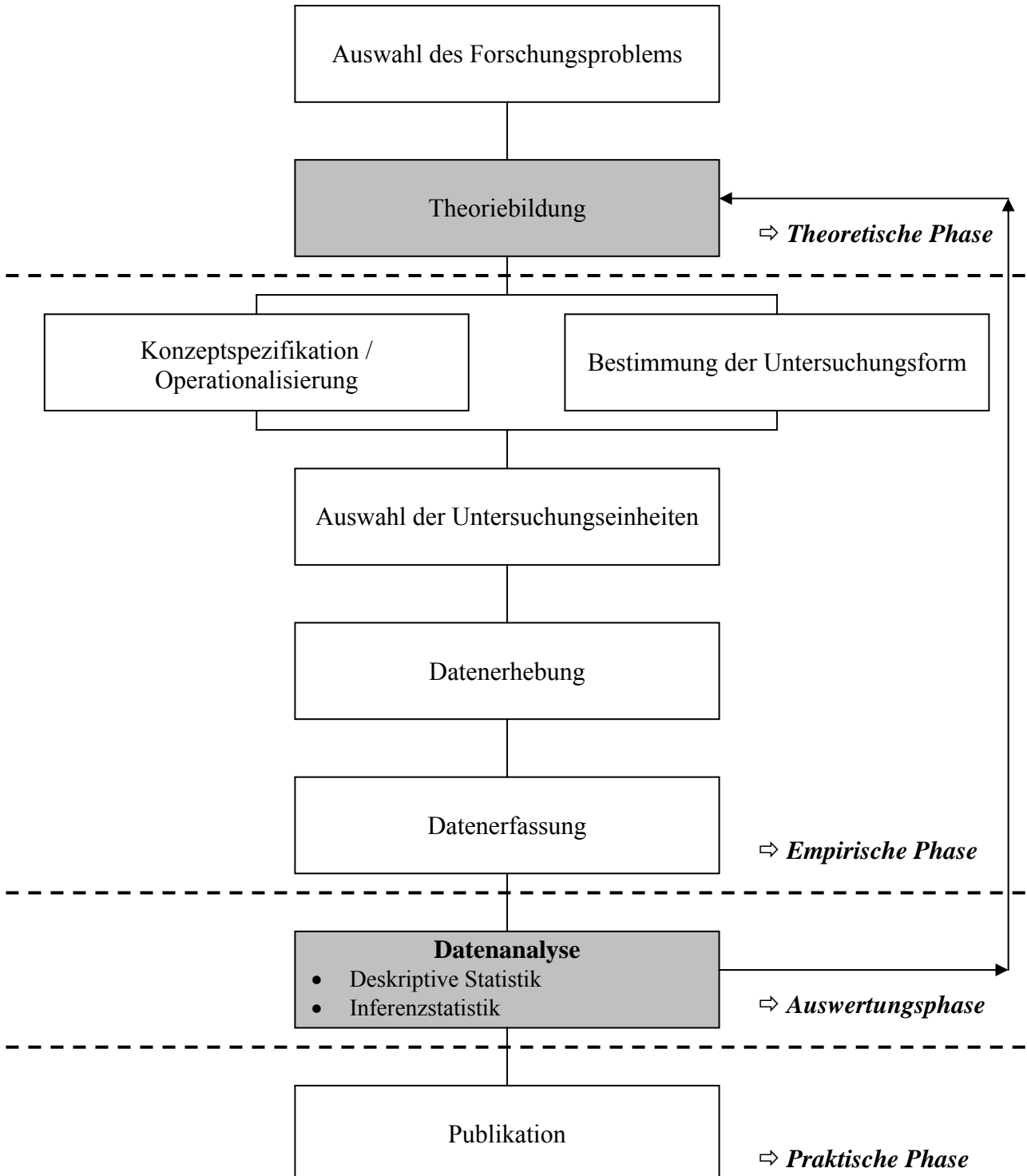


Skript zur Übung: Grundlagen der empirischen Sozialforschung - Datenanalyse

Phasen des Forschungsprozesses



Deskriptive Statistik (Beschreibende Statistik)

Die deskriptive Statistik verfolgt das Ziel, die Untersuchungsergebnisse darzustellen. Die Darstellung erfolgt durch tabellarische und graphische Darstellungen, Maßzahlen der zentralen Tendenz, Streuungsmaße und Korrelationskoeffizienten.

- Die deskriptive Statistik gliedert sich in drei Bereiche auf:
 - a. **Univariate Analyse**
 - b. **Bivariate Analyse**
 - c. **Multivariate Analyse**

Inferenzstatistik (schließende bzw. analytische Statistik)

Die Inferenzstatistik verfolgt das Ziel, von den Ergebnissen einer Stichprobe auf die Grundgesamtheit (Population) zu schließen. Die Inferenzstatistik prüft, in welcher Weise die Ergebnisse der deskriptiven Statistik verallgemeinert werden können. D. h., ob die Ergebnisse einer Stichprobe auf die Grundgesamtheit übertragen werden kann.

Grundbegriffe der Statistik:

Untersuchungseinheit:

Untersuchungseinheit ist das Objekt, an dem Messungen vorgenommen werden.

Die Untersuchungseinheit, ganz allgemein auch Beobachtung oder Fall genannt, ist als **Merkmalsträger** das Bezugsobjekt der Forschung. (z.B. Personen, Schulen, Texte) Die Untersuchungseinheit ist die Einheit, auf die sich die Untersuchung bezieht.

Variable:

Variablen sind **Merkmale** oder **Eigenschaften**, die von Untersuchungseinheit zu Untersuchungseinheit variieren können. Eine Variable ist die Eigenschaft der jeweiligen Untersuchungseinheit. Eine Variable besitzt mindestens zwei Merkmalsausprägungen. Die kleinste Ausprägung wäre „das Vorhandensein“ oder „das Nicht Vorhandensein“ eines Merkmals.

Merkmalsausprägung:

- Merkmalsausprägungen sind Werte, die eine Variable annehmen kann.

Dichotome Variable (Dichotomie):

Besitzt eine Variable **zwei Merkmalsausprägungen**, so spricht man von einer dichotomen (zweiteiligen, zweistufigen) Variablen.

Beispiel:

- Die Variable Geschlecht besitzt zwei Merkmalsausprägungen (1) männlich und (2) weiblich.
- Zu Fragen wie „Sind Sie berufstätig?“ mit den dazugehörige Antwortkategorien (1) ja und (2) nein.

Trichotome Variable (Trichotomie):

Besitzt eine Variable **drei Merkmalsausprägungen**, so spricht man von einer trichotomen (dreistufigen) Variablen.

Beispiel:

- Die Variable Schichtzugehörigkeit besitzt drei Merkmalsausprägungen: (1) Unterschicht, (2) Mittelschicht und (3) Oberschicht.
- Die Variable Augenfarbe besitzt die Merkmalsausprägungen (1) grün, (2) blau und (3) braun.

Polytome Variable (Polytomie):

Besitzt eine Variable **mehr als drei Merkmalsausprägungen**, so spricht man von einer polytomen (mehrstufigen) Variablen.

Beispiel:

- Die Variable Nationalität besitzt mehr als drei Merkmalsausprägungen (1) Belgier, (2) Brite, (3) Franzose, (4) Italiener usw. (5, 6...)
- Die Variable Berufsstatus kann ebenfalls mehr als drei Merkmalsausprägungen besitzen, nämlich: (1) Arbeiter, (2) Angestellter, (3) Beamter und (4) Selbstständiger.
- Zusätzliche Beispiele hierfür sind die Variablen Familienstand, Konfessionszugehörigkeit, Körpergewicht usw.

Objekte können sich der **Qualität** oder der **Quantität** nach unterscheiden. Demgemäß gibt es eine Unterscheidung **qualitativer** und **quantitativer Variablen**. Die beiden Eigenschaftsmöglichkeiten einer Variablen schließen sich gegenseitig aus. D.h., wenn eine Variable qualitativ ist, dann kann sie nicht quantitativ sein.

Quantitative Variable:

Kann man Objekte im Hinblick auf eine bestimmte Eigenschaft **der Größe nach unterscheiden**, d.h. können Objekte hoch oder niedrig, größer oder kleiner, mehr oder weniger sein, so spricht man von einer quantitativen Variablen. Diese Variablen können mit mannigfaltigen **Messeinheiten** in Verbindung gebracht werden.

Beispiel:

Variable Lebensalter (Tag, Monat, Jahr),
Körpergröße (mm, cm, m),
Körpergewicht (Gramm, Kilogramm),
Einkommen (Cent, Euro),
Haushaltsgröße (0, 1, 2 und Personen),
Geburtenrate (0%, 1%, 2%, 3% usw. Geburten),
Anzahl vollendeter Schuljahre (0, 1, 2 usw. Schuljahre).

Qualitative Variable:

Kann man Objekte im Hinblick auf eine bestimmte Eigenschaft **der Art nach unterscheiden**, so spricht man von einer qualitativen Variablen.

Beispiel:

Variable Geschlecht (Merkmalsausprägung: männlich, weiblich),
Nationalität (Däne, Deutscher, Amerikaner usw.),
Konfessionszugehörigkeit (evangelisch, katholisch usw.),
Familienstand (ledig, verheiratet, geschieden usw.),
Gewerkschaftszugehörigkeit (ja, nein).

Da die Kategorien qualitativer Variablen *nicht* größenmäßig geordnet sind, kann ein Objekt bezüglich einer qualitativen Variablen nicht höher, größer oder mehr sein als ein anderes Objekt; die Objekte sind entweder gleich oder ungleich.

Diskrete (*diskontinuierliche*) Variable:

Eine diskrete Variable kann **nur ganz bestimmte Werte annehmen (d.h. nur endlich viele oder abzählbar unendlich viele Werte)**. Obwohl die Werte einen großen Bereich abdecken können, sind sie stets isolierte Werte, zwischen denen **Lücken bzw. Sprungstellen** existieren. Sie beruhen auf einem Zählvorgang.

Beispiel:

Die Variable Haushaltsgröße besteht aus 0, 1, 2, 3, 4 usw. Personen, die in einem Haushalt wohnen. Aber es existiert kein Haushalt, der aus 3,5 Personen besteht. Es liegen hierbei größere Lücken zwischen den Merkmalsausprägungen vor.

Stetige (kontinuierliche) Variable:

Eine stetige Variable kann in einem bestimmten Bereich **jeden beliebigen Wert aus der Menge der reellen Zahlen annehmen**. Sie kann stets in noch feineren Einheiten gemessen werden, so dass zwischen den Werten **keine oder sehr kleine Lücken bzw. Sprungstellen** bestehen. Zwischen den Messwerten sind beliebig viele Zwischenwerte möglich. Sie beruhen auf einem Messvorgang.

Beispiel:

- Die Variable Lebensalter kann in Jahren, Monaten, Wochen, Tagen, Stunden, Sekunden, Millisekunden usw. gemessen werden.
- Das gilt auch beispielsweise für die Variablen Körpergröße, Körpergewicht, Temperatur usw.

Die Variableneigenschaften stetig und diskret schließen sich gegenseitig aus. D.h., wenn eine Variable stetig ist, dann kann sie nicht diskret sein.

Manifeste Variable:

Manifeste Variablen sind **direkt beobachtbar** oder können direkt gemessen werden (in Form einer Frage wie z.B. „Gehören Sie einer Gewerkschaft an?“)

Beispiel:

Die Variablen Körpergröße, Haarfarbe, Gewerkschaftszugehörigkeit, Geschlecht, Schulnoten usw. sind manifeste Variablen

Latente Variable:

Latente Variablen sind **nicht direkt beobachtbar**. Sie werden mit Hilfe von **Indikatoren** erfasst.

Beispiel:

Die Variablen Arbeitszufriedenheit, Vertrauen in die Regierung, Religiosität, Diskriminierung, politische Einstellung usw. sind latente Variablen, also theoretische Begriffe.

Indikatoren:

Indikatoren sind **manifeste Variablen** (beobachtbare Sachverhalte), die als **Ersatz für die latente Variable** fungieren. Indikatoren sind Variablen, die als Hinweis auf eine nicht sichtbare Variable dienen.

Beispiel:

Die latente Variable Arbeitszufriedenheit kann durch folgende Indikatoren messbar gemacht werden.

Mögliche Indikatoren

- für Arbeitszufriedenheit:
- Fernbleiben vom Arbeitsplatz
 - Häufigkeit des Arbeitsplatzwechsels
 - Verbale Zufriedenheitsbekundung usw.

Mögliche Indikatoren

- für Religiosität:
- Gebetshäufigkeit
 - Kirchengangshäufigkeit usw.

Messen:

Der Prozess der Datenerhebung kann auch als Messen bezeichnet werden, denn im Prozess der Datenerhebung messen wir Merkmalsausprägungen von Untersuchungseinheiten.

Messen ist die **strukturtreue Zuordnung von Zahlen zu Objekten nach festgelegten Regeln**. Strukturtreue bedeutet, wenn man eine Variable nehme (z.B. Körpergröße), dann definiert die Variable zwischen den Objekten (z.B. Personen) eine Beziehung (Relation). Person A ist größer/kleiner als Person B.

Skalenniveau bzw. Messniveau:

1. **Nominalskala**
 2. **Ordinalskala**
 3. **Intervallskala**
 4. **Ratio- bzw. Verhältnisskala**
- } metrisch skaliert

Die Nominalskala stellt das niedrigste Messniveau und die Ratio- bzw. Verhältnisskala das höchste Messniveau dar. Ratioskalierte Variablen beinhalten im Vergleich den höchsten Informationsgehalt, daher kann man mit ihnen die meisten Rechenoperationen durchführen.

1) Nominalskala:

- Klassifikation von Untersuchungseinheiten hinsichtlich ihres **Besitzens** oder **Nicht-Besitzens** einer bestimmten Merkmalsausprägung.
- Das Messen auf einer Nominalskala bedeutet nichts anderes als die Einordnung von Untersuchungseinheiten in Merkmalskategorien.
- Die Merkmalsausprägungen bzw. Kategorien müssen sich gegenseitig ausschließen, d.h., kein Fall darf in mehr als eine Kategorie gelangen.

Beispiel:

- Variable Geschlecht, Gewerkschaftszugehörigkeit, Telefonnummern, Hobby von Studierenden usw.

2) Ordinalskala:

- Klassifikation von Untersuchungseinheiten hinsichtlich ihres Besitzens oder Nicht-Besitzens einer bestimmten Merkmalsausprägung. (siehe Nominalskala)

Zusätzlich:

- Bei einer Ordinalskala **werden die Objekte im Hinblick auf den Grad, in dem sie eine Merkmalsausprägung besitzen, geordnet**.
- Eine größer-kleiner-Relation wird zwischen den Merkmalsausprägungen aufgestellt.
- Ordinales Messen informiert aber nicht über die Größe der Differenzen zwischen den Kategorien.

Beispiel:

- Variable Schulnoten, Schichtzugehörigkeit (Unterschicht, Mittelschicht, Oberschicht), Arbeitszufriedenheit (sehr zufrieden - gar nicht zufrieden) usw.

3) Intervallskala:

- Klassifikation von Untersuchungseinheiten hinsichtlich ihres Besitzens oder Nicht-Besitzens einer bestimmten Merkmalsausprägung. (siehe Nominalskala)
- Die Objekte werden im Hinblick auf den Grad, in dem sie eine Merkmalsausprägung besitzen, geordnet (siehe Ordinalskala).
- Eine größer-kleiner-Relation wird zwischen den Merkmalsausprägungen aufgestellt. (siehe Ordinalskala)

Zusätzlich:

- Bei einer Intervallskala können die **exakten Abstände zwischen den Ausprägungen** angegeben werden.
- Der Abstand zwischen zwei beliebig aufeinander folgenden Objekten ist gleich groß, d.h., die Intervalle müssen die gleiche Größe besitzen.
- Bei der Intervallskala liegt ein **willkürlicher Nullpunkt** vor.

Beispiel:

- Variable Temperaturmessung (z.B. in Celsius oder Fahrenheit), Kalenderrechnung usw.

4) Ratioskala:

- Die Ratioskala ist eine **Intervallskala mit einem absoluten Nullpunkt**.
- Der Messwert Null entspricht der tatsächlichen Abwesenheit des gemessenen Merkmals.
- Es sind Aussagen über Quotienten zweier beliebiger Objekte möglich (z.B. Objekt A besitzt doppelt so viel X wie Objekt B).

Beispiel:

- Variable Einkommen, Lebensalter, Haushaltsgröße, Körpergewicht, Körpergröße usw.

Univariate Analyse

Univariate Verteilung:

Die univariate Verteilung ist eine eindimensionale Verteilung, bei der lediglich eine einzelne Variable betrachtet wird.

Häufigkeitsverteilung:

Die Beobachtungsdaten werden so organisiert, dass die in ihnen enthaltenen Informationen in gedrängter Form zum Ausdruck gebracht werden können. Um Einsicht in die Struktur der Daten zu gewinnen, werden die Rohdaten daraufhin untersucht, wie viele Untersuchungseinheiten auf jede Variablenausprägung entfallen. Die aus dieser Operation resultierende Zusammenstellung der Ausprägungen mit den dazugehörigen Häufigkeiten heißt Häufigkeitsverteilung.

Beispiel:

- 10 Kinder einer Schulklasse werden in Hinblick auf die Variable Geschlecht betrachtet. Die Messung ergab folgende Ergebnisse:
 - 6 Kinder = männlich
 - 4 Kinder = weiblich

Tabellarische Darstellung der univariaten Verteilung:

Geschlecht (x_i)	Häufigkeit (f_i)
männlich (x_1)	6 (f_1)
weiblich (x_2)	4 (f_2)
Anzahl der Fälle	10 (N)

Allgemein:

Variable X mit den

- **Messwerten** $x_i = x_1, x_2, x_3, \dots, x_n$
- **(absoluten) Häufigkeiten** $f_i = f_1, f_2, f_3, \dots, f_n$
- **N bzw. n** = Anzahl der Fälle bzw. Untersuchungseinheiten / Stichprobengröße

Die folgenden Rechenoperationen können **ab nominalem Messniveau** angewendet werden:

- **f_n = relative Häufigkeiten** (Betrachtung des i-ten Wertes einer Verteilung)

$$f_n = \frac{f_i}{N}$$

- **$\% f_i$ = prozentuale Häufigkeiten bzw. absolute Prozentwerte**

$$\%f_i = \frac{f_i}{N} \cdot 100$$

Die folgenden Rechenoperationen dürfen erst **ab ordinalem Messniveau** angewendet werden:

- **$\text{cum } f_i$ bzw. f_{ci} = kumulierte (= addierte) absolute Häufigkeiten**

$$\begin{aligned} f_{c1} &= f_1 \\ f_{c2} &= f_1 + f_2 \\ f_{c3} &= f_1 + f_2 + f_3 \\ &\dots \end{aligned}$$

- **$\text{cum } f_n$ = kumulierte relative Häufigkeiten**

$$\text{cum } f_n = \frac{\text{cum } f_i}{N}$$

- **$\text{cum } f_i \%$ bzw. $\% f_{ci}$ = kumulierte prozentuale Häufigkeiten**

$$\text{cum } \%f_i = \frac{\text{cum } f_i}{N} \cdot 100$$

Beispiel:

x_i	f_i	f_n	$\% f_i$	$\text{cum } f_i$	$\text{cum } f_n$	$\text{cum } f_i \%$
1 (x_1)	2	0,2	20%	2	0,2	20%
2 (x_2)	1	0,1	10%	3	0,3	30%
3 (x_3)	5	0,5	50%	8	0,8	80%
4 (x_4)	2	0,2	20%	10	1	100%
Σ	$N = 10$	1,00	100%			

Umgang mit klassierten (gruppierten) Variablen:Beispiel:

Variable „Lebensalter“		
Klassenintervall	Exakte Grenzen	Klassenmitte (x_i)
21 - 25	20,5 - 25,5 bzw. 25,49	23
26 - 30	25,5 - 30,5 bzw. 30,49	28
31 - 35	30,5 - 35,5 bzw. 35,49	33

- Exakte Grenzen \Rightarrow exakte untere und exakte obere Grenze
- Klassenmitte \Rightarrow der Punkt, der das Intervall in Hälften teilt

Berechnung der Klassenmitte:

1. Möglichkeit: $x_i = \frac{(21 + 25)}{2} = \frac{46}{2} = 23$

2. Möglichkeit: $x_i = \frac{(20,5 + 25,5)}{2} = \frac{46}{2} = 23$