Verteilungsformen

Verschiedene Verteilungen

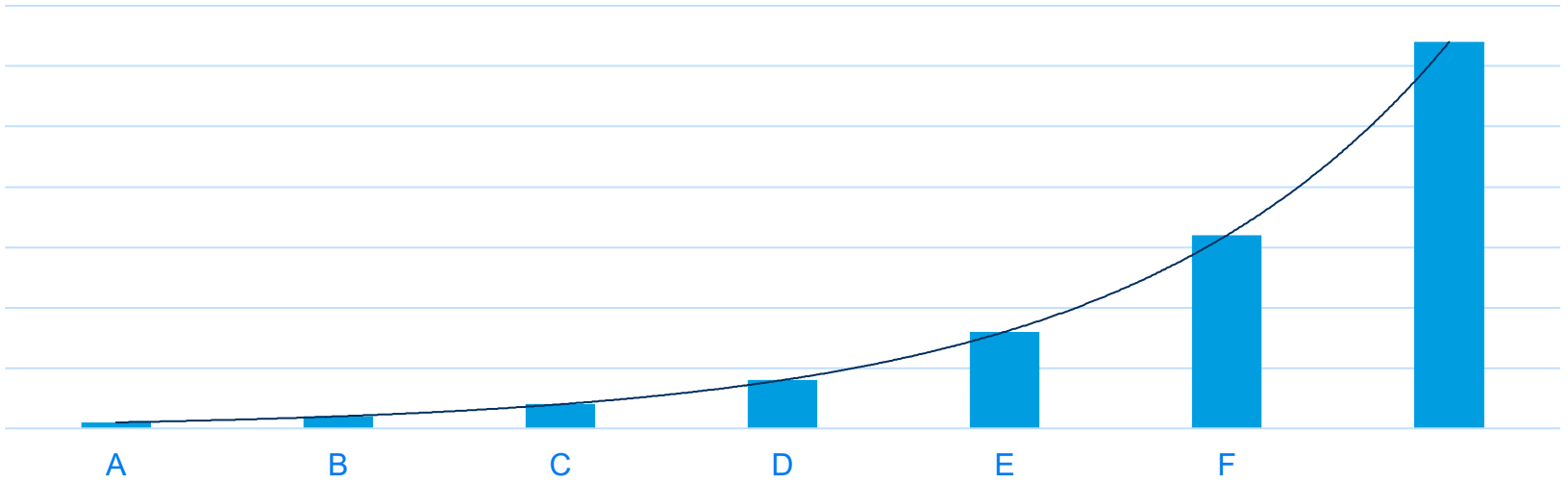
Bereits behandelt:

- **Zentraler Grenzwertsatz:** Normalverteilung (schiefe, gewöbte Verteilungen)
- **Häufigkeitsverteilung:** Gleichverteilung, Zweigipflige Verteilung
- **Wahrscheinlichkeitsrechnung:** Hypergeometrische-, Binominal- und Geometrische Verteilung
- **Testverteilungen:** t-Verteilung, Chi²-Verteilung, F-Verteilung, Normalverteilung

Weitere Verteilungen:

- Exponentielle Verteilung

Exponentielle Verteilung, Poissonverteilung

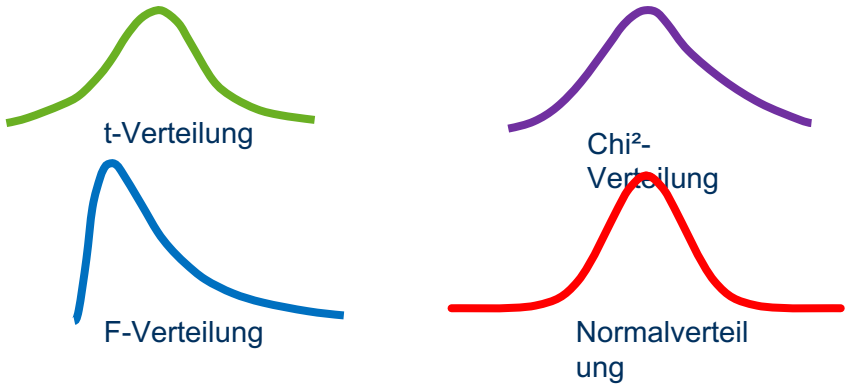


Exponentialverteilung

Eigenschaften von Verteilungen

Jede dieser (idealen) Verteilung kann mittels Varianz, Erwartungswert, Verteilungsfunktion und Dichtefunktion beschreiben werden:

https://de.wikipedia.org/wiki/Liste_univariater_Wahrscheinlichkeitsverteilungen



In empirische Verteilungen in den Sozialwissenschaften erhalten wir solche theoretischen Verteilungen übrigens … nie. Oder sagen wir so gut wie nie.

Entscheidungsbäume für die Auswahl von Statistik

Der gesamte Prozess für einen statistischen Test

Einen guten Fragebogen entwickeln und einsetzen. Evtl. deskriptive Statistiken oder Diagramme anschauen

Forschungshypothese festlegen → Nullhypothese auswählen → richtigen Test dafür ermitteln (z.B. anhand der Skalenniveaus und Anzahl der Variablen) → Formel für den Test ermitteln und alle Werte einsetzten und den Wert der Teststatistik ausrechnen →Freiheitsgrade bestimmen → Mit dem Wert der Teststatistik und Freiheitsgraden den kritischen Wert in einer Tabelle bestimmen

→ A. ist der kritische Wert kleiner als Wert der Teststatistik = signifikantes Ergebnis → Nullhypothese zurückweisen und sich für die Alternativhypothese entscheiden → Interpretation: es gibt einen Zusammenhang / Unterschied → nun auch weitere Werte analysieren

→ A. Mittelwert, Häufigkeiten für die Interpretation herangezogen werden.

→ B. Korrelations- oder Regressionskoeffizienten näher interpretiert werden

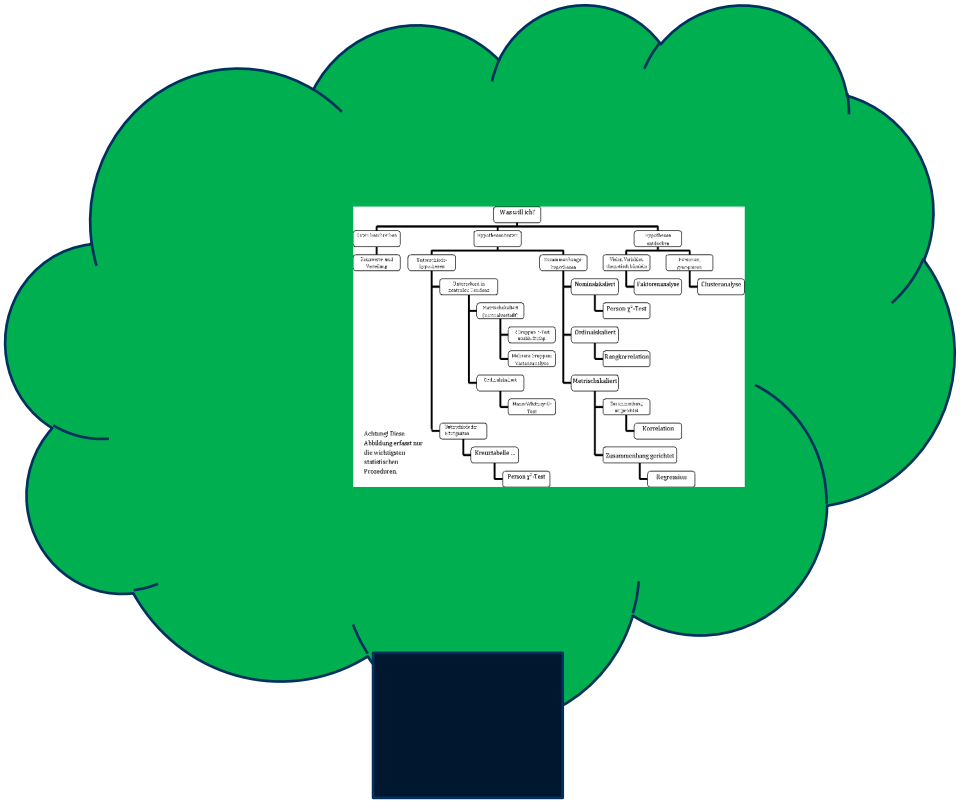
→ Vorzeichen (positiv oder negativ) → Wert: gering, mittel oder starker Zusammenhang

→ B. ist der kritische Wert größer als Wert der Teststatistik = nicht signifikantes Ergebnis → Nullhypothese beibehalten → Interpretation: es gibt keinen Zusammenhang / Unterschied

Entscheidungsbaum für die Statistik und quantitative Methoden Veranstaltung

Aufgabe: Wie ordnen Sie die fehlenden Tests ein (ANOVA; Chi²-Anpassungstest; Chi²-Streuungstest)

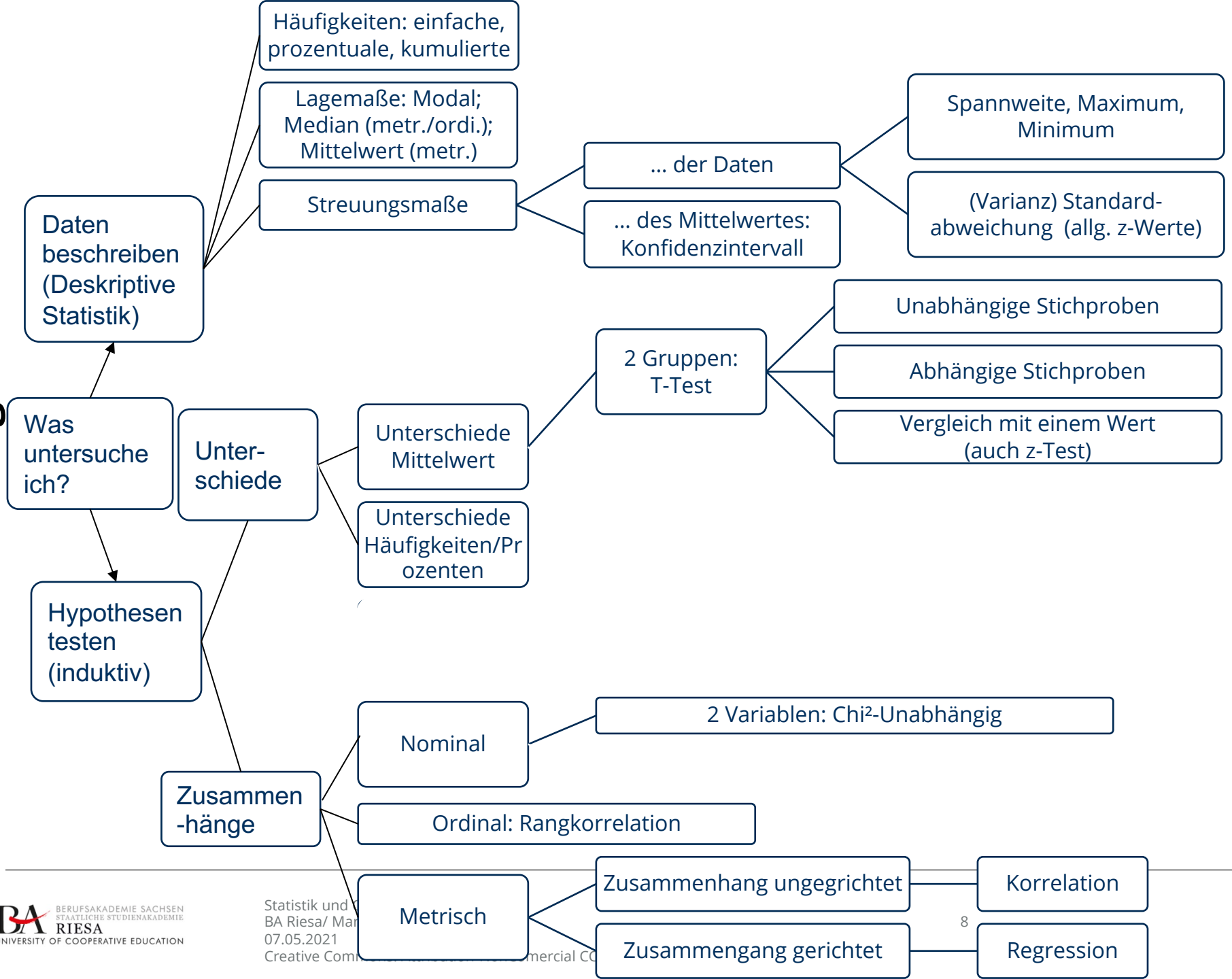
Hinweise:
Nutzen Sie die nächste Folie.
Bedenken Sie, dass es mehrere Möglichkeiten gibt für einen endgültigen Entscheidungsbaum.
Beispiel:
https://www.methodenberatung.uzh.ch/de/datenanalyse_spss.html



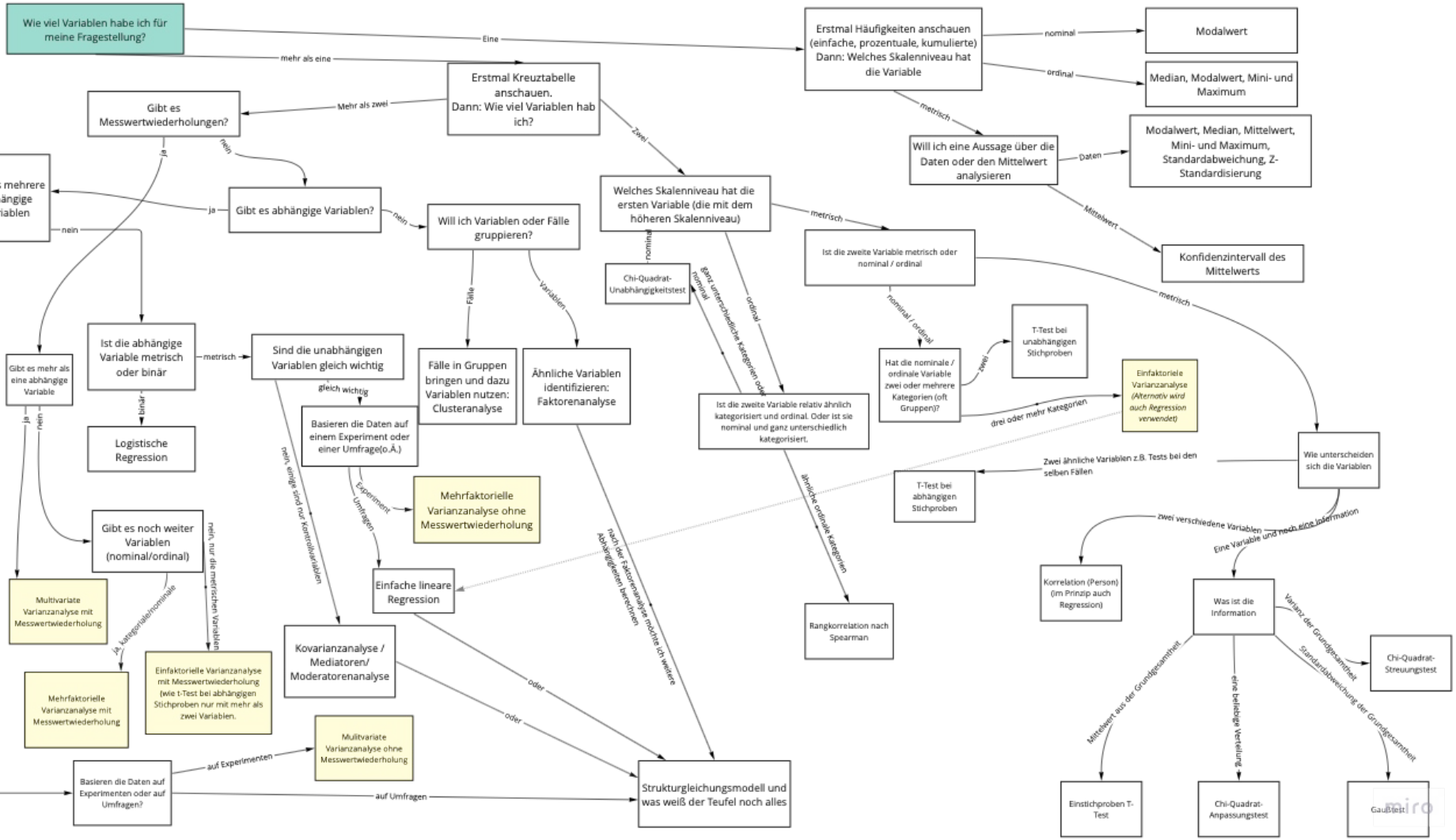
Zwischenschritte/Teilfragen:

- Haben wir alle statistischen Verfahren erfasst?
- Gefällt Ihnen die Struktur, wie wären Sie evtl. herangegangen?
- Nach welchen Kriterien möchten Sie den Baum einteilen

Unser Entscheidungsbaum



Ein noch komplexerer Entscheidungsbaum



Ergebnisse veröffentlichen und Reflektieren

Ethik und Statistik

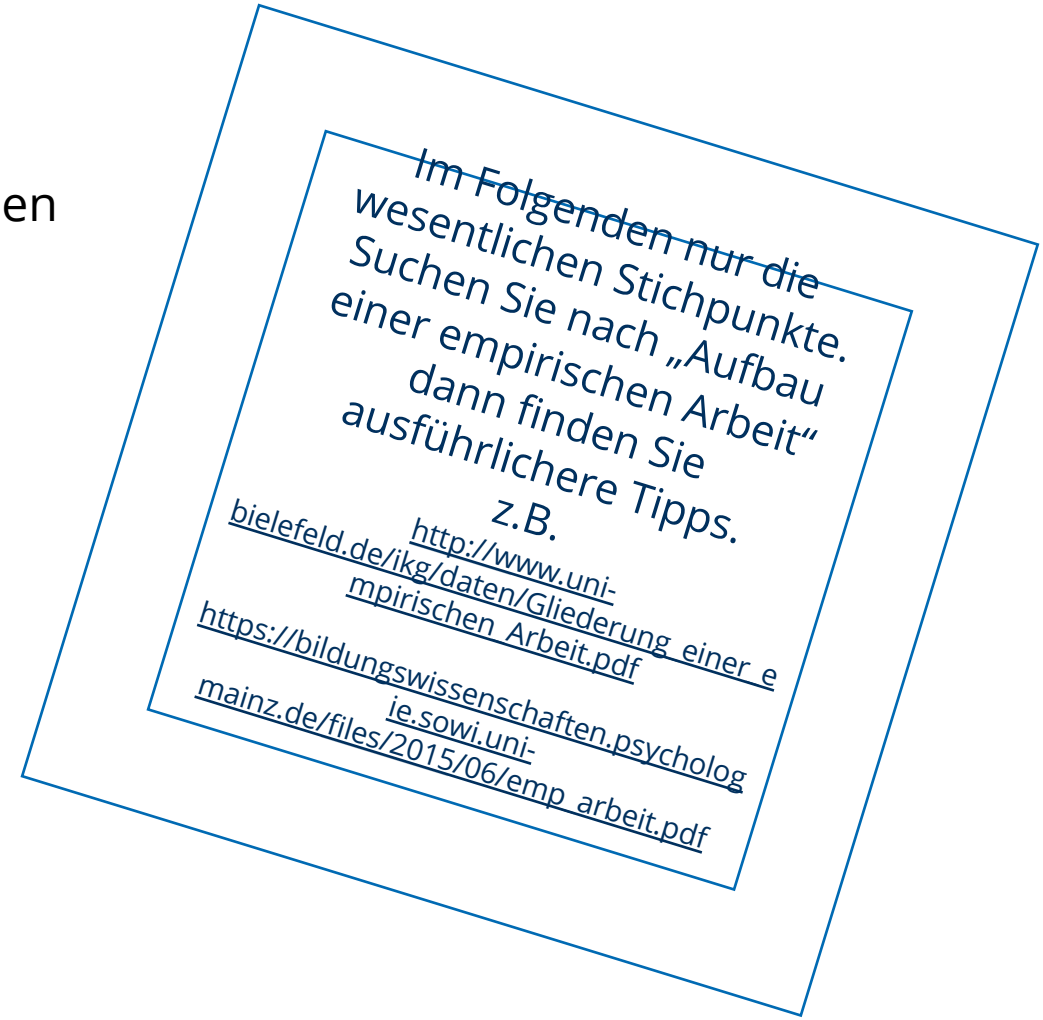
„Glauben Sie nur der Statistik, die Sie selbst gefälscht haben.“

Was bedeutet es wenn Sie mit Statistik „lügen“ für

1. Interne unternehmerische Ziele
2. Für Marketingzwecke
3. Für medizinische Zwecke

Veröffentlichung einer empirischen und statistischen Arbeit: Übersicht

- Einleitung
- Forschungsstand und Hypothesen
- Methoden und Daten
- Ergebnisse
- Ausblick/Diskussion
- Zusammenfassung
- Anhang



Veröffentlichung: Einleitung

Verortung, Bedeutung des Themas

Evtl. Historie des Themas

Was macht der Artikel (wie und wo)

„In dieser Arbeit wurde untersucht wie A und B auch ohne C auskommt. Dafür habe ich ABC und XYZ in einer Gruppe von Menschen gemessen. Im Absatz Methoden wird der Fragebogen vorgestellt … die Arbeit schließt mit einer Zusammenfassung.“

Abgrenzung: Was macht die Arbeit nicht (was man eigentlich Erwarten könnte)

Veröffentlichung: Forschungsstand und Hypothesen

(Hinweise: Hier nur zusätzliche Informationen für Arbeiten in denen methodisch und statistisch gearbeitet wurde, Weiteres entnehmen Sie vermutlich auch anderen Vorlesungen)

Zentrale Begriffe oder Zusammenhänge erläutern

(Der Theorieteil ist in empirischen Arbeiten kurz und es reicht, wenn Sie auf andere Literatur verweisen.)

Relevante Ergebnisse anderer kurz vorstellen.

Auf Forschungslücken und Widersprüche verweisen.

Hinüberleiten zu eigenen Forschungshypothesen (die sie testen werden)

Evtl. theoretische Begründung Ihres vermuteten Zusammenhangs erläutern.

Evtl. Ihre neuen Begriffe einführen.

Veröffentlichung: Methoden und Daten I

Begründung der Erhebungsmethoden (Fragebogen), welche Alternativen gäbe es.

Beschreibung der Erhebungsmethode. Welche Pretests, Woher kommen Fragen, Wie wurden Fragen entwickelt.

Operationalisierung der Hypothesen: Welche Frage misst welches theoretischen Konstrukt aus Ihren Hypothesen.

Beschreibung der Stichprobe (hier evtl. wichtige deskriptive Statistiken). Wer und wann und wo?

Erwähnung von Besonderheiten während der Erhebung.

... Fortsetzung nächste Folie

Veröffentlichung: Methoden und Daten II

Verwendete Variablen beschreiben: Was bedeuten die Werte. Welche Datentransformationen wurden vorgenommen. Wie berechnen sich neue Variablen (Summen, Indizes)

Verwendetes statistisches Verfahren nennen mit Bezug auf die Hypothese:

„Um die Hypothese XY zu testen habe ich eine Regression mit der abhängigen Variable Y verwendet und der unabhängigen Variablen X1 und X2. Die Nullhypothese der Regression besagt, dass die unabh. Var. Keinen Einfluss haben. Eine Betätigung meiner Hypothese XY ist dann gegeben, wenn die Koeffizienten von X1 und X2 signifikant auf dem Niveau von 0,05 sind.“

Nur spezielle Statistiken erklären

Veröffentlichung: Ergebnisse I

Mit wichtigen einfachen Ergebnissen beginnen. Auch spannende Grafiken.
Jede Tabelle, jede Grafik im Text beschreiben.

Große Tabellen mit nebensächlicher Information nur in Anhang.

Evtl. Tabellen mit wichtigen Ergebnissen für statistische Tests, falls mehr als 3 Zahlen gezeigt werden sollen.

Veröffentlichung: Ergebnisse II

Signifikante Ergebnisse: Darauf eingehen, ob eine Hypothese abgelehnt wird oder nicht. Auch nicht signifikante Ergebnisse sind erwähnenswert!

„Die Ergebnisse der Zweifaktoriellen Varianzanalyse mit Messwertwiederholung finden sich in Abbildung X. Die Berechnung der unterschiede in den Stichproben und in den Spalten (Früh, Mittags) ist nicht signifikant. Das bedeutet, die Nullhypothesen – es gibt keine Unterschiede zwischen den Stichproben und den Spalten - wird beibehalten. Hingegen ist der p-Wert für die Wechselwirkung der Faktoren (Gruppen/Tageszeit) signifikant. Das bedeutet es gibt eine Wechelwirkung zwischen Tageszeit und Gruppen. Ein Blick auf das Häufigkeitsdiagramm zeigt, … “

Evtl. zusätzliche Analysen vorstellen und begründen warum sie notwendig sind

Veröffentlichung: Ausblick/Diskussion

Inhaltliche Interpretation der Ergebnisse (eigene Interpretation und mit Hilfe der Literatur); Bedeutung für Fragestellung.

Eventuell neue Hypothesen formulieren für zukünftige Forschung

Andeuten was man in zukünftigen Untersuchungen anders machen könnte

Schlussfolgerungen für die Praxis ziehen

Grenzen der Untersuchung klar machen (es ist nur eine Stichprobe, Besonderheit der untersuchten Personen).

Selbstkritik an den verwendeten Prozeduren/Skalen

Veröffentlichung: Zusammenfassung

Die Ziele des Artikels wiederholen, von den Ergebnissen nur die Bedeutendsten erwähnen. Nichts Neus hier!

In wissenschaftlichen Artikeln gibt es meist noch eine zweite Zusammenfassung die aus weniger als 10 Sätzen besteht. Anhand dieses „Abstracts“ (engl.) können Leser schnell einschätzen, ob der Artikel für sie relevant ist oder nicht. (Für den Autor ist er auch gut, weil er so seine Kerngedanken und Ergebnisse noch mal hervorbringt.)

Da Leser wenig Lust haben viele Aufsätze vollständig zu lesen, sollte die Zusammenfassung tatsächlich alles zusammenfassen.

Veröffentlichung: Anhang

Jede Tabelle und Grafik sollte aus sich heraus so gut es geht allein verständlich sein.

Nur Tabellen hier die zu groß sind für den Text und dort auch nur am Rande erwähnt werden.

Oft werden deskriptive Statistiken (Häufigkeiten, Prozentangaben) hier veröffentlicht.

Im Anhang ist auch Platz für verwendete Fragebögen

Komplizierte Berechnungen (neue Variablen, unbekannte statistische Verfahren) können hier auch gezeigt werden.

Weitere Tipps für empirische Arbeit

- Protokoll führen (so lange bis man es nicht mehr brauch)
- Sicherheitskopie vom Originaldatensatz
- Zwischenziele aufschreiben
- Notieren, welche Fälle/Variable ausgewählt worden
- Erwähnenswerte (Zwischen-)ergebnisse notieren oder aus der Auswertung in eine extra (Excel-) Datei kopieren
- Konfuse Ergebnisse löschen
- Ergebnisse die veröffentlichungswert erscheinen, in Sprache umformen. Dann erkennt man was vielleicht falsch gemacht wurde bzw. welche weiteren Schritte man unternehmen könnte.
- Bei komplizierten Vorgängen genau aufschreiben was man vor hat
- Eigene Ideen/Kritik an der Methode festhalten

Tipps für Diagramme

Diagramme sollten meist bearbeitet werden (so wenig wie möglich Informationen aber alle nötigen Informationen: Skalentitel, Legende, Einheiten).

Das Anklicken einzelner Elemente erlaubt detaillierte Bearbeitung.

Es geht einfacher ein kompliziertes Diagramm zu erstellen, wenn man vorher auf Papier skizziert wie es im Endeffekt aussehen soll.

Excel macht im Vergleich zu andere Statistikprogrammen sehr gute Diagramme.

Kleine Fehler finden

Schlechte Beispiele: Tabelle

	Unternehmen A	Unternehmen B	N
Frauen	700	500	1200
	31,111%	22,222%	53,333%
Männer	400	650	1050
	17,777%	28,888%	46,666%
n	1100	1150	2250

Welche Kritikpunkte sehen Sie?:

-
-
-
-
-
-

Bezugspunkte ändern

Ein Sportverein hat die Gruppen Junioren und Senioren. Leute mittleren Altern dürfen frei wählen. Was passiert, wenn der 35-jährige wechselt?

	Junioren	Senioren
	5	
	10	40
	15	45
	20	50
	25	55
	30	65
	35	70
Mittelwert **vor** Wechsel:	20	57,1

	Junioren	Senioren
Mittelwert **nach** Wechsel:		

Fehlschluss: Große Stichproben sind immer gut

Was passiert hier durch die größere Stichprobe?

Gruppe 1	Gruppe 2
1	2
2	3

p-Wert des t-Tests: 0,293

Gruppe 1	Gruppe 2
1	2
2	3
1	2
2	3

p-Wert des t-Tests: 0,050

Trends zu gewagt

Körpergröße von 18 jährigen Männern in Metern. Welcher Trend lässt sich anhand der Daten formulieren. Wäre das sinnvoll?

1970: 1,80

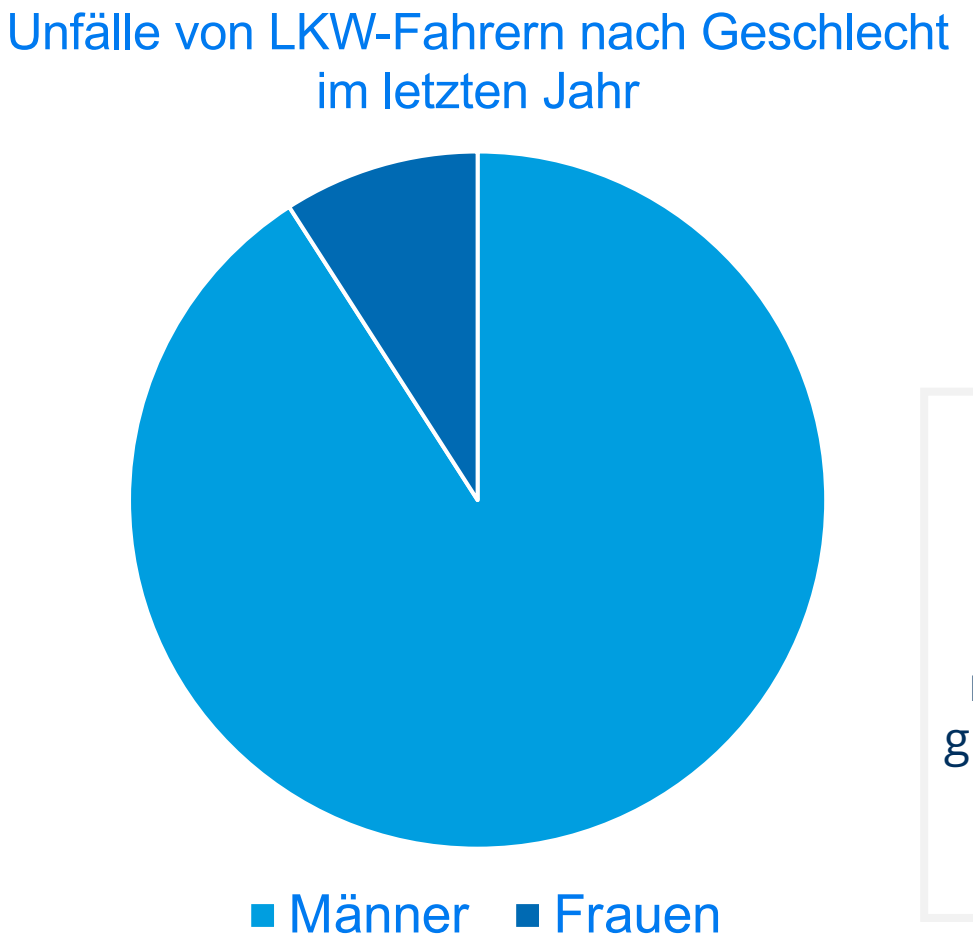
1990: 1,85

2010: 1,90

2030:

Quelle: Statista. Lügen mit Statistiken https://de.statista.com/statistik/lexikon/definition/8/luegen_mit_statistiken/
Abgerufen: 10.12.2018

Bezugsrahmen verheimlichen



Ihr Unternehmen analysiert die Unfallstatistik bei den angestellten LKW-Fahrenden nach Geschlecht. Jemand zeiht den Schluss: es müssen mehr Frauen eingestellt werden, weil die weniger Unfälle machen.

Was fehlt hier? Erfinden Sie Daten, die genau die gegenteilige Aussage nahelegen würden, aber zum gleichen Kreisdiagramm führen könnten.

Prozente können schön klingen

Partei X 2010:
Partei X 2015:

Verdopplung des Frauenanteils in einer Partei X

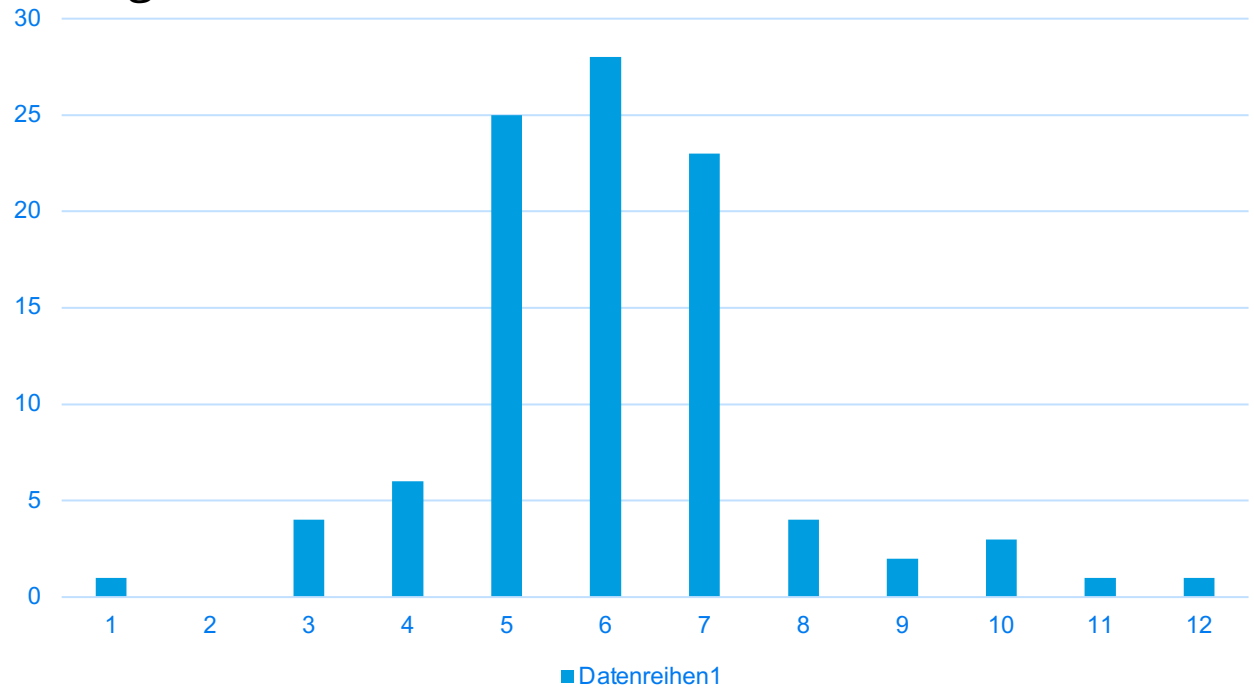
Finden Sie Datenbeispiele die die nebenstehenden Aussagen zulassen.

Anstieg um nur 20 Prozent in Partei Y

Partei Y 2010:
Partei Y 2015:

„Normalerweise passiert das nicht“

Abgebildet sind die Lieferzeiten Ihres Möbelwarengeschäfts. Ein Kunde beschwert sich, er hat eine Lieferzeit von 12 Tagen und meint, das ist doppelt so lang wie der Durchschnitt. Das ist doch nicht normal.

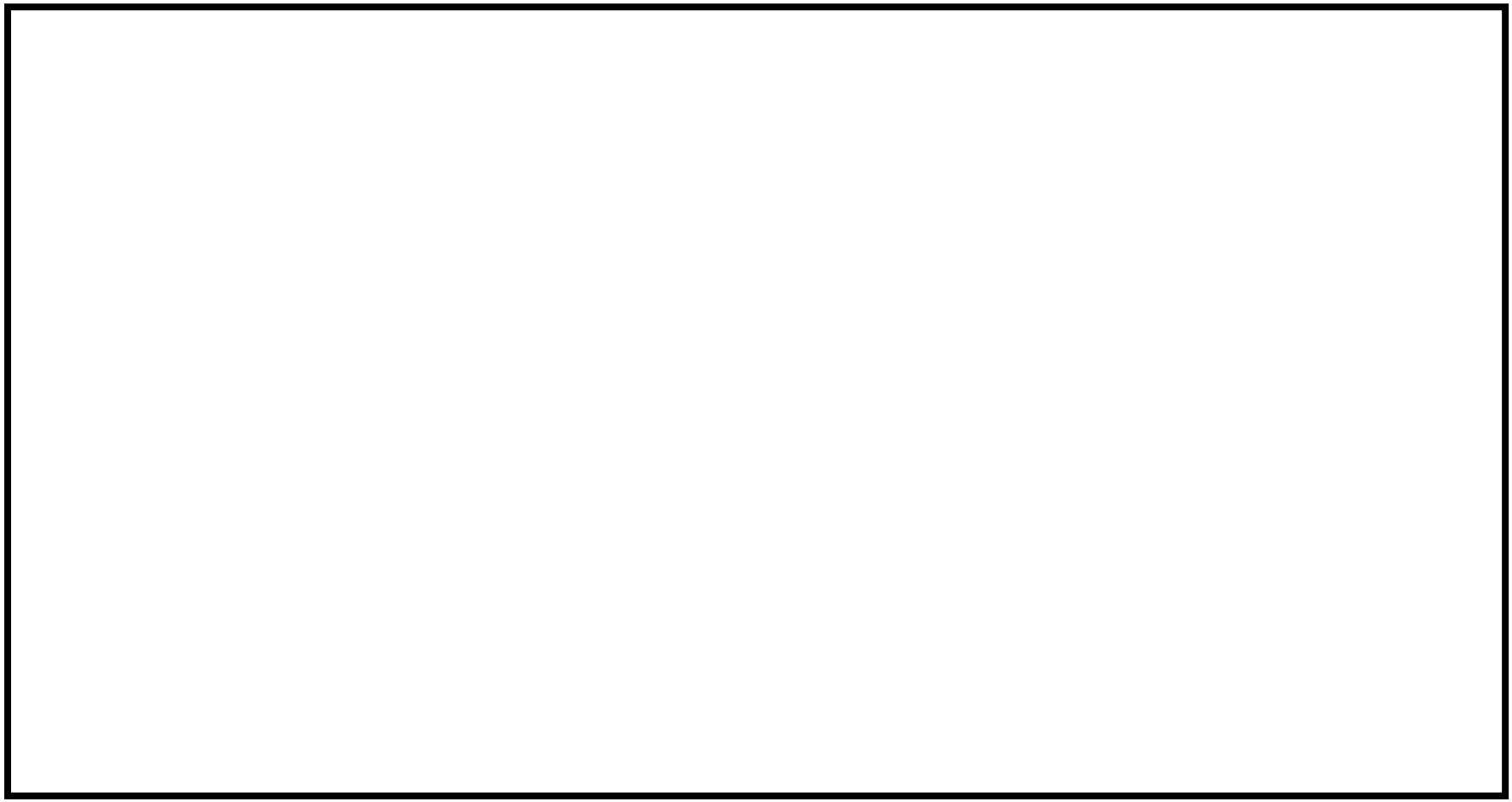


Was antworten Sie dem Kunden?

Datenbanken

Woher nehme ich Daten, wenn ich oder meine Organisation sie nicht selbst erhoben haben?

Kurze Geschichte der Statistik



(Daten-)Datenbanken

Wer stellt sie bereit?

- Staatliche Institutionen
- Forschungseinrichtungen und nichtstaatliche Organisationen (Themengebunden)

Warum werden sie bereitgestellt

- Berichtspflichten von staatlichen Einrichtungen
- Interesse von Organisationen

Was wird dort angeboten?

- Inhalt
- Variablen
- Fälle
- Anleitung zum Lesen von Daten
- Evtl. verwendete Fragebögen

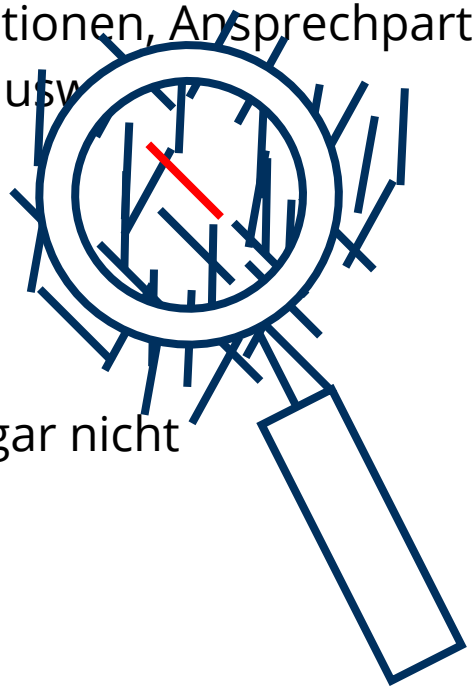
Problem bei Suche nach Datenbanken

Unter dem Begriff „Datenbanken“ versteckt sich sehr viel:

- Datenbanken für Zeitschriften
- Datenbanken für Listen mit Organisationen, Institutionen, Ansprechpartner
- Datenbanken für Informationsseiten, News-Artikel usw.
- Datenbanken für Projekte, Geldgeber
- Datenbanken für …
- Datenbanken für Daten

Daten zu finden ist nicht leicht.

Daten zu spezifischen Fragestellungen gibt es evtl. gar nicht



Wie findet man DATEN-Datenbanken: Metasuchen

Registry of Research Data Repositories

https://www.re3data.org

Suche einschränken Hummanitiers and Social Sciences und Empirical Social Research

 Browse

 Browse by subject

Weitere Einschränkung z.B. „country“ Germany und „subjects“ auf der linken Seite

Fachdatenbanken in Bibliotheken

http://rzblx10.uni-regensburg.de/dbinfo/fachliste.php?bib_id=slub

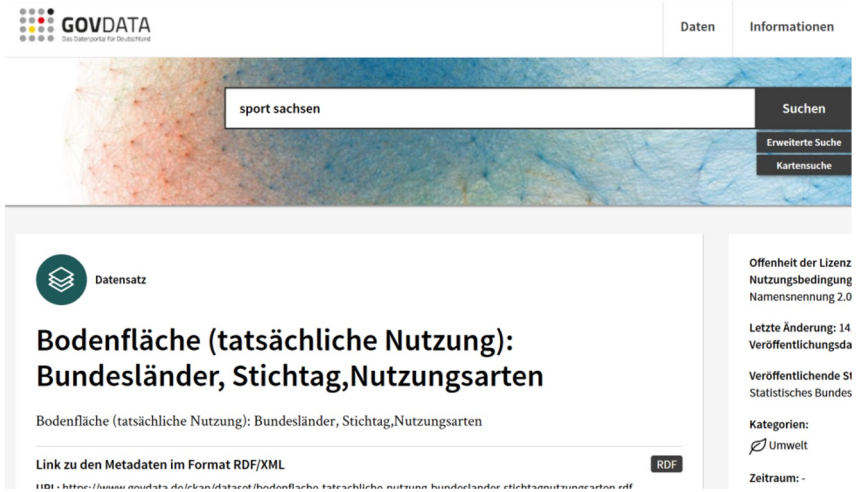
Beispiele für eine Datenbanken

Amadeus (Finanzdaten von Unternehmen)

https://amadeus.bvdinfo.com/version-2019919/Search.QuickSearch.serv?_CID=1&context=36H0I3VCVYADEY6

GOVDATA: Beispiel Bodenfläche Nutzung. Gefunden über die Suchbegriffe „Sport Sachsen“

https://www.govdata.de/web/guest/suchen/-/details/bodenflache-tatsachliche-nutzung-bundeslander-stichtagnutzungsarten



Format von Datensätzen

Datensätze werden in verschiedenen Formaten angeboten.

.pdf: meist Beschreibung von Metadaten, also Informationen zum Datensatz

.csv: „Comma seperatet Value“ kann einfach mit Excel oder anderen Programmen zur Datenverarbeitung geöffnet werden.

.xlsx: Exceldatei

.sav: Datensätze für die Statistikprogramme SPSS oder PSPP

.dta: Datensätze für das Statistikprogramm STATA

.xml: für automatisches Auslesen mit einem Programm

Abschluss

Spiele

3er/2er Gruppen: eine Person sagt einen Begriff aus der Statistik und die andere aus dem Alltag. Zusammen finden sie die Gemeinsamkeit. Beispiel: Regression und Bahnhof → Der Bahnhof sollte so geplant werden, dass er nah genug an allen Häusern ist…

Kurzreferate halten: Studierende dürfen ein Zettel mit Begriffen füllen und ich muss daraus eine Stehgreifrede machen.

Klausurvorbereitung Fallstudie

Klausurvorbereitung

Quiz Teil I

Frage	Richtig	Falsch
Bei der einfachen linearen Regression beeinflusst ein oder mehrere Variablen die abhängige Variable.		
Das konstante Glied der einfachen linearen Regressionsanalyse entspricht dem Wert des Y-Achsenschnittpunktes der linearen Regressionsgeraden.		
Der 1. Schritt in jedem stat. Testverfahren besteht in der Entscheidung, ob die Nullhypothese oder die Alternativhypothese getestet werden soll.		
Der Korrelationskoeffizient kann nur Wert zwischen 0 und 1 annehmen		
Der Median wird von Ausreißern beeinflusst.		
Der Mittelwert ist gegenüber Ausreißern robust.		

Quiz Teil II

Frage	Richtig	Falsch
Der Modal wert ist der Wert, der genau in der Mitte der geordneten Verteilung liegt.		
Der Regressionskoeffizient entspricht der Steigung der linearen Regression		
Die Spannweite wird nie von Ausreißern beeinflusst.		
Die Standardabweichung berechnet sich als positive Wurzel aus der Varianz.		
Die Wahrscheinlichkeiten aller möglichen Elementarereignisse eines Zufallsvorgang ergeben zusammenaddiert den Wert 2.		
Diskrete Variablen mit sehr vielen Ausprägungen gelten auch als quasi stetig.		

Quiz Teil III

Frage	Richtig	Falsch
Ein korrekt durchgeführter stat. Test gestattet eine definitive Aussage über die Korrektheit von Null- und Alternativhypothese.		
Ein Korrelationskoeffizient von -0,85 deutet auf eine starke lineare Korrelation hin?		
Ein Zufallsexperiment ist die beliebig häufige Wiederholung eines Zufallsvorgang unter gleichen Rahmenbedingungen.		
Eine zufällig gezogene Stichprobe mit hoher Rücklaufquote ist unabhängig von ihrem Umfang stets repräsentativ.		
Erwartungswert, Median und Modus einer normalverteilten Variablen alle ungleich groß.		
Es gibt keine Zufallsvariablen die diskret sind.		

Quiz Teil IV

Frage	Richtig	Falsch
Je mehr Hypothesen man an einem Datensatz testet, desto höher wird die Wahrscheinlichkeit, dass eine davon fehlerhaft als zutreffend angenommen wird.		
Kreisdiagramme eignen sich eher für stetige als für diskrete Daten.		
Nominalskalierte Daten können in eine natürliche Reihenfolge gebracht werden.		
Ordinalskalierte Daten können in eine natürliche Reihenfolge gebracht werden.		
Stetige Daten sollten vor der Erstellung von Säulendiagrammen klassiert werden.		
Streudiagramm zeigen die Verteilung von zwei Variablen.		