test
11	28.06.2023	Korrelation (1)
12	05.07.2023	Korrelation (2)
13	12.07.2023	Effektstärke und Stichprobenumfangsplanung
14	19.07.2023	Klausurvorbereitung

# Material (bitte mitbringen)

Es werden händische Berechnungen durchgeführt.

- Taschenrechner
- Lineal
- Bleistift
- kariertes Papier

Interaktion während der Lehrveranstaltung:

- Note-Pad
- Folien ersetzen nicht den Vorlesungsbesuch

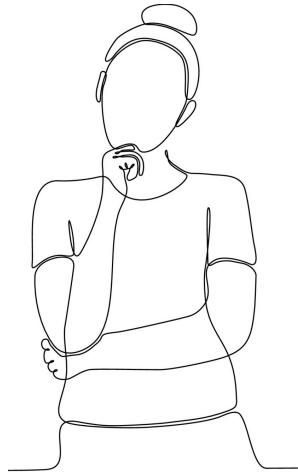
# Warum Quantitative Methoden und Statistik?

Eine Forschungsgeschichte:

Das ist **Dr. Charlotte**.

Sie ist ambitionierte Forscherin und möchte etwas bahnbrechendes entdecken.

Nur was?



# Warum Quantitative Methoden und Statistik?

Eine Forschungsgeschichte:

**Charlotte** macht sich an die **Recherche**.

Sie sucht nach einer Fragestellung, die noch nicht gelöst wurde und ihr machbar erscheint.

Sie liest eine Menge Bücher und die neuesten Paper..



# Warum Quantitative Methoden und Statistik?

Eine Forschungsgeschichte:

**Charlotte** findet heraus, dass Menschen sich im Durchschnitt 5 - 7 Dinge merken können.

Das ist die so genannte **Gedächtniskapazität**.

Das erscheint ihr etwas wenig...

Sie entschließt sich, ein **Mittel zur Erhöhung der Gedächtniskapazität** zu entwickeln.



# Warum Quantitative Methoden und Statistik?

Eine Forschungsgeschichte:

Nach einer Menge Arbeit...



Ist **Charlotte** mit dem Resultat zufrieden.

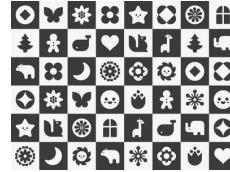


# Warum Quantitative Methoden und Statistik?

## Eine Forschungsgeschichte:

Nun muss sie das Mittel testen. Ihre Versuchsperson ist ihre Freundin **Anna**.

Annas Gedächtnisleistung im Spiel Memory: **5 Paare**



Dann gibt **Charlotte** Anna das Mittel zum Trinken.



**Nach dem Trinken** spielt Anna noch einmal.



Das Ergebnis: Anna merkt sich **8 Paare**.

# Warum Quantitative Methoden und Statistik?

Eine Forschungsgeschichte:

Charlotte ist begeistert! Ihr Mittel erhöht die Leistung um **3 Punkte**.

Mit dem Ergebnis geht sie zum Chef ihres Labors.

"Das Mittel wirkt vielleicht nur bei Anna. Zufall?"



# Warum Quantitative Methoden und Statistik?

## Eine Forschungsgeschichte:

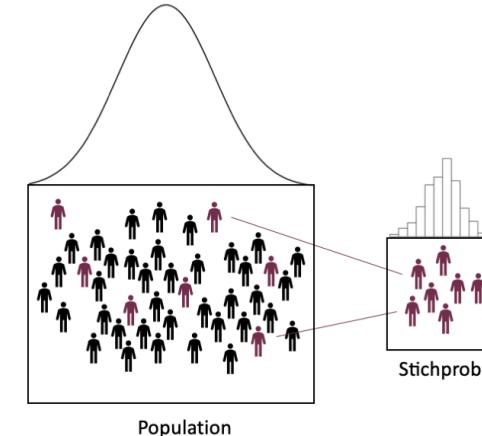
Charlotte sieht ein, dass ihre Entdeckung nur Bestand hat, wenn es bei mehr Leuten als Anna funktioniert.

Aber sie kann unmöglich prüfen, ob das Mittel bei jedem funktioniert

Dazu gibt es zu viele Menschen...

Wenn Charlotte nicht alle Menschen testen kann..

..dann zumindest eine kleinere Gruppe.



# Warum Quantitative Methoden und Statistik?

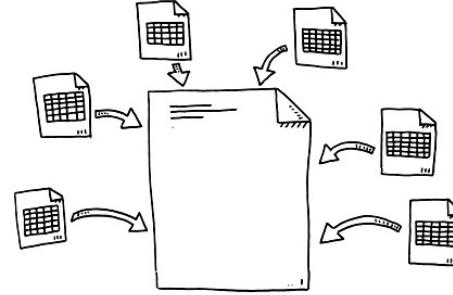
Eine Forschungsgeschichte:

Charlotte macht sich an die Arbeit...

Sie wiederholt ihr Experiment bei 100 Leuten.



Und trägt alle Daten in eine Tabelle ein.

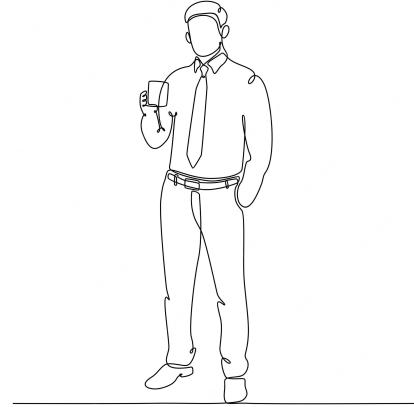


# Warum Quantitative Methoden und Statistik?

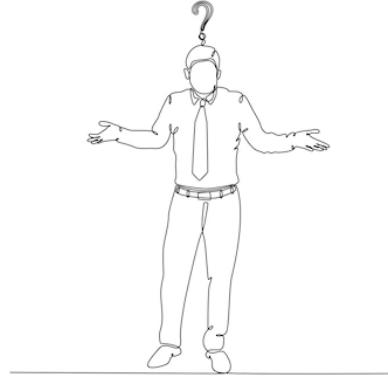
## Eine Forschungsgeschichte:

Nach langer Arbeit hat Charlotte ihre Stichprobe fertig erhoben.

Mit den 100 Zahlen geht sie zu ihrem Chef.



Die vielen Zahlen verwirren den Chef.



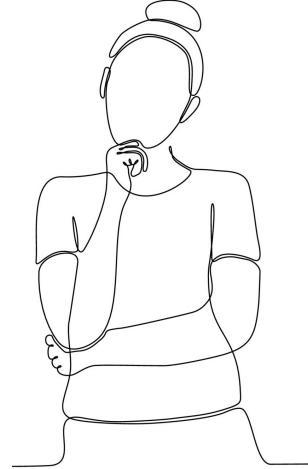
Bei einigen wird das Gedächtnis besser, bei anderen nicht.. Wirkt das Mittel nun, oder nicht?

# Warum Quantitative Methoden und Statistik?

Eine Forschungsgeschichte:

Charlotte macht sich an die Arbeit...

Sie muss die Daten irgendwie zusammenfassen.



Aus den 100 Werten berechnet sie den Durchschnitt.



Im Schnitt wird ihre Stichprobe um 2.5 Worte besser.

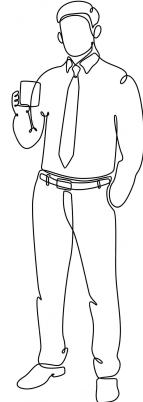
# Warum Quantitative Methoden und Statistik?

## Eine Forschungsgeschichte:

Mit dem Ergebnis geht Charlotte zu ihrem Chef.

"Chef, die Leute werden 2.5 Worte besser."

Der Chef ist sich unsicher, wie verlässlich die Schätzung aus der Stichprobe ist.



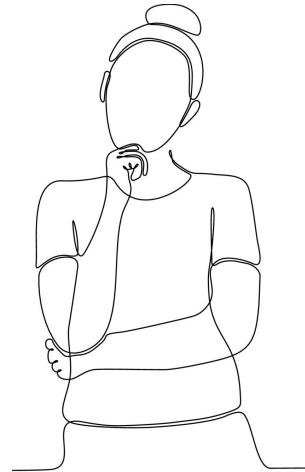
Wäre es nicht besser, sie hätten mehr Leute gefragt?

# Warum Quantitative Methoden und Statistik?

Eine Forschungsgeschichte:

Jetzt wird es kompliziert...

Wann ist ein Stichprobenwert eine verlässliche Schätzung für alle Menschen?



Charlotte trifft einige Entscheidungen

1. Sie will zumindest 95% sicher sein
2. Je mehr Leute sie testet, desto sicherer ist sie (Stichprobengröße)
3. Je näher die Ergebnisse der Testpersonen an 2.5 dran sind, desto sicherer ist sie (Streuung)

In einer Formel kann sie ein Konfidenzintervall berechnen, in das diese Infos eingehen:

$$\bar{x} \pm z \cdot \frac{\sigma}{\sqrt{n}} = 1.5; 3$$

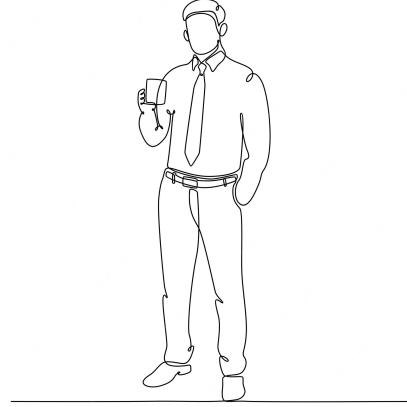
# Warum Quantitative Methoden und Statistik?

## Eine Forschungsgeschichte:

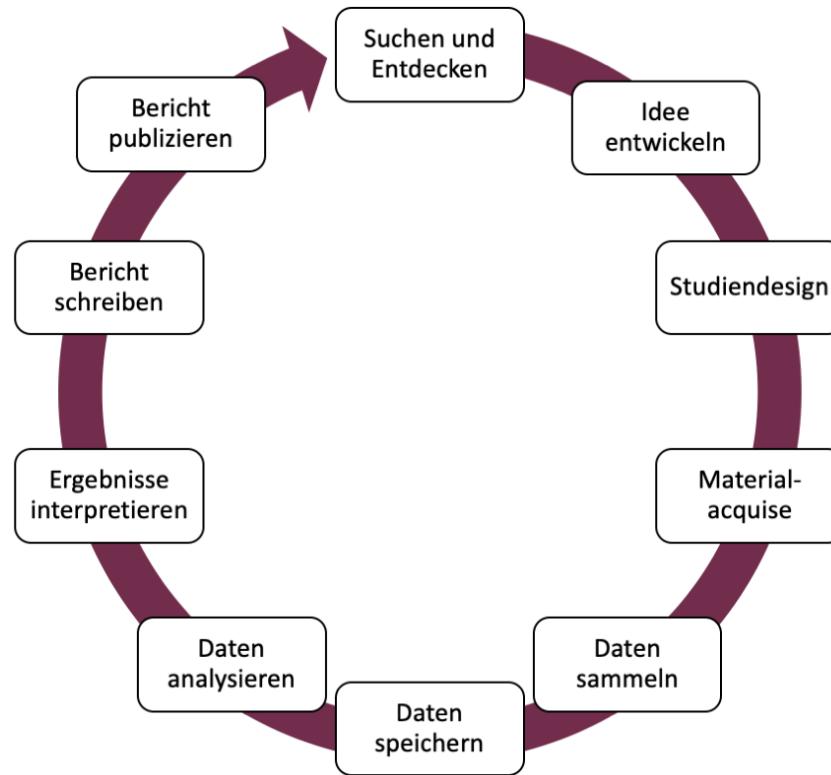
Mit dem Ergebnis geht Charlotte zu ihrem Chef.

"Chef, die Leute werden 2.5 Worte besser und dieser Wert liegt mit 95% Sicherheit zwischen 1.5 und 3."

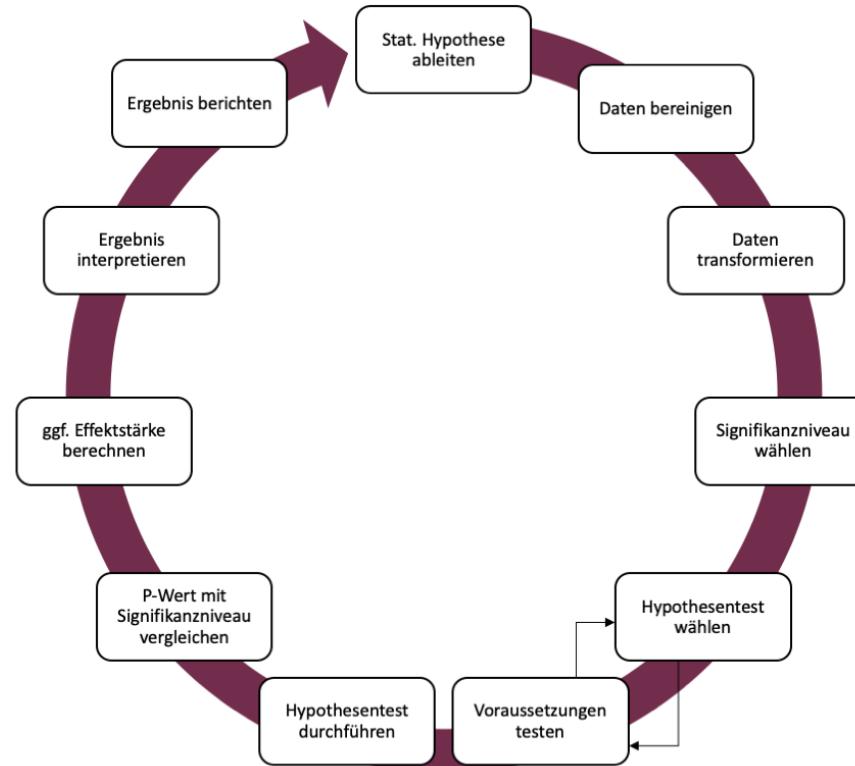
Der Chef ist zufrieden. Er kann sich zu 95% sicher sein, dass die Verbesserung nicht 0 oder schlechter ist.



# Quantitative Methoden im Forschungsprozess



# Prozess statistische Studienauswertung



# Begriffsklärung

- **Daten:** Beobachtungen der gesamten zu interessierenden Eigenschaften
- **Datensatz:** Zusammenstellung der beobachteten Daten (häufig im Matrixformat)
- **Datenbank:** bereits existierende archivierte Datenansammlungen (z.B. UK Biobank)
- **Statistik:** Verfahren zur Datenzusammenfassung & -analyse

Ausschnitt Datensatz:

Filter   Cols: << 1 - 50 >>														
ID	Age	Gender	OS	Education	Relationship_status	Job_status	NEO_1_N	NEO_2_E	NEO_3_O	NEO_4_V	NEO_5_G	NEO_6_N		
1	22	male	MacOs	Realschule	Single	in Ausbildung	4	4	0	4	0	4		
2	45	female	Android	Hauptschule	Single	erwerbstätig	4	4	0	4	0	3		
3	58	female	Android	Hochschulabschluss	Verheiratet/in Beziehung	erwerbstätig	4	4	1	4	4	0		
4	18	male	Android	Abitur	Single	arbeitslos	4	4	0	4	0	0		
5	38	female	Android	Realschule	Verheiratet/in Beziehung	erwerbstätig	4	4	0	4	0	0		
6	38	male	Android	Hauptschule	Single	erwerbstätig	4	4	4	4	0	0		
7	18	female	MacOs	Hochschulabschluss	Verheiratet/in Beziehung	in Ausbildung	4	4	0	1	0	0		
8	25	female	MacOs	Hochschulabschluss	Verheiratet/in Beziehung	erwerbstätig	4	4	0	4	0	4		
9	57	female	MacOs	Hochschulabschluss	Verheiratet/in Beziehung	erwerbstätig	4	4	0	4	1	2		
10	19	male	Android	Hochschulabschluss	Single	in Ausbildung	0	4	0	4	0	0		

# Anwendung von Statistik

- In dieser Übung werden wir auch Statistik **per Hand** rechnen
- Notwendigkeit zum mathematischen Verständnis der Verfahren
- In der Praxis wird beinahe nur noch digital gearbeitet (**Statistiksoftware**)
  - schneller
  - fehlerfreier
  - manche Verfahren zu komplex
  - Datensätze zu groß (Big Data)
  - bestimmte Verfahren basieren auf **Iteration** (z.B. Monte-Carlo Simulation teils 10.000x wiederholt)

→ würde per Hand Jahre dauern

- Bekannte Statistiksoftware:
  - R (wird an der CFH genutzt)
  - SPSS
  - JASP
  - Python

# Voraussetzung für Quantitative Methoden: wissenschaftliche Hypothesen

## Was sind Hypothesen?

Hypothese (griech.) = Unterstellung, Vermutung

Eine Vermutung/Annahme ist dann als wissenschaftliche Hypothese zu verstehen wenn sie folgende 4 Kriterien erfüllt:

Wenn sie...

1. ...sich auf reale Sachverhalte bezieht, die empirisch untersuchbar sind. (**Empirie**)
2. ...allgemein gültig ist und über den Einzelfall bzw. ein singuläres Ereignis hinausgeht (**All-Satz**)
3. ...zumindest implizit die Form eines Konditionalsatzes hat (**wenn-dann, je-desto**)
4. ...durch Erfahrungen potenziell widerlegbar ist (**Falsifizierbarkeit**)
5. ...(theoretisch begründbar ist)

## Beispiele für Hypothesen:

- Frauen sind kreativer als Männer
- Mit zunehmender Müdigkeit sinkt die Konzentrationsfähigkeit
- Je schöner das Wetter, desto besser die Stimmung
- Buben und Mädchen lesen unterschiedlich viel in ihrer Freizeit

Behauptungen erfüllen alle genannten Kriterien: sind daher Hypothesen

## UNIVERSELLE HYPOTHESE



# Alltagsvermutungen und wissenschaftliche Hypothesen

## Hypothesen: JA oder NEIN?

- Bei starkem Zigarettenkonsum kann es zu Herzinfarkt kommen
- Wenn es regnet, kann die Sonne scheinen.
- Es gibt Kinder, die niemals weinen.
- SchülerInnen aus Gymnasien zeigen gute Leistungen

# Alltagsvermutungen und wissenschaftliche Hypothesen

## Hypothesen: JA oder NEIN?

- Bei starkem Zigarettenkonsum kann es zu Herzinfarkt kommen

→ Kann-Sätze sind nicht falsifizierbar

- Wenn es regnet, kann die Sonne scheinen.

→ Kann-Sätze sind nicht falsifizierbar

- Es gibt Kinder, die niemals weinen.

→ kein All-Satz, nicht falsifizierbar

- SchülerInnen aus Gymnasien zeigen gute Leistungen

→ Wenn-dann Struktur nicht gegeben, daher nicht falsifizierbar

# Alltagsvermutungen und wissenschaftliche Hypothesen

## Hypothesen: JA oder NEIN?

- Die Konzentrationsfähigkeit hängt mit der Blutalkoholkonzentration zusammen.
- Positive Verstärkung durch Lehrer/innen kann zu guten Leistungen bei Schüler/innen führen.
- Positives Feedback beeinflusst die Arbeitsleistung.
- Viele Studierende mögen Methodenlehrveranstaltungen.

# Alltagsvermutungen und wissenschaftliche Hypothesen

## Hypothesen: JA oder NEIN?

- Die Konzentrationsfähigkeit hängt mit der Blutalkoholkonzentration zusammen.

→ JA

- Positive Verstärkung durch Lehrer/innen kann zu guten Leistungen bei Schüler/innen führen.

→ NEIN

- Positives Feedback beeinflusst die Arbeitsleistung.

→ JA

- Viele Studierende mögen Methodenlehrveranstaltungen.

→ NEIN

## Richtung von Hypothesen

Je nach Erkenntnisstand kann eine ungerichtete oder eine gerichtete Hypothese formuliert werden.

### **ungerichtete Hypothese:**

Die Konzentrationsfähigkeit hängt mit der Blutalkoholkonzentration zusammen.

→ Eher wenig theoretisches Vorwissen.

### **gerichtete Hypothese:**

Je höher die Blutalkoholkonzentration, desto niedriger die Konzentrationsfähigkeit.

→ Mehr theoretisches Vorwissen notwendig.

## Ableitung von statistisch-prüfbaren Hypothesen aus wiss. Hypothese

- Die Hypothese muss in numerische Ausdrücke umgewandelt werden.
- Man spricht von einem **Hypothesenpaar**:
  - **Nullhypothese ( $H_0$ )**: Der hypothetisierte Effekt besteht nicht.
  - **Alternativhypothese ( $H_1$ )**: Der hypothetisierte Effekt besteht
- In der Forschung hofft man oft, dass die  $H_1$  zutrifft (Hier steckt der angenommene Effekt drin)
- Man versucht "**Die  $H_0$  zu verwerfen**" (Hypothesentest kommt später)

## Ableitung von statistisch-prüfbaren Hypothesen aus wiss. Hypothese

Beispiel: **(Ungerichtete) Forschungshypothese:**

"Männer und Frauen sind im Schnitt unterschiedlich groß."

- $H_0$  Es besteht **kein Unterschied** zwischen der durchschnittlichen Größe der Männer (Mittelwert) und der durchschnittlichen Größe der Frauen.
- $H_1$  Es besteht **ein Unterschied** zwischen der durchschnittlichen Größe der Männer (Mittelwert) und der durchschnittlichen Größe der Frauen.

## Ableitung statistische Hypothese:

- $H_0$  Mittelwert Männer - Mittelwert Frauen = 0

→ **In Zahlen:** Wenn kein Unterschied besteht ist Differenz = 0 (z.B. 10 - 10 = 0)

- $H_1$  Mittelwert Männer - Mittelwert Frauen  $\neq$  0

→ **In Zahlen:** Wenn ein Unterschied besteht ist Differenz  $\neq$  0 (z.B. 15 - 10 = 5)

## Ziel empirischer Forschung:

Registrierte Merkmalsunterschiede (= **Variabilität**) zu analysieren und zu erklären.

- **Variable** = Interessierendes Merkmal, das unterschiedliche Ausprägungen annehmen kann
- Beispiele für Variablen:
  - Geschlecht
  - Lieblingsfarbe
  - Länge
- **Merkmalsausprägung** = konkrete Erscheinungsform einer Variable
- Beispiele für Merkmalsausprägungen:
  - Geschlecht [männlich, weiblich, divers]
  - Lieblingsfarbe [rot, gelb, grün, blau]
  - Länge [1 cm, 1.5 cm, 3 cm,...]

## Arten von Variablen

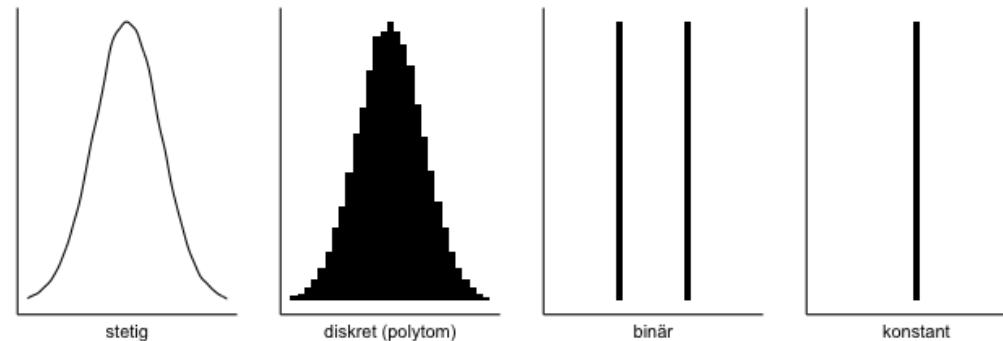
Variablen können anhand unterschiedlicher Eigenschaften unterteilt werden:

- nach Art der Merkmalsausprägungen
- nach empirischer Zugänglichkeit
- nach Stellenwert in der Untersuchung

Arten von Variablen – nach Art der Merkmalsausprägungen

- **stetig** (kontinuierlich): jedes Intervall besitzt unendlich viele Merkmalsausprägungen (z.B. Länge, Zeit, Masse)
- **diskret** (diskontinuierlich): Intervall mit endlich vielen Ausprägungen z.B. Geschlecht, Lieblingsfarbe
  - **dichotom** (binär) = 2 Abstufungen (0, 1)
  - **polytom** = mehrfach gestuft
  - **konstant** = nur 1 Merkmalsausprägung

→ Art der Variable bestimmt das statistische Verfahren (z.B. stetig → Regression, binär → logistische Regression)



Arten von Variablen – nach empirischer Zugänglichkeit

- **manifest** = direkt beobachtbar (Bsp. Raucher sein, Alter)
- **latent** = nicht unmittelbar beobachtbar; hypothetisches Konstrukt (Bsp. Intelligenz)

Arten von Variablen – nach Stellenwert in der Untersuchung

Variablen haben im empirischen Forschungskontext unterschiedliche funktionale Bedeutungen:

- abhängige Variable
- unabhängige Variable
- Störvariable
- Kontrollvariable
- Moderatorvariable
- Mediatorvariable

## Abhängige & unabhängige Variable (AV & UV)

Die Veränderung einer AV soll durch den Einfluss der UV erklärt werden.

**Beispiel:**

Dosis des Schlafmittel (**UV**) → Schlafdauer (**AV**)



**UV** gehört zum „Wenn-Teil“ bzw. dem „Je-Teil“ einer Hypothese

**AV** gehört zum „Dann-Teil“ bzw. „Desto-Teil“

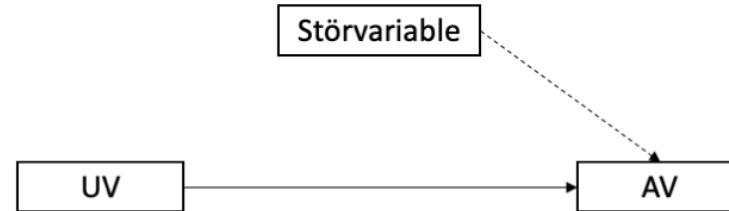


Wenn man mehr Schlafmittel nimmt, schläft man länger.



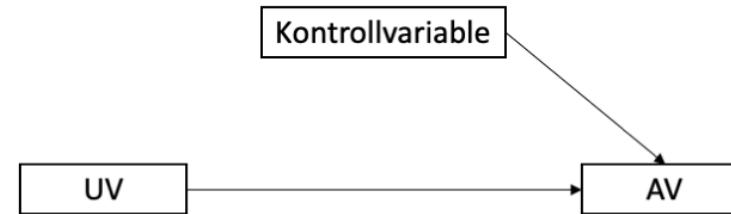
## Störvariable

- alle Einflussgrößen auf die AV, die in einer Untersuchung nicht erfasst werden
- egal ob nicht bekannt oder vergessen



## Kontrollvariable

- Störvariable deren Ausprägungen erhoben (gemessen) wurde
- Einfluss kann kontrolliert wird (z.B. mittels statistischer Methoden)

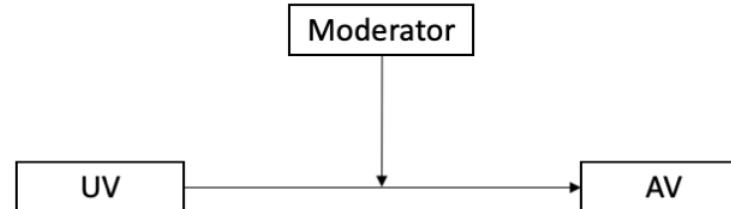


## Moderatorvariable

- **Moderator** verändert den Einfluss der UV auf die AV
- Moderationsanalyse prüft Interaktionen
- Frage: Variiert der Effekt von UV auf AV in Abhängigkeit einer weiteren Variable

### Beispiel:

Schlafmitteldosis (**UV**) erhöht die Schlafdauer (**AV**); Straßenlärm (**Moderator**) wirkt zusätzlich auf die **AV**

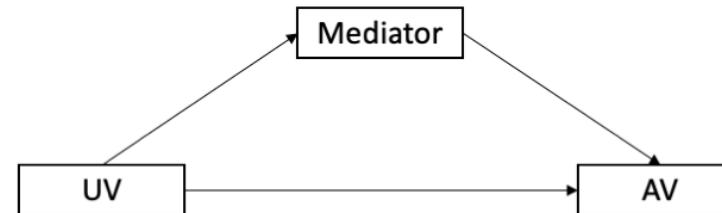


## Mediatorvariable

- **Mediator** vermittelt (**mediert**) den Einfluss der **UV** auf die **AV**
- **Indirekter Effekt:** **UV** beeinflusst **Mediator**, dies führt zur einem Effekt des **Mediators** auf **AV**
- **Direkter Effekt:** Effekt von **UV** auf **AV** (in Anwesenheit des Mediators)
- **Keine Mediation:** indirekter Effekt nicht signifikant
- **Partielle Mediation:** indirekter Effekt signifikant und direkter Effekt auch signifikant
- **Totale Mediation:** indirekter Effekt signifikant und direkter Effekt nicht mehr signifikant.

### Beispiel:

Schulnote (**UV**) beeinflusst Selbstwertgefühl (**Mediator**); Selbstwertgefühl (**Mediator**) beeinflusst Lebenszufriedenheit **AV**



# Begriffsklärungen

## Population:

- Menge aller potenziellen Untersuchungsobjekte (z.B. Personen), über die etwas ausgesagt werden soll
- **Grundgesamtheit**
- Ist im Allgemeinen zu groß
- Vollerhebung praktisch nicht möglich
- z.B. Grundgesamtheit aller Studierenden (auf der Welt)

## Stichprobe:

- Auswahl aus einer Population, die untersucht wird
- Wird nach verschiedenen Prinzipien **gezogen** (idealerweise **Zufall**)
- die Auswahl hat massive Konsequenzen auf die Ergebnisse (Bsp. **Verzerrungseffekte**)
- **Repräsentativität:** Stichprobe ist der Population in entscheidenden Charakteristika möglichst ähnlich
- je repräsentativer, um so gesicherter die **Verallgemeinerung** auf die Population

## Was macht Deskriptive Statistik?

- Bereitet Informationen über erfasste Merkmale auf
- Einzelwerte werden zu statistischen Kennwerten **zusammengefasst**
- Ziel: Beschreibung der Daten mittels Kennwerten, Graphiken, Tabellen, Diagrammen
- Bezieht sich auf die konkret untersuchte Stichprobe

## Beispiel:

Daten: 1, 23, 5, 6, 7, 8, 4, 43, 4, 5, 6, 5, 43, 34, 20, 6, 5, 34, 5, 6, 7, 7, 5, 4, 5, 3, 4, 5, 54, 43, 5, 54, 3, 4, 5, 5



Mittelwert (aka. Durchschnitt) = 13.42

## Beispiele

- Betrachtung der Verteilung von Häufigkeiten
- Maße der zentralen Tendenz
  - können Informationen über die durchschnittliche Ausprägung in einer Stichprobe liefern
  - Modalwert (diejenige Merkmalausprägung, die die meisten Elemente enthält)
  - Median (Wert bis zu dem 50% aller Werte liegen)
  - Arithmetisches Mittel (Mittelwert)
- Streuungsmaße
  - Varianz
  - Standardabweichung
  - Quartilabstand

## Was macht Inferenzstatistik?

- Synonym: schließende oder induktive Statistik
- "Schließen" von Stichprobe auf die Population
- Ziel: In der Stichprobe berechnete Statistik (z.B. Deskriptivstatistik) soll auf Population **verallgemeinert** werden
- Achtung: Habe ich die Möglichkeit, jedes Individuum meiner Population zu messen brauche ich keine Inferenzstatistik
- Methode: **Statistische Hypothesenprüfung** (Grundidee der **Signifikanztestung**)

## Beispiele

- Binomialtest: Auftretenshäufigkeit
- $\chi^2$ -Test: Unterschiede in Verteilungen
- unabhängiger t-Test: Unterschiede in Gruppenmittelwerten
- abhängiger t-Test: Unterschiede in Mittelwerten zwischen Zeitpunkten
- F-Test (ANOVA): Unterschiede in Varianzen
- Korrelation: Zusammenhänge
- Regression: Vorhersagen

- **N** Anzahl der Personen in der Stichprobe (Stichprobenumfang).
- Notation von Stichprobenvariablen: hier nutzt man oft Großbuchstaben, z.B.  $X$
- Hat man 2 Variablen nutzt man für die UV oft  $X$  und für die AV  $Y$
- Merkmalsausprägungen der einzelnen Personen in der Stichprobe bei Variable X (beobachtete Werte)

$x_1, \dots, x_2, \dots, x_n$

- $i$  steht für **Index**: Platzhalter für beliebige Zahl  $x_1, \dots, x_i, \dots, x_n$
- $k$  wird oft für die Anzahl von Variablen oder Gruppen genutzt

## Nur 1 Merkmal erhoben

- Notation der Befragungsergebnisse in einer Beobachtungsreihe (**Urliste**)
- Eine Zahlenreihe nennt man auch **Vektor** (eindimensional)

Beispiel:

- Erhobene Variable  $X$ : Alter in Jahren
- Stichprobenumfang  $N = 10$

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
50	34	70	33	22	61	69	73	62	56

→  $x3 = 70$  bedeutet, dass die Urliste an dritter Stelle eingetragene Person 70 Jahre alt ist.

## Nur 1 Merkmal erhoben

- Notation der Befragungsergebnisse in einer Beobachtungsreihe (**Urliste**)
- Eine Zahlenreihe nennt man auch **Vektor** (eindimensional)

Beispiel:

- Erhobene Variable  $X$ : Geschlecht
- Kategorien werden in der Statistik **codiert**:  $1 = \text{weiblich}$ ,  $2 = \text{männlich}$
- Stichprobenumfang  $N = 10$

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
1	1	1	2	1	2	2	2	1	1

→  $x3 = 1$  bedeutet, dass die Urliste an dritter Stelle eingetragene Person weiblich ist.

## Nur 1 Merkmal erhoben

- Notation der Befragungsergebnisse in einer Beobachtungsreihe (**Urliste**)
- Eine Zahlenreihe nennt man auch **Vektor** (eindimensional)

Beispiel:

- Erhobene Variable  $X$ : Name
- Wörter, die keine Kategorien sind müssen nicht **codiert** werden
- Stichprobenumfang  $N = 5$

x1	x2	x3	x4	x5
Max	Anna	Leo	Nina	John

→  $x3 = \text{Leo}$  bedeutet, dass der Name der in der Urliste an dritter Stelle eingetragenen Person "Leo" ist.

## Mehrere Merkmale erhoben

- Notation der Befragungsergebnisse in einer **Datenmatrix**
- **Matrix:** Reihen und Spalten (zweidimensional)
- **Variablen in Spalten** dargestellt (engl. Columns)
- **Personen in Reihen** dargestellt (engl. Rows)
- Die Matrix besteht aus  $n$  Zeilen und  $p$  Spalten (aka  $n \times p$ -Matrix)
- Einzelwerte (z.B. 1. Spalte, 1. Reihe) stehen in **Zellen**

# Notation von Daten

## Mehrere Merkmale erhoben

Beispiel Datenmatrix (3 Variablen):

- Alter
- Geschlecht [1 = *weiblich*, 2 = *männlich*]
- Anzahl Kinder
- Stichprobenumfang  $N = 10$

ID	Geschlecht	Alter	Kinder
1		1	71
2		1	33
3		1	73
4		2	44
5		1	45
6		2	46
7		2	24
8		2	70
9		1	46
10		1	76

## Aufgabe 1

Erstellen Sie eine **Datenmatrix** mit den folgenden Variablen

- Vorname
- Haarfarbe [ $1 = \text{schwarz}$ ,  $2 = \text{braun}$ ,  $3 = \text{blond}$ ,  $4 = \text{rot}$ ]
- Lieblingspsycholog:in (Nachname)
- Minitest: Wie viele Psycholog:innen kann die Person in 10 Sekunden nennen

Stichprobe: Personen in Ihrer Sitzreihe  $N = \dots$

# Summenzeichen

- In der Statistik benötigt man sehr oft die Summe von Messwerten
- z.B. Gesamtwert der Stichprobe (oder Teilstichprobe auf einer Variable)

Beispiel:

Summe aller Messwerte  $x_i$  für  $i = 1$  bis  $n$ .

$$x_1 + x_2 + x_3 + \dots + x_n$$

Hat eine Summe sehr viele Summanden, ist es zweckmäßig, das Summenzeichen (griech. Sigma) zu verwenden.

$$\sum_{i=1}^n x_i$$

- $i$  = Startwert
- $n$  = Endwert

## Beispiel 1:

Summe der Variable "Anzahl Kinder" aller Personen aus unserer Datenmatrix:

$$\sum_{i=1}^n x_i = 3 + 0 + 0 + 0 + 2 + 3 + 1 + 2 + 1 + 0 = 12$$

## Beispiel 2:

Summe der Variable "Anzahl Kinder" für die ersten 5 Personen aus unserer Datenmatrix:

$$\sum_{i=1}^5 x_i = 3 + 0 + 0 + 0 + 2 = 5$$

## Beispiel 3:

Summe der Variable "Anzahl Kinder" für die letzten 5 Personen aus unserer Datenmatrix:

$$\sum_{i=6}^{10} x_i = 3 + 1 + 2 + 1 + 0 = 7$$

# Summenzeichen

Es gelten die allgemeinen Rechenregeln für Additionen

## **Beispiel 1:**

$$\sum_{i=1}^n a - x_i = (a - x_1 + a - x_2 + \dots + a - x_n)$$

## **Beispiel 2:**

$$\sum_{i=1}^n ax_i = (ax_1 + ax_2 + \dots + ax_n) = a(x_1 + x_2 + \dots + x_n) = a \sum_{i=1}^n x_i$$

## Aufgabe 2:

Berechnen Sie mit korrekter Notation (Summenzeichen) die Gesamtzahl der genannten Psychologen (Mehrfachnennungen erlaubt) in Ihrer selbst erstellten Datenmatrix.

## Aufgabe 3:

$$X = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10]$$

Berechnen Sie folgenden Ausdruck:

$$\sum_{i=1}^n 11 - x_i$$

# Häufigkeitstabelle

- Ziel: Daten effizient zusammenfassen
- **Häufigkeit:** Anzahl der Ausprägungen eines Merkmals
- z.B. zur Beschreibung der Stichprobe in klinischer Studie
- Man unterscheidet **absolute** ( $n$ ) vs. **relative** (%) Häufigkeit

Beispiel: Urliste Lieblingstier  $N = 10$

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
Hund	Katze	Hund	Elefant	Hund	Adler	Katze	Delphin	Maulwurf	Goldfisch

Häufigkeitstabelle (absolute Häufigkeiten):

Adler	Delphin	Elefant	Goldfisch	Hund	Katze	Maulwurf
1	1	1	1	3	2	1

# Häufigkeitstabelle

Beispiel: Urliste Lieblingstier  $N = 10$

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
Hund	Katze	Hund	Elefant	Hund	Adler	Katze	Delphin	Maulwurf	Goldfisch

Häufigkeitstabelle (absolute Häufigkeiten):

Adler	Delphin	Elefant	Goldfisch	Hund	Katze	Maulwurf
1	1	1	1	3	2	1

Häufigkeitstabelle (relative Häufigkeiten ( $absolut/N$ ) \* 100):

10%	10%	10%	10%	30%	20%	10%
-----	-----	-----	-----	-----	-----	-----

# Häufigkeitstabelle

- Absolute und relative Häufigkeit **beide wichtig** für das Verständnis von Daten
- **In Publikationen** werden i.d.R. beide angegeben und oft im Format  $N(%)$  berichtet

Beispiel: Urliste Lieblingstier  $N = 10$

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
Hund	Katze	Hund	Elefant	Hund	Adler	Katze	Delphin	Maulwurf	Goldfisch

Häufigkeitstabelle:

Adler	Delphin	Elefant	Goldfisch	Hund	Katze	Maulwurf
1 (10)	1 (10)	1 (10)	1 (10)	3 (30)	2 (20)	1 (10)

## Aufgabe 4

- Erstellen Sie eine Urliste für folgende Variable
- Vorliebe Psychotherapie Leitlinienverfahren [1 = *Psychoanalyse*, 2 = *Verhaltentherapie*, 3 = *Tiefenpsychologisch*, 4 = *Systemisch*]
- Erstellen Sie aus der Urliste eine Häufigkeitstabelle für die Ausprägungen der Variable in der Schreibweise  $N(\%)$ .

# Take-aways

- Unterscheidung: **Deskriptiv- vs. Inferenzstatistik**
- Deskriptivstatistik: Stichprobendaten **zusammenfassen** (z.B. Mittelwert)
- Inferenzstatistik nutzen wir, um Ergebnisse unter Wahrscheinlichkeitsannahmen auf Population zu **verallgemeinern** (z.B. Konfidenzintervall)
- Quantitative Methoden: Erkenntnisgewinn durch **Hypothesentesten**
- Variablen werden i.d.R. in einer **Datenmatrix** erfasst
- Auprägungen von Variablen können mit **Häufigkeitstabelle** zusammengefasst werden
- Wichtige Deskriptivstatistiken: **Relative und absolute Häufigkeiten**