

Statistik1

Modellgüte

Sebastian Sauer

Einstieg

Lernziele

- Sie kennen gängige Maße der Streuung einer Stichprobe und können diese definieren und anhand von 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.

Die Schlankheitspille von Prof. Weiss-Ois



Was er sagt: “Meine Pille reduziert das Gewicht im Schnitt um 1 kg!”

Was er NICHT sagt: Die Werte streuten um 10 kg um den Mittelwert.

Wie sehr die Werte eines Modells streuen, ist eine wichtige Information — bei der Pille kann es sein, dass Sie 10 kg *zunehmen*.

Wie man seine Kuh über den Fluss bringt

Fritz will mit seiner Kuh einen Fluss überqueren. Er fragt Karla, ob der Fluss tief ist.

Karla: Nö, im Schnitt nur einen Meter.

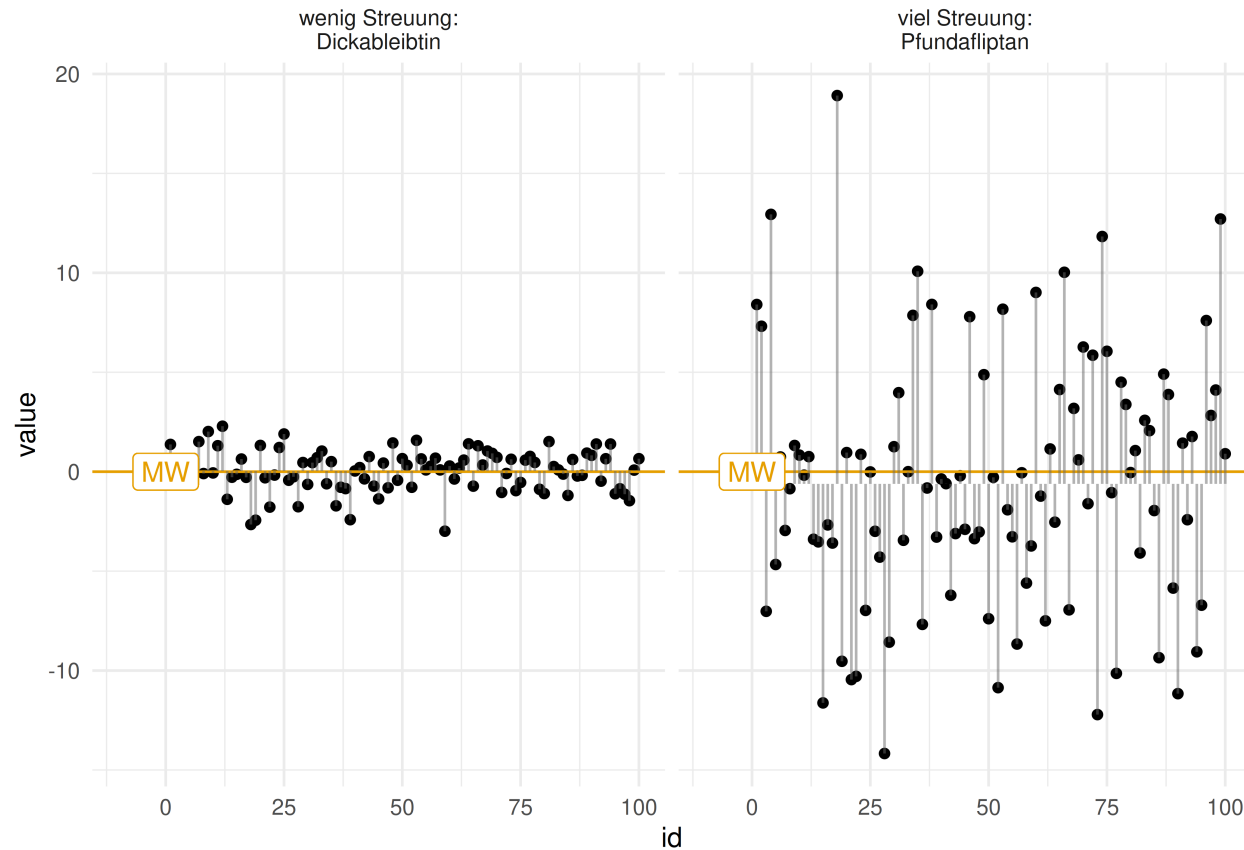
Fritz führt seine Kuh durch den Fluss — die Kuh ertrinkt.

! Wichtig

Die Streuung der Daten zu kennen, ist eine wesentliche Information — ein Lagemaß (wie der Mittelwert) allein reicht nicht.

Woran erkennt man ein gutes Modell?

Wenig vs. viel Streuung



Bei wenig Streuung liegen beobachtete Werte (y) nah an den Modellwerten ($\hat{y}$) — die Abweichungen $e = y - \hat{y}$ sind gering.

Woran erkennt man ein gutes Modell?

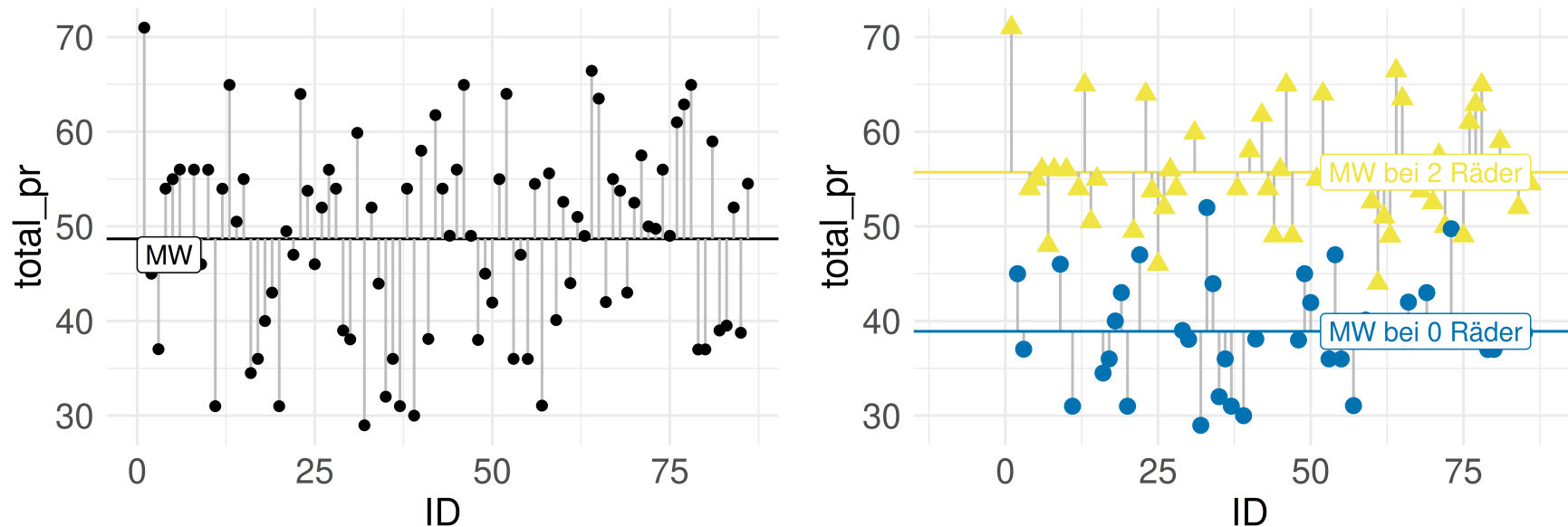
Modellgüte = Präzision des Modells

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

! Wichtig

Ein Modell ohne Angabe der Modellgüte (Präzision der Schätzung) ist wenig nütze.

Ein komplexeres Modell reduziert die Streuung



Innerhalb der Gruppen (0 vs. 2 Lenkräder) sind die Fehlerbalken im Schnitt kürzer als im ungruppierten Modell.

Streuungsmaße

Definition: Streuungsmaß

Definition: Streuungsmaß

Ein Streuungsmaß quantifiziert die Variabilität (Unterschiedlichkeit, Streuung) eines Merkmals.

Definition: Spannweite

Definition: Spannweite

Die Spannweite (Range) R ist die Differenz von größtem und kleinstem Wert eines Merkmals X : $R := X_{max} - X_{min}$.

Beispiel: Werte 1, 23, 42, 100 $\rightarrow R = 100 - 1 = 99$.

Die Spannweite ist nicht robust gegenüber Extremwerten und sollte nur eingeschränkt verwendet werden.

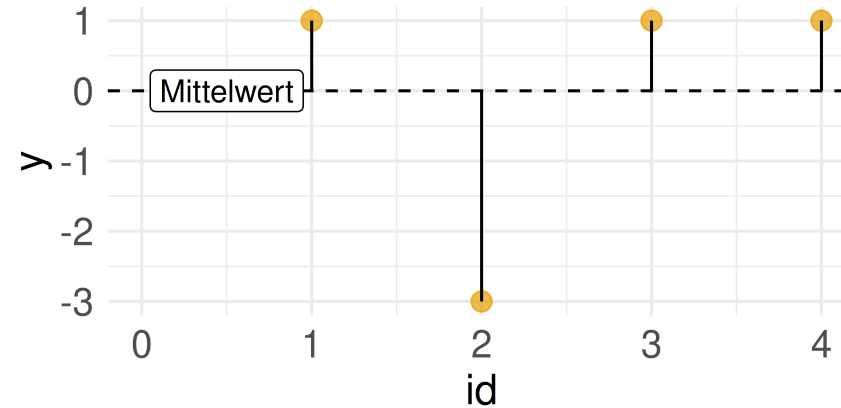
Der mittlere Abweichungsbalken (MAE)

Legt man alle Abweichungsbalken e aneinander und teilt durch die Anzahl n , erhält man den “mittleren Abweichungsbalken” $\bar{e}$ — die *Mittlere Absolutabweichung* (MAE, MAA).

i Definition: Mittlere Absolutabweichung

$$\text{MAE} := \frac{\sum_{i=1}^n |y_i - \bar{y}|}{n} = \frac{\sum_{i=1}^n |e_i|}{n} = \bar{e}$$

MAE: Beispiel & R-Berechnung



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

Den Mittelwert kann man als lineares Modell (Gerade) auffassen: `lm(y ~ 1, data = d)`.

```
1 lm_ohne_x_var <- lm(y ~ 1, data = d)
2 mae(lm_ohne_x_var) # aus dem Paket easystats
```

Der Interquartilsabstand (IQR)

 Definition: Interquartilsabstand

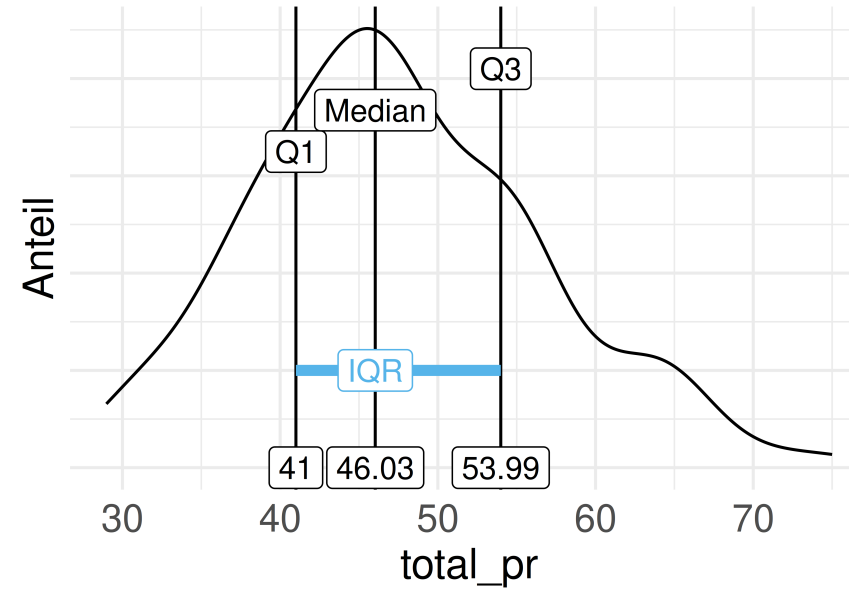
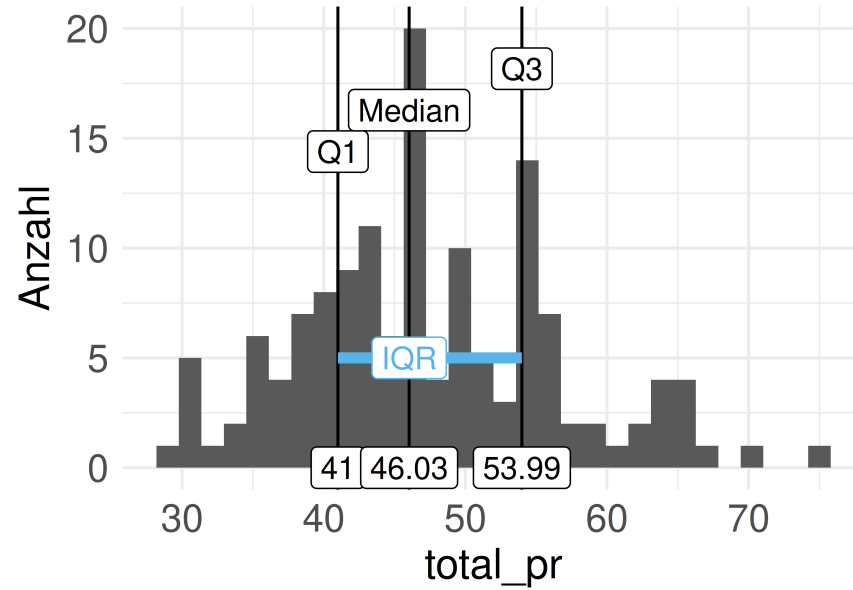
$$IQR := Q_3 - Q_1$$

Der IQR ist robuster als MAA, Varianz und Standardabweichung — er baut nicht auf dem Mittelwert auf.

Beispiel: $Q_1 = 1.65 \text{ m}$, $Q_3 = 1.75 \text{ m} \rightarrow$

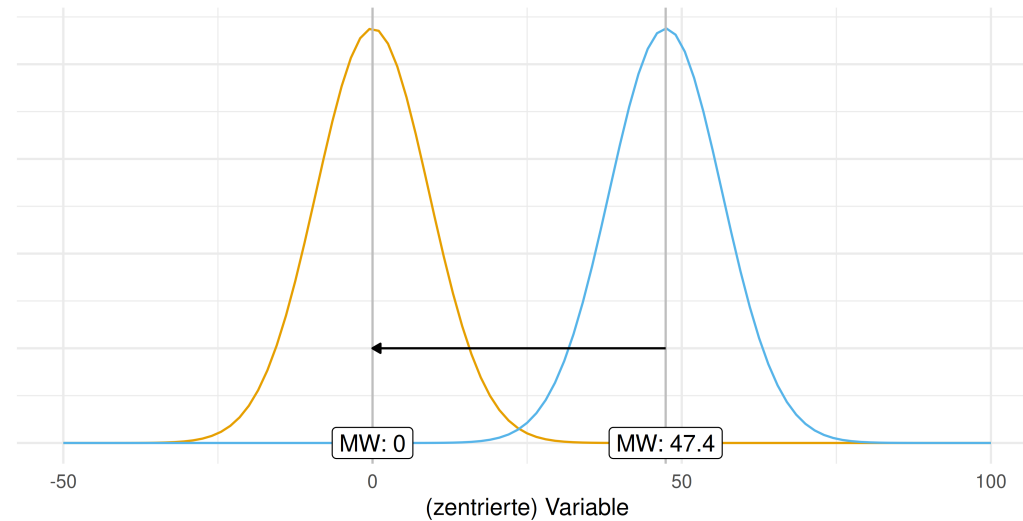
$$IQR = 1.75 \text{ m} - 1.65 \text{ m} = 0.10 \text{ m} (10 \text{ cm}).$$

IQR am Beispiel Mariokart



Streuungsmaße für Normalverteilungen

Eine Normalverteilung lässt sich komplett über *Mittelwert* und *Standardabweichung* beschreiben.



Eine Konstante addieren “verschiebt den Berg”, ohne die Form zu verändern.

Definition: Varianz

Die Varianz ist der mittlere quadrierte Abstand jedes Werts vom Mittelwert.

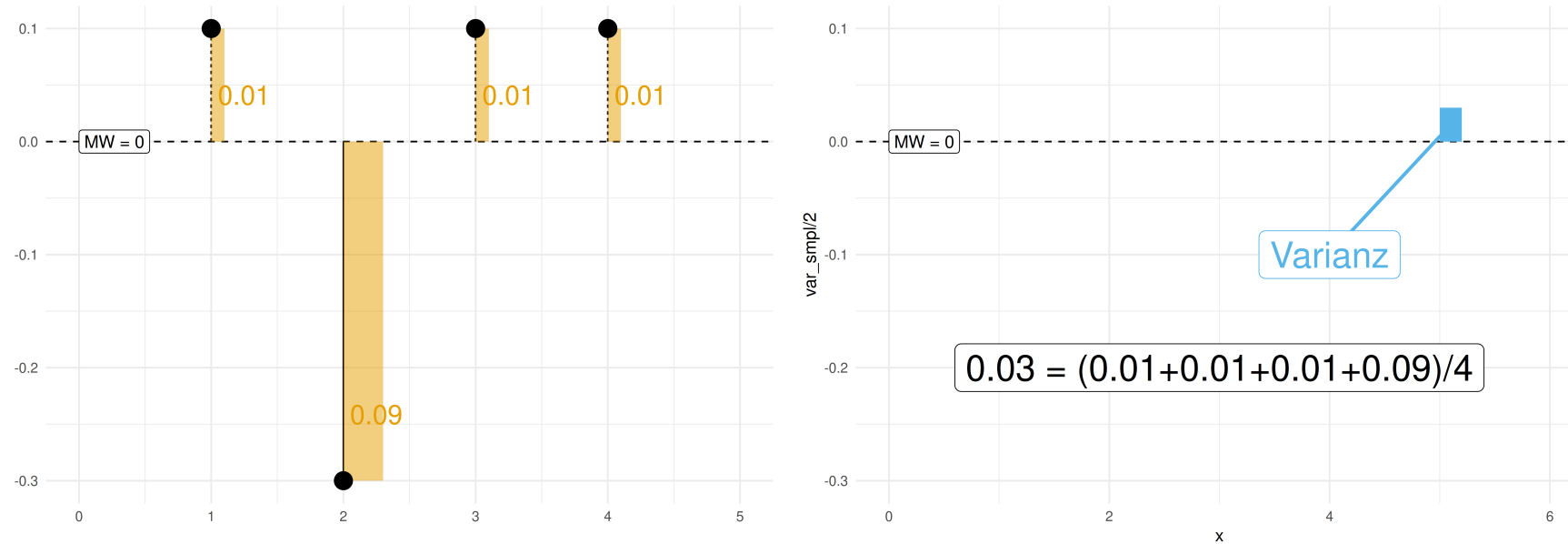
Definition: Varianz

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

Sie entspricht dem *Mean Squared Error* (MSE) eines Modells:

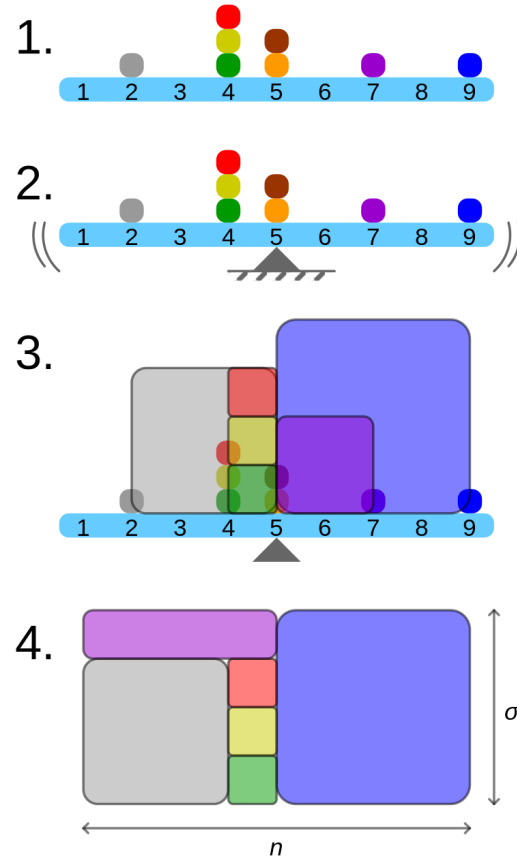
$$MSE := \frac{1}{n} \sum_{i=1}^N (y_i - \hat{y})^2$$

Varianz als “mittlerer Quadratfehler”



Die Varianz fasst die typische *quadrierte* Abweichung der Beobachtungen vom Mittelwert in eine Zahl.

Varianz geometrisch



Varianz als Fläche eines Rechtecks aus umgeformten Abweichungsquadraten ([Cmglee, 2015](#))

Man ordnet die Abweichungsquadrate zu einem Rechteck mit Seitenlängen n und σ^2 .

Die Standardabweichung

Definition: Standardabweichung

Die Standardabweichung (SD, s , σ) ist die Quadratwurzel der Varianz: $s := \sqrt{s^2}$

Aus Modellierungssicht ist die SD die Wurzel des MSE — man nennt sie *Root Mean Squared Error*: $RMSE := \sqrt{MSE}$

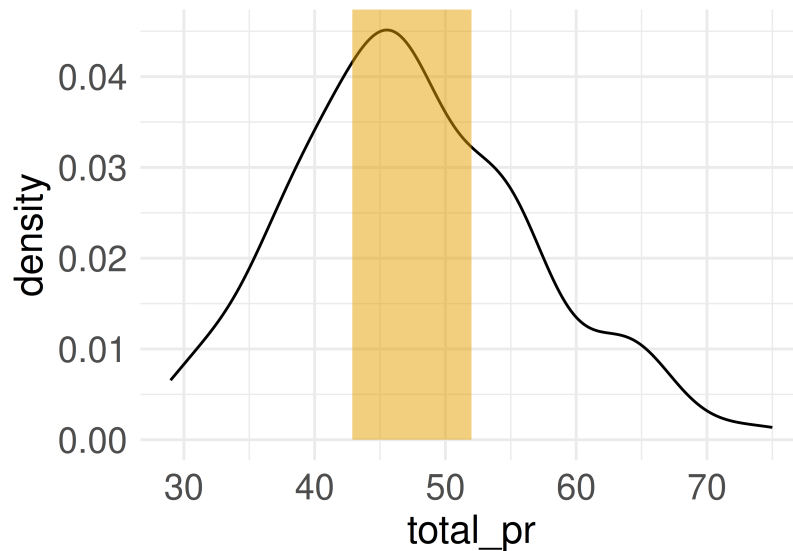
Hinweis

Die SD ist i. d. R. *ungleich* zur MAE, aber (fast) gleich zur RMSE.

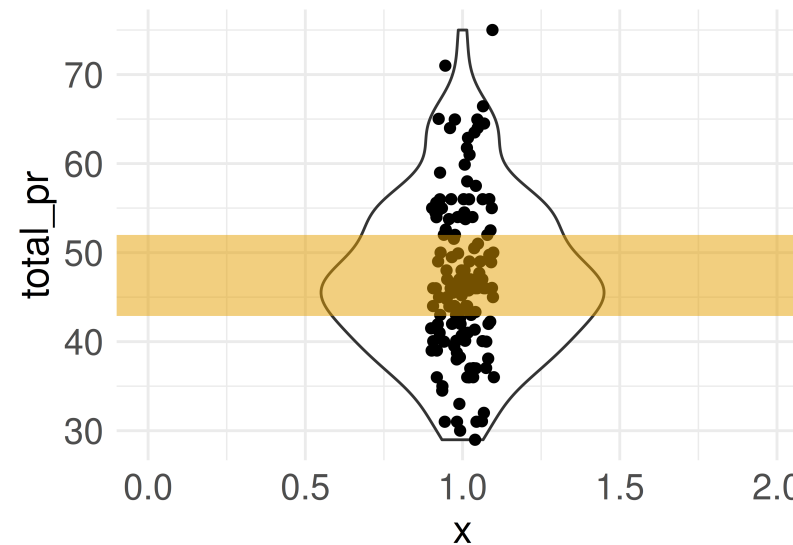
Streuungsmaße im Vergleich: Mariokart

Modellgüte des Mittelwerts anhand der Streuung im Verkaufspreis (`mariokart`, Extremwerte entfernt):

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



Streuungsmaße



Streuung als Modellfehler

Nullmodell und Modellgüte in R

Fassen wir den Mittelwert als Punktmodell des Verkaufspreises auf, sind die Streuungskennwerte zugleich Kennwerte der Modellgüte. Ein Modell ohne UV heißt *Nullmodell*.

```
1 lm_mario_ohne_x_var <- lm(total_pr ~ 1, data = mariokart)
2
3 mae(lm_mario_ohne_x_var) # Mean absolute error
4 mse(lm_mario_ohne_x_var) # Mean squared error
5 rmse(lm_mario_ohne_x_var) # Root mean squared error
```

Ein komplexeres Modell (z. B. `total_pr ~ wheels`) senkt i. d. R. den MAE gegenüber dem Nullmodell.

Die z-Transformation

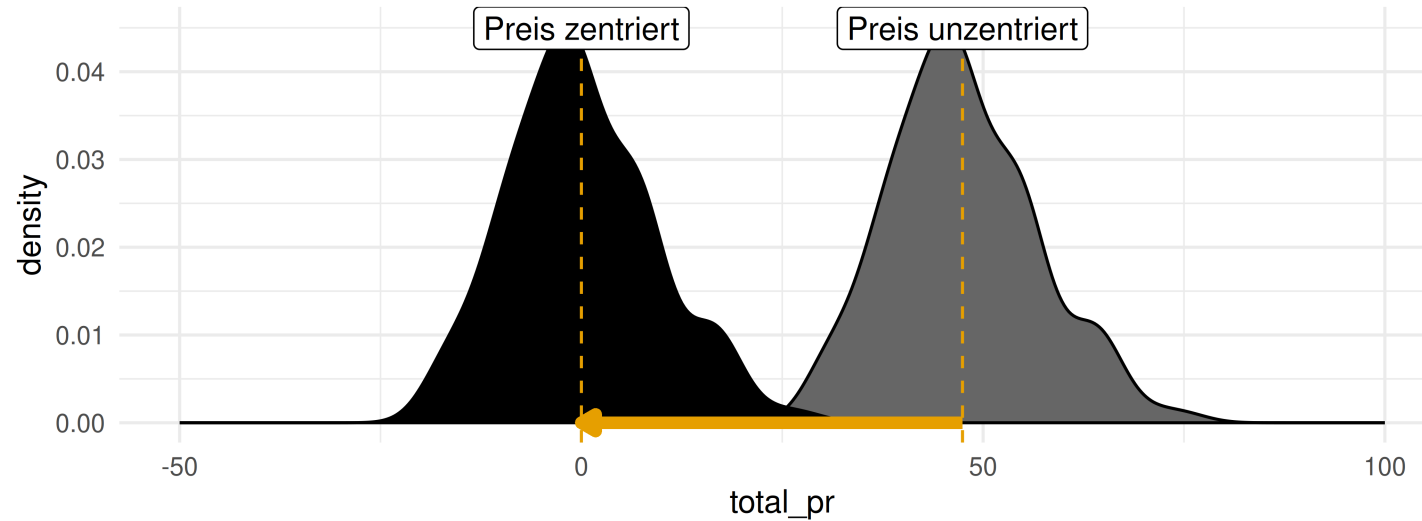
Wozu die z-Transformation?

Bei einem Datensatz ist oft schwer zu sagen, welche Ausprägungen “groß” oder “klein” sind. Die z-Transformation transformiert Werte in bekannte, verständliche Bereiche:

- Mittelwert wird auf Null transformiert
- Streuung (SD) wird auf 1 transformiert

Sie besteht aus zwei Schritten: **Zentrieren** und **Streuung standardisieren**.

Zentrieren



Die vertikalen gepunkteten Linien zeigen die Mittelwerte der Verteilungen.

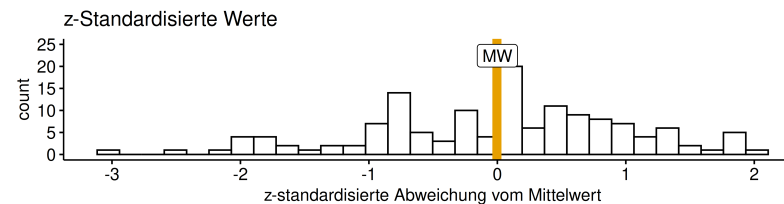
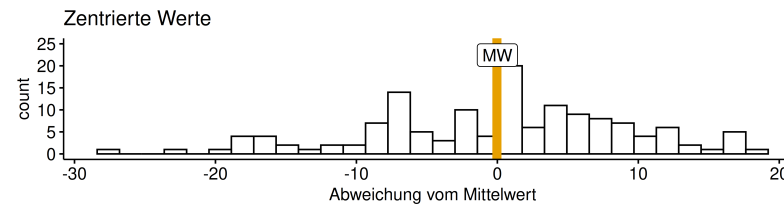
i Definition: Zentrieren

Von jedem Wert einer Verteilung X wird der Mittelwert subtrahiert: $X_c = X - \bar{X}$. Der neue Mittelwert ist Null.

Streuung standardisieren

Wie viel Abweichung vom Mittelwert ist “viel”? Man setzt die Abweichung in Bezug zur typischen Abweichung (SD): weicht ein Spiel 10 € vom Mittelwert ab und beträgt die SD 10 €, ist die standardisierte Abweichung 1 ($10/10 = 1$).

```
1 mariokart_no_extreme <- mariokart_no_extreme %>%  
2   mutate(abw_std = abw / sd(abw))
```



Die z-Transformation

Definition: z-Werte

Kochrezept: (1) Abweichung vom Mittelwert berechnen, (2) durch die SD teilen.

Definition: z-Werte

Für die Variable X berechnet sich der z-Wert der i -ten Beobachtung: $z_i := \frac{x_i - \bar{x}}{sd_x}$

Faustregel: Extremwerte (Ausreißer), wenn $z_i \geq 2.5$ ([Shimizu, 2022](#)).

Definition: Standardnormalverteilung

Definition: Standardnormalverteilung

Eine *Standardnormalverteilung* ist eine Normalverteilung mit Mittelwert 0 und Standardabweichung 1: $X \sim \mathcal{N}(0, 1)$

Jede Normalverteilung lässt sich mit der z-Transformation in eine Standardnormalverteilung überführen.

```
1 mariokart_no_extreme_z <- mariokart_no_extreme |>
2   mutate(total_pr_zentriert = as.numeric(center(total_pr)),
3           total_pr_z = as.numeric(standardise(total_pr)))
```

Zusammenfassung

- **Streuungsmaße:** Spannweite, MAE, IQR, Varianz, SD — quantifizieren die Modellgüte eines Punktmodells
- **MAE:** mittlerer absoluter Fehler; **Varianz/MSE:** mittlerer quadrierter Fehler; **SD/RMSE:** Wurzel davon
- Der IQR ist robuster gegenüber Extremwerten als MAE, Varianz und SD
- Ein komplexeres (sinnvolles) Modell reduziert die Streuung/den Modellfehler
- Die **z-Transformation** (zentrieren + standardisieren) macht Abweichungen vergleichbar:

$$z_i = \frac{x_i - \bar{x}}{sd_x}$$

Literaturhinweise

- Downey ([2023](#)) — kurzweilige Einführung in die Statistik, auch zu Streuungsmaßen
- Çetinkaya-Runde & Hardin ([2021](#)) — mehr “Lehrbuch-Feeling”, frei online verfügbar

Referenzen

Çetinkaya-Runde, M., & Hardin, J. (2021). *Introduction to Modern Statistics*. <https://openintro-ims.netlify.app/>

Cmglee. (2015). *English: Geometric Visualisation of the Variance of the Example Distribution (2, 4, 4, 4, 5, 5, 7, 9) on w:Standard Deviation*. [Artwork]. <https://commons.wikimedia.org/w/index.php?curid=39472834>

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

Shimizu, Y. (2022). Multiple Desirable Methods in Outlier Detection of Univariate Data With R Source Codes. *Frontiers in Psychology*, 12, 819854. <https://doi.org/10.3389/fpsyg.2021.819854>