WZB

Wissenschaftszentrum Berlin
für Sozialforschung

Einführung in die Quantitative Datenanalyse

Sitzung 2: Skalenniveau und Datensatzexploration

Proseminar an der Freien Universität Berlin
08.05.2017 - Marcus Spittler

Inhalt der 2. Sitzung

- Skalenniveaus
- Datensatz, Tidy Data
- Datentypen in R

Nominalskala

- 2 oder mehr Ausprägungen
- Unterscheidungskriterium ist gleich / ungleich
- Keine Ordnung zwischen den Ausprägungen
- **Beispiele**: *Konfession, Wahlabsicht, Familienstand*

- Dennoch zahlenmäßige Kodierung möglich: `1 = "ledig"`, `2 = "verpartnert"`, etc.

Nominalskala: Die Untersuchungseinheiten müssen im Hinblick auf die interessierende Eigenschaft nur als gleich oder ungleich klassifiziert werden können. Dabei müssen die Kategorien vollständig sein (alle Fälle müssen in irgendeine Klasse eingeordnet werden können) und sich gegenseitig ausschließen (ein Fall darf nur in eine Klasse fallen). Zur Kennzeichnung der Kategorien können Zahlen verwendet werden. Diese dürfen aber nicht zur Berechnung statistischer Maßzahlen verwendet werden, da sie lediglich Etiketten für die Ausprägungen darstellen.

Ordinalskala

- Ausprägungen lassen sich ordnen
- Intervalle/Abstände sind nicht zu interpretieren
- **Beispiele**: Gesellschaftsschicht, Schulnoten, Airbnb-Bewertungen
- mit theoretischer Begründung bedingt als **quasi-**metrische Variablen nutzbar

Ordinalskala: Das Standardbeispiel eines ordinalskalierten Merkmals ist die Schulnote. Mit der Note in einem bestimmten Fach können wir die Schüler nach ihrer Leistung in dem betreffenden Fach einstufen. Wir wissen allerdings nichts über den Grad der Unterschiede zwischen zwei Schülern. Wir können nicht sagen, dass der Unterschied zwischen einem Schüler mit einer Eins und einem Schüler mit einer Zwei gleich ist zu dem Unterschied zwischen einem Schüler mit einer Zwei und einem Schüler mit einer Drei. Daraus folgt, dass wir Zahlen zu den Ausprägungen eines ordinalskalierten Merkmals so zuordnen müssen, dass die Rangordnung der Kategorien (empirisches Relativ) erhalten bleibt. Sieht man also die Schulnoten als ordinalskaliertes Merkmal an, so könnte man genausogut die Notenstufen 1,17,35,36,100 vergeben.

Intervallskala

- Äquivalenz-, Ordnungs- und Abstandsrelation
- Kein fester Nullpunkt, wodurch keine Aussagen über Verhältnisse möglich
- Wird zusammen mit ratioskalierten Merkmalen als **metrisch** bezeichnet
- **Beispiele**:
  - Temperatur in Grad Celsius,
  - Links-Rechts-Dimension,
  - 10-Punkte-Skala zur Erfassung einer rechtsextremistischen Ideologie

Intervallskala: Können wir über die Möglichkeit der Rangordnung hinaus noch zusätzlich die Abstände zwischen den Ausprägungen angeben und die Abstände als gleich ansehen, so sprechen wir von einer Intervallskala.

Ratioskala

- Höchstes Skalenniveau
- Alle elementaren Rechenoperationen möglich
- Wird als **metrisch** bezeichnet
- **Beispiele**:
  - Einkommen,
  - Alter,
  - Entfernungen,
  - Flächen, Preise, Prozentsätze

Ratioskala: Kann auch noch ein absoluter, unveränderlicher Nullpunkt angegeben werden, so liegt eine Ratioskala vor. Bei einer Ratioskala (Verhältnisskala) können - im Gegensatz zu den anderen Skalen - Quotienten (Verhältnisse) zweier Ausprägungen sinnvoll interpretiert werden.

Bedeutung der Skalenniveaus:

- Erst ab Intervallskalenniveau können die *elementaren Rechenoperationen* eingesetzt werden.

- Intervall- und ratioskalierte Merkmale werden auch als **metrische** Merkmale bezeichnet.

- Eine gewisse Sonderrolle kommt denjenigen Merkmalen zu, die nur zwei Ausprägungen besitzen (**dichtom**), wie etwa das Merkmal Parteimitgliedschaft. Kodiert man die Mitgliedschaft mit `1` und die Nichtmitglieder mit `0`, so kann die durch die Anzahl `n` dividierte Summe der Ausprägungen als ein spezieller Mittelwert, im Beispiel als Anteil der Mitglieder, interpretiert werden.

- Es wird sogar die Position vertreten, dichtome Merkmale seien ratioskaliert, da die eine Ausprägung als natürlicher Nullpunkt zur anderen gesehen werden könne.

Zusammenfassung

	Rangordnung möglich	Abstände interpretierbar	Fester Nullpunkt	Beispiel
Nominalskala	nein	nein	nein	Wahlabsicht
Ordinalskala	ja	nein	nein	Schichtzugehörigkeit
Intervallskala	ja	ja	nein	Temperatur in Grad Celsius
Ratioskala	ja	ja	ja	Einkommen

Diskrete und stetige Merkmale

Diskrete Merkmale:

- Man spricht von einem diskreten Merkmal, wenn es endlich oder abzählbar unendlich viele Ausprägungen hat.
- Beispiele: Kinderzahl, monatliches Gehalt.

Stetige Merkmale:

- Ein Merkmal ist stetig, wenn die Menge seiner Ausprägungen ein Kontinuum darstellt, wenn also X *überabzählbar* viele Ausprägungen hat.
- Beispiele: Körpergröße, Temperatur, Haarfarbe.

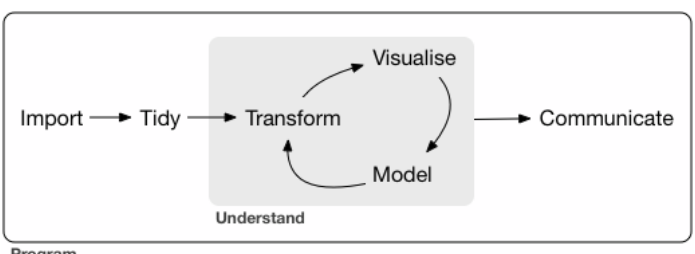
Untersuchungseinheit und Merkmal

- Die **Untersuchungseinheit** e_i ist das Objekt einer statistischen Untersuchung. Sie ist der Träger von Merkmalen, für deren Verteilungen oder etwa deren Zusammenhänge untereinander man sich interessiert.
  - Alternative Begriffe: *Merkmalsträger*, *Fall*, *Element*
  - Symbolik: $1, \ldots i, \ldots . n$
- Die Eigenschaften einer Untersuchungseinheit bezeichnet man als **Merkmal**.
  - Alternativer Begriff: *Variable*
  - Symbolik: x, y, z
- Die möglichen Werte eines Merkmals sind die **Merkmalsausprägungen**.
  - Symbolik: $x_1, x_2, \ldots, x_i, \ldots, x_n$

Merkmale

- **Merkmale** enthalten die veränderlichen Ausprägungen eines Merkmals.
  - Regel: Jeder Untersuchungseinheit e_i wird eine *Zahl* (numerisches Relativ) zugeordnet, die der *realen Merkmalsausprägung* (empirisches Relativ) entspricht.
  - Bsp.: Person antwortet *"stimme voll und ganz zu"*, wir kodieren 5.
  - Wichtig: Ausprägungen müssen sich **wechselseitig ausschließen** und **vollständig** sein.

Alternative Darstellung des Arbeitsprozess



Weitere Begriffe:

- **Data Wrangling**
  - Die Datenaufbereitung, ein zeitintensiver Schritt des Forschungsprozesses.

- **Tidy data**
  - Ein Konzept zum Aufbau von (Analyse-) Datensätzen.
  - **Zeilen** präsentieren *Fälle / (cases)*, die spezifisch, einzigartig, und doch ähnlich zueinander sind.
  - **Spalten** präsentieren **Merkmale**.
  - Daten sind *tidy*, wenn die Variablen aufeinander **abgestimmte Variablentypen** haben, d.h. die gleiche Art von Wert für jede Spalte.

Datentypen in R

In R gibt es verschiedene Datentypen in denen sich Informationen speichern lassen. Abhängig vom Typ lassen sich unterschiedliche Rechenoperationen ausführen.

- **characters**
  - Enthalten typischerweise Text-Informationen
  - Keine Rechenoperationen möglich
  - Keine Rangordnung der Ausprägungen
  - auch `strings` genannt

```
topic <- "Getränkekonsum"
topic
```

```
## [1] "Getränkekonsum"
```

```
class(topic)
```

```
## [1] "character"
```

Datentypen in R

- **numerics**
  - Enthalten Zahlen
  - Alle Rechenoperationen möglich
  - *integers* und *floats* sind Subgruppen
  - Bieten sich besonders für *metrische* Variablen an

```
beer.consumption <- 25
class(beer.consumption)
```

```
## [1] "numeric"
```

Faktoren

- Haben Zahlen als *Werte* und normalerweise Text als *Label*
- *ordered.factors* können verwendet werden um ordinalskalierte Merkmale darzustellen.

```
rating <- c("Low", "High", "Med", "Med", "High")
rating <- factor(rating)
rating
```

```
## [1] Low  High Med  Med  High
## Levels: High Low Med
```

```
rating <- factor(rating, levels=c("Low", "Med", "High"), ordered=TRUE
rating
```

```
## [1] Low  High Med  Med  High
## Levels: Low < Med < High
```

Datentypen in R

- **Vektoren**
  - Aneinanderreihung mehrerer Werte

```
beer.consumption <- c(4,8,20,45,5)
class(beer.consumption)
```

```
## [1] "numeric"
```

```
consumer <- c("Tina", "Stefan", "Aylin", "Horst", "Sergej")
class(consumer)
```

```
## [1] "character"
```

Datentypen in R

- **Data Frames**
  - Das sind gebündelte *vectoren*, die alle die gleiche Länge haben

```
data.frame(
  consumer = consumer,
  beer.consumption = beer.consumption,
  wine.consumption = c(3,1,5,1,0),
  beer.rating = rating
)
```

```
##   consumer beer.consumption wine.consumption beer.rating
## 1     Tina                4                3         Low
## 2   Stefan                8                1        High
## 3    Aylin               20                5         Med
## 4    Horst               45                1         Med
## 5   Sergej                5                0        High
```