

# Vorlesung: Statistik I

Prof. Dr. Simone Abendschön

Grundbegriffe

Am Ende der heutigen Vorlesung ...

- ...kennen Sie wichtige Grundbegriffe der Statistik
- ...kennen Sie grundlegende Anwendungen statistischer Kenntnisse
- ...kennen Sie unterschiedliche Skalenniveaus

# Was ist Statistik?

- „im **materiellen Sinn** die geordnete Menge von Informationen in Form empirischer Zahlen (‚Statistiken‘); im **instrumentalen Sinn** (statistische Methoden) der Inbegriff der Verfahren, nach denen empirische Zahlen gewonnen, dargestellt, verarbeitet, analysiert und für Schlußfolgerungen, Prognosen und Entscheidungen verwendet werden“

Der große Brockhaus Kompaktausgabe (1984), Band 21: S. 29 (eigene Hervorhebungen)

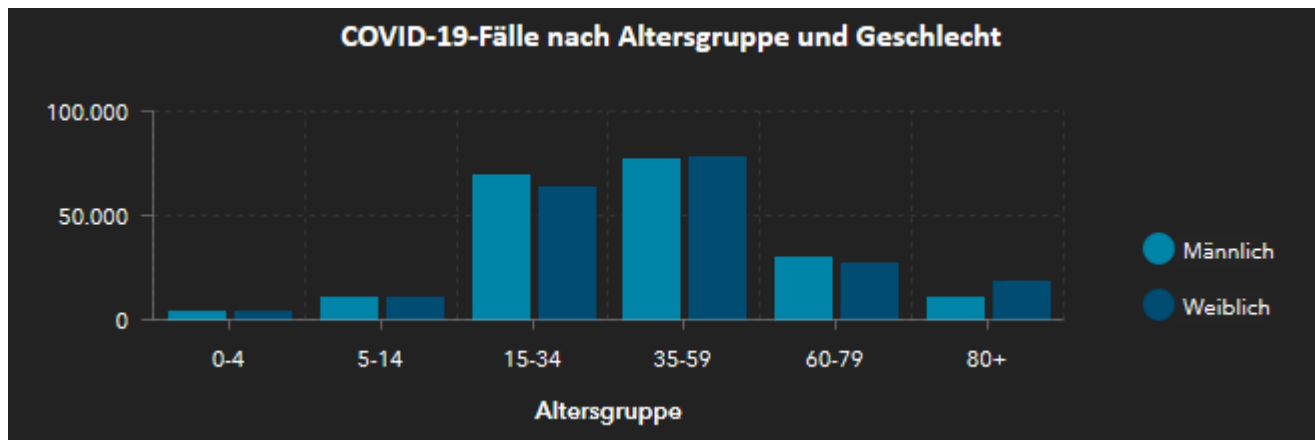
# Was ist Statistik?

- „umfassendes methodisch-quantitatives Instrumentarium zur **Charakterisierung und Auswertung empirischer Befunde bei gleichartigen Einheiten („Massenphänomenen“)** mit **universellen Einsatzmöglichkeiten** in Politik, Wirtschaft und Gesellschaft und allen Geistes-, Sozial- und Naturwissenschaften einschließlich Medizin und Technik, in denen mit Zahlen oder Bewertungen gearbeitet wird“ (<https://wirtschaftslexikon.gabler.de/definition/statistik-45267>, eigene Hervorhebungen)

# Was ist Statistik?

„im **materiellen Sinn** die geordnete Menge von Informationen in Form empirischer Zahlen („Statistiken“)“

Beispiel:



RKI Dashboard: <https://experience.arcgis.com/experience/478220a4c454480e823b17327b2bf1d4>  
(Stand: 23.10.)

# Was ist Statistik?

„im **materiellen Sinn** die geordnete Menge von Informationen in Form empirischer Zahlen („Statistiken“)“

Beispiel 2:



[https://www.destatis.de/DE/Themen/Wirtschaft/Konjunkturindikatoren/\\_inhalt.html](https://www.destatis.de/DE/Themen/Wirtschaft/Konjunkturindikatoren/_inhalt.html)(Stand: 4.11.21)

# Was ist Statistik?

„im **instrumentalen Sinn** (statistische Methoden) der Inbegriff der Verfahren, nach denen empirische Zahlen gewonnen, dargestellt, verarbeitet, analysiert und für Schlußfolgerungen, Prognosen und Entscheidungen verwendet werden“

- In den Sozialwissenschaften dient Statistik zur Analyse empirischer (= erfahrungswissenschaftlicher) numerischer Informationen
- Im Vordergrund stehen aber nicht die mathematischen Grundlagen
- Im Vordergrund steht der Bezug zu sozialwissenschaftlichen Fragestellungen

## **Angewandte sozialwissenschaftliche Statistik $\neq$ mathematische Statistik**

- Arbeitet anwendungsorientiert auf Basis mathematisch statistischer Grundlagen
- Teilgebiet der sozialwissenschaftlichen Methodenlehre
- Befasst sich mit der Anwendung von Analyseverfahren auf sozialwissenschaftliche Fragestellung
- Dient der Beschreibung und Auswertung von sozialwissenschaftlich relevanten Daten
- Schnittstelle zwischen sozialwissenschaftlichen Theorien und empirischer Forschung → Überprüfung von Hypothesen



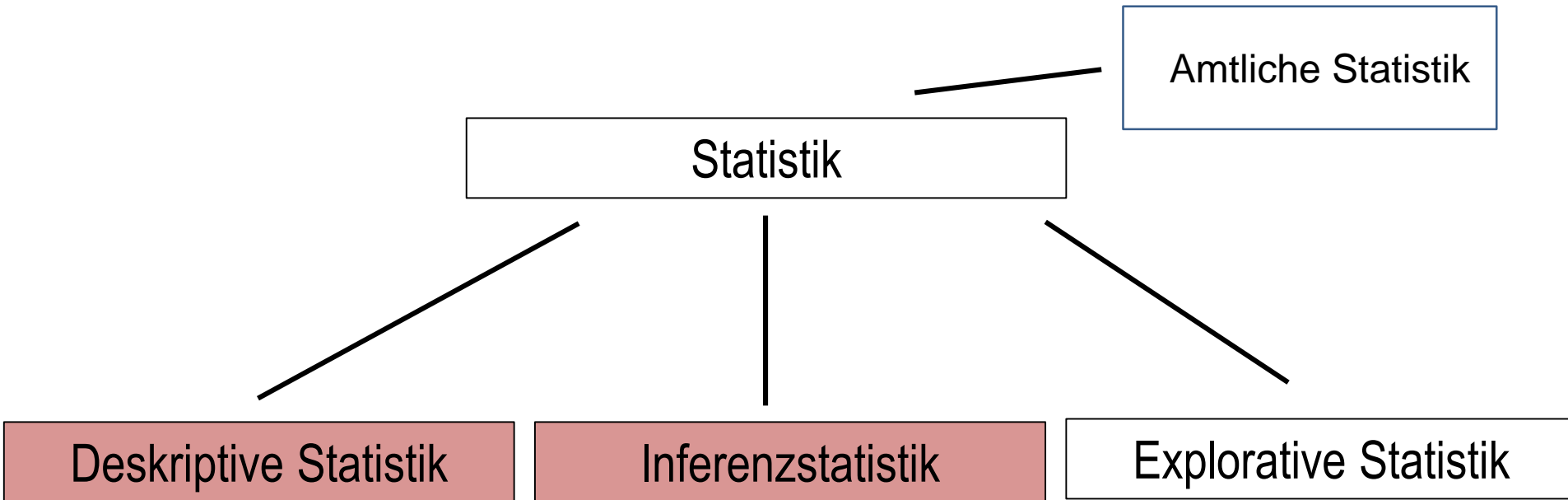
# Warum soll ich das lernen?

- **Statistische Informationen verstehen**
- **Sozialwissenschaftliche Forschung verstehen und selbst durchführen**
- **Berufsperspektiven**
  - Firmen, Organisationen, Politik, Medien, Verbände etc. brauchen Absolvent\*innen...
    - ...die mit grundlegenden statistischen Konzepten vertraut sind
    - ...auf Daten und Statistiken basierende Inhalte verstehen, bewerten, kommunizieren können

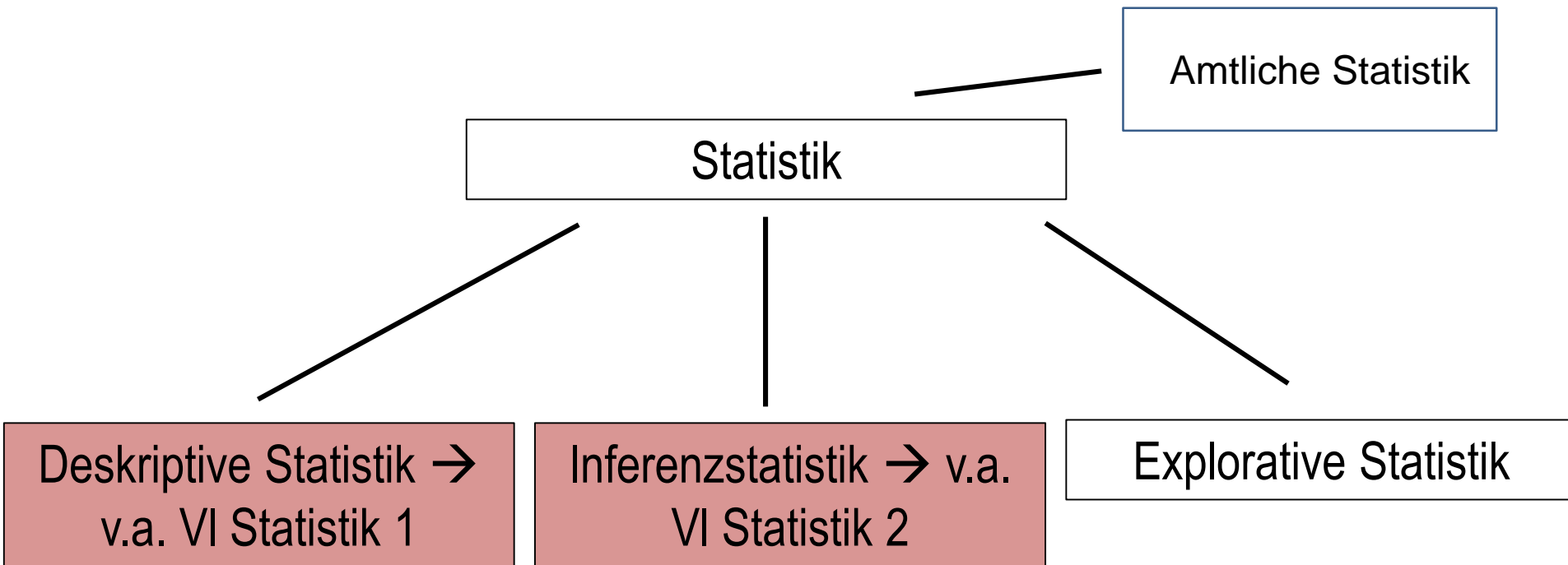
# Statistische Begriffe

- **Statistik nutzt (wie alle Fächer) eigene Fachbegriffe**
- **kein Selbstzweck**
- **Erleichtert die Kommunikation von Wissensinhalten**

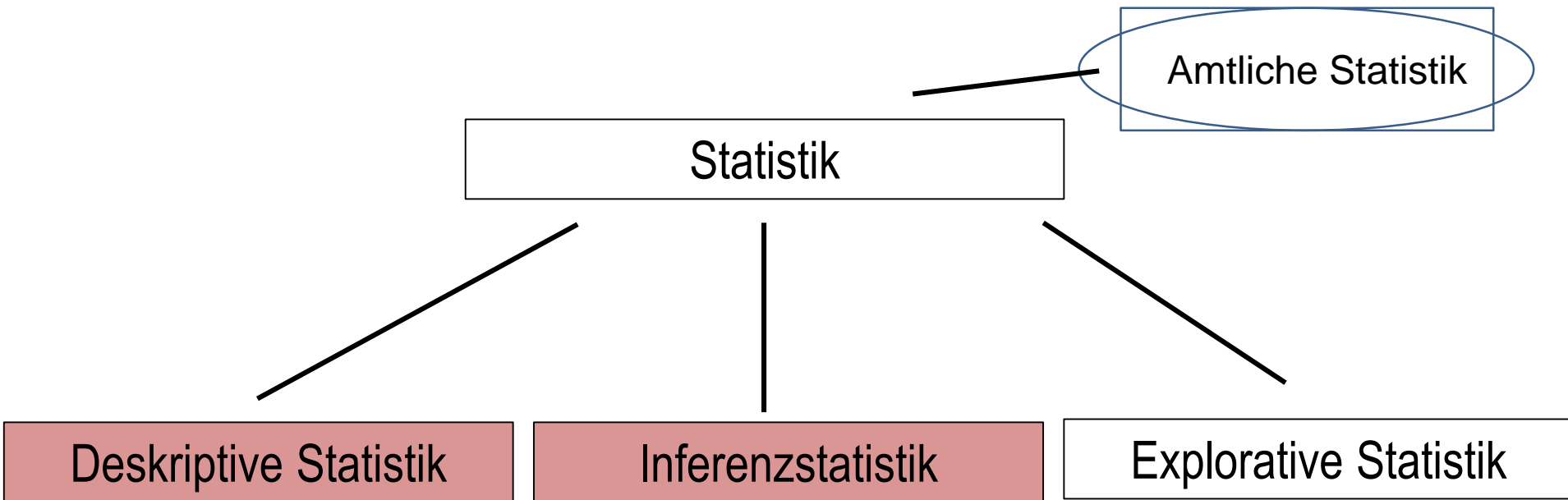
# Bereiche der Statistik



# Bereiche der Statistik



# Bereiche der Statistik



Covid-19

## Robert Koch-Institut legt neue Zahlen zum Coronavirus vor

12. April 2020



[https://www.deutschlandfunk.de/covid-19-robert-koch-institut-legt-neue-zahlen-zum.2932.de.html?drn:news\\_id=1119889](https://www.deutschlandfunk.de/covid-19-robert-koch-institut-legt-neue-zahlen-zum.2932.de.html?drn:news_id=1119889) (Stand: 14.04.2020)



[https://www.destatis.de/DE/Home/\\_inhalt.html](https://www.destatis.de/DE/Home/_inhalt.html) (Stand: 27.10.2020)

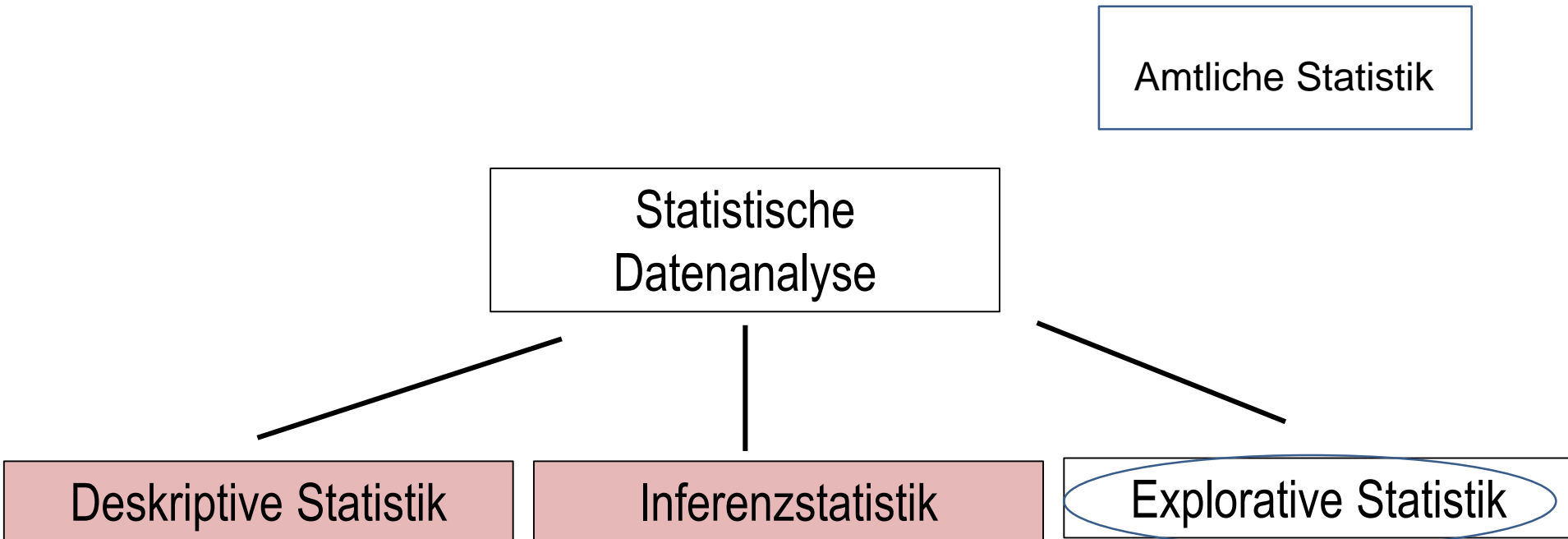


<https://ec.europa.eu/eurostat> (Stand: 27.10.2020)

- **Datenerhebung, Datenanalyse und Datenbereitstellung durch öffentliche Institutionen (z.B. Statistisches Bundesamt)**
- **Trägt zur politischen Willensbildung und politischen Entscheidungsprozessen bei**
- **Verständliche Aufbereitung, damit politische und gesellschaftliche Akteure sowie Bürger\*innen sie verstehen können**



# Bereiche der Statistik



# Aktuelles Beispiel explorativ, Big Data

Zur Eindämmung des Coronavirus

03.04.2020, 14:19 Uhr

## Google veröffentlicht anonymisierte Bewegungsdaten von Nutzern

Die Statistiken zeigen laut dem Konzern, ob das Besuchsaufkommen in Parks, Geschäften, Häusern oder Arbeitsplätzen zu- oder abnehme.



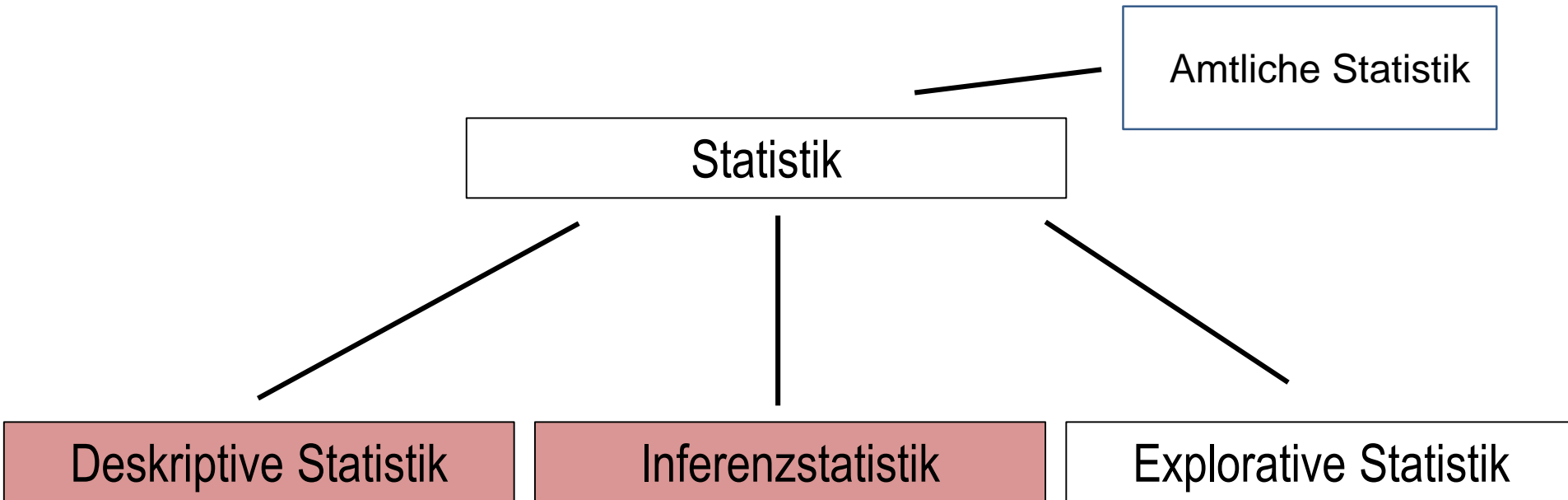
<https://www.tagesspiegel.de/wirtschaft/zur-eindaemmung-des-coronavirus-google-veroeffentlicht-anonymisierte-bewegungsdaten-von-nutzern/25712306.html> (Stand: 14.04.2020)

# Explorative Statistik

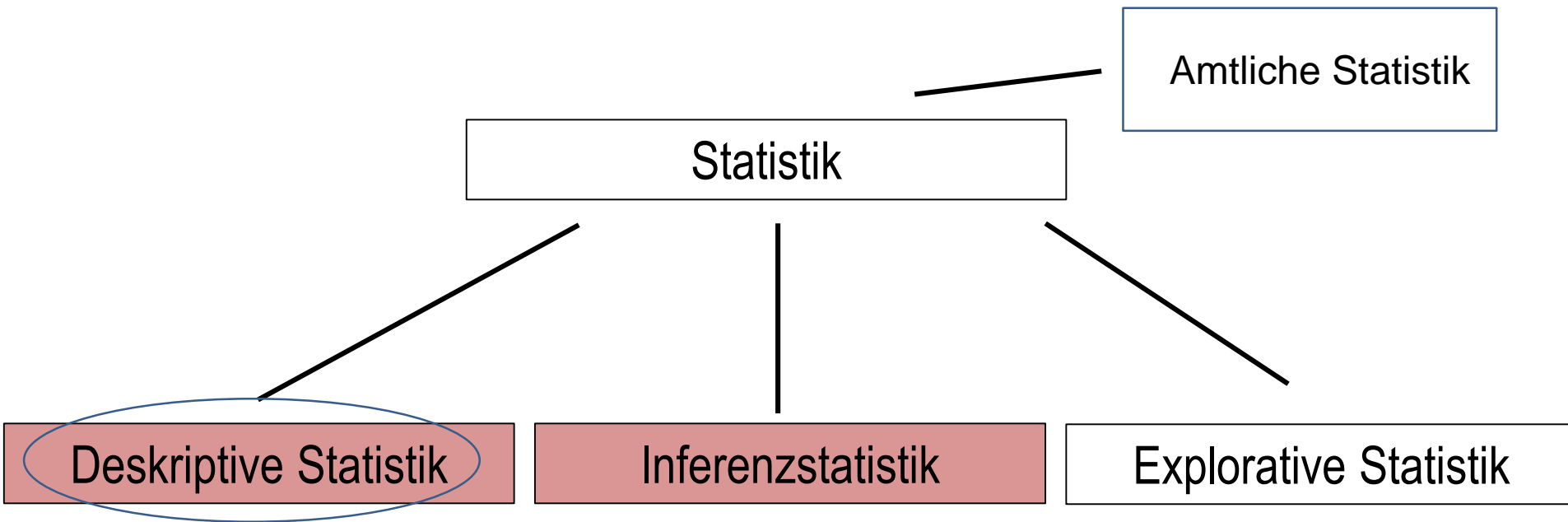
- **„erkundende“ Statistik („Data Mining“)**
- **Ziel: (bislang) nicht bekannte Strukturen und Zusammenhänge in den Daten zu finden, Formulierung neuer Hypothesen**
- **Zwischen beschreibender und schließender Statistik angesiedelt**
- **Wichtig für viele Anwendungen in der Wirtschaft**
- **v.a. im Kontext „Big Data“ und zunehmender Datafication des Lebens**

**In den Sozialwissenschaften (noch) weniger verbreitet (aber im Kommen)**

# Bereiche der Statistik



# Bereiche der Statistik

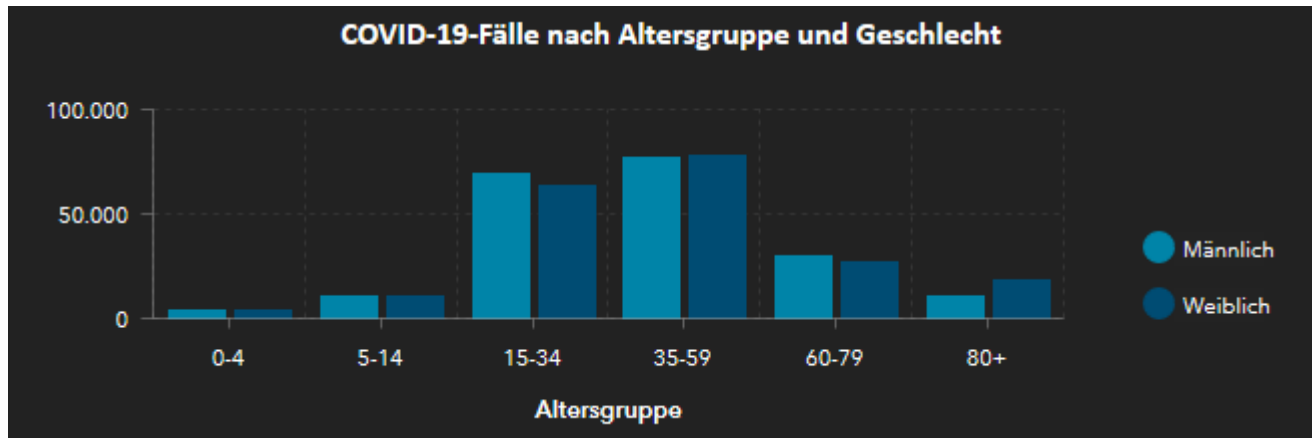


# Deskriptive Statistik

- **Beschreibende Statistik**
- **Zentrale Idee: Informationsverdichtung**
- **Information über eine oder mehrere Variablen wird verdichtet, statt mehrerer 100 oder 1000 Zahlen eine Zahl**
- **Grafische Darstellungen, Tabellen als Mittel der Veranschaulichung**
- **Merkmale & Zusammenhänge zwischen Merkmalen werden grundlegend beschrieben**

## Kennwerte und Parameter („Zahlen“/Maße) für

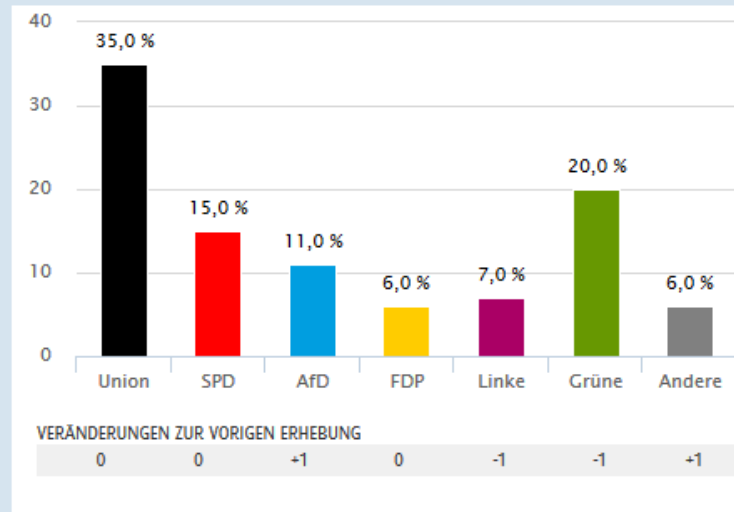
- Häufigkeiten
- Verteilungen (von Merkmalen bzw. Variablen)
- Streuung und Konzentration von Variablen
- Zusammenhänge (zwischen zwei oder mehreren Variablen)



<https://experience.arcgis.com/experience/478220a4c454480e823b17327b2bf1d4>  
(Stand: 23.10.2020)

# Deskriptive Statistik

Sonntagsfrage (bundesweit) vom 16.10.2020



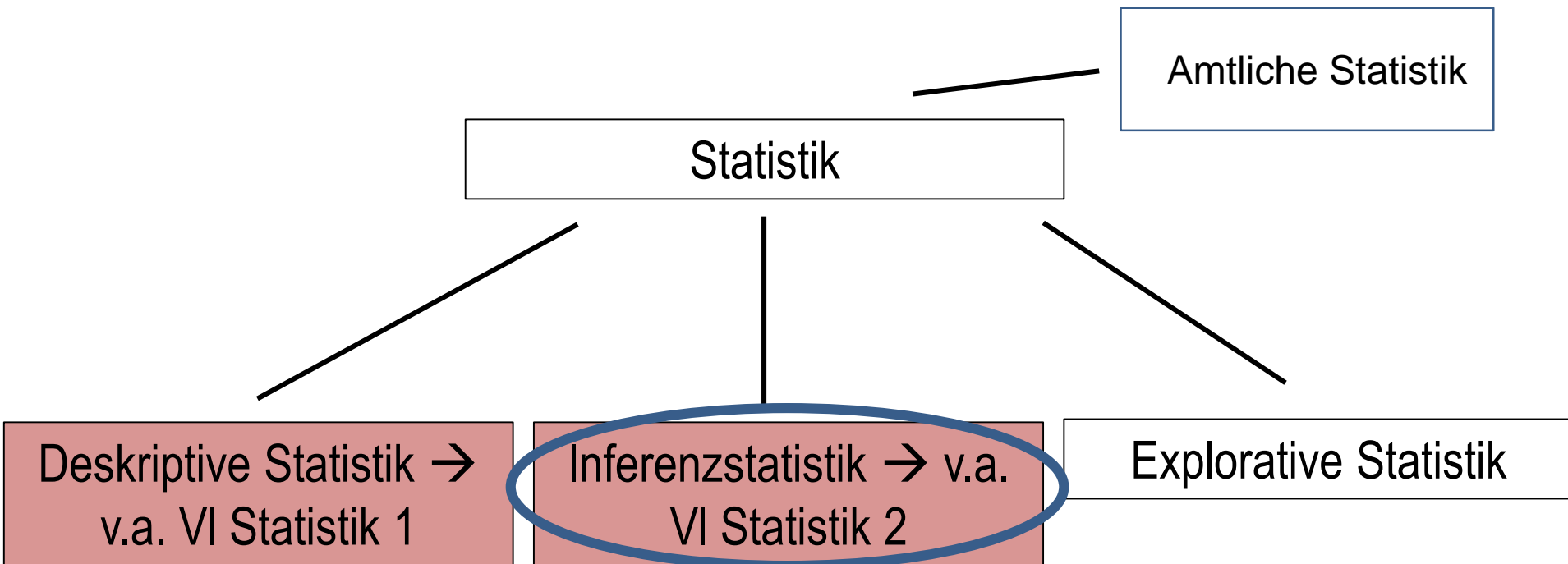
infratest dimap

<https://www.infratest-dimap.de/umfragen-analysen/bundesweit/sonntagsfrage/> (Stand: 16.10.2020)

**Antworten auf die Frage: „Welche Partei würden Sie wählen, wenn am Sonntag Bundestagswahl wäre?“**



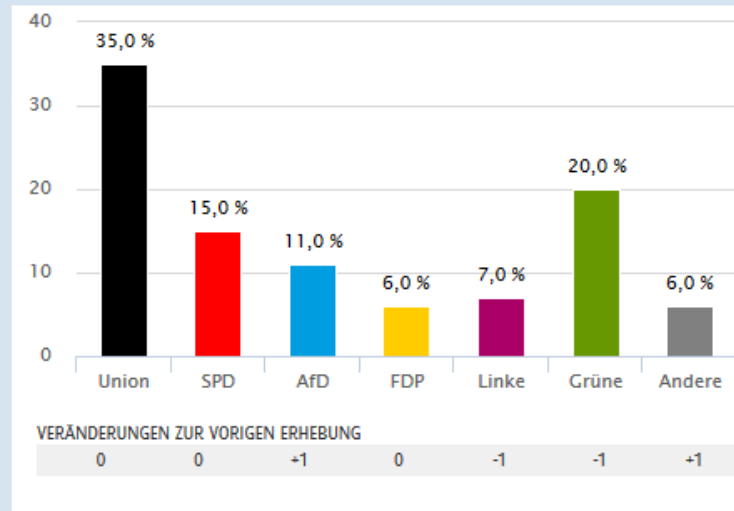
# Bereiche der Statistik



- Schließende Statistik
- Forschungsfragen der quantitativen Forschung beziehen sich i.d.R. auf Grundgesamtheiten (Ziel: Verallgemeinerung)
- Datenerhebung und Datenanalysen werden aber i.d.R. anhand einer Stichprobe durchgeführt

# Deskriptive Statistik

Sonntagsfrage (bundesweit) vom 16.10.2020



infratest dimap

<https://www.infratest-dimap.de/umfragen-analysen/bundesweit/sonntagsfrage/> (Stand: 16.10.2020)

**Stichprobe: ca. 1000 wahlberechtigte Bürger\*innen**

**Grundgesamtheit: deutsche Wahlbevölkerung**

**Häufig können nicht alle relevanten Objekte untersucht werden**

- > 80 Mio. Deutsche, 330 Mio. US-Amerikaner...
- 8 657 deutsche, 8 207 US-amerikanische, ...Befragte

Grundgesamtheit



Stichprobe

Wie bekomme ich eine  
„gute“ Stichprobe?

**Häufig können nicht alle relevanten Objekte untersucht werden**

- > 80 Mio. Deutsche, 60 Mio. Italiener...
- 8 657 deutsche, 1 207 italienische, ...Befragte

Grundgesamtheit



Stichprobe

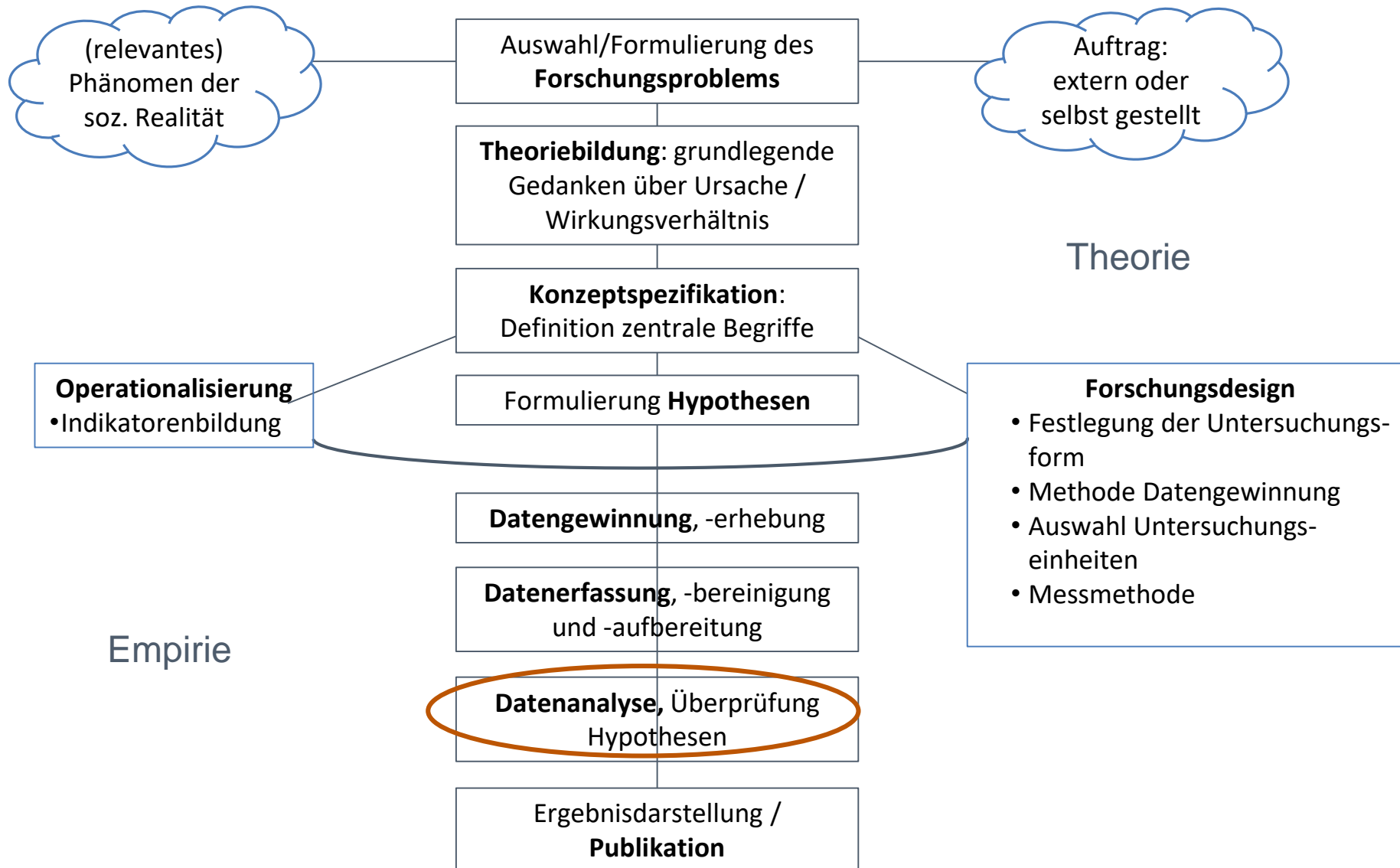
Wie kann ich von der  
Stichprobe auf  
Grundgesamtheit  
schließen?

# Worum geht es bei der Inferenzstatistik?

- Wie kommt man von einer Stichprobe zu einem (gültigen) Schluss auf die Grundgesamtheit?
- Bzw. wie groß ist der Fehler, den man dabei wahrscheinlich macht? (→ **sample error, Stichprobenfehler**)
- Übertragung von Kennwerten der Stichprobe auf die Grundgesamtheit (z.B. Mittelwert, Standardabweichung, Zusammenhangsmaße, etc.)

**Wo würden Sie im sozialwissenschaftlichen Forschungsprozess die Statistik und ihre Anwendungen einordnen?**

# Forschungsprozess, „klassisch“





# Übung: zentrale Begriffe

Bitte formulieren Sie eine kurze Antwort auf folgende Fragen

- Was sind (sozialwissenschaftliche) **Daten**?
- Was ist eine **Grundgesamtheit** bzw. eine **Population**?
- Was ist eine **Stichprobe**?
- Was sind **Beobachtungseinheiten** bzw. **Merkmalsträger**?
- Was ist ein **Merkmal**? Was ist eine **Variable**?
- Was ist eine **Messung**?
- Welche **Skalenniveaus** werden i.d.R. unterschieden?

# Zentrale Begriffe

Was sind sozialwissenschaftliche **Daten**?

**Allgemein:** mithilfe sozialwissenschaftlicher Methoden und Techniken gewonnene Informationen (quantitativ und qualitativ)

**Statistik mit quantitativer Perspektive:**

- Empirische Beobachtungen/Messungen an Objekten (z.B. Personen, Staaten, Parteien, als mögliche Merkmalsträger)
- Werden quantifiziert und auf einem Computer oder auf Papier gespeichert
- Verschiedene Skalenniveaus werden unterschieden

# Übung: zentrale Begriffe

- **Grundgesamtheit oder Population**
  - umfasst die Menge aller möglichen Beobachtungseinheiten, über die man eine Aussage treffen will
  - in der Regel sehr groß, kann meist nicht untersucht werden
- **Stichprobe**
  - eine Teilmenge der Grundgesamtheit. Sie soll ein möglichst getreues Abbild der Grundgesamtheit sein – „repräsentativ“

- **Merkmale/Variablen**
  - Eigenschaften von Merkmalsträgern bzw. Beobachtungseinheiten
- **Beobachtungseinheiten**
  - Merkmalsträger
  - In der quantitativen Sozialforschung häufig Individuen bzw. andere Objekte, an denen etwas beobachtet wird (Wahlkreise, Kommunen, etc.)



## Hintergrund:

quantitative empirische Sozialforschung beschreibt und erklärt soziale Sachverhalte mittels Messungen an sozialen Objekten

gemessen werden nicht die Objekte selbst, sondern **Merkmale**, also Eigenschaften von Objekten

## Definition Messung:

- Zuordnung von Zahlen zu Objektmerkmalen nach bestimmten Regeln (Stevens 1946)
- Zuweisung von Zahlen „entsprechend den Ausprägungen der an diesen Objekten betrachteten Merkmale“ (Kromrey 2009, S. 202)

**Ziel: möglichst strukturtreue Abbildung der Merkmale**

# Beispiel: „Item“ aus Fragebogen

39

 **Liste 203 vorlegen!**

Wie wichtig sind für Sie persönlich die einzelnen Lebensbereiche auf dieser Liste?

Der Wert 1 bedeutet **überhaupt nicht wichtig**, der Wert 7 **sehr wichtig**.

Mit den Werten dazwischen können Sie die Wichtigkeit der Lebensbereiche abstufen.

		Überhaupt nicht wichtig				Sehr wichtig		
		1	2	3	4	5	6	7
A	Eltern und Geschwister .....	<input type="checkbox"/>	...	<input type="checkbox"/>	....	<input type="checkbox"/>	....	<input type="checkbox"/>
B	Freizeit und Erholung .....	<input type="checkbox"/>	...	<input type="checkbox"/>	....	<input type="checkbox"/>	....	<input type="checkbox"/>
C	Schul- und Berufsausbildung .....	<input type="checkbox"/>	...	<input type="checkbox"/>	....	<input type="checkbox"/>	....	<input type="checkbox"/>
D	Partnerschaft .....	<input type="checkbox"/>	...	<input type="checkbox"/>	....	<input type="checkbox"/>	....	<input type="checkbox"/>
E	Politik .....	<input type="checkbox"/>	...	<input type="checkbox"/>	....	<input type="checkbox"/>	....	<input type="checkbox"/>
		<input type="checkbox"/>		<input type="checkbox"/>		<input type="checkbox"/>		<input type="checkbox"/>

Frage des DJI Jugendsurvey 2003

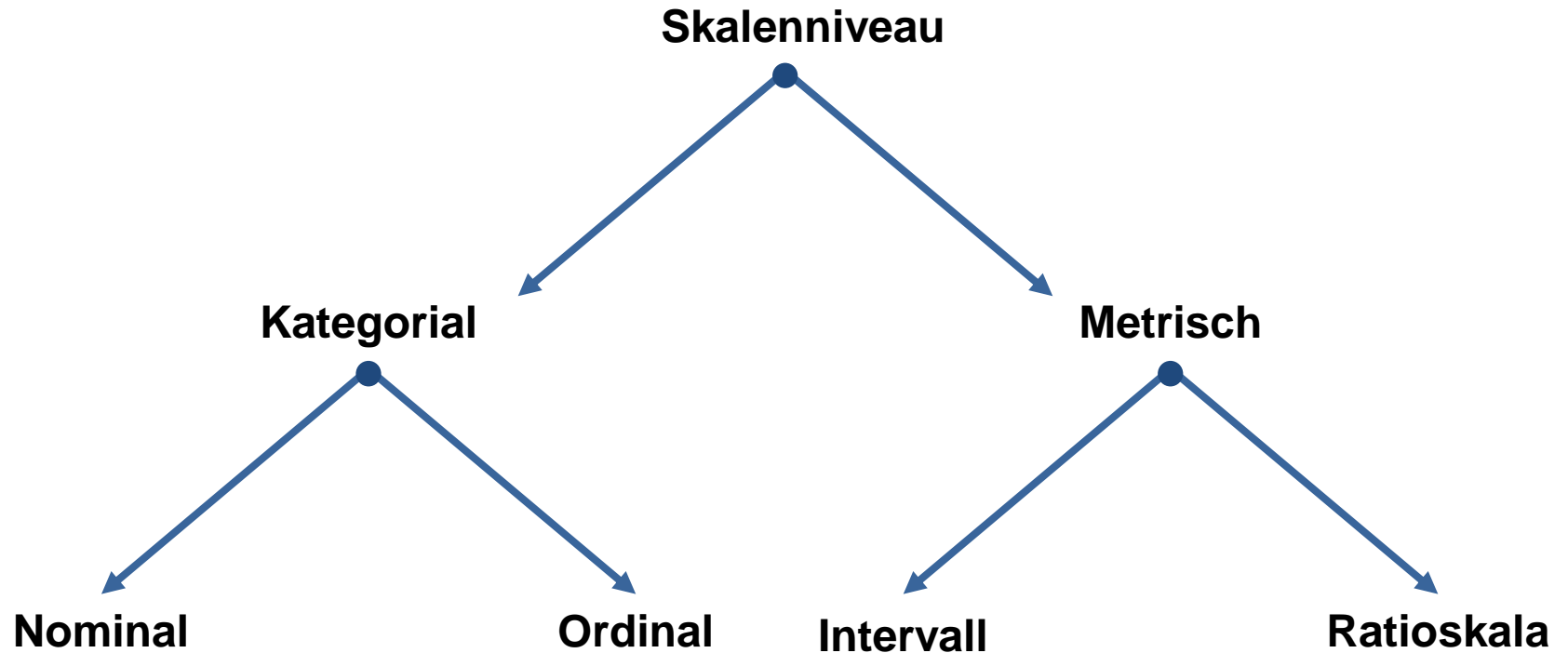
- Nicht jedes Merkmal lässt sich gleich gut in Zahlen darstellen (z.B. Einkommen vs. Parteipräferenz)
- → je nach Art der erhobenen Daten der Merkmale werden verschiedene „Messlatten“ bzw. **Skalenniveaus** verwendet.



- Je nach Beschaffenheit eines Merkmals lassen sich verschiedene Stufen der Skalierbarkeit unterscheiden
- Grundunterscheidung: kategoriale vs. metrische Daten
- Klassischerweise **4 Skalenniveaus** (Stevens 1946):
  - **Nominalskala**
  - **Ordinalskala**
  - **Intervallskala**
  - **Verhältnis-/Ratioskala**
- Das jeweilige Skalenniveau entscheidet darüber, welche statistischen Verfahren zulässig sind!

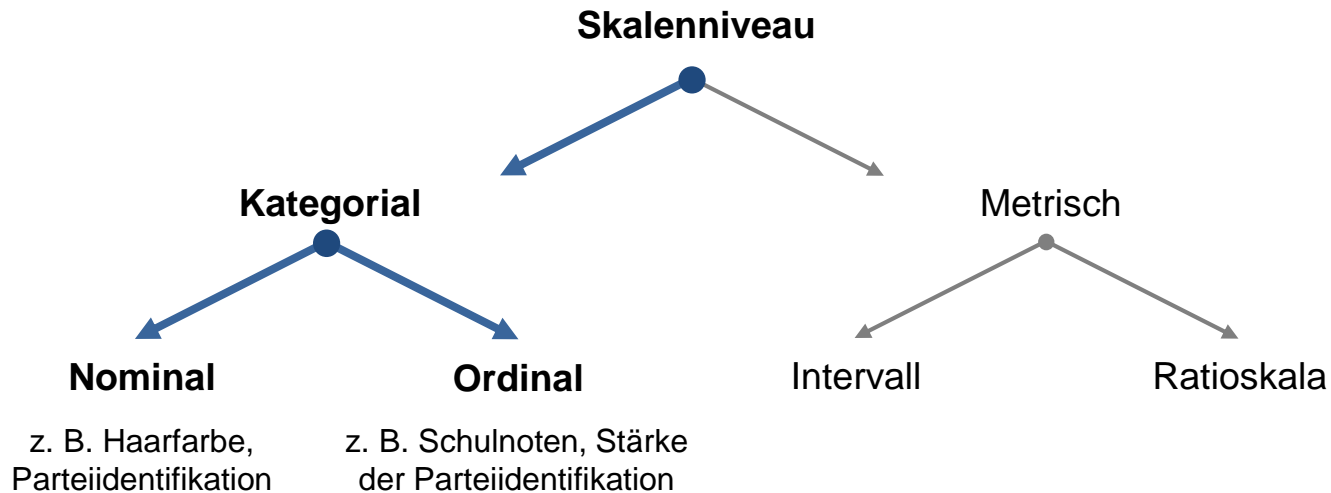
# Zentrale Begriffe: Skalenniveaus

Grundunterscheidung: **kategoriales** vs. **metrisches Skalenniveau**



# Kategoriale Skalenniveaus

- Ergebnis einer Messung erfolgt durch Klassifikation / Einteilung in Kategorien
- Können Wörter, Zahlen oder andere Zeichen sein
- **nominales** und **ordinales** Skalenniveau
  - \* **nominal**: ungeordnete Merkmale
  - \* **ordinal**: geordnete Merkmale



## 1. Nominales Skalenniveau

- „niedrigstes“ Skalenniveau
- Ausprägungen lassen sich nur danach unterscheiden, ob sie gleich oder ungleich sind
- Keine Rangfolge der Kategorien
- Beispiele: Augenfarbe, Automarke, Konfessionszugehörigkeit, Parteipräferenz
- Werden in der quantitativen Forschung dennoch durch Zahlen dargestellt
- Diese Zahlen bieten jedoch keine quantitative Information, man weiß nur, dass bspw. die Parteipräferenz CDU mit einer „1“ „kodiert“ ist

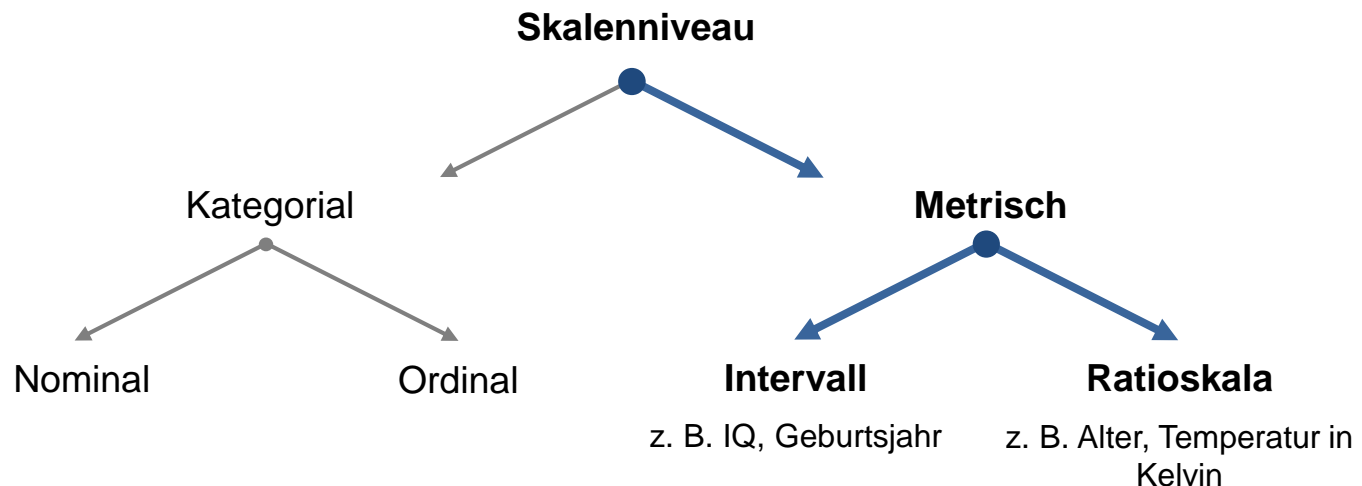
# Wie war das mit den Skalenniveaus?

## 2. Ordinales Skalenniveau

- Nächst höheres Skalenniveau
- Ausprägungen lassen sich auch danach unterscheiden, ob sie ein „mehr“ oder „weniger“ anzeigen
- Abstände nicht konstant, aber klare Reihenfolge
- Bsp: Schulbildung (Hauptschulabschluss, Mittlere Reife, Abitur), Kleidergröße (S, M, L, XL)

# Metrische Skalenniveaus

- Ergebnis einer Messung durch „Zählen“, Messung quantitativer Eigenschaften
- Unterscheidung zwischen Intervall- und Ratioskalenniveau für die sozialwissenschaftliche Statistik kaum relevant:
  - \* **Intervall:** unterteilt die Skala in gleich große Abschnitte
  - \* **Ratio:** hat zusätzlich natürlichen Nullpunkt



## 3. Intervallskalenniveau

- Nächst höheres Skalenniveau
- Ausprägungen lassen sich nicht nur danach unterscheiden, ob sie ein „mehr“ oder „weniger“ anzeigen, sondern auch, wie groß der Unterschied ist
- Kein natürlicher Nullpunkt, aber konstante Abstände
- Bsp: Temperatur in Celsius, Zeitrechnung nach Christi Geburt


## 4. Ratio-Skalenniveau

- „höchstes“ Skalenniveau
- Natürlicher Nullpunkt, Verhältnis zwischen Messwerten interpretierbar
- Bsp: Stimmenanteil eines Kandidaten, Lebensalter in Jahren



# Skalenniveau

- Skalenniveaus sind abwärtskompatibel: Alle Eigenschaften einer niederwertigen Skala gelten auch für eine höherwertige Skala
- In der sozialwissenschaftlichen Forschungspraxis werden ordinale Skalenniveaus häufig als intervallskaliert behandelt (z.B. Fragebogendaten, Einstellungsmessung) → „pseudo-/quasi-“metrisches Skalenniveau

 **Liste 203 vorlegen!**

Wie wichtig sind für Sie persönlich die einzelnen Lebensbereiche auf dieser Liste?

Der Wert 1 bedeutet **überhaupt nicht wichtig**, der Wert 7 **sehr wichtig**.

Mit den Werten dazwischen können Sie die Wichtigkeit der Lebensbereiche abstufen.

		Überhaupt nicht wichtig							Sehr wichtig	
		1	2	3	4	5	6	7		
A	Eltern und Geschwister .....	<input type="checkbox"/>	...	<input type="checkbox"/>	....	<input type="checkbox"/>	....	<input type="checkbox"/>	....	<input type="checkbox"/>
B	Freizeit und Erholung .....	<input type="checkbox"/>	...	<input type="checkbox"/>	....	<input type="checkbox"/>	....	<input type="checkbox"/>	....	<input type="checkbox"/>
C	Schul- und Berufsausbildung .....	<input type="checkbox"/>	...	<input type="checkbox"/>	....	<input type="checkbox"/>	....	<input type="checkbox"/>	....	<input type="checkbox"/>
D	Partnerschaft .....	<input type="checkbox"/>	...	<input type="checkbox"/>	....	<input type="checkbox"/>	....	<input type="checkbox"/>	....	<input type="checkbox"/>
E	Politik .....	<input type="checkbox"/>	...	<input type="checkbox"/>	....	<input type="checkbox"/>	....	<input type="checkbox"/>	....	<input type="checkbox"/>
		<input type="checkbox"/>		<input type="checkbox"/>		<input type="checkbox"/>		<input type="checkbox"/>		<input type="checkbox"/>

# Skalenniveau

Skalenniveau	Relation zwischen Ausprägungen	Beispiele
Nominal	Klassifikation (gleich/ungleich)	Familienstand, Parteipräferenz
Ordinal	Rangordnung (Reihenfolge möglich, Abstände nicht interpretierbar)	Bildungsabschluss, Schulnoten
Intervall	Abstand (Abstände sind äquidistant und deshalb inhaltlich interpretierbar)	Aktuelle Zeitrechnung
Ratio/Verhältnis	Verhältnis (Skala hat echten Nullpunkt, Verhältnisse können berechnet werden)	Einkommen, Alter

## Diskrete und stetige Variablen

- Eine Variable ist **diskret**, wenn sie endlich viele verschiedene Ausprägungen hat; keine „Zwischenwerte“ (Bsp.: Anzahl koreanische Automarken, politische Parteien im Parlament; Kinder in Schulklasse)
- Eine Variable ist **stetig**, wenn zwischen zwei Ausprägungen unendliche viele weitere Ausprägungen vorstellbar sind (Zeit gemessen in Stunden, Minuten, Sekunden, Zehntelsekunden...)

→ Nominale und ordinale Variablen sind immer diskret, metrisch skalierte Variablen können diskret oder stetig sein (Anzahl Kinder einer Person vs. Körpergröße)

# Übungen: Skalenniveau?

**Über welches Skalenniveau verfügen folgende Merkmale aus einer Studierendenbefragung:**

- 1) Semesteranzahl
- 3) Studiengang
- 4) Dauer der Anreise vom Wohnort zur Uni in Minuten?
- 5) Verkehrsmittel für den Weg zur Uni?

## A1 Welchen Abschluss erzielen Sie in Ihrem jetzigen Studiengang? (bei Doppelstudium nur Abschluss des für Sie wichtigsten Studienfaches)

	1	2	3	4	5	6	7	8	9	10
1	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	Bachelor	Master	Diplom	Magister	L1	L2	L3	L5	Medizin, Jura, Veterinärmedizin (Staatsexamen)	Promotion

## A11 Wann werden Sie voraussichtlich Ihr Studium abschließen?

- 1 ☐ noch dieses Semester
- 2 ☐ im kommenden Semester
- 3 ☐ in zwei Semestern
- 4 ☐ in drei Semestern
- 5 ☐ in vier Semestern
- 6 ☐ in fünf Semestern
- 7 ☐ in sechs Semestern
- 8 ☐ in sieben oder mehr Semestern

**Bitte fügen Sie beispielhaft ein**

- **Merkmalsträger:**
- **3 Merkmale:**
- **Merkmalsausprägung:**
- **Skalenniveau:**

- Kenntnis der Organisation und Durchführung der Vorlesung
- Kenntnis und Verständnis statistischer Grundbegriffe
- Kenntnis und Anwendung von Skalenniveaus