BERUFSAKADEMIE SACHSEN
STAATLICHE STUDIENAKADEMIE
RIESA
UNIVERSITY OF COOPERATIVE EDUCATION

Einführung in die Statistik

Ziel von Statistik

„Statistiken lügen zwar nicht, vertuschen aber sehr gern.“

(Martin Gerhard Reisenberg)

„Mit Statistiken kann man alles beweisen, nur nicht die Wahrheit.“

(Unbekannt)

Aufgabe Zitate:

1. Wählen Sie eines der beiden Zitate
2. Stellen Sie so viel Fragen wie möglich zu dieser Behauptung auf! Bewerten Sie dabei keine Ihrer entwickelten Fragen.
3. Versuchen Sie nun alle offenen Fragen in geschlossene und alle geschlossene in offene Fragen umzuwandeln.
4. Entscheiden Sie sich für 2-4 interessanteste / wichtigste / dringendsten Fragen.

Dauer: 20 Minuten

Abgabeformat: Textdatei

Warum Statistik? Sie analysieren Prozesse / Strukturen / Veränderungen / Zusammenhänge in Ihrem Unternehmen / Kunden / Konkurrenten

Risikoeinschätzung eines Kreditnehmers

Mitarbeiterzufriedenheit in Abhängigkeit von Stress analysieren

Verbesserungen durch neu Maßnahmen auch überprüfen

Fehlerquote bei Produkten überwachen durch Stichproben

Warum Statistik? Um Aussagen zu generieren

Statistische Tests anwenden um praxisrelevante Fragen zu beantworten z.B.:

- Wie gut misst der Mittelwert?
- Unterscheidet sich Gruppe A von Gruppe B.
- Gibt es einen Zusammenhang zwischen X und Y.

Dazu nutz man statistische Tests. Solche Tests beruhen auf **Wahrscheinlichkeitstheorie** und diese teilweise auf **Kombinatorik**.

Warum Statistik? Varianz verheimlichen

Aussage: Im Durchschnitt machen die Deutschen 32 Überstunden im Jahr.

Das Problem hier: das trifft nicht für jeden zu, besser wäre die Angabe eines Intervalls.

Es gibt ein Intervall: Das Konfidenzintervall, es umfasst den „wahren Wert“ zu 95 Prozent %.

Warum Statistik? Den Begriff der Signifikanz richtig verstehen

Typische Formulierungen in wissenschaftlichen Texten verstehen:

Das A Medikament ist signifikant wirksamer als B.

→ Einfach: A ist besser als B.

Von 100 getesteten Mitteln hatten fünf signifikante Wirkung.

→ Achtung! Am Ende der Statistik Veranstaltung werden Sie verstehen, dass dies kein gutes Zeichen ist.

Aufgaben zur Einführung

Die Aufgaben auf den folgenden Folien können Sie teilweise „aus dem Bauch heraus“ beantworten. Es macht nichts, wenn Sie Ihre Antworten falsch sind.

Aufgabe Creme: Sie haben eine Feuchtigkeitscreme gekauft. Auf der Verpackung lesen Sie „Sie hilft zu 78 %“ (getestet an 16 Konsumenten). Wie interpretieren Sie dies?
Dauer: 2 Minuten **Abgabeformat:** Textdatei

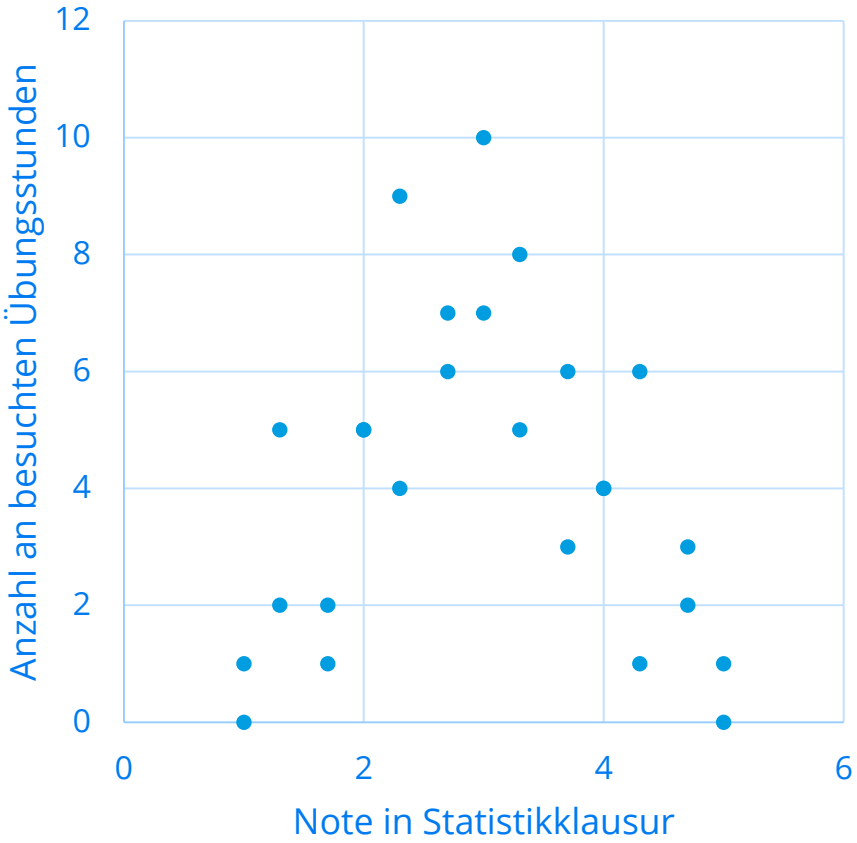
Aufgabe Einkommen der Nachbarn: Drei Ihrer Nachbarn verdienen monatlich 1000 Euro und einen der 10000 Euro verdient. Würden Sie nun sagen, dass ihre Nachbarn im Durchschnitt 3250 Euro verdienen?
Dauer: 2 Minuten **Abgabeformat:** Textdatei

Aufgabe Film: Sie möchten einen Film anschauen. Zuvor betrachten sie die Bewertung des Films. Die mögliche Bewertung geht von 1 (sehr schlecht) bis 10 (sehr gut). Unter Männern vergeben 6,9 % 10 Punkte und 16,7 % nur einen Punkt. 19 % der Frauen geben 10 Punkte und 9 % einen Punkt. Wer ist die Zielgruppe des Films?
Dauer: 2 Minuten **Abgabeformat:** Textdatei

Warum Statistik?

Beispiel: Auswertung einer Statistikklausur

Aufgabe Übungsstunden und Note: 1. Wie interpretieren Sie die nebenstehende Grafik? 2. Was können Sie den Zusammenhang auch erklären?
Dauer: 5 Minuten
Abgabeformat: Textdatei



Warum Statistik? Lügen aufdecken

Aufgabe Mitarbeiter prüfen:
Ein Mitarbeiter behauptet, dass er täglich 100 Stück eines Produktes herstellen kann. Nach dem sie 10 Tage lang seine Produktion überprüft haben, stellen Sie fest, dass der Mitarbeiter im Durchschnitt nur 99 Stück herstellt. Hat der Mitarbeiter gelogen?
Dauer: 1 Minuten **Abgabeformat:** keine

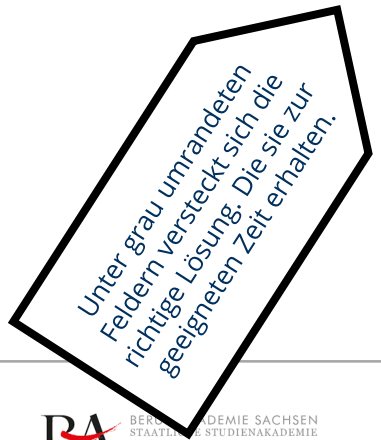
Warum Statistik? Amazons Algorithmus zur Einstellung von Bewerbern bevorzugte Männer

Aufgabe Amazon:
Lesen Sie den Bericht über die Amazons (verworfene) Einstellungspraxis bei Bewerbern: https://www.golem.de/news/machine-learning-amazon-verwirft-sexistisches-ki-tool-fuer-bewerber-1810-137060.html
Dauer: 3 Minuten **Abgabeformat:** keine

Warum Statistik? Definition / Messung von Konstrukten

Die Zahl der Armen fällt!

Die Zahl an Krebsfällen steigt!



Aufgabe Arme und Krebsfälle:
In einer Zeitung stehen die Überschriften „Die Zahl der Armen fällt!“ und „Die Zahl an Krebsfällen steigt!“
Wie kann es zu beiden Ergebnissen kommen, ohne das in Wirklichkeit sich etwas geändert hat?
(Die Überschrift dieser Folie gibt Ihnen einen Hinweis)
Dauer: 5 Minuten
Abgabeformat: Textdatei

Warum Statistik? Formulierung von Fragen

Statistik braucht Daten. Daten müssen einwandfrei sein. Dazu gehören auch gute Fragen in einem Fragebogen.

Sind Sie nicht auch der Meinung, dass man mehr für den Umweltschutz tuen sollte?

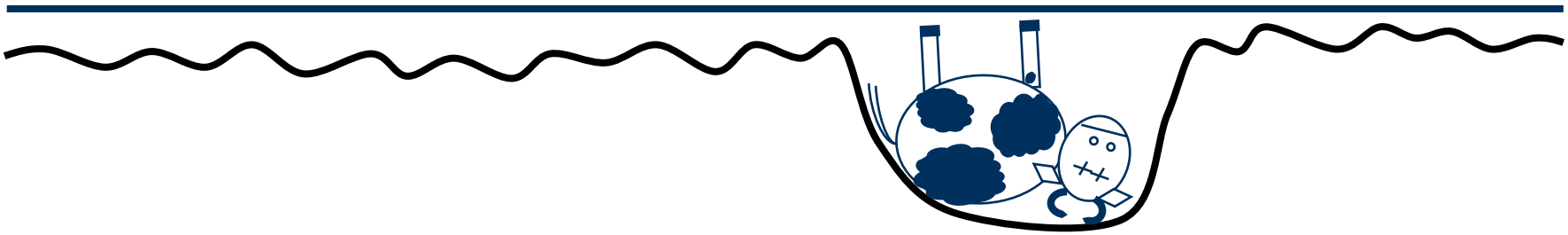
O Ja

O Nein

Aufgabe Frageformulierung:
Was ist an dieser Frage falsch formuliert? Das ist keine rein statistische Frage.
Dauer: 2 Minuten **Abgabeformat:** keine Abgabe

Folien zur Einführung

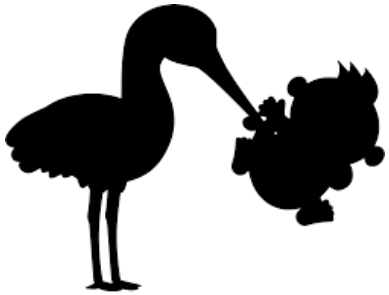
Die folgenden Folien sind eine heitere Art des Umgangs mit Statistik und schließen die Einführung ab.



Im Durchschnitt war der See einen Meter tief, und trotzdem ist die Kuh ertrunken. (Man muss die Verteilung verstehen.

Warum Statistik? Scheinkausalität entdecken

Scheinkausalität kennen Sie auch von dem Zusammenhang von Störchen und Babys



Aufgabe Störche:
Welche Scheinerklärung kennen Sie, die den Zusammenhang zwischen Babys und Störchen erklärt? Geht es um Raum oder Zeit?
Lesen Sie auf Wikipedia die Erklärung: https://de.wikipedia.org/wiki/Scheinkorrelation
Dauer: 5 Minuten **Abgabeformat:** keine Abgabe

Warum Statistik? Scheinkausalität entdecken

"I used to think correlation implies causation. Then i took a statistics class. Now i don't."

"Sounds like the class helped"

"Well, maybe."

Weitere verrückte Scheinkorrelationen finden Sie hier:
https://scheinkorrelation.jimdofree.com/

Warum Statistik? Umgang mit Schlussfolgerungen in der Sozialwissenschaft, Physik, Mathematik

Ein Sozialwissenschaftler, ein Physiker und ein Mathematiker fahren mit einem Zug durch die Schweiz. Als sie aus dem Fenster schauen, entdecken sie auf einem Acker ein schwarzes Schaf. Der Sozialwissenschaftler, der noch nie in der Schweiz war und hier das erste und bisher einzige Schaf dieses Landes kennen lernt, folgert messerscharf: “Aha – in der Schweiz sind alle Schafe schwarz!” Der Physiker denkt, er wäre schlauer und macht sich sogleich über den SoWi lustig: “Das ist eine völlig unerlaubte Verallgemeinerung – das einzige, was du sagen kannst, ist: Es gibt in der Schweiz ein schwarzes Schaf.” Der Mathematiker, der sich bisher nicht an der Diskussion beteiligt hatte, kann daraufhin nur müde lächeln und meint: “Auch das ist völliger Unsinn. Du kannst nur behaupten: Es gibt in der Schweiz ein Schaf, das von einer Seite schwarz ist!”

Warum Statistik? Illusion von Präzession und Umgang mit Zahlen

1. 55,2304358% aller Statistiken täuschen eine viel zu hohe Genauigkeit vor und sind damit immun vor Kritik.
2. Und 115 % aller Statistiken beruhen auf übertriebenen Zahlen.
3. Es werden **immerhin** in 1 von 6 Fällen Grafiken richtig dargestellt, während **nur** in 20 Prozent Tabellen richtig sind.
4. Die Krebsrate der Bewohner um Atomkraftwerke 500 % größer als normalerweise, sie beträgt 0,0005 %.
5. Die Statistik beweist, dass 97% aller Statistiken uns nichts sagen. (Erhard Blanck)
6. Achtung 99,999 % alle Statistiken auf dieser Folie sind ad hoc entwickelt.

BERUFSAKADEMIE SACHSEN
STAATLICHE STUDIENAKADEMIE
RIESA
UNIVERSITY OF COOPERATIVE EDUCATION

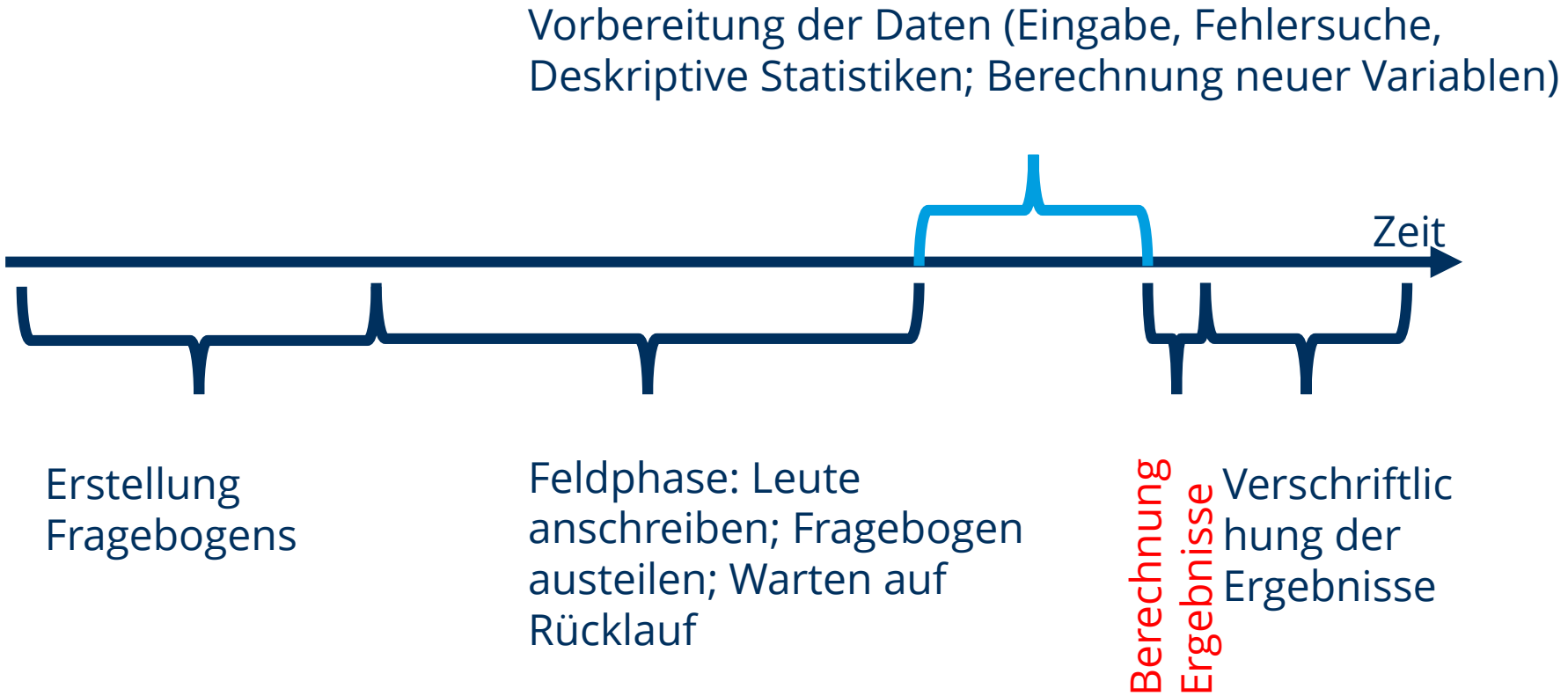
Datenmanagement

Bevor Sie Statistik betreiben können, benötigen Sie Daten, diese müssen eingegeben warden. Es gibt verschiedene Arten von Daten.

Datenmanagement: Bedeutung

Wir beginnen nun mit dem Datenmanagement. Denn Statistik beruht auf Daten.

Benötigte Zeit für das Datenmanagement im Vergleich zu anderen Schritten:



Klassische Situation für die Nutzung von Statistik

Sie haben in Ihrem Unternehmen z.B. eine neue Kommunikationsplattform probiert, das z.B. die Mitarbeiterzufriedenheit heben soll. Nach und vor der Einführung haben Sie eine **Mitarbeiterbefragung** gemacht mit Hilfe eines Fragebogens aus Papier.

Sie wollen wissen,

- gibt es eine Verbesserung der Zufriedenheit.
- unterscheiden sich die Zufriedenheit in Abteilung A und B?

Folgende drei Schritte bieten sich nun an:

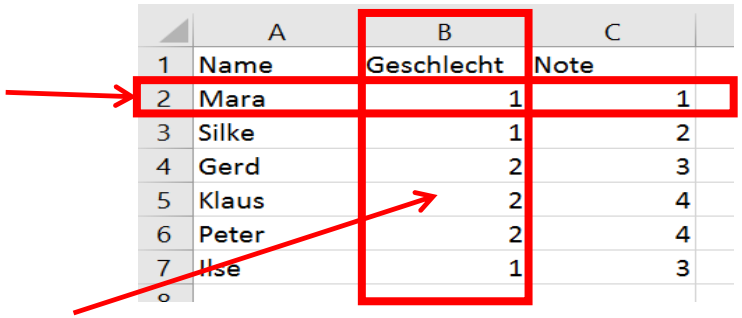
1. Schritt: Bringen Sie die Daten in Tabellenform.
2. Schritt: Analysieren Sie die Daten deskriptiv (z.B. Mittelwert)
3. Schritt: Analysieren Sie die Daten mittels schließender Statistik

Dateneingabe: Fälle und Variablen

Bei der Übertragung von Daten in (Excel-)Tabellen sollten Sie so vorgehen, wie es auf dieser Folie gezeigt wird. Dies ist der Standard.

Zeilen = Fälle (auch Beobachtung genannt) Personen, Staaten, Schulklassen, allgemein auch als Untersuchungsobjekt bezeichnet

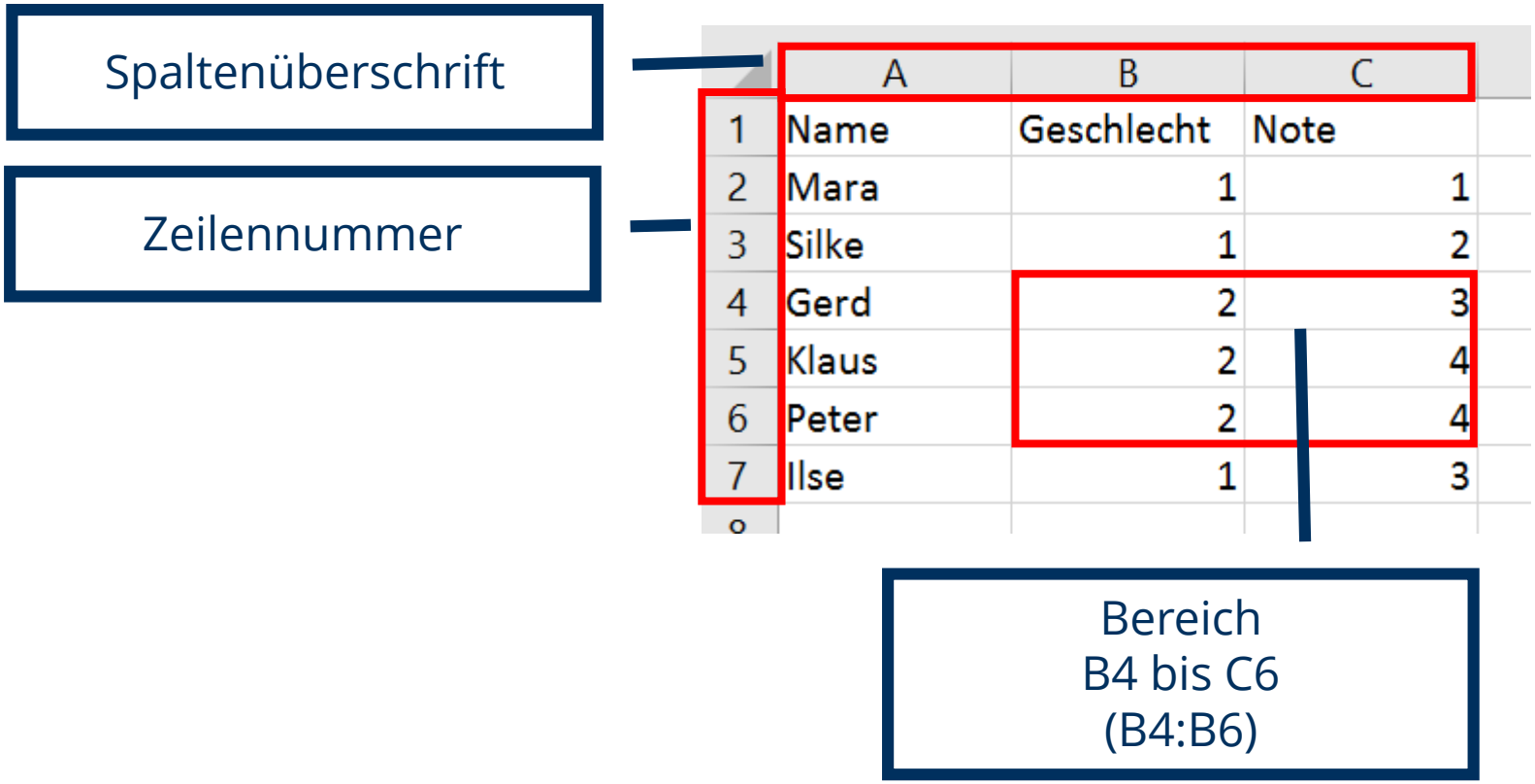
Spalten = **Variablen**, Merkmale/Eigenschaften von Personen/Untersuchungsobjekten (Größe, Name) Frage im Fragebögen



	A	B	C
1	Name	Geschlecht	Note
2	Mara	1	1
3	Silke	1	2
4	Gerd	2	3
5	Klaus	2	4
6	Peter	2	4
7	Ilse	1	3

Zellennamen

Sie können jedem Bereich in Excel auch einen Zellennamen vergeben.



Aufgabe Fälle/Variablen:

1. Was kann von dem Folgendem alles ein **Fall** sein?: Menschen, Personen, Gruppen, Staaten, Organisationen, Planeten
2. Was von dem Folgendem kann eine **Variable** sein? Alter, Geschlecht, IQ oder Größe, Jahresgewinn eines Unternehmens

Dauer: 5 Minuten **Abgabeformat:** Textdatei

Skalenniveaus

Variablen können verschiedene Arten/Typen oder besser „Skalen“ haben.

Symbol bei SPSS	**Skala**	**Merkmale**
	Nominal-Skala	Klassifizierung qualitativer Merkmale
	Ordinal-Skala	Rangwerte mit Ordinalzahlen *Diskrete Daten*
	Metrische-Skala	Gleiche Abstände *(Stetig Daten)*

Zur Vollständigkeit: die metrische wird noch in Intervallskala unterschieden (dann hat die Variable einen natürlichen Nullpunkt. In der statistischen Praxis spielt das aber selten eine Rolle.

Zusätzliche Informationen: Ordinale Daten können manchmal als metrisch interpretiert werden

Die Information auf dieser Seite können Sie sich mal überfliegen.
Sie kann für Ihre praktische Tätigkeit relevant werden.
Tatsächlich gibt es in der Wissenschaft heftige Diskussionen.

Es gibt statistische Tests für ordinale und statistisches Tests für metrische Daten. Theoretisch und praktisch sind die Ergebnisse meist sehr ähnlich.

Es gibt nun empirische Belege, dass einige ordinale Skalen auch als metrische Skalen interpretiert werden können: z.B. wenn es sich z.B. um Einschätzungsfragen auf eine Skala von 1 bis 10.

Bei ordinalen Daten wie der Schultyp (Hauptschule, Realschule, Gymnasium) – wo die Abstände zwischen den einzelnen Kategorien nicht gleich definiert sind, sollte die Skala als ordinal interpretiert werden.

Im Notfall können Sie die ordinalen Skalen auch als nominale interpretieren und z.B. als Gruppen behandeln. Dann können Sie je nach Art der Daten einen t-Test / Anova oder Chi-Quadrat-Test durchführen.

Aufgabe: Dateneingabe 1:

Sie haben einen Fragebogen (für Jugendliche) erstellt der folgende Fragen enthält. Erstellen Sie eine Tabelle, die alle Variablen enthält. Achtung einige Variablen sind einfach andere schwerer umzusetzen.

- Name;
- Alter;
- Geschlecht;
- Abitur (ja/nein);
- Wie oft machen Sie Fitnesstraining (sehr selten, eher selten, eher häufig, sehr häufig)
- Wie hoch ist Ihre Motivation zur Schule zu gehen? (Skala von 1 bis 5)
- Ihr Berufswunsch? (offene Frage)
- Was sind Ihre **drei** Lieblingsfächer? (Liste der Fächer: Sport, Deutsch, Mathe, Physik, Biologie, Sozialwissenschaft, Musik)

Erfinden Sie vier Fälle und erfinden Sie dafür die Daten. Welches Skalenniveau liegt jeweils vor?

Dauer: 15 Minuten **Abgabeformat:** (Excel-)Tabelle

Übung Dateneingabe 1: Lösungen

Name	Alter	Geschlecht	Notendurchschnitt	Wohnort (Stadtteil)	Abitur	Wie oft Fitnesstraining	Motivation für Schule (1-5)	Berufswunsch

Aufgabe Dateneingabe 2:

Fügen Sie folgende zwei Fragebogenfragen auch in die Tabelle ein:

Welche Sportarten übst Du aus (Joggen, Farhrradfahren, Fußballspielen, Sonstiges und zwar _____:

Sortieren Sie folgende Dinge nach Beliebtheit?

Äpfel, Birnen, Pflaumen, Bananen?

Dauer: 10 Minuten **Abgabeformat:** (Excel-)Tabelle

Aufgabe Dateneingabe 3:

Fügen Sie folgende zwei Fragebogenfragen auch in die Tabelle ein:

Wissensfrage 1: Wann ist Weihnachten? Im November, im Dezember, Im Januar.

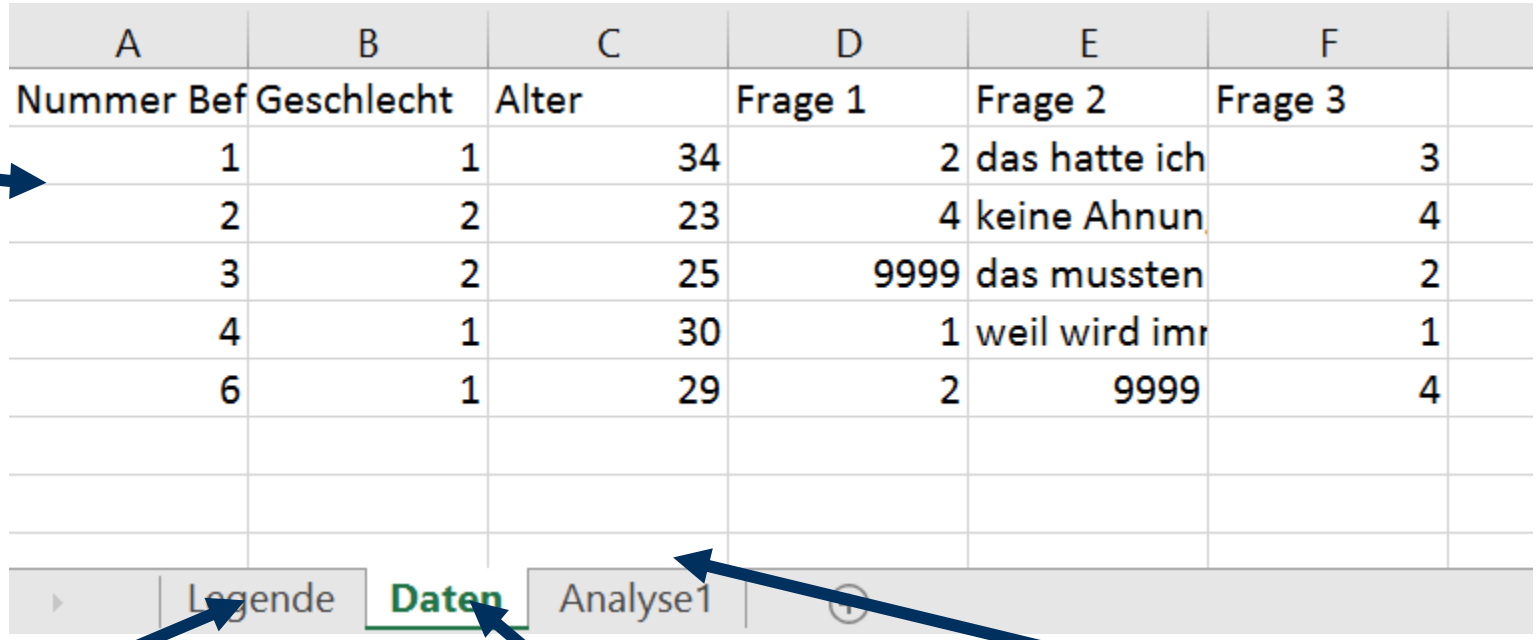
Wissensfrage 2: Wann ist Ostern?

Dauer: 10 Minuten **Abgabeformat:** (Excel-)Tabelle

Tipps: Datenmanagement in Excel

Sortieren Sie Ihre Daten so: Erste Zeile = Überschriften-der einzelnen Fragen.

Nummerieren Sie jeden Papierfragebogen auch auf dem Papier und nutzen Sie die Nummer im Datenblatt



A	B	C	D	E	F
Nummer Bef	Geschlecht	Alter	Frage 1	Frage 2	Frage 3
1	1	34	2	das hatte ich	3
2	2	23	4	keine Ahnun	4
3	2	25	9999	das mussten	2
4	1	30	1	weil wird imr	1
6	1	29	2	9999	4

Legende | Daten | Analyse1

Im Tabellenblatt Legende beschreiben Sie für jede Variable die Codierung u.Ä. z.B. was bedeutet eine 9999.

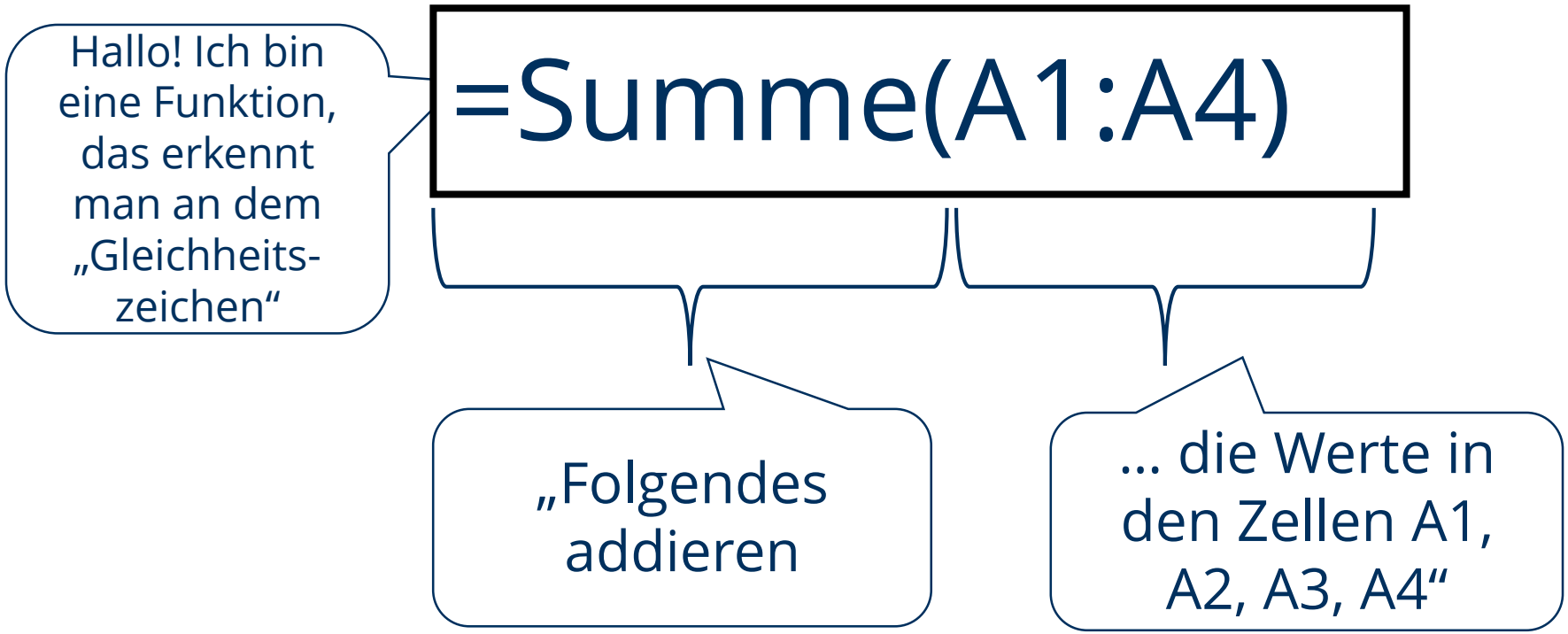
Im Tabellenblatt Daten/Rohdaten etc. ist die Originalversion ihrer Daten.

In weiteren Tabellenblättern können Sie Ihre Analysen eingeben. Dort können sie ggf. Kopien der Variablen einfügen, für eine leichtere Analyse.

Umcodieren Excel: Die Funktionen

Für Anfänger etwas schwierig ist das „Ausdenken und Neuformulieren“ von Variablen und das Anpassen von Variablen an gewünschte statistische Tests daher üben wir das auf den nächsten Folien.

Wenn Sie erprobt im Umgang mit Excel sind, dann ist das eine Widerholung für Sie, wie in Excel Funktionen ausgedrückt werden. Wir benötigen einfache Funktionen um neue Variablen zu berechnen. Hier stellt sich eine Formel vor:



Aufgabe Neuberechnung Formel:

A. Erstellen Sie eine Kundendatei: Erfinden Sie vier Variablen mit der Summe in Euro an bestellten Waren für 1. bis 4. Quartal
B. Tragen Sie die Antworten für fünf fiktive Kunden ein.
C. Erstellen und errechnen Sie vier neue Variablen für alle Kunden mit

- 5. Variable: Durchschnittlichen Summe in Euro an bestellten Waren
- 6. Variable: Totale Summe in Euro an bestellten Waren
- 7. Variable: Differenz zwischen maximaler und minimaler Summe in Euro an bestellten Waren
- 8. Variable: Erfinden Sie eine weitere Variable, die auf (1-7) vorangegangen Variablen basiert

Dauer: 20 Minuten **Abgabeformat:** (Excel-)Tabelle (nutzen Sie einfach ein neues Tabellenblatt in Ihrer bereits erstellten Datei.

Umcodieren

Wenn Sie einen Fragebogen haben, bei dem die Daten eine andere Form haben, als Sie brauchen müssen Sie die Daten oft umcodieren.

Simple Möglichkeiten:

- Sortieren und manuell mit Werten ersetzen
- Suchen und Ersetzen (evtl. in eigenem Tabellenblatt)
- Einfache Berechnungen (siehe nächste Folie)

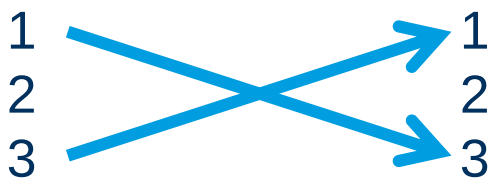
Fortgeschrittene Möglichkeiten:

- In Excel die Funktion „=Wenn()“ nutzen.

Tipp: Wenn Sie Papierfragebögen haben, dann sollten Sie die Daten gleich in das ideale Datenformat überbringen.

Umcodieren: Beispiele

Beispiel 1: Aus dem Wert 1 soll der Wert 3 gemacht werden



In dem Fall ist das relativ einfach mit einer Funktion: =4-Wert

Beispiel 2: Aus den Originaldaten („Motivation“) wurde eine neue Variable generiert („MotivationZahlen“).

ID	Motivation	MotiviertZahlen
1	Sehr motiviert	1
2	Eher motiviert	2
3	Sehr unmotiviert	4
4	Sehr motiviert	1
5	Sehr unmotiviert	4
6	Sehr motiviert	1
7	Eher motiviert	2
8	Eher unmotiviert	3
9	Eher unmotiviert	4
10	Sehr motiviert	2

Aufgabe Umcodieren:

Sie haben eine Online-Befragung mit 15 Teilnehmern durchgeführt. Nun haben Sie eine vollständig ausgefüllte Excel-Datei zum Thema Motivation auf einer Skala von 1-4 (sehr motiviert, eher motiviert, eher unmotiviert, sehr unmotiviert). Leider hat die Onlinesoftware die Zahlen anders codiert, als Sie es möchten. Sie wollen überall dort wo originale eine 1 ist eine 5, wo eine 2 ist eine 4 usw. Was wäre der schnellste Schritt in Excel für diese Umcodierung?

1 → 5
2 → 4
3 → 3
4 → 2
5 → 1

Dauer: 10 Minuten **Abgabeformat:** Textdatei oder Excel.

Fehlende Werte

Wenn ein Teilnehmer auf eine Frage keine Antwort gegeben hat, nenn man das fehlende Werte.

Nr.	Motivation	Arbeitsleistung
1	2	3
2	3	4
3	4	2
4	5	
5	2	2

Aufgabe fehlende Werte:

Was machen Sie mit fehlenden Daten? Nennen Sie mindestens zwei Lösung.

Dauer: 5 Minuten **Abgabeformat:** Textdatei

Weitere Tipps für das Datenmanagement

Grundlage der für die Analyse sind beispielsweise Fragebögen. Diese sollten einwandfrei sein (Pretest, mehrmals Korrektur lesen lassen, Literatur zur Fragebogenkonstruktion). Das ist nicht Thema dieser Veranstaltung.

Für eine bessere Übersicht:

- neu berechnete Variablen immer wieder löschen, wenn sie nicht gebraucht werden.
- Aber: Niemals Stammvariablen löschen
- (In SPSS Syntax verwenden, damit jederzeit die Variablen schnell neu berechnet werden können.)

Das ist das Ausdenken und Neuformulieren von neuen Variablen und das Anpassen von Variablen an gewünschte statistische Tests erfordert Erfahrung und/oder Kreativität.

BERUFSAKADEMIE SACHSEN
STAATLICHE STUDIENAKADEMIE
RIESA
UNIVERSITY OF COOPERATIVE EDUCATION

Häufigkeiten

Häufigkeiten

Verkaufte Produkte pro Monat

	Absolute Häufigkeit	Relative Häufigkeit	Kumulierte Häufigkeit
1. Quartal	23		
2. Quartal	34		
3. Quartal	32		
4. Quartal	12		

Aufgabe Häufigkeiten:
a) Bilden Sie die relativen und kumulierten Häufigkeiten.
b) Welche der drei Häufigkeitsauswertungen sind leichter zu interpretieren? Begründen Sie Ihre Antwort.

Dauer: 10 Minuten **Abgabeformat:** Textdatei/Exceldatei

Übungen

Aufgabe Häufigkeiten Studierende:

Wir haben einen Datensatz der folgende Variablen für 10 Fälle beinhaltet:

ID Nummer des Befragten

A) Studiendauer in Semester

B) Engagement im Studium: Skala 1 (sehr engagiert) bis 5 (gar nicht engagiert)

C) Note der Abschlussarbeit

I. Erstellen Sie eine Häufigkeitstabelle mit der Variable Note (absolute Häufigkeit, relative Häufigkeit, kumulierte Häufigkeit)

II. Erstellen Sie ein Säulendiagramm der Variable Note.

III. Erstellen Sie ein Box-Plot für die Studiendauer

IV. Erstellen Sie ein Histogramm für die Variable Engagement im Studium.

Dauer: 45 Minuten **Abgabeformat:** Textdatei/Exceldatei

ID	B	C	E
1	5	3	3
2	5	3	2
3	7	4	1,2
4	7	5	3
5	9	2	2,5
6	5	4	1,7
7	6	1	2,1
8	7	2	1
9	8	3	3,2
10	6	1	4
11	7	4	
12	6	3	2,4
13	8	5	1,5
14	7	6	2,7
15	6	3	3

Zweidimensionale Häufigkeitsverteilung: Kontingenztabelle

Zweidimensionale Häufigkeitsdarstellungen sind Sie sicher schon begegnet. Wir werden Sie später noch mal brauchen als Grundlage für einen statistischen Test.

	Kartenzahler	Barzahler	Gesamt
Frauen	23	13	36
Männer	33	31	64
Gesamt	56	44	100