Statistik

[MAT.301UF]

gelesen von: Siegfried Hörmann, Univ.-Prof. Mag.rer.nat. Dr.rer.nat.
am Institut für Statistik
Technische Universität Universität Graz

verfasst von: Laura Philomena Mossböck
11820925

Wintersemester 2023/24

Inhaltsverzeichnis

1 Einführung

Mehr und mehr Daten werden in einer Vielzahl an Gebieten gesammelt, das umfasst unter anderem

- Forschung allgemein (Genetik, Physik, ...)
- Klima und Umweltverschmutzung (Temperatur, Niederschlag, CO_2-Werte,...)
- Wirtschaft (hochfrequenter Aktienhandel, Arbeitsmarkt, ...)
- Internet (Suchverhalten, Likes, ...)
- Sensorik (Smart-Home, autonomes Fahren,...)
- Medizin (MRT, ...)
- uvm.

Das Ziel der Statistik ist es, Dazu zu sammeln, zu verarbeiten, zusammenzufassen und daraus Schlüsse zu ziehen. Um überhaupt erst Statistik betreiben zu können, müssen wir eine zugrundeliegende Population festlegen, die wir erforschen wollen. Das kann z.B. eine Charge eines Medikaments sein, alle Österreicher, oder die Preise von Häusern im Jahr 2023. Sammeln wir systematisch alle Daten, an denen wir interessiert sind, so nennen wir diesen Sammelprozess einen Zensus. Wird man z.B. in Österreich geboren, so wird ein Eintrag im zentralen Melderegister erstellt. Damit haben die entsprechenden Stellen Zugriff auf Daten wie das Alter oder Geschlecht.

Ein Zensus ist in vielen Fällen nicht durchführbar, da Faktoren wie Kosten Grenzen setzen. Stattdessen sammelt man also nur Daten von einer Teilmenge der Population. Die gesammelte Menge an Daten nennen wir eine Stichprobe, und ein einzelnes Element einer Stichprobe nennt man eine Beobachtung.

Gesammelte Daten kategorisieren wir in verschiedene Datentypen. Wir arbeiten mit drei verschiedenen Datentypen. Der erste Typ sind numerische Daten. Diese können wir mittels Zahlen darstellen. Beispiele sind etwa Alter, Körpergröße, Preis oder Länge. Als zweiten Typen betrachten wir kategorische Daten. Dabei handelt es sich z.B. um Job oder Familienstand. Als dritten Typ nennen wir ordinale Daten. Das sind kategorische Daten, die geordnet sind, wie etwa Noten in einer Schule.

Erheben wir von jedem Subjekt nur eine Variable, so arbeiten wir mit univariaten Daten. Erheben wir mehrere Variablen, so sprechen wir von multivariaten Daten. Multivariate Daten müssen nicht zwingen aus nur einem Datentypen bestehen.

Mit wachsenden technischen Möglichkeiten, insbesondere größere Speicher und höhere Rechenleistung, werden wir mit neuen Datentypen konfrontiert. Hier gibt es etwa auch funktionale Daten (wie z.B. Bilder), objektorientierte Daten (z.B. nicht euklidische Geometrie), oder Netzwerkdaten. Diese seien hier nur der Vollständigkeit halber erwähnt.

Typischerweise starten wir eine statistische Betrachtung mit einer empirischen Datenanalyse. Deren Ziel ist es, einfache Merkmale der Daten zusammenzufassen und zu beschreiben. Dabei handelt es sich oftmals um graphische Werkzeuge wie etwa

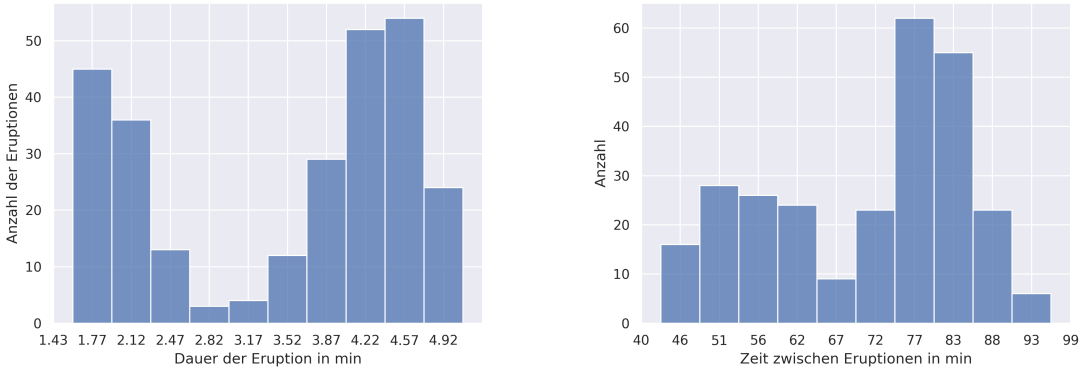
- Histogramme
- Liniendiagramme
- Kreisdiagramme
- Fehlerbalken
- Boxplots
- Streudiagramme
- Quantil-Quantil-Diagramme
- etc.

Zusätzlich können wir Daten durch verschiedene Kenngrößen zusammenzufassen. Diese umschließen etwa

- Mittelwert
- Median
- Quantil
- Korrelation
- Standardabweichung
- Schiefe
- Kurtosis (Wölbung)
- ...

Diese Kenngrößen bestimmen wir in der Praxis mit Softwaresystemen wie R, Python, SPSS, SAS oder Excel.

Beispiel 1.1. *Der Geysir „Old Faithful“ befindet sich im Yellowstone Nationalpark. Das Softwarepaket R stellt ein bivariates Datenset zur Verfügung, in dem sich 272 Messungen befinden. Gemessen wurden die Dauer der Eruptionen und die Zeit zwischen Eruptionen. Mittels Histogrammen können wir einen Eindruck der Randverteilungen der zwei Variablen erhalten.*



(a) Histogramm der Eruptionsdauern

(b) Histogramm der Wartezeiten

Abbildung 1: Beispiele für Histogramme

Man beachte die bimodale Form der beiden Histogramme. Eine Frage, die sich hier auftut ist, ob die Dauer der Eruption einen Einfluss auf die folgende Wartezeit hat. Betrachten wir das Streudiagramm. Dieses deutet

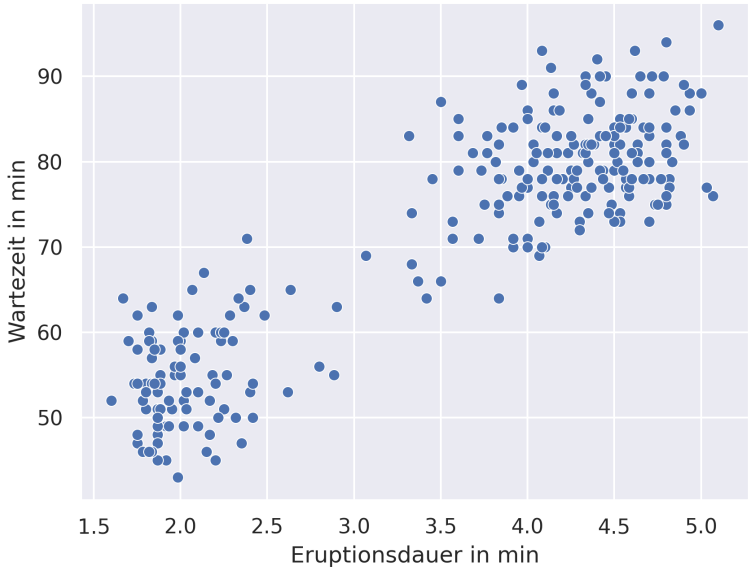


Abbildung 2: Streudiagramm zu den Eruptionen von „Old Faithful“

auf eine positive Korrelation der beiden Variablen hin.

Fortschreitend wollen wir ausgehend von unserer Stichprobe Rückschlüsse machen. Dazu brauchen wir eine geeignete mathematische Theorie, wozu wir die Wahrscheinlichkeitstheorie verwenden werden. Beobachtungen werden somit als Zufallsvariablen modelliert, deren Verteilungsfunktionen als Eigenschaften der Population betrachtet werden.