

Statistik1

Sebastian Sauer

2024-08-29

Inhaltsverzeichnis

| | |
|--|---------------|
| 1. Willkommen! | 3 |
| 1.1. Es geht um Ihren Lernerfolg | 3 |
| 1.1.1. Lernziele | 3 |
| 1.1.2. Was lerne ich hier und wozu ist das gut? | 4 |
| 1.1.3. Was ist hier das Erfolgsgeheimnis? | 5 |
| 1.1.4. Motivieren Sie mich! | 5 |
| 1.1.5. Voraussetzungen | 6 |
| 1.1.6. Überblick | 6 |
| 1.2. Software | 7 |
| 1.2.1. Installation | 7 |
| 1.2.2. Viel R (?) | 7 |
| 1.3. Zum Autor | 7 |
| 1.4. Nomenklatur | 7 |
| 1.4.1. Griechische Buchstaben | 7 |
| 1.5. Zitation | 8 |
| 1.6. Reproduzierbarkeit | 8 |
| I. Organisatorisches | 9 |
| II. Modellieren | 10 |
| 2. Modellgüte | 11 |
| 2.1. Lernsteuerung | 11 |
| 2.1.1. Standort im Lernpfad | 11 |
| 2.1.2. Lernziele | 11 |
| 2.1.3. Benötigte R-Pakete | 11 |
| 2.1.4. Benötigte Daten | 11 |
| 2.1.5. Zum Einstieg | 12 |
| 2.2. Warum Sie die Streuung Ihrer Daten kennen sollten | 12 |
| 2.2.1. Die Schlankkeitspille von Prof. Weiss-Ois | 12 |
| 2.2.2. Wie man seine Kuh über den Fluss bringt | 13 |
| 2.3. Woran erkennt man ein gutes Modell? | 14 |
| 2.4. Streuungsmaße | 15 |
| 2.4.1. Der mittlere Abweichungsbalken | 16 |
| 2.4.2. Der Interquartilsabstand | 17 |
| 2.4.3. Histogramm | 18 |

| | | |
|--------|--|----|
| 2.4.4. | Dichtediagramm | 18 |
| 2.4.5. | Streuungsmaße für Normalverteilungen | 19 |
| 2.4.6. | Varianz | 19 |
| 2.4.7. | Die Standardabweichung | 23 |
| 2.5. | Streuung als Modellfehler | 24 |
| 2.6. | z-Transformation | 25 |
| 2.7. | Fazit | 28 |
| 2.8. | Aufgaben | 28 |
| 2.8.1. | Datenwerk | 28 |
| 2.8.2. | Aufgabe | 29 |
| 2.8.3. | Lösung: Daten importieren | 29 |
| 2.8.4. | Lösung: Daten aufbereiten | 29 |
| 2.8.5. | Lösung: Komplett | 30 |
| 2.8.6. | Fallstudie zur Lebenszufriedenheit | 30 |
| 2.9. | Literaturhinweise | 30 |
| | Literatur | 31 |

1. Willkommen!

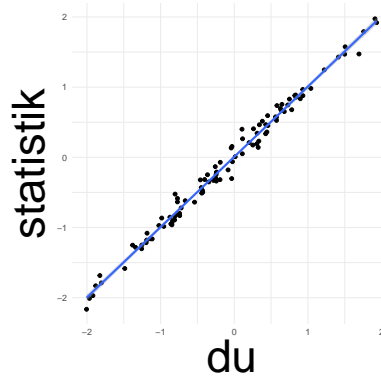


Abbildung 1.1.: Statistik und Du: Guter Fit!

1.1. Es geht um Ihren Lernerfolg

Meister Yoda rät: Lesen Sie die Hinweise (Abbildung 1.2).

Quelle: [Imgflip Memengenerator](#)

1.1.1. Lernziele

- Die Studentis sind mit wesentlichen Methoden der explorativen Datenanalyse vertraut und können diese selbständig anwenden.
- Die Studentis können gängige Forschungsfragen in lineare Modelle übersetzen, diese auf echte Datensätze anwenden und die Ergebnisse interpretieren.

Kurz gesagt: Das ist ein Grundkurs in Daten zähmen.

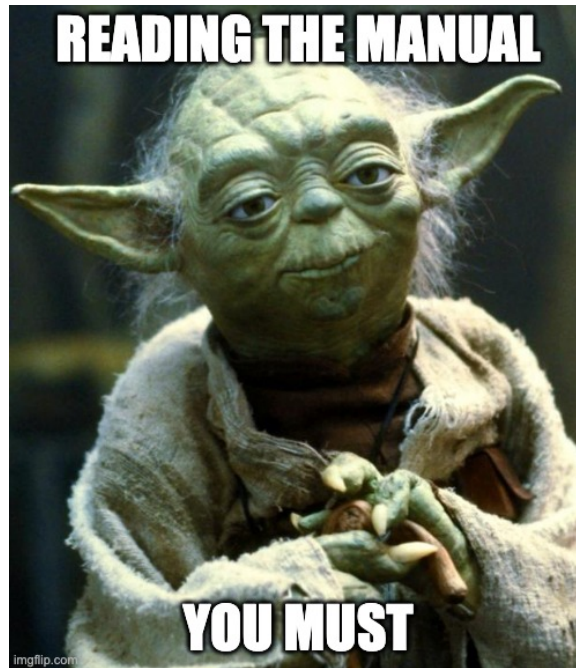


Abbildung 1.2.: Lesen Sie die folgenden Hinweise im eigenen Interesse

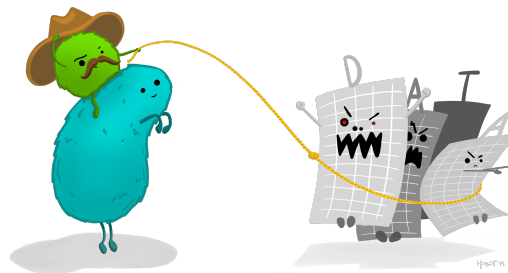


Abbildung 1.3.: Daten zähmen

Bildquelle: Allison Horst, CC-BY

1.1.2. Was lerne ich hier und wozu ist das gut?

Was lerne ich hier?

Sie lernen das *Handwerk der Datenanalyse* mit einem Schwerpunkt auf Vorhersage. Anders gesagt: Sie lernen, *Daten aufzubereiten* und aus Daten *Vorhersagen* abzuleiten. Zum Beispiel: Kommt ein Student zu Ihnen und sagt “Ich habe 42 Stunden für die Klausur gelernt, welche Note kann ich in der Klausur erwarten?”. Darauf Ihre Antwort: “Auf Basis meiner Daten und meines Modells müsstest du

eine 2.7 schreiben!”¹. Außerdem lernen Sie, wie man die Güte einer Vorhersage auf Stichhaltigkeit prüft. Denn Vorhersagen kann man ja in jeder Eckkneipe oder beim Wahrsager bekommen. Wir wollen aber belastbare Vorhersagen und zumindest wissen, wie gut die Vorhersagen (von jemanden) bisher waren.

Warum ist das wichtig?

Wir wollen nicht auf Leuten vertrauen, die behaupten, sie wüssten, was für uns richtig und gut ist. Wir wollen selber die Fakten prüfen können.

Wozu brauche ich das im Job?

Datenanalyse spielt bereits heute in vielen Berufen eine Rolle. Tendenz stark zunehmend.

Wozu brauche ich das im weiteren Studium?

In Forschungsarbeiten (wie in empirischen Forschungsprojekten, etwa in der Abschlussarbeit) ist es üblich, statistische Ergebnisse hinsichtlich quantitativ zu analysieren.

Ist Statistik nicht sehr abstrakt?

Der Schwerpunkt dieses Kurses liegt auf Anwenden und Tun; ähnlich dem Erlernen eines Handwerks. Theorien und Abstraktionen stehen nur am Rand.

Gibt es auch gute Jobs, wenn man sich mit Daten auskennt?

Das Forum (2020) berichtet zu den “Top 20 job roles in increasing and decreasing demand across industries” (S. 30, Abb. 22):

1. Data Analysts und Scientists
2. AI and Machine Learning Specialists
3. Big Data Specialists

1.1.3. Was ist hier das Erfolgsgeheimnis?

! Wichtig

Dran bleiben ist der Schlüssel zum Erfolg. Üben Sie regelmäßig. Geben Sie bei Schwierigkeiten nicht auf.



1.1.4. Motivieren Sie mich!

Schauen Sie sich das Video mit einer [Ansprache zur Motivation](#) an.²

¹Darauf dis Studenti: “Hpmf.”

²<https://youtu.be/jtNlzpcPr5Y>

1.1.5. Voraussetzungen

Um von diesem Kurs am besten zu profitieren, sollten Sie Folgendes mitbringen:

- Bereitschaft, Neues zu lernen
- Bereitschaft, nicht gleich aufzugeben
- Kenntnis grundlegender Methoden wissenschaftlichen Arbeitens

Was Sie *nicht* brauchen, sind besondere Mathe-Vorkenntnisse.

1.1.6. Überblick

Abb. Abbildung 1.4 gibt einen Überblick über den Verlauf und die Inhalte des Buches. Das Diagramm hilft Ihnen zu verorten, wo welches Thema im Gesamtzusammenhang steht.

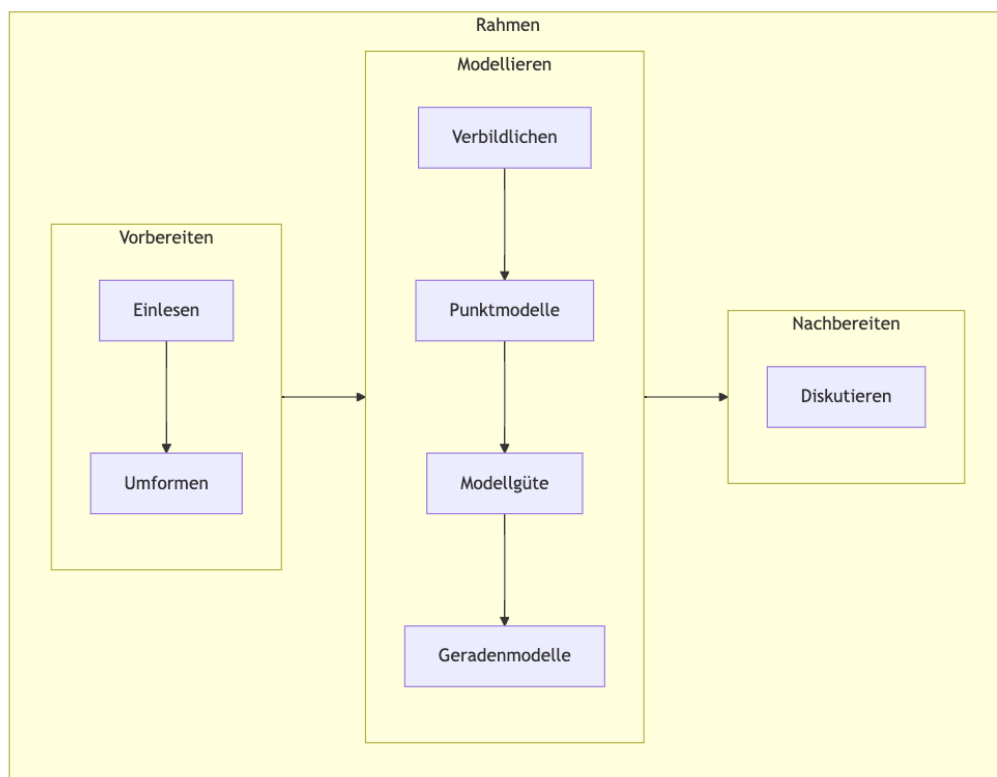


Abbildung 1.4.: Überblick über den Inhalt und Verlauf des Buches

Das Diagramm zeigt den Ablauf einer typischen Datenanalyse. Natürlich kann man sich auch andere sinnvolle Darstellungen dieses Ablaufs vorstellen.

1.2. Software

Sie benötigen R, RStudio und einige R-Pakete für diesen Kurs.

1.2.1. Installation

[Hier](#) finden Sie *Installationshinweise*.³

1.2.2. Viel R (?)

Dieses Buch enthält “mittel” viel R. Auf fortgeschrittene R-Techniken wurde aber komplett verzichtet. Dem einen oder der anderen Anfänger:in mag es dennoch “viel Code” erscheinen. Es wäre ja auch möglich gewesen, auf R zu verzichten und stattdessen eine “Klick-Software” zu verwenden. [JASP](#) oder [Jamovi](#) sind Beispiele für tolle Software aus dieser Kategorie. Ich glaube aber, der Verzicht auf eine Skriptsprache (R) wäre ein schlechter Dienst an den Studentis. Mit Blick auf eine “High-Tech-Zukunft” sollte man zumindest mit etwas Computer-Code vertraut sein. Auf Computercode zu verzichten erschiene mir daher fahrlässig für die “Zukunftsfestigkeit” der Ausbildung.

1.3. Zum Autor

Nähere Hinweise zum Autor dieses Buch, Sebastian Sauer, finden Sie [hier](#).⁴ Dort gibt es auch einen Überblick über [weitere Bücher des Autors zum Themenkreis Datenanalyse](#).⁵

1.4. Nomenklatur

1.4.1. Griechische Buchstaben

In diesem Buch werden ein paar (wenige) griechische Buchstaben verwendet, die in der Statistik üblich sind. Häufig werden *griechische* Buchstaben verwendet, um eine Grundgesamtheit (Population) zu beschreiben (die meistens unbekannt ist). Lateinische (“normale”) Buchstaben werden demgegenüber verwendet, um eine Stichprobe (Datensatz, vorliegende Daten) zu beschreiben. Tabelle 1.1 stellt diese Buchstaben zusammen mit ihrer Aussprache und Bedeutung vor.

Tabelle 1.1.: Griechische Buchstaben, die in diesem Buch verwendet werden.

| Zeichen | Aussprache | Buchstabe | Bedeutung in der Statistik |
|---------|------------|-----------|----------------------------|
| β | beta | b | Regressionskoeffizient |

³<https://hinweisbuch.netlify.app/hinweise-software>

⁴<https://sebastiansauer-academic.netlify.app/>

⁵<https://sebastiansauer-academic.netlify.app/#ebooks>

| Zeichen | Aussprache | Buchstabe | Bedeutung in der Statistik |
|----------|------------|-----------|----------------------------|
| μ | mü | m | Mittelwert |
| σ | sigma | s | Streuung |
| Σ | Sigma | S | Summenzeichen |
| ρ | rho | r | Korrelation (nach Pearson) |

Mehr griechische Buchstaben finden sich [z.B. in Wikipedia](#).⁶

1.5. Zitation

Bitte zitieren Sie dieses Buch wie folgt:

Sauer, S. (2024). *Statistik1*. <https://statistik1.netlify.app/>

Hier sind die maschinenlesbaren Zitationsinfos (Bibtex-Format), die Sie in Ihre Literatursoftware importieren können:

```
@book{sauer_statistik1,
  title = {Statistik1},
  rights = {CC-BY-NC},
  url = {https://statistik1.netlify.app/},
  author = {Sauer, Sebastian},
  date = {2024},
}
```

Hier ist die DOI:

[10.5281/zenodo.10082517](https://doi.org/10.5281/zenodo.10082517)

1.6. Reproduzierbarkeit

Die verwendeten R-Pakete sind mit [renv](#) dokumentiert.⁷

Der Quellcode ist [in diesem Github-Repo](#) dokumentiert.⁸

Dieses Dokument wurde erzeugt am/um: 2024-08-29 09:33:27.

⁶https://de.wikipedia.org/wiki/Griechisches_Alphabet

⁷<https://rstudio.github.io/renv/index.html>

⁸<https://github.com/sebastiansauer/statistik1>

Teil I.

Organisatorisches

Teil II.

Modellieren

2. Modellgüte

2.1. Lernsteuerung

2.1.1. Standort im Lernpfad

Abbildung 1.4 zeigt den Standort dieses Kapitels im Lernpfad und gibt damit einen Überblick über das Thema dieses Kapitels im Kontext aller Kapitel.

2.1.2. Lernziele

- Sie kennen gängige Maße der Streuung einer Stichprobe und können diese definieren und mit Beispielen erläutern.
- Sie können gängige Maße der Streuung einer Stichprobe mit R berechnen.
- Sie können die Bedeutung von Streuung für die Güte eines Modells erläutern.

2.1.3. Benötigte R-Pakete

In diesem Kapitel benötigen Sie folgende R-Pakete.

```
library(tidyverse)
library(easystats)
library(DataExplorer)
```

2.1.4. Benötigte Daten

```
mariokart_path <- paste0(
  "https://vincentarelbundock.github.io/Rdatasets/",
  "csv/openintro/mariokart.csv")

mariokart <- read.csv(mariokart_path)
```

2.1.5. Zum Einstieg

Übungsaufgabe 2.1 (Freiwillige vor!). Für diese kleine Live-Demonstration brauchen wir einige Freiwillige. Die Lehrkraft teilt die Freiwilligen in zwei Gruppen, Gruppe *Gleich-Groß* und Gruppe *Verschieden-Groß*. Erkennen Sie, dass die *Unterschiedlichkeit* der Größe in Gruppe *Gleich-Groß* gering ist, aber in Gruppe *Verschieden-Groß* hoch? ☐

2.2. Warum Sie die Streuung Ihrer Daten kennen sollten

2.2.1. Die Schlankheitspille von Prof. Weiss-Ois

Prof. Weiss-Ois hat eine Erfindung gemacht, eine Schlankheitspille ...

2.2.1.1. Was er sagt



Abbildung 2.1.: “Ich habe eine Schlankheitspille entwickelt, die pro Einnahme das Gewicht im Schnitt um 1kg reduziert!”

2.2.1.2. Was er NICHT sagt



Abbildung 2.2.: “Allerdings streuten die Werte der Gewichtsveränderung um 10kg um den Mittelwert herum.”

Icon unter Flaticon licence, Autor: iconixar

Würden Sie die Pille von Prof. I. Ch. Weiss-Ois nehmen?¹

- a) ja
- b) nein
- c) Nur wenn ich 100 Euro bekomme
- d) Okay, für 1000 Euro ☐

! Wichtig

Wie sehr die Werte eines Modells streuen, ist eine wichtige Information. ☐

2.2.2. Wie man seine Kuh über den Fluss bringt

Treffen sich zwei Bauern, Fritz Furchenzieher und Karla Kartoffelsack. Fritz will mit seiner Kuh einen Fluss überqueren, nur kann die Kuh nicht schwimmen².

👨(Fritz): Sag mal, Karla, ist der Fluss tief?

👩(Karla): Nö, im Schnitt nur einen Meter.

Also führt Fritz seine Kuh durch den Fluss, leider kam die Kuh nicht am anderen Ufer an, im Floß erstickte, s. Abbildung 2.3.

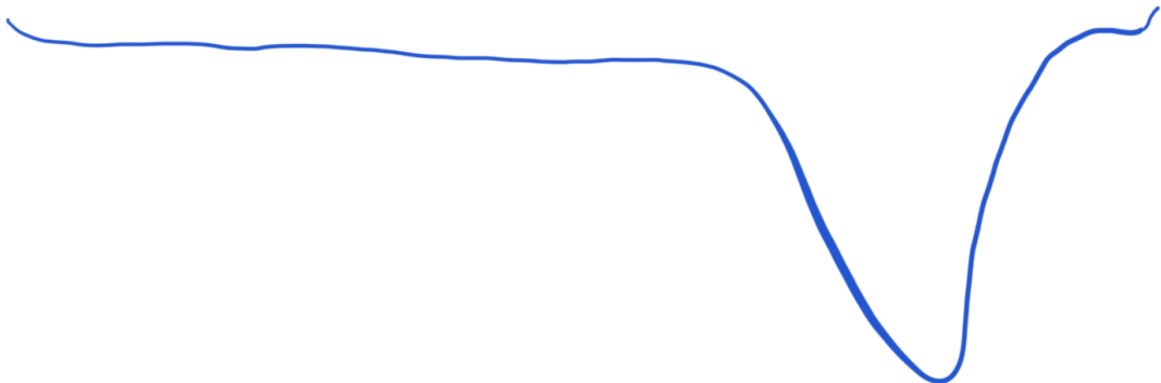


Abbildung 2.3.: Der Fluss ist im Schnitt nur einen Meter tief, trotzdem ist die Kuh erstickte.

👩(Karla): Übrigens, Lagemaße sagen nicht alles, Fritz.

👨: Lläuft die Kuh durch den Fluss, kann sie schwimmen oder 's ist Schluss.

! Wichtig

Die Streuung ihrer Daten zu kennen ist eine wesentliche Information. ☐

¹Ich auf keinen Fall.

²ob es Fritz kann, ist nicht überliefert.

2.3. Woran erkennt man ein gutes Modell?

Abbildung 2.4 zeigt ein einfaches Modell (Mittelwert) mit wenig Streuung (links) vs. ein einfaches Modell mit viel Streuung (rechts). Links ist die Streuung der Schlankheitsspielle *Dicktableitin* und rechts von der Schlankheitsspielle *Pfundafliptan* abgetragen.

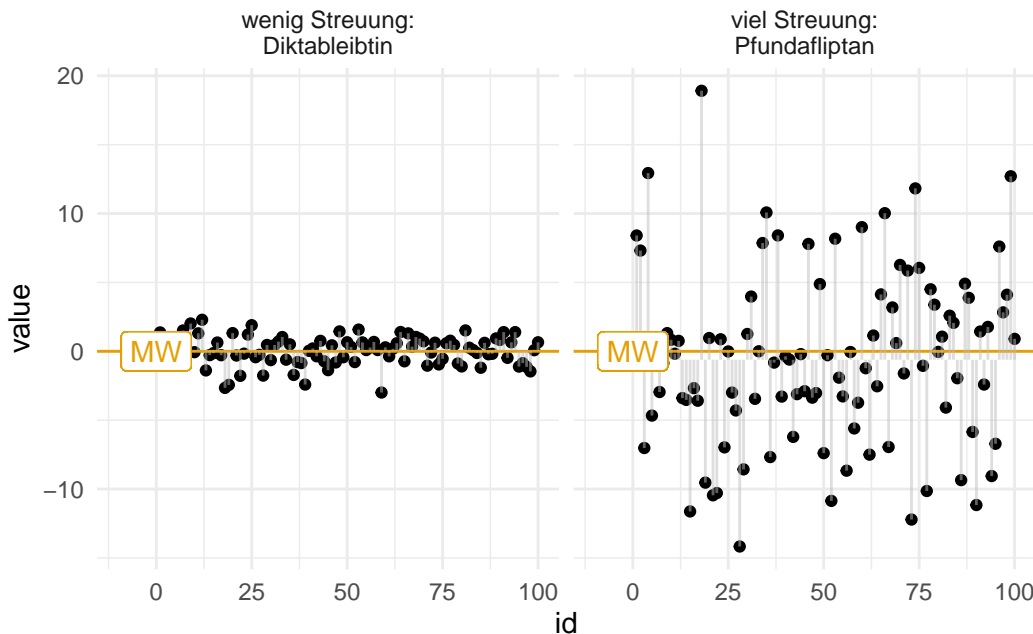


Abbildung 2.4.: Ein Modell mit wenig Streuung (links) vs. ein Modell mit viel Streuung (rechts). Die vertikalen grauen Balken kennzeichnen den (absoluten) Abstand von jeweils einem Datenpunkt zum Mittelwert (horizontale orange Linie). Je länger die ‘Abstandsbalken’, desto größer die Streuung.

Bei einem Modell mit *wenig* Streuung liegen die tatsächlichen, beobachteten Werte (y) nah an den Modellwerten (vorhergesagten Werten, \hat{y}); die Abweichungen $e = y - \hat{y}$ sind also gering (der Modellfehler ist klein). Bei einem Modell mit *viel* Streuung ist der Modellfehler e (im Vergleich dazu) groß.

Beispiel 2.1 (Daten zur Schlankheitskur von Prof. Weiss-Ois). In Abbildung 2.4 sind die Daten zu der Gewichtsveränderung nach Einnahme von “Schlankheitsspillen” zweier verschiedener Präparate. Wie man sieht unterscheidet sich die typische (vorhergesagte) Gewichtsveränderung zwischen den beiden Präparaten kaum. Die Streuung allerdings schon. Links sieht man die Gewichtsveränderungen nach Einnahme des Präparats “Dickableibtin extra mild” (c) und rechts das Präparat von Prof. Weiss-Ois “Pfundafliptan Forte”. Welches Präparat würden Sie lieber einnehmen?□

! Wichtig

Wir wollen ein präzises Modell, also kurze Fehlerbalken: Das Modell soll die Daten gut erklären, also wenig vom tatsächlichen Wert abweichen. Jedes Modell sollte Informationen über die

Präzision des Modellwerts bzw. der Modellwerte (Vorhersagen) angeben. Ein Modell ohne Angaben der Modellgüte, d.h. der Präzision der Schätzung des Modellwerts, ist wenig nütze. □

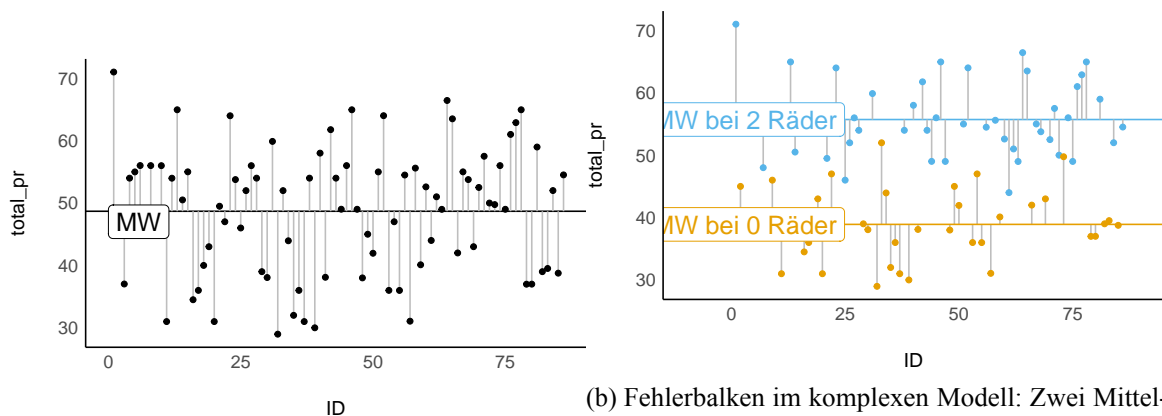
🤖 Ich frage mich, ob man so ein Modell nicht verbessern kann?

🤖 Die Frage ist, was wir mit “verbessern” meinen?

🤖 Naja, kürzere Fehlerbalken, ist doch klar!

Da die Anzahl der Lenkräder mit dem Verkaufsgebot zusammenhängt, könnte es vielleicht sein, dass wir die Lenkräder-Anzahl da irgendwie nutzen könnten. Das sollten wir ausprobieren.

Abbildung 2.5 zeigt, dass die Fehlerbalken *kürzer* werden, wenn wir ein (sinnvolles) komplexeres Modell finden. Innerhalb jeder der beiden Gruppen (mit 2 Lenkrädern vs. mit 0 Lenkrädern) sind die Fehlerbalken jeweils im Durchschnitt kürzer (rechtes Teildiagramm) als im Modell ohne Gruppierung (linkes Teildiagramm).³



(a) Fehlerbalken im einfachen Modell: Ein Mittelwert; viel Streuung insgesamt

(b) Fehlerbalken im komplexen Modell: Zwei Mittelwerte; weniger Streuung in jeder Gruppe. Das erkennt man daran, dass die vertikalen, grauen Abstandsbalken im Schnitt kürzer sind als im einfachen Modell (links)

Abbildung 2.5.: Fehlerbalken in einem einfachen und komplexeren Modell

! Wichtig

Durch sinnvolle, komplexere Modelle sinkt die Fehlerstreuung eines Modells. □

2.4. Streuungsmaße

Definition 2.1 (Streuungsmaße). Ein Streuungsmaß quantifiziert die Variabilität eines Merkmals. □

³Aus Gründen der Übersichtlichkeit wurden nur Autos mit Verkaufsgebot von weniger als 100 Euros berücksichtigt und nur Spiele mit 0 oder mit 2 Lenkrädern.

Ein einfaches Streuungsmaß ist der *Range*, definiert als Abstand von größtem und kleinsten Wert eines Merkmals. Dieses Mermals ist aber nicht robust (gegenüber Extremwerten) und sollte daher nur mit Einschränkung verwendet werden.

2.4.1. Der mittlere Abweichungsbalken

👉 Wir müssen jetzt mal präziser werden! Wie können wir die Streuung berechnen?

👉 Gute Frage! Am einfachsten ist es, wenn wir die mittlere Länge eines Abweichungsbalkens ausrechnen.

Legen wir (gedanklich) alle Abweichungsbalken e aneinander und teilen durch die Anzahl n der Balken, so erhalten wir den “mittleren Abweichungsbalken”, den wir mit \bar{e} bezeichnen könnten. Diesen Kennwert bezeichnet man als *Mean Absolute Error* (MAE) bzw. als *Mittlere Absolutabweichung* (MAA). Er ist so definiert, s. Gleichung 2.1.

$$MAE = \frac{\sum_{i=1}^n |y_i - \bar{y}|}{n} = \frac{\sum_{i=1}^n |e_i|}{n}. \quad (2.1)$$

Definition 2.2 (Mittlere Absolutabweichung). Die Mittlere Absolutabweichung (MAA, MAE) ist definiert als die Summe der Absolutwerte der Differenzen eines Messwerts zum Mittelwert, geteilt durch die Anzahl der Messwerte.⁴□

Beispiel 2.2. Abbildung 2.6 visualisiert ein einfaches Beispiel zum MAE. Rechnen wir den MAE für das Beispiel von Abbildung 2.6 aus:

$$MAE = \frac{1+|-3|+1+1}{4} = 6/4 = 1.5$$

Natürlich können wir R auch die Rechenarbeit überlassen.

👉 Loving it!!

Schauen Sie: Den Mittelwert (s. Abbildung 2.6) kann man doch mit Fug und Recht als ein *lineares Modell*, eine Gerade, betrachten, oder nicht? Schließlich erklären wir y anhand einer Gerade (die parallel zur X-Achse ist).

In R gibt es einen Befehl für ein *lineares Modell*, er heißt `lm`.

Die Syntax von `lm()` lautet:

```
lm(y ~ 1, data = meine_daten).
```

In Worten:

Hey R, berechne mit ein lineares Modell zur Erklärung von Y. Aber verwende keine andere Variable zur Erklärung von Y, sondern nimm den Mittelwert von Y.

⁴Wenn man solche Sätze liest, fühlt sich die Formel fast einfacher an.

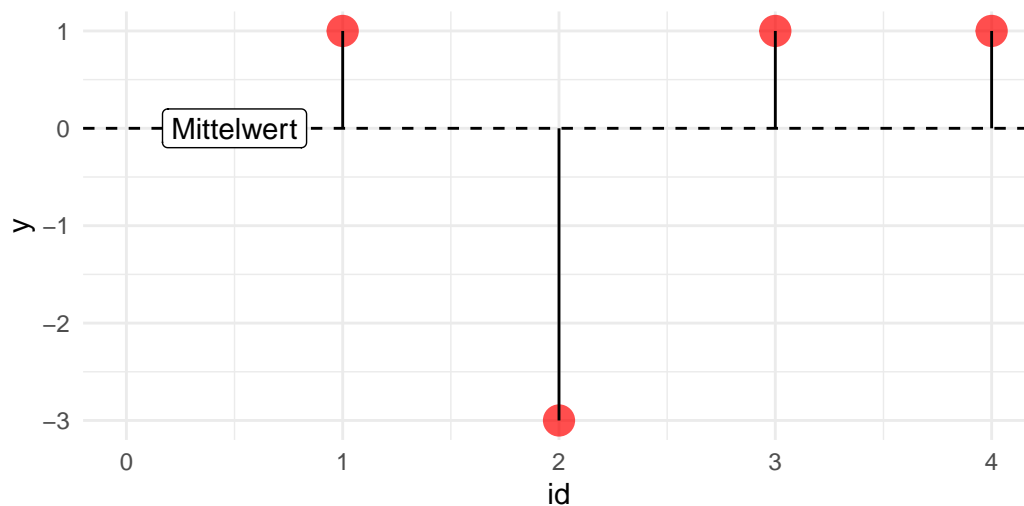


Abbildung 2.6.: Abweichungsbalken und der MAE

```
lm1 <- lm(y ~ 1, data = d)
```

Den MAE können wir uns jetzt so ausgeben lassen:

```
mae(lm1)
## [1] 1.5
```

2.4.2. Der Interquartilsabstand

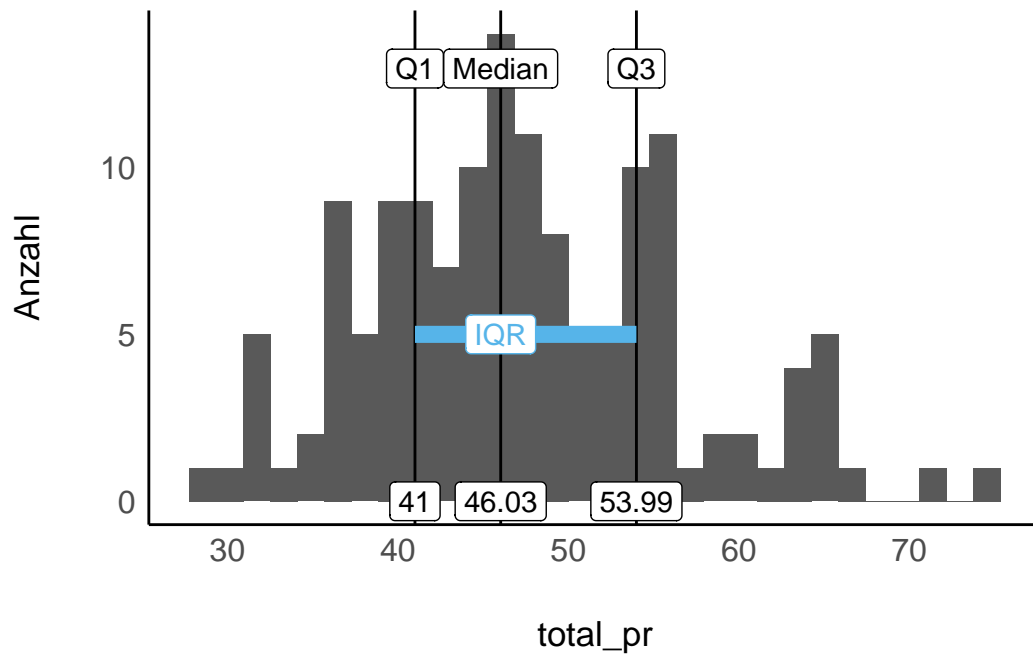
Der Interquartilsabstand (IQA; engl. inter quartile range, IQR) ist ein Streuungsmaß, das nicht auf dem Mittelwert aufbaut. Der IQR ist robuster als z.B. der MAA oder die Varianz und die Standardabweichung.

Definition 2.3 (Interquartilsabstand). Der Interquartilsabstand ist definiert als der die (absolute) Differenz vom 3. Quartil und 1. Quartil. \square

Beispiel 2.3 (IQR im Hörsaal). In einem Statistikkurs betragen die Quartile der Körpergröße: Q1: 1.65m, Q2 (Median): 1.70m, Q3: 1.75m. Der IQR beträgt dann: $IQR = Q3 - Q1 = 1.75m - 1.65m = 0.10m$, d.h. 10 cm. \square

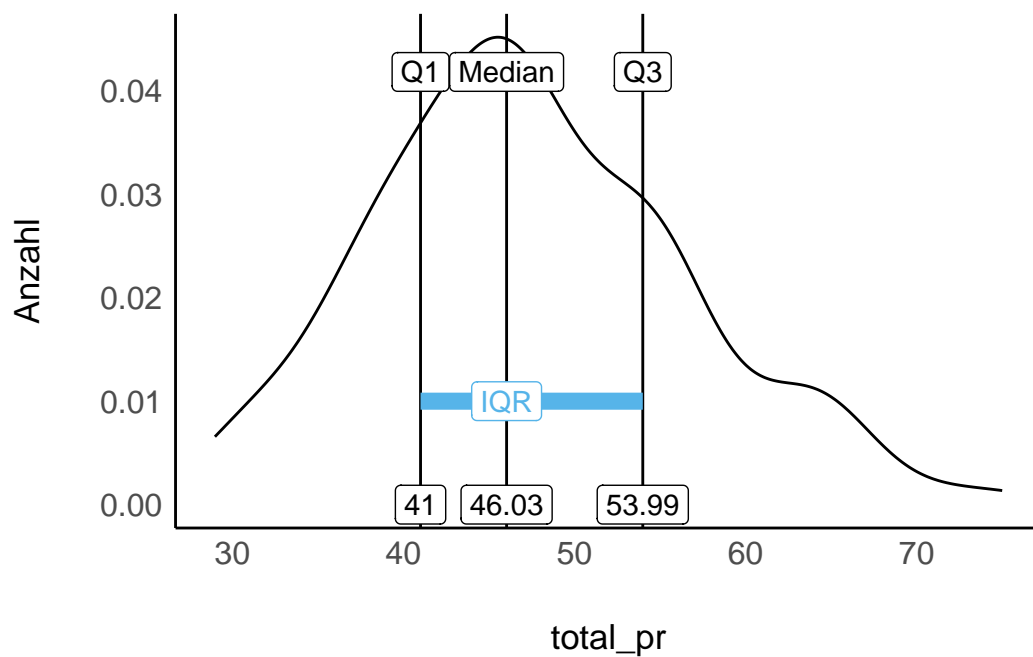
Abbildung 2.7 stellt den IQR (und einige Quantile) für den Verkaufspreise von Mariokart-Spielen dar.

2.4.3. Histogramm



(a) IQR, Q1, Q2 und Q3 für das Schlussgebot (nur Spiele für weniger als 100 Euro)

2.4.4. Dichtediagramm



(b) IQR, Q1, Q2 und Q3 für das Schlussgebot (nur Spiele für weniger als 100 Euro)

Abbildung 2.7.: Der IQR für den Verkaufspreis von Mariokart-Spielen.

2.4.5. Streuungsmaße für Normalverteilungen


Normalverteilungen sind recht häufig anzutreffen in der Praxis der Datenanalyse. Daher lohnt es sich, zu überlegen, wie man diese Verteilungen gut zusammenfasst. Man kann zeigen, dass eine Normalverteilung sich komplett über ihren *Mittelwert* sowie ihre *Standardabweichung* beschreiben lässt. Außerdem gilt: Sind Ihre Daten normalverteilt, dann sind die Abweichungen vom Mittelwert auch normalverteilt. Denn wenn man eine Konstante zu einer Verteilung addiert (bzw. subtrahiert), “verschiebt man den Berg” ja nur zur Seite, ohne seine Form zu verändern, s. Abbildung 2.12.

Hinweis

Hat man normalverteilte Variablen/Abweichungen/Residuen, so ist die *Standardabweichung* (engl. standard deviation, SD, σ , s) eine komfortable Maßeinheit der Streuung, denn damit lässt sich die Streuung (Abweichung vom Mittelwert, Residuen) der Normalverteilung gut beschreiben.□

 Aber wie berechnet man jetzt diese Standardabweichung?

 Moment, noch ein kurzer Exkurs zur Varianz ...

 (seufzt)

2.4.6. Varianz

2.4.6.1. Intuition

Hinweis

Die Varianz einer Variable (z.B. Verkaufspreis von Mariokart) ist, grob gesagt, der typische Abstand eines Verkaufspreis vom mittleren Verkaufspreis.□

Abbildung 2.10 illustriert die Varianz:

1. Man gehe von der Häufigkeitsverteilung der Daten aus.
2. Betrachtet man die Daten als Gewichte auf einer Wippe, so ist der Schwerpunkt der Wippe der Mittelwert.
3. Man bilde Quadrate für jeden Datenpunkt mit der Kantenlänge, die dem Abstand des Punktes zum Mittelwert entspricht.
4. Die Quadrate quetscht man jetzt wo nötig in rechteckige Formen (ohne dass sich die Fläche ändern darf) und verschiebt sie, bis sich alle Formen zu einem Rechteck mit Seitenlänge n und σ^2 anordnen.

By Cmglee - Own work, CC BY-SA 3.0

Abbildung 2.9 visualisiert die Varianz für Beispiel 2.2.⁵

⁵Die Abweichungsquadrate wirken optisch nicht quadratisch, da die X-Achse breiter skaliert dargestellt ist als die Y-Achse. Trotzdem sind es Quadrate, nur nicht optisch, wenn Sie wissen, was ich meine...

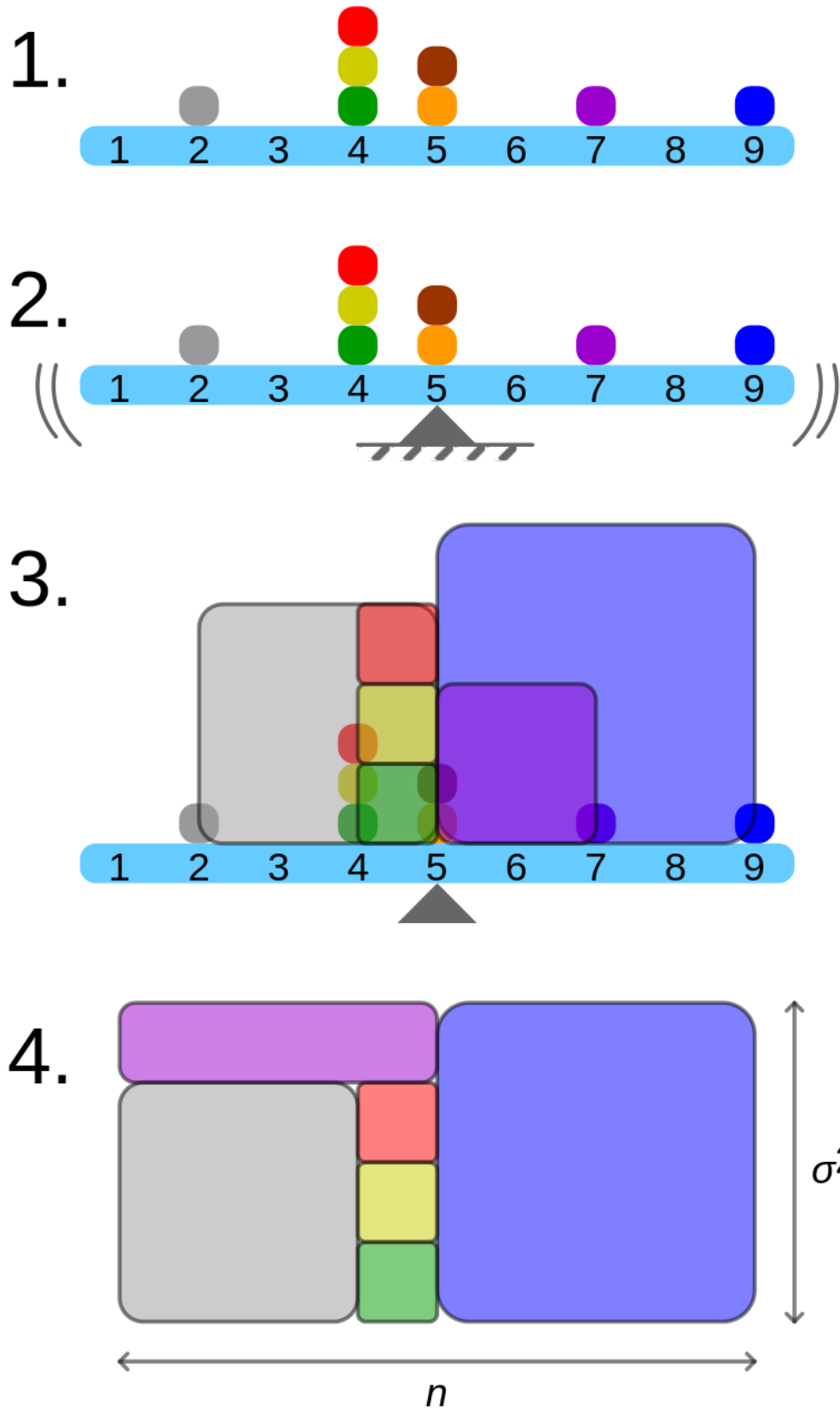


Abbildung 2.8.: Illustration zur Varianz als “mittlerer Quadratfehler”

Links sind die *Abweichungsquadrate* dargestellt, rechts die Varianz als “*typisches Abweichungsquadrat*”.

i Hinweis

Die Varianz ist also ein Maß, das die typische Abweichung der Beobachtungen vom Mittelwert in eine Zahl fasst. □

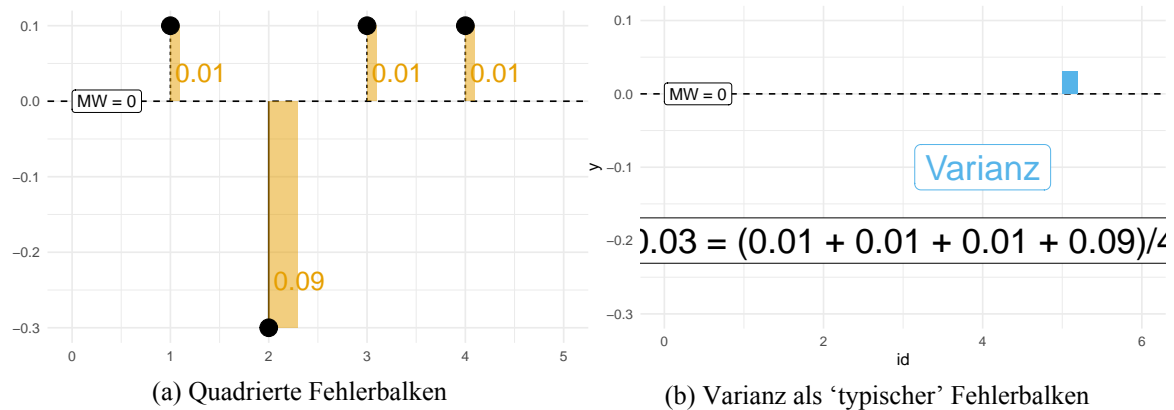


Abbildung 2.9.: Sinnbild zur Varianz als typischer Fehlerbalken

Beispiel 2.4. Sie arbeiten immer noch bei einem Online-Auktionshaus und untersuchen den Verkauf von Videospielen. Natürlich mit dem Ziel, dass Ihre Firma mehr von dem Zeug verkaufen kann.

Dazu berechnen Sie die Streuung in den Verkaufspreisen, s. Listing 2.1. □

Listing 2.1 Berechnung der Streuung des Verkaufspreises als Indikatoren für die Modellgüte des Mittelwerts.

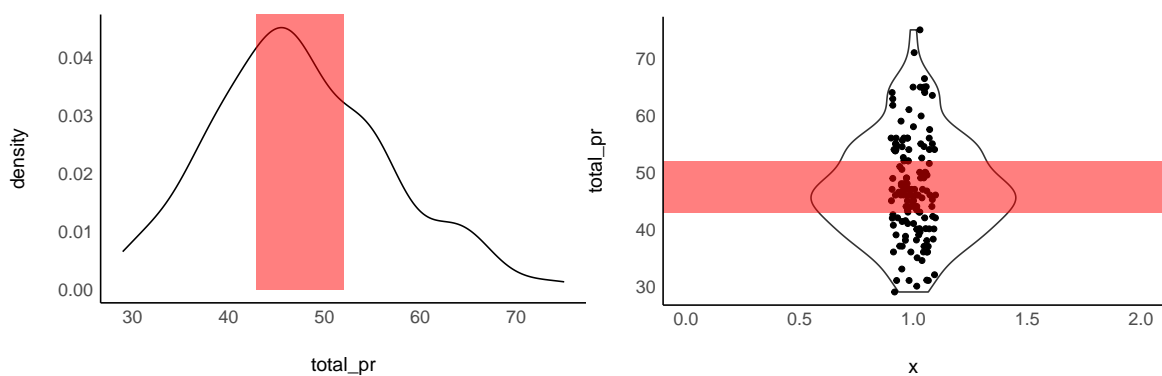
```
mariokart_no_extreme <-
  mariokart %>%
  filter(total_pr < 100) # ohne Extremwerte

m_summ <-
  mariokart_no_extreme %>%
  summarise(
    pr_mw = mean(total_pr),
    pr_iqr = IQR(total_pr),
    pr_maa = mean(abs(total_pr - mean(total_pr))),
    pr_var = var(total_pr),
    pr_sd = sd(total_pr))
```

| pr_mw | pr_iqr | pr_maa | pr_var | pr_sd |
|-------|--------|--------|--------|-------|
| 47.43 | 12.99 | 7.20 | 83.06 | 9.11 |

Statistiken sind ja schön ... aber Bilder sind auch gut, s. Abbildung 2.10. Datendiagramme eignen sich gut, um (grob) die Streuung einer Variable zu erfassen.

```
mariokart %>%
  mariokart %>%
  select(total_pr) %>%
  filter(total_pr < 100) %>% # ohne Extremwerte
  plot_density()
```



(a) Dichtediagramm mit MW±SD in roter Farbe

(b) Violindiagramm mit MW±SD in roter Farbe

Abbildung 2.10.: Die Verteilung des Verkaufspreises von Mariokart-Spielen

Wer sich die Berechnung von Hand für `pr_maa` sparen möchte (s. Listing 2.1), kann die [Funktion MeanAD aus dem Paket DescTools](#) nutzen.

2.4.6.2. Kochrezept für die Varianz

Um die Standardabweichung zu berechnen, berechnet man zunächst die *Varianz*, s^2 abgekürzt. Hier ist ein “Kochrezept”⁶ zur Berechnung der Varianz:

1. Für alle Datenpunkte x_i : Berechne die Abweichungen vom Mittelwert, \bar{x}
2. Quadriere diese Werte
3. Summiere dann auf
4. Teile durch die Anzahl N der Werte

⁶Algorithmus

Als Formel ausgedrückt, lautet die Definition der Varianz⁷ einer Stichprobe wie folgt, s. Gleichung 2.2.

$$s^2 = \frac{1}{N} \sum_{i=1}^n (y_i - \bar{y})^2 = \frac{1}{N} \sum_{i=1}^n e_i^2. \quad (2.2)$$

Definition 2.4 (Varianz). Die Varianz (s^2, σ^2) ist definiert als der Mittelwert der quadrierten Abweichungen, e_i^2 , (vom Mittelwert).□

Die Varianz steht im engen Verhältnis zur Kovarianz, s. [?@sec-cov](#). Die Varianz kann auch verstanden als den *mittleren Quadratfehler* (Mean Squared Error, MSE) eines Modells, s. Gleichung 2.3.

$$MSE = \frac{1}{N} \sum_{i=1}^N (y_i - \hat{y})^2. \quad (2.3)$$

Im Fall eines Punktmodells ist der Mittelwert der vorhergesagte Wert eines Modells.

2.4.7. Die Standardabweichung

Kennt man die Varianz, so lässt sich die Standardabweichung einfach als Quadratwurzel der Varianz berechnen.

Definition 2.5 (Standardabweichung). Die Standardabweichung (SD, s, σ) ist definiert als die Quadratwurzel der Varianz, s. Gleichung 2.4.

$$s := \sqrt{s^2} \quad (2.4)$$

Durch das Wurzelziehen besitzt die Standardabweichung wieder *in etwa* die gleiche Größenordnung wie die Daten (im Gegensatz zur Varianz, die durch das Quadrieren sehr groß werden kann).

Aus einem Modellierungsblickwinkel kann man die SD definieren als die Wurzel von MSE. Dann nennt man sie *Root Mean Squared Error* (RMSE): $RMSE := \sqrt{MSE}$.

i Hinweis

Die SD ist i.d.R. *ungleich* zur MAE, aber (fast) gleich zur RMSE. Entsprechend ist die Varianz (fast) gleich zur MSE.□

Beispiel 2.5. Sie arbeiten weiter an Ihrem Mariokart-Projekt. Da Sie heute keine Lust auf viel Tippen haben, nutzen Sie das R-Paket `easystats` mit der Funktion `describe_distribution`.

⁷sog. unkorrigierte Stichprobenvarianz; um anhand einer Stichprobe die Varianz der zugehörigen Population zu schätzen, teilt man nicht durch N , sondern durch $N - 1$


```
library(easystats)

mariokart %>%
  select(total_pr) %>%
  describe_distribution()
```

| Variable | Mean | SD | IQR | Min | Max | Skew- ness | Kurtosis | n | n_Miss- ing |
|----------|----------|----------|-------|-------|--------|---------------|----------|-----|----------------|
| total_pr | 49.88049 | 25.68856 | 12.99 | 28.98 | 326.51 | 9.035897 | 96.14414 | 143 | 0 |

Ah! Das war einfach. Wird auch langsam Zeit für Feierabend.□

Beispiel 2.6. Ihr Job als Datenanalyst ist anstrengend, aber auch mitunter interessant. So auch heute. Bevor Sie nach Hause gehen, möchten Sie noch eine Sache anschauen. In einer früheren Analyse (s. Abbildung 2.5) fanden Sie heraus, dass die Fehlerbalken kürzer werden, wenn man ein geschickteres und komplexeres Modell findet. Das wollen Sie natürlich prüfen. Sie überlegen: “Okay, ich will ein einfaches Modell, in dem der Mittelwert das Modell des Verkaufspreis sein soll.”

Das spezifizieren Sie so:

```
lm1 <- lm(total_pr ~ 1, data = mariokart)
mae(lm1)
## [1] 10.01811
```

Im nächsten Schritt spezifizieren Sie ein Modell, in dem der Verkaufspreis eine Funktion der Anzahl der Lenkräder ist (ähnlich wie in Abbildung 2.5):

```
lm2 <- lm(total_pr ~ wheels, data = mariokart)
mae(lm2)
## [1] 7.375873
```

Ah! Sehr schön, Sie haben mit `lm2` ein besseres Modell als einfach nur den Mittelwert gefunden. Ab nach hause!□

2.5. Streuung als Modellfehler

Wenn wir den Mittelwert als Punktmodell des Verkaufspreises auffassen, so kann man die verschiedenen Kennwerte der Streuung als verschiedene Kennwerte der Modellgüte auffassen.

Definieren wir zunächst als Punktmodell auf Errisch:

```
lm_mariol <- lm(total_pr ~ 1, data = mariokart)
```

Zur Erinnerung: Wir modellieren `total_pr` ohne Prädiktoren, sondern als Punktmodell, und zwar schätzen wir den Mittelwert mit den Daten `mariokart`.

Das (Meta-)Paket `easystats` bietet komfortable Befehle, um die Modellgüte zu berechnen:

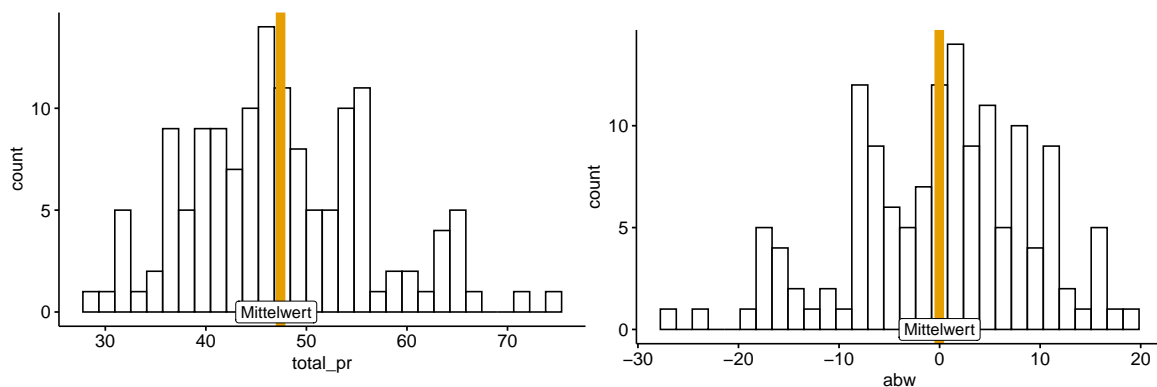
```
mae(lm_mariol) # Mean absolute error
## [1] 10.01811
mse(lm_mariol) # Mean squared error
## [1] 655.2874
rmse(lm_mariol) # Root mean squared error
## [1] 25.59858
```

2.6. z-Transformation

Sie arbeiten immer noch als Datenknecht, Moment, *Datenhecht* bei dem Online-Auktionshaus. Heute untersuchen Sie die Frage, wie gut sich die Verkaufspreise mit einer einzigen Zahl, dem mittleren Verkaufspreis, beschreiben lassen. Einige widerspenstige Werte haben Sie dabei einfach des Datensatzes verwiesen. Schon ist das Leben leichter, s. `mariokart_no_extreme`.

```
mariokart_no_extreme <-
  mariokart %>%
  filter(total_pr < 100)
```

Abbildung 2.11 (links) zeigt, dass es einige Streuung um den Mittelwert herum gibt. Abbildung 2.11 (rechts) zeigt die (um den Mittelwert) *zentrierten* Daten.



(a) Wie nah drängen sich die Verkaufspreise um ihren Mittelwert? (b) Abweichungen vom Mittelwert: zentrierte Daten

Abbildung 2.11.: Verteilung von `mariokart_no_extreme`

Tja, das ist doch etwas Streuung um den Mittelwert herum.

! Wichtig

Je weniger Streuung um den Mittelwert (ca. 47 Euro) herum, desto besser eignet sich der Mittelwert als Modell für die Daten, bzw. desto höher die Modellgüte. □

Ja, es ist *etwas* Streuung, aber wie viel? Kann man das genau angeben? Sie überlegen ... und überlegen. Da! Eine Idee!

Man könnte vielleicht angeben, wie viel Euro jedes Spiel vom Mittelwert entfernt ist. Je größer diese Abweichung, desto schlechter die Modellgüte! Also rechnen Sie diese Abweichung aus.

```
mariokart_no_extreme <-  
  mariokart_no_extreme %>%  
  mutate(abw = 47.4 - total_pr)
```

Anders gesagt: Wir haben die Verkaufspreise *zentriert*.

Definition 2.6 (Zentrieren). Zentrieren bedeutet, von jedem Wert einer Verteilung X den Mittelwert abzuziehen. Daher ist der neue Mittelwert (der zentrierten Verteilung) gleich Null. □

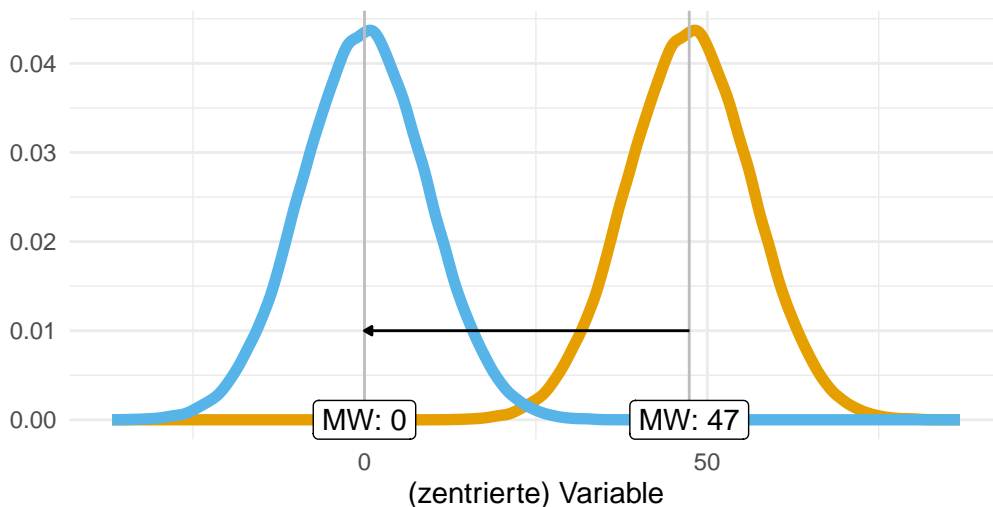


Abbildung 2.12.: Die Abweichungen zum Mittelwert (MW) einer normalverteilten Variable sind selber normalverteilt

Aber irgendwie sind Sie noch nicht am Ziel Ihrer Überlegungen: Woher weiß man, ob 10 Euro oder 20 Euro “viel” Abweichung vom Verkaufspreis ist? Man müsste die Abweichung eines Verkaufspreis zu irgendetwas in Bezug setzen. Wieder! Ein Geistesblitz! Man könnte doch die jeweilige Abweichung in Bezug setzen zur *mittleren (absoluten) Abweichung* (MAA)! Ein alternativer, ähnlicher Kennwert zur mittlerer absolute Abweichung ist die SD. Sie haben gehört, dass die SD gebräuchlicher ist als die

MAA. Um sich als Checker zu präsentieren, berechnen Sie also auch die SD; die beiden Koeffizienten sind ja ähnlich.

Also: Wenn ein Spiel 10 Euro vom Mittelwert abweicht und die SD 10 Euro betragen sollte, dann hätten wir eine “standardisierte”⁸ Abweichung von 1, weil $10/10=1$.

Begeistert über Ihre Schlaueit machen Sie sich ans Werk.

```
mariokart_no_extreme <-  
  mariokart_no_extreme %>%  
  mutate(abw_std = abw / sd(abw), # std wie "standardisiert"  
         abw_std2 = abw / mean(abs(abw)))
```

Zufrieden betrachten Sie Ihr Werk, s. Abbildung 2.13. In Abbildung 2.13 sieht man oben die Rohwerte und unten die transformierten Werte, die wir hier als *standardisiert* bezeichnen, da wir sie in Bezug zur “typischen Abweichung”, der SD, gesetzt haben.

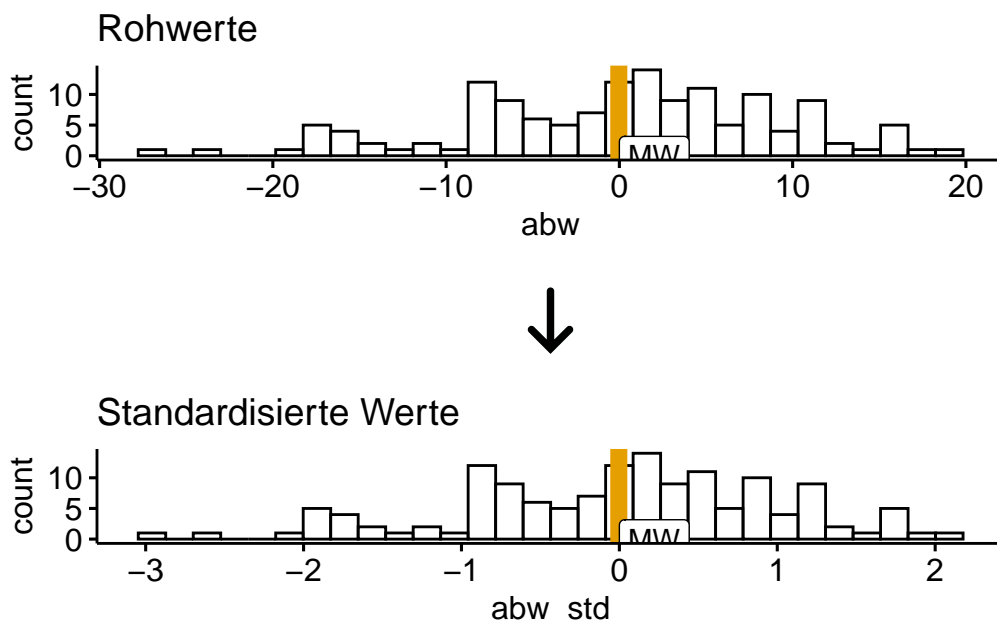


Abbildung 2.13.: Standardisierung von Abweichungswerten bzw. einer Verteilung; der vertikale Balken zeigt den Mittelwert

Wir fassen die Schritte unserer Umrechnung (“Transformation”) zusammen wie in einem Kochrezept:

1. Nimm die Verteilung der Verkaufspreise
2. Berechne die Abweichungen vom mittleren Verkaufspreis (Differenz Mittelwert und jeweiliger Verkaufspreis)
3. Teile die Abweichungen (Schritt 2) durch die SD

⁸abgekürzt manchmal mit *std*

Diese Art von Transformation bezeichnet man als *z-Transformation* und die resultierenden Werte als *z-Werte*.

Definition 2.7 (z-Werte). z-Werte sind das Resultat der z-Transformation. Für die Variable X berechnet sich der z-Wert der i -ten Beobachtung so: $z_i = \frac{x_i - \bar{x}}{sd_x}$. \square

z-Werte sind nützlich, weil sie die “relative” Abweichung einzelner Beobachtungen vom Mittelwert anzeigen.

Nach einer *Faustregel* spricht man von extremen Abweichungen (Extremwerten, Ausreißern), wenn $z_i > 2$ oder $z_i > 3$.

2.7. Fazit

Der „gesunde Menschenverstand“ würde spontan den mittleren Absolutabstand (MAA oder MAE) der Varianz (oder der Standardabweichung, SD) vorziehen. Das ist vernünftig, denn die MAA ist anschaulicher und damit nützlicher als die Varianz und die SD.

Warum sollte man überhaupt ein unanschauliches Maß wie die Varianz verwenden? Wenn es nur um deskriptive Statistik geht, braucht man die Varianz (oder die SD) nicht unbedingt. Gründe, warum Sie die Varianz (bzw. SD) kennen und nutzen sollten, sind:⁹

- Die SD ist sehr nützlich zur Beschreibung der Normalverteilung
- Die Varianz wird häufig verwendet bzw. in Forschungsarbeiten berichtet, also müssen Sie die Varianz kennen.

Liegen Extremwerte vor, kann es vorteilhafter sein, den IQR vorzuziehen gegenüber Mittelwert basierten Streuungsmaßen (MAA, Varianz, SD).

2.8. Aufgaben

2.8.1. Datenwerk

Die Webseite datenwerk.netlify.app stellt eine Reihe von einschlägigen Übungsaufgaben bereit. Sie können die Suchfunktion der Webseite nutzen, um die Aufgaben mit den folgenden Namen zu suchen:

- [mariokart-sd2](#)
- [mariokart-sd3](#)
- [Kennwert-robust](#)
- [summarise04](#)
- [summarise05](#)

⁹Ich wollte noch hinzufügen, dass die Varianz eng verknüpft mit der linearen Algebra, aber ich war nicht sicher, ob das Argument allgemein überzeugen würde.

- [vis-mariokart-variab](#)
- [sd-vergleich](#)
- [nasa01](#)
- [Streuung-Histogramm](#)
- [mariokart-sd1](#)
- [summarise06](#)
- [mariokart-desk01](#)

Übungsaufgabe 2.2 (Analysieren Sie den Datensatz zur Handynutzung).

2.8.2. Aufgabe

Sind Sie händysüchtig? Das ist die Forschungsfrage [dieser Umfrage](#). Nehmen Sie ggf. an dieser Umfrage teil (sie ist anonym und dauert drei Minuten). Laden Sie den [Datensatz zur Handynutzung](#) von Google-Docs herunter.¹⁰ Berechnen Sie dann gängige deskriptive Statistiken und visualisieren Sie sie. □

2.8.3. Lösung: Daten importieren

Sie können die Daten entweder selber herunterladen oder aber die folgende Version des Datensatzes verwenden. In beiden Fällen ist es nützlich, den (absoluten oder relativen) Pfad anzugeben:

```
data_path <-  
  ↪ "https://raw.githubusercontent.com/sebastiansauer/statistik1/main/daten"
```

Dann können Sie die Daten wie gewohnt importieren:

```
smartphone_raw <- read.csv(data_path)
```

2.8.4. Lösung: Daten aufbereiten

Die Spaltennamen sind sehr unschön. Lassen Sie uns daher die Spaltennamen umbenennen (aber vorab sichern):

```
item_labels <- names(smartphone_raw)  
names(smartphone_raw) <- paste0("item", 1:ncol(smartphone_raw))
```

Check:

¹⁰https://docs.google.com/spreadsheets/d/1SWMj4rIIIJdAsfsSKQHSg8jHr_OuKLpJx_0XV4LGnH0/edit?usp=sharing

```
glimpse(smartphone_raw)
## Rows: 70
## Columns: 18
## $ item1 <chr> "21/03/2024 15:36:52", "05/04/2024 10:24:58~
## $ item2 <chr> "15:31:00", "10:23:00", "10:40:00", "11:14:~
## $ item3 <int> 3, 4, 3, 3, 5, 5, 5, 5, 1, 2, 5, 3, 2, 2, 2~
## $ item4 <int> 5, 3, 3, 3, 4, 3, 3, 6, 2, 4, 5, 1, 1, 2, 3~
## $ item5 <int> 3, 3, 1, 5, 1, 3, 2, 4, 3, 2, 1, 1, 1, 4, 1~
## $ item6 <int> 4, 2, 4, 3, 5, 4, 6, 3, 2, 5, 6, 4, 2, 6, 5~
## $ item7 <int> 4, 3, 2, 3, 3, 1, 3, 2, 1, 2, 1, 1, 1, 3, 2~
## $ item8 <int> 1, 3, 1, 2, 3, 1, 1, 2, 2, 2, 1, 1, 2, 4, 1~
## $ item9 <int> 2, 6, 1, 3, 6, 5, 5, 2, 2, 5, 6, 1, 1, 5, 4~
## $ item10 <int> 2, 5, 5, 3, 4, 3, 1, 5, 1, 5, 3, 4, 3, 5, 4~
## $ item11 <int> 5, 6, 6, 5, 6, 6, 5, 6, 4, 3, 6, 4, 4, 5, 3~
## $ item12 <int> 1, 3, 1, 2, 5, 2, 4, 2, 1, 1, 3, 1, 1, 1, 1~
## $ item13 <int> 4, 3, 4, 2, 4, 2, 5, 3, 1, 1, 4, 1, 3, 4, 1~
## $ item14 <chr> "", "", "", "", "", "", "", "", "", "", "", ~
## $ item15 <chr> "", "", "", "", "", "", "", "", "", "", "", ~
## $ item16 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
## $ item17 <chr> "", "", "", "", "", "", "", "", "", "", "", ~
## $ item18 <int> NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, NA, ~
```

2.8.5. Lösung: Komplet

□

2.8.6. Fallstudie zur Lebenszufriedenheit

Die OECD führt eine [weltweite Studie zur Lebenszufriedenheit](#) durch.¹¹

Arbeiten Sie die die [Fallstudie “OECD Wellbeing”](#) durch, um ein tieferes Verständnis für die Lebenszufriedenheit in verschiedenen Ländern der Welt zu bekommen.

2.9. Literaturhinweise

Allen Downey (2023) stellt in seinem vergnüglich zu lesenden Buch eine kurzweilige Einführung in die Statistik vor; auch Streuungsmaße haben dabei einen Auftritt. Wer mehr “Lehrbuch-Feeling” sucht, wird bei ([cetinkaya-rundel_introduction_2021-1?](#)) fündig (das Buch ist online frei verfügbar). Es ist kein Geheimnis, dass Streuungsmaße keine ganz neuen Themen in der Statistik sind. Aber hey, Oldie is Goldie, ohne Streuungsmaße geht’s nicht. Jedenfalls werden Sie in jedem Statistik-Lehrbuch, dass Sie in der Bib (oder sonstwo) aus dem Regal ziehen, fündig werden zu diesem Thema. Die

¹¹<https://www.oecd.org/wise/measuring-well-being-and-progress.htm>

Bücher unterscheiden sich meist “nur” in ihrem Anspruch bzw. der didaktischen Aufmachung; für alle ist da was dabei.

Literatur

Downey, Allen. 2023. *Probably Overthinking It: How to Use Data to Answer Questions, Avoid Statistical Traps, and Make Better Decisions*. Chicago ; London: The University of Chicago Press.

Forum, World Economic. 2020. „The Future of Jobs Report 2020“. CH-1223 Cologny/Geneva Switzerland: World Economic Forum. https://www3.weforum.org/docs/WEF_Future_of_Jobs_2020.pdf.