Deskriptive Statistik

by Woche 3

> **i** Hinweis
>
> Dieses Kapitel ist ein Exkurs in das generelle Verständnis von Statistik. Es hätte prinzipiell auch früher oder später im Kurs platziert werden können, ist also nicht zwingend mit NumPy verbunden. Tatsächlich ist es nicht mal direkt mit Python verbunden. Da wir aber endlich die Werkzeuge haben, um Daten zu verarbeiten, soll es hier eine Abwechslung zum reinen Programmieren bieten.

Dieses Kapitel soll einen Überblick über die wichtigsten Kennzahlen der deskriptiven Statistik geben. Dazu werden zum Einen die **NumPy** Funktionen aus dem letzten Kapitel herangezogen und zum Anderen erste Abbildungen mit **Matplotlib** und **Seaborn** erzeugt. Das Kapitel lehrt also mehrere Dinge auf einmal, was aber auch sinnvoll ist, da die Themen eng miteinander verknüpft sind und sich gut ergänzen.

Die **deskriptive Statistik** (auch: beschreibende Statistik) ist ein Teilgebiet der Statistik und hat zum Ziel Daten durch Tabellen, Kennzahlen und Grafiken übersichtlich darzustellen und zu ordnen. Ihr gegenüber steht die **schließende Statistik** (auch: mathematische/inferentielle/induktive Statistik) die aufgrund von Stichproben auf die Gesamtheit schließt und Hypothesen testet.

Bei der deskriptiven Statistik gibt es also demnach keine Hypothesen, die mit Signifikanztests geprüft werden und folglich auch keine p-Werte oder Konfidenzintervalle. Stattdessen geht es ersteinmal darum, ein Grundverständnis für die Daten zu bekommen. Das mag auf den ersten Blick banal klingen, ist aber in jedem Fall der erste Schritt, auch wenn anschließend schließende Statistik betrieben werden soll. Und tatsächlich ist deskriptive Statistik mit zunehmender Komplexität der Daten auch nicht mehr so banal, wie es auf den ersten Blick scheint, sondern ein Handwerk, das gelernt sein will.

Die wichtigsten Kennzahlen der deskriptiven Statistik sind:

- **Lagemaße**
  - geben an wo die Daten liegen, also in welchem Bereich sie sich befinden.
  - **Mittelwert** (auch: Durchschnitt, arithmetisches Mittel)
  - **Median** (auch: Zentralwert)
  - **Minimum** und **Maximum**
  - **Quantile** (bzw. **Quartile**)

- **Streuungsmaße**
  - geben an wie weit die Daten auseinander liegen, also wie stark sie streuen.
  - **Spannweite** und **Interquartilsabstand**
  - **Varianz** und **Standardabweichung**
  - **Variationskoeffizient**
- **Zusammenhangsmaße**
  - geben an wie stark zwei Variablen miteinander zusammenhängen.
  - **Korrelation**

Natürlich dürften zumindest einige dieser Begriffe bereits bekannt sein. Dennoch sollte in einem Data Analyst Kurs sichergestellt sein, dass alle ein Verständnis auch für die bestimmte Feinheiten dieser Kennzahlen haben.

Im gleichen Zug lernen wir wie gesagt auch **Matplotlib** kennen, das uns die Möglichkeit gibt, die Daten grafisch darzustellen.