

1. Allgemeines

1.1. Was ist Statistik?

Statistik besteht aus zwei Grundbestandteilen:

1. **Zusammenstellung von Daten**, die bestimmte Aspekte der menschlichen Zivilisation oder der Umwelt beschreiben: z.B.
 - Bevölkerungsstatistik oder Entwicklung des BIP eines Landes,
 - Umsatzstatistik eines Unternehmens,
 - Klimastatistik (Temperatur, Niederschläge, ...) einer Region.
2. **Gesamtheit der Methoden** zur Analyse, Beschreibung und Interpretation dieser Daten, z.B.
 - Bestimmung eines Mittelwertes oder Trends,
 - Ermittlung von bestimmten Wahrscheinlichkeiten,
 - Ermittlung von Korrelationen, Abhängigkeiten, Zusammenhängen,
 - Methoden zur Hochrechnung von Stichproben.

Statistik = *methodisches* Vorgehen zur Beschaffung und Auswertung von *quantitativen* Informationen über *Massenphänomene*.

1.2 Arten der Statistik

Deskriptive oder beschreibende Statistik.

Es werden *alle* relevanten Daten der zu untersuchenden statistischen Gesamtheit gesammelt. Alle Aussagen beziehen sich nur auf diese Daten: Hochrechnungen oder Verallgemeinerungen auf eine größere Datenmenge sind nicht erlaubt.

⇒ Dieses Semester.

Induktive oder schließende Statistik.

Daten werden nur von einem *repräsentativen Teil* der Gesamtheit beschafft. Von dieser **Stichprobe** schließt man mit mathematischen Methoden sowie mit der **Wahrscheinlichkeitsrechnung** auf die statistische Gesamtheit, z.B. bei

- Meinungsumfragen,
- zerstörenden Materialprüfungen.

⇒ Statistik II, nächstes Semester.

“Mathematik ist die Wissenschaft der reinen Zahl, Statistik die der empirischen Zahl”

1.3 Wozu nutzt die Statistik?

Ziel ist eine *quantitative, vorurteilsfreie* Beschreibung nahezu aller Bereiche der Zivilisation und Umwelt, bei der eine größere Menge an Daten anfällt, zum Beispiel

- **Struktur:** Wie verteilt sich das gesamte private Vermögen auf die Haushalte? Steigt oder sinkt die Bevölkerungszahl Deutschlands?
- **Wirtschaft:** Steigt oder fällt das BIP im Vergleich zum Vorjahr?
- **Risikoanalyse:** Wieviel Prozent der Bevölkerung sterben durch einen Verkehrsunfall, wieviel durch Lungenkrebs? Was ist gefährlicher beim Genuss eines Steaks: (i) das BSE-Risiko, (ii) die Fahrt zum Steakrestaurant?
- **Ermittlung und Beurteilung von Zusammenhängen,** z.B. Lungenkrebsrisiko in Abhängigkeit vom Zigarettenkonsum.
- Erstellen von **Prognosen:** Gibt es nächstes Jahr eine Rezession? Um wieviel ändert sich die Weltbevölkerung in den nächsten 10 Jahren? Gibt es in 30 Jahren noch Benzin als Treibstoff?

1.4 Fehlerquellen der Statistik

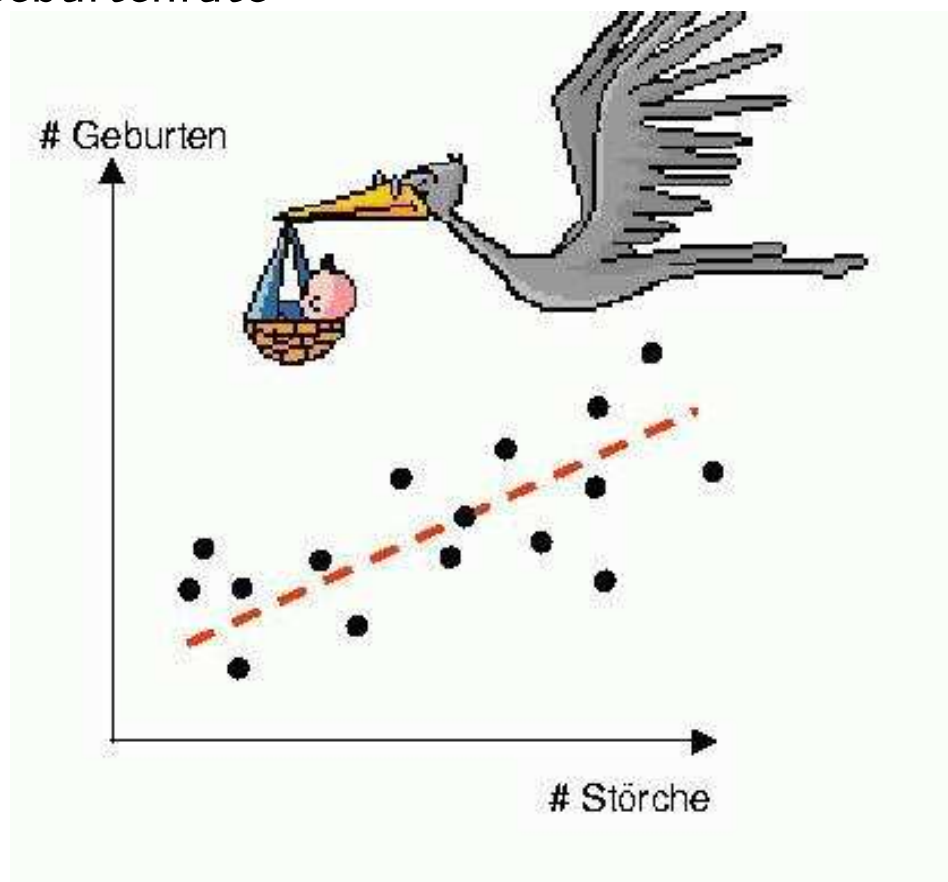
Es gibt drei Arten von Lügen: Lügen, infame Lügen und Statistik.

Benjamin Disraeli (1804-81), engl.

Politiker und Romanschriftsteller, zitiert in Mark Twain

Mögliche Fehlerquellen

- Fehler in den Daten
- Statistische Fehler beim Schließen vom Teil aufs Ganze
- Bewusste "Verdrehung" der Daten
- Falsche Interpretation der Ergebnisse, z.B. Störche vs. Geburtenrate



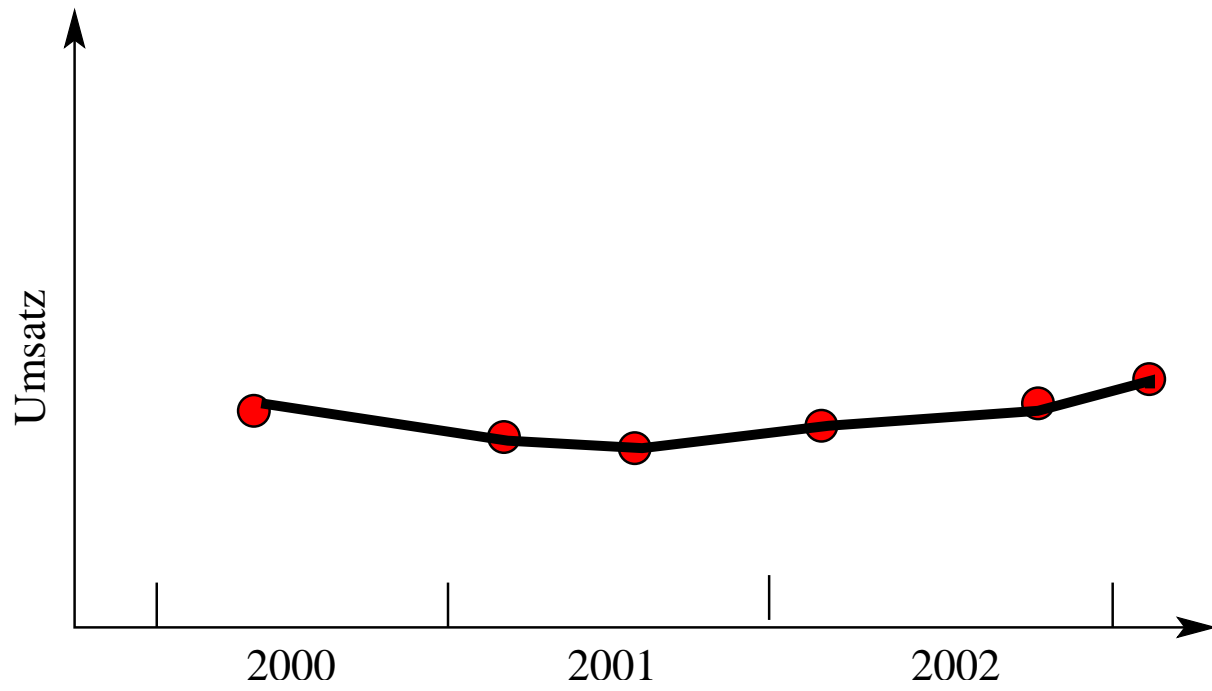
Beispiele von falscher Interpretation/Manipulation

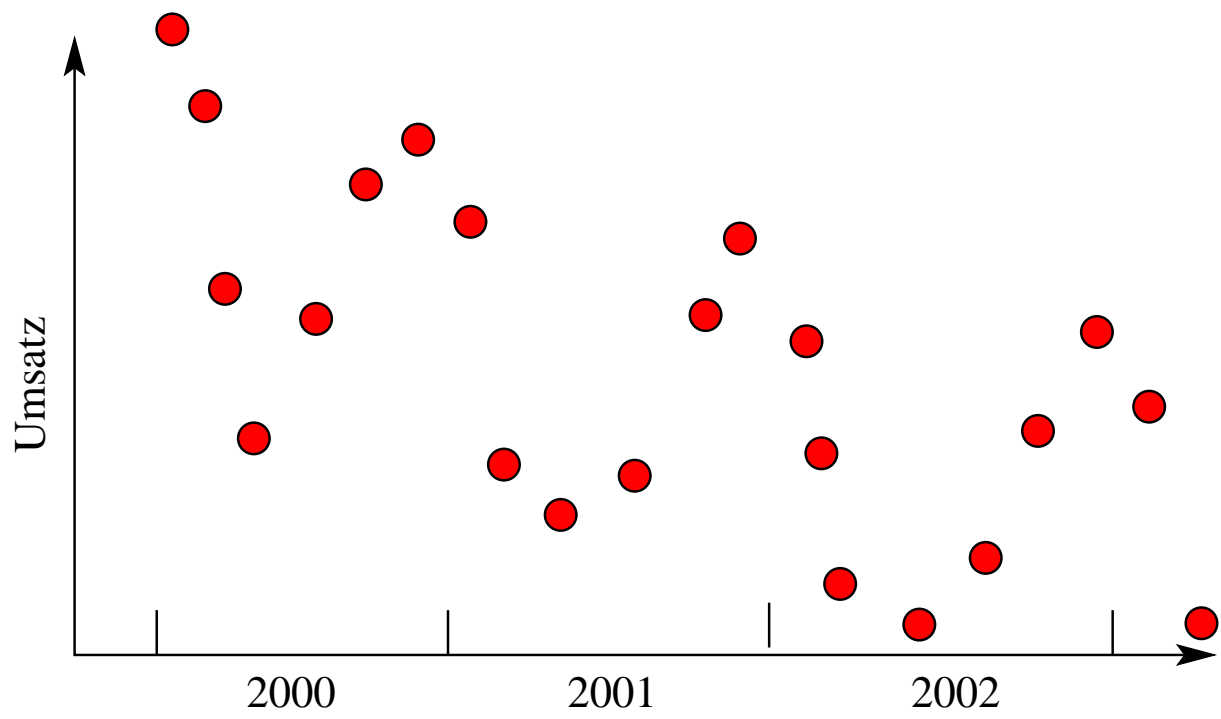
Trau keiner Statistik, die Du nicht selbst gefälscht hast

Churchill

1. Falsche Auswahl der Daten bzw. Wahl der statistischen Gesamtheit
2. Wahl von relativen bzw. absoluten Größen, je nachdem, was "besser passt". Beispiel: Entwicklung der Energieproduktion aus erneuerbaren Quellen.
3. Die statistische Gesamtheit ist nicht homogen.
 - Paradoxon von Simpson (\Rightarrow Rechenbeispiel)
 - Leute mit hohem Einkommen haben statistisch signifikant größere Füße ("sie leben auf großem Fuße")
 - Verkehrszählung: Bei sehr geringer Verkehrsdichte nimmt die mittlere Geschwindigkeit mit der Dichte *ab!*
4. Statistischer Zusammenhang wird mit Kausalität verwechselt (je höher die Klapperstorch-Dichte, desto höher die Geburtenrate)
5. Fallen bei der Interpretation. Beispiel: Zeitreihe des BIP als Indiz für Wohlstand. Fehler durch (i) Nichtberücksichtigung der nichtformellen Arbeit, (ii) Schwarzarbeit, (iii) Inflationskorrektur.

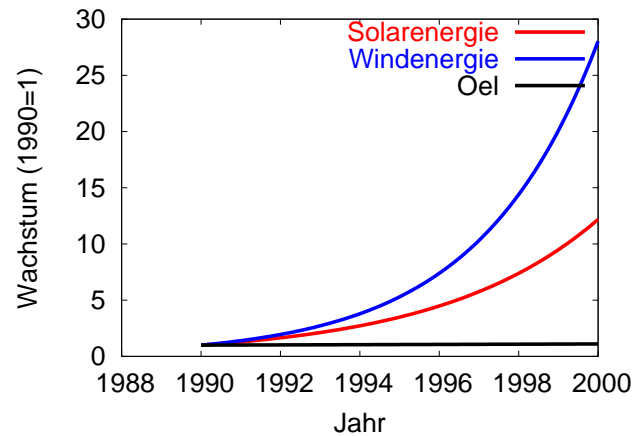
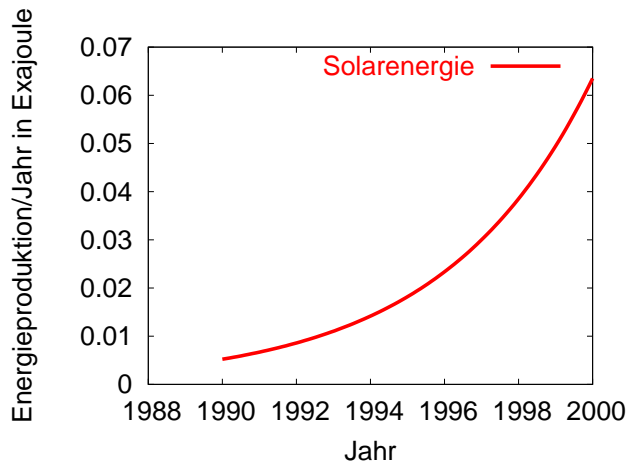
Falsche Auswahl der Daten



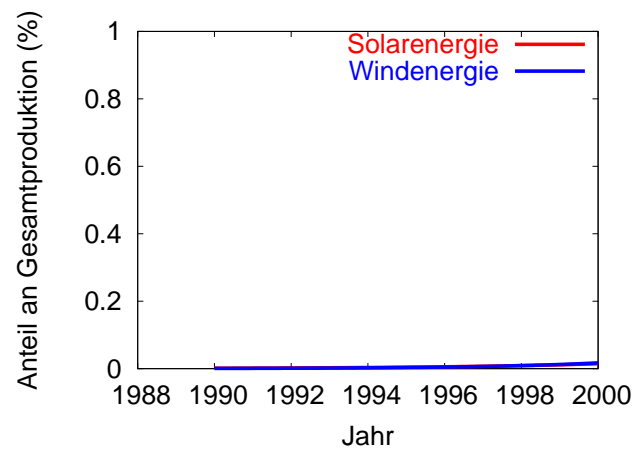
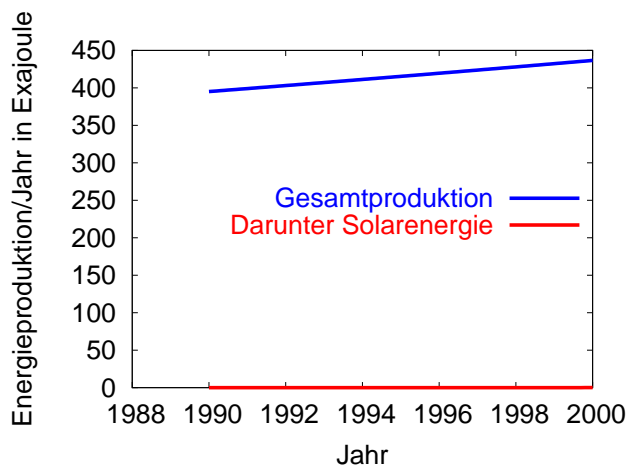


Relative bzw. absolute Größen

Entwicklung der Energieproduktion aus erneuerbaren Quellen

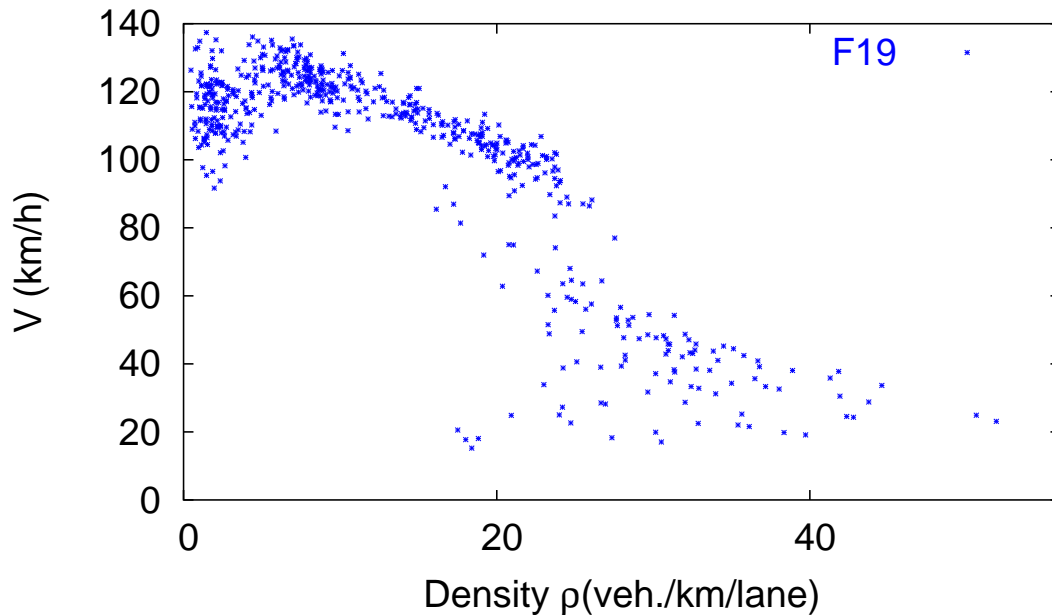


Präsentationsvariante 1



Präsentationsvariante 2

Die statistische Gesamtheit ist nicht homogen



Aufgetragen ist die in Minutenabständen gemittelte Geschwindigkeit auf der A5-Süd in der Nähe von Frankfurt als Funktion der Verkehrsdichte. Jeder Punkt entspricht einer Minute. Wie zu erwarten, nimmt die Geschwindigkeit bei hoher Verkehrsdichte ab (ab etwa 30 Fahrzeuge/km herrscht zähfließender bzw. gestauter Verkehr) *Warum nimmt bei sehr kleinen Dichten die Geschwindigkeit aber mit der Dichte zu anstatt abzunehmen oder konstant zu bleiben??*

1.5. Ablauf einer statistischen Untersuchung

1. Planung

- Formulierung des Zwecks der Untersuchung
- !Definition der statistischen Gesamtheit
- Auswahl der statistischen Verfahren

2. Erhebung

- Stichprobe oder Vollerhebung?
- Wie beschaffe ich die Daten? Sind schon welche vorhanden?

3. **Aufbereitung.** Das Urmaterial wird verdichtet und geordnet (z.B. indem man gewisse Merkmalsklassen definiert und aus der Urliste ein Histogramm erstellt) sowie auf Fehler untersucht.

4. **Analyse**, z.B. mit den im Verlauf der Vorlesung besprochenen Methoden

5. **Interpretation.** Vergleiche u.a. Abschnitt 1.4!

2. Grundbegriffe

1. **Statistische Einheit:** Das zu untersuchende Einzelobjekt, welches Gegenstand der statistischen Untersuchung ist: Einwohner, Unternehmen, Land, Kraftfahrzeug, Tag, etc.
2. **Statistische Masse bzw. Gesamtheit:** Gesamtheit der zu untersuchenden statistischen Einheiten. Die Festlegung der statistischen Masse beinhaltet sehr viele Fehlermöglichkeiten! auf jeden Fall muss sie
 - sachlich,
 - räumlich,
 - und zeitlichabgegrenzt werden.
3. **Merkmale:** Die zu untersuchenden Eigenschaften der statistischen Einheiten
4. **Merkmalsausprägungen:** Die konkreten Ergebnisse der Messung bzw. Beobachtung an den statistischen Einheiten.

2a. Beispiele

1. Ermittlung des Anteils an TUD-Studenten, die mit dem Kfz zur Uni fahren
2. Ermittlung des Wahlverhaltens vor einer Bundestagswahl
3. SrV (System repräsentativer Verkehrsbefragungen)
4. MiD (Mobilität in Deutschland)

Nr	Einheit	Masse	Merkmal	Auspräg.
1				
2				
3				
4				

Nr	Sachliche Abgrenzung	räumliche Abgrenzung	zeitliche Abgrenzung
1			
2			
3			
4			

2.2 Bestands- und Bewegungsmassen

- **Bestandsmassen:** Erfassung zu gewissen *Zeitpunkten*. Die entsprechenden statistischen Einheiten weisen eine gewisse Lebensdauer auf.
- **Bewegungsmassen:** Erfassung in gewissen *Zeiträumen*. Die entsprechenden statistischen Einheiten wachsen mit der Zeit an und verschwinden, wenn das Zeitintervall der Erfassung gegen Null geht.

Oft beschreiben Bewegungsmassen einfach Veränderungs*raten* (Zu- oder Abgänge) einer Bestandsmasse. Solche Paare von statistischen Massen heißen **korrespondierende Massen**; die Änderung einer Bestandsmasse durch korrespondierende Bewegungsmassen heißt **Fortschreibung**.

Beispiele:

Was sind Bewegungsmassen zu folgenden Bestandsmassen?

Bestandsmasse	Bewegungsmassen
zugelassene Kfz	
Fahrzeuge im Stau	
Einwohnerzahl von Dresden	
Kapitaldecke einer Firma	

2.3 Skalierung der Merkmalsausprägungen

Bei der Angabe von Merkmalsausprägungen unterscheidet man drei Skalenarten, die für verschiedene Arten bzw. Qualitäten von Merkmalsausprägungen stehen:

1. **Nominalskala:** Angabe einer *qualitativen* Verschiedenartigkeit, z.B. Geschlecht, Nationalität, Gesellschaftsform
2. **Ordinalskala:** Qualitative Verschiedenartigkeit, bei der zusätzlich eine *natürliche Rangordnung* gegeben ist, z.B. Noten, Güteklassen, Schadensfreiheitsklassen, Rangplätze.
3. **Kardinalskala:** *Quantitative* Verschiedenartigkeit: Neben einer natürlichen Rangordnung sind auch die *Abstände* zwischen je zwei Merkmalsausprägungen zahlenmäßig vergleichbar. Unterteilung in
 - (a) **Intervallskala:** Kein natürlicher Nullpunkt, z.B. Temperatur in Celsius, Jahreszahlen
 - (b) **Verhältnisskala:** Es gibt einen natürlicher Nullpunkt; neben einer Differenzenbildung ist auch eine *Quotientenbildung* sinnvoll (z.B. Körpergewicht)
 - (c) **Absolutskala:** Eine Verhältnisskala, die nicht von den gewählten Einheiten abhängt: Ausnahmslos Stückzahlen.

2.4 Weitere Eigenschaften statistischer Merkmale

Quasistetige Merkmale: Diese sind zwar prinzipiell diskret, werden aber *de facto* als stetig behandelt. *Beispiel:* alle größeren Geldbeträge. Umgekehrt werden stetige Merkmale durch **Klassifizierung** (\Rightarrow Kap. 4) häufig diskret behandelt.

Häufbarkeit: Ein (notwendigerweise nominalskaliertes) Merkmal heißt **häufbar**, wenn ein- und dieselbe statistische Einheit mehrere Ausprägungen dieses Merkmals haben kann

- Man kann z.B. mehrere Berufe haben:
 - Statistische Einheit: Berufstätiger
 - Merkmal: Beruf
 - Ausprägungen: z.B. Physiker *und* Ingenieur
- Andere Beispiele: Freunde, Krankheiten, Bücher etc

Frage: Warum können nur nominalskalierte Merkmale häufbar sein?

Dichotome Merkmale: Ein (notwendigerweise nominalskaliertes) Merkmal mit nur zwei möglichen Ausprägungen heißt **dichotom** (gr. *dicha*=zweifach, *to*=mein=teilen)

Beispiel: Geschlecht.

2.5. Statistischer Merkmale: Aufgaben

1. Ist das Geschlecht häufbar?
2. Warum ist ein dichotomes Merkmal immer nominalskaliert (die Antwort ist nicht so trivial wie sie scheint)
3. Füllen Sie folgende Tabelle aus:

Merkmal	Skalierung	stetig?	häufbar?	dichotom?
Ursachen von Verkehrsunfällen				
Schadenshöhe von Verkehrsunfällen				
Zahl der Verletzten bei Verkehrsunfällen				
Ergebnis (Zeit) beim Abfahrtslauf				
Ergebnis (Note) beim Eiskunstlauf				
erlernte Fremdsprachen				
Kraftstoffverbrauch (l/100 km)				

2.6 Konsequenzen der verschiedenen Skalierungen

- Die weiter unten stehenden Skalierungen sind “höherwertiger”: Jedes absolutskalierte Merkmal ist auch verhältnisskaliert, jedes verhältnisskalierte Merkmal ist auch intervallskaliert, ...
- Die Anwendbarkeit von statistischen Verfahren hängt von der Skalierung ab. Je “quantitativer” (weiter untenstehend) die Skalierung ist, desto mehr Verfahren sind möglich.
- Kardinalskalierte Merkmale heißen auch **metrisch** oder **quantitativ**, die anderen sind **nichtmetrisch** bzw. **qualitativ**.
- Desweiteren wird unterschieden zwischen
 - **diskreten Merkmalen**: Nur bestimmte isolierte Werte sind möglich, sowie
 - **stetigen Merkmalen**: Zumindest innerhalb eines Intervalls sind beliebige Werte möglich.

Stetige Merkmale sind i.A. verhältnisskaliert, während diskrete Merkmale beliebig skaliert sein können.