Häufigkeiten und deren Visualisierung

Michael Linke

Table of contents

Figure 1: Die Skyline als Barplot, Bild generiert von Midjourney

Zunächst ein paar Begriffserläuterungen: Bei der quantitativen Datenerhebung misst man die Ausprägung eines bestimmten Merkmals. Ein Merkmal kann zum Beispiel das Alter einer Per-

son sein. Die Person ist dann der Merkmalsträger. Die Merkmalsausprägung bzw. der Messwert ist die konkrete Altersangabe, z.B. 23 Jahre. Die Menge aller Merkmalsträger, über die man durch die Untersuchung zu Erkenntnissen kommen will, heißt Grundgesamtheit. Oft kann man aber gar nicht oder nur sehr schwer die ganze Grundgesamtheit heranziehen und muss sich mit einer Teilmenge begnügen: Der Stichprobe. Die Daten sind erhoben, die Arbeitsumgebung ist eingerichtet, wir haben die Datensätze geladen und wissen, wie wir sie bearbeiten und visualisieren können. Nun wird es Zeit, sich einen ersten Eindruck von den statistischen Eigenschaften unserer Daten zu verschaffen. Von diesen Eigenschaften hängt ab, welche Methoden wir auf sie anwenden können und somit, welche Fragen wir mit ihnen überhaupt beantworten können. Dabei interessieren uns zunächst zwei Aspekte, erstens: Welche Methoden passen zur Beschaffenheit der Daten und zweitens: Welche inhaltlichen Eigenschaften können wir darauf aufbauend in den Daten erkennen? Ersteres ist abhängig vom Skalenniveau, für letzteres schauen wir uns einige der wichtigsten Maßzahlen (Parameter) an.

1 Häufigkeiten

Zur Beschreibung insbesondere nominaler Merkmale ist der Begriff der Häufigkeit wichtig.

Absolute Häufigkeit

Die absolute Häufigkeit gibt an, wie oft eine bestimmte Merkmalsausprägung im Datensatz vorkommt. Die absolute Häufigkeit kann folglich nur eine natürliche Zahl sein.

Relative Häufigkeit

Die relative Häufigkeit gibt den Anteil eines bestimmten Messwertes im Datensatz an. Sie berechnet sich, indem man die absolute Häufigkeit des jeweiligen Messwertes durch die Gesamtgröße des Datensatzes teilt. Die relativen Häufigkeiten summieren sich also zu 1 auf.

2 Maßzahlen

Wir werden hier, abhängig vom Skalenniveau, zwei verschiedene Arten von Maßzahlen betrachten. Das ist zum einen das Lagemaß, das uns etwas über die zentrale Tendenz der Daten sagt, also wo sich besonders viele Messwerte häufen bzw. wo der zentrale Punkt ist, um den sich die Messwerte gruppieren. Zum anderen schauen wir uns verschiedene Streuungsparameter an, mit deren Hilfe wir einschätzen können, wie stark unsere Messwerte vom Lageparameter abweichen.

2.1 Data Management

Im Folgenden werden wir aus dem Allbus-2021-Datensatz ein paar Beispiele herausgreifen, um die Berechnung und Visualisierung von Häufigkeiten und Parametern zu demonstrieren. Dazu installieren und laden wir zunächst die nötigen Pakete mit Hilfe von Pacman und dem `p_load`-Befehl:

```
if(!require("pacman")) {install.packages("pacman");library(pacman)}
p_load(tidyverse, ggplot2, haven, dplyr)
```

Dann legen wir den Visualisierungshintergrund fest:

```
theme_set(theme_classic())
```

Nun laden wir den Allbus-Datensatz:

```
daten = haven::read_dta("Datensatz/Allbus_2021.dta")
```

Bei den Häufigkeiten beschränken wir uns auf einen nominal skalierten Datensatz: Das Geschlecht ("sex").
Anschließend konvertieren wir die Daten zu Zahlenwerten und entfernen fehlerhafte Daten:

```
allbus_messniveau_bsp = daten %>%                          ①
  select("sex") %>%                                        ②
  na.omit() %>%                                            ③
  mutate(sex = haven::as_factor(sex))                      ④
```

① Wir erstellen ein neues Objekt basierend auf dem Datensatz `allbus_messniveau_bsp`,
② Mit `select` wählen wir die Variable `sex`aus
③ Wir entfernen mit der Funktion `na.omit` fehlende Werte aus dem Datensatz
④ Wir kodieren die Variable `sex` als Faktor und übernehmen die Label.

3 Berechnung und Visualisierung von Häufigkeiten

Die absoluten Häufigkeiten lassen sich einfach mit der `table`-Funktion abfragen. Im folgenden Codebeispiel schauen wir uns dazu die Häufigkeiten im Datensatz "sex" an. Uns werden zwei Zeilen ausgegeben: Die erste Zeile enthält den Tabellenkopf (1: männlich, 2: weiblich, 3: divers). Die zweite Zeile enthält zu jeder Merkmalsausprägung die zugehörige Anzahl (= absolute Häufigkeit):

```r
geschlecht_haeufigkeit_abs = table(allbus_messniveau_bsp$sex)
stichprobengroesse = length(allbus_messniveau_bsp$sex)

geschlecht_haeufigkeit_abs                                                    ①
```

① Ausgabe der Tabelle

```
KEINE ANGABE         MANN         FRAU       DIVERS
          20         2614         2705            3
```

Insgesamt summieren sie sich (wie erwartet) zu 5322 Merkmalsträgern auf, was auch der Größe des Datensatzes entspricht:

```r
sum(geschlecht_haeufigkeit_abs)  #<1>
stichprobengroesse                                                            ②
```

① Ausgabe der Summe der absoluten Häufigkeiten
② Ausgabe der Gesamtanzahl an Merkmalsträgern im Datensatz

```
[1] 5342
[1] 5342
```

Die relativen Häufigkeiten können wir uns ausgeben lassen, indem wir den Inhalt der Tabelle durch die Gesamtanzahl der Merkmalsträger teilen:

```r
geschlecht_haeufigkeit_rel = geschlecht_haeufigkeit_abs / stichprobengroesse  #<1>
geschlecht_haeufigkeit_rel     #<2>
```

① Berechnung der relativen Häufigkeiten
② Ausgabe der Tabelle mit den relativen Häufigkeiten

```
KEINE ANGABE         MANN         FRAU       DIVERS
0.0037439161 0.4893298390 0.5063646574 0.0005615874
```

Erwartungsgemäß summieren sich die relativen Häufigkeiten zu 1 auf:

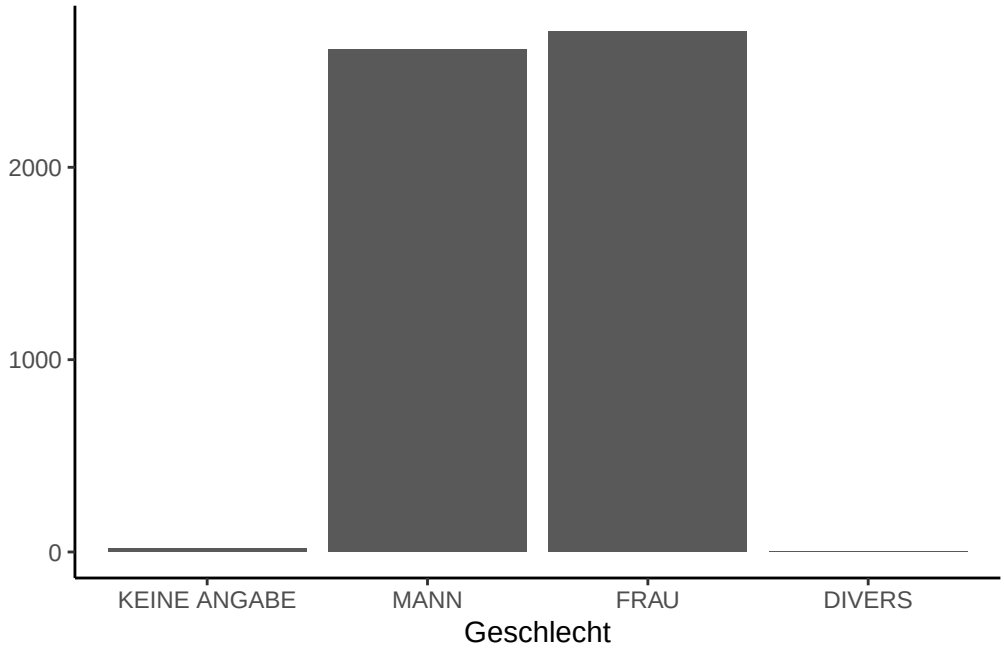
```r
sum(geschlecht_haeufigkeit_rel) #<1>
```

① Ausgabe der Summe der relativen Häufigkeiten

```
[1] 1
```

Eine Möglichkeit der Darstellung von Häufigkeiten ist das Balkendiagramm:

```r
ggplot(allbus_messniveau_bsp, aes(x=sex)) +
  geom_bar() +
  xlab("Geschlecht") +
  ylab("")
```



Auf den ersten Blick lässt sich so ein generelles Bild des Datensatzes machen: Es sind in etwa so viele Männer wie Frauen im Datensatz, wobei Frauen etwas stärker vertreten sind. Personen, die sich weder als Mann noch als Frau identifizieren, kommen nur in sehr geringer Zahl vor (3 Fälle, die leider nicht korrekt angezeigt werden).