

Statistik1

Zusammenfassen

Sebastian Sauer

Einstieg

Lernziele

- Sie können gängige Arten von Lagemaßen definieren.
- Sie können erläutern, inwiefern man ein Lagemaß als ein Modell verstehen kann.
- Sie können Lagemaße mit R berechnen.

Mittelwert als Modell

Beispiel: Notendurchschnitt

Ein Statistikkurs besteht aus drei Studentinnen: Anna, Berta und Carla. Anna hat eine 1, Berta eine 2, Carla eine 3.

Die Rechenregel zum Mittelwert lautet:

1. Addiere alle Werte.
2. Teile durch die Anzahl der Werte.
3. Fertig!

Der Durchschnitt (das arithmetische Mittel) beträgt: 2.

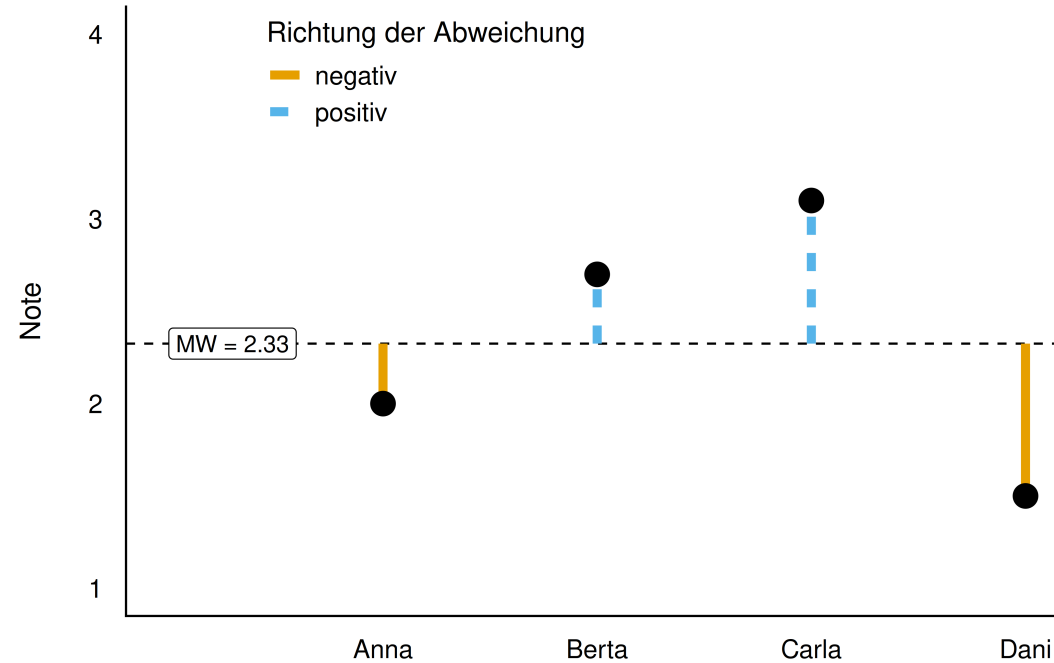
Definition: Mittelwert

Definition: Mittelwert

Der Mittelwert (MW, mean) von X ist definiert als die Summe der Elemente von X geteilt durch deren Anzahl n . Man bezeichnet ihn auch mit $\bar{x}$.

$$\bar{x} := \frac{1}{n} \sum_{i=1}^n x_i = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

Mittelwert und Residuum



Die Abweichung (“Fehler”, “Rest”, “Residuum”) e_i einer Beobachtung zum Mittelwert:

Daten = Modell + Rest, formal: $y_i = \bar{x} + e_i$

Mittelwert als Modell

Summe der Abweichungen ist Null

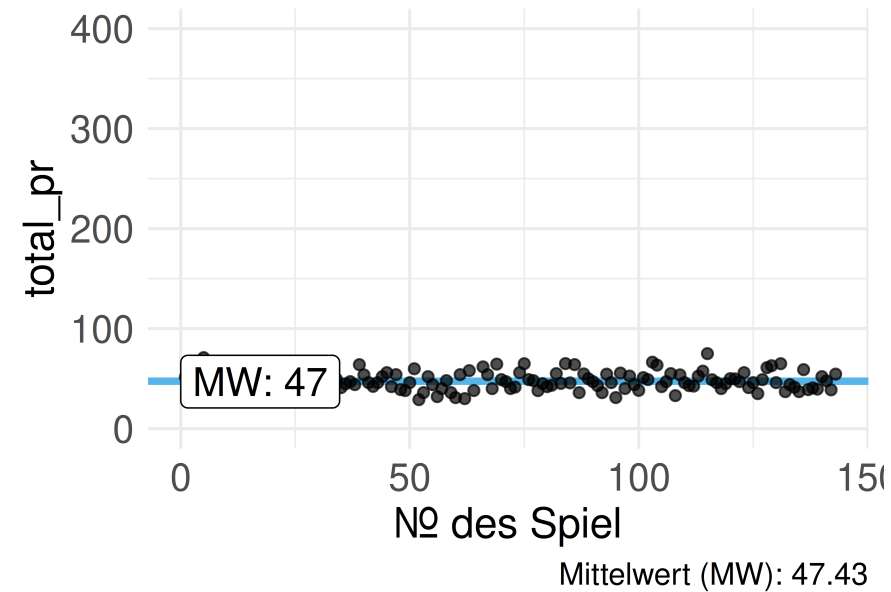
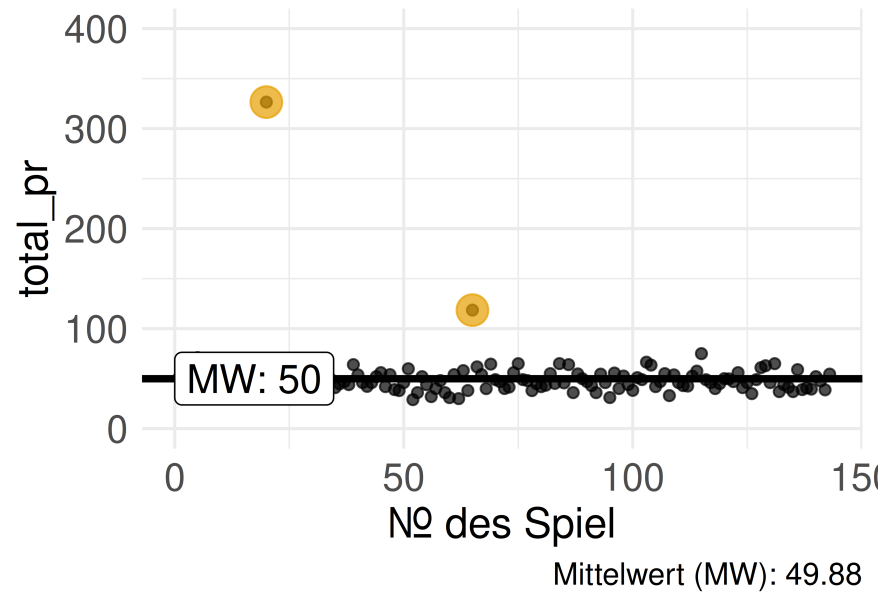
$$\begin{aligned}\sum_{i=1}^n (x_i - \bar{x}) &= \sum_{i=1}^n x_i - \sum_{i=1}^n \bar{x} \\ &= n \cdot \bar{x} - n \cdot \bar{x} = 0\end{aligned}$$

Die nach oben ragenden Abweichungen gleichen die nach unten ragenden exakt aus.

Mittelwert als lineares Modell

Definition: Lineares Modell

Ein lineares Modell beschreibt die Daten durch eine Gerade.



In R: der Befehl `lm()`

```
1 lm(einkommen ~ 1) # lm wie "lineares Modell"
```

- `lm()` gibt als `Coefficient` (Koeffizient) den Mittelwert zurück — den Achsenabschnitt (*intercept*).
- Die Gerade des Mittelwerts schneidet die Y-Achse genau an diesem Punkt.
- Mehr zur `lm()`-Syntax folgt in einem späteren Kapitel.

Der wertvollste Fußballer im Hörsaal

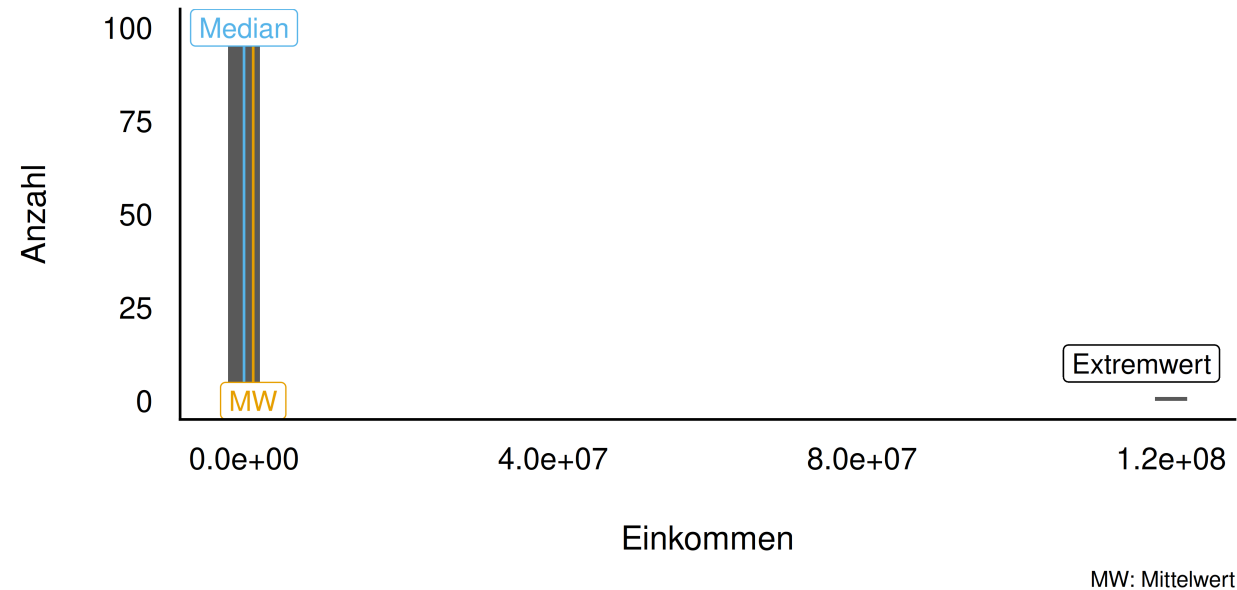
Kommt Kylian Mbappé (Jahreseinkommen ca. 120 Mio. Euro) in einen Hörsaal mit 100 Studis, die je 1000 Euro/Jahr verdienen:

```
1 einkommen_studis <- rep(x = 1000, times = 100)
2 einkommen <- c(einkommen_studis, 120*1e6)
3 mean(einkommen)
```

Der Mittelwert liegt bei über einer Million Euro — kein gutes Modell für das typische Vermögen im Hörsaal!

Der Median als Modell

Einkommensverteilung im Hörsaal



Der Mittelwert ist anfällig für Extremwerte — man sagt, er ist nicht *robust*.

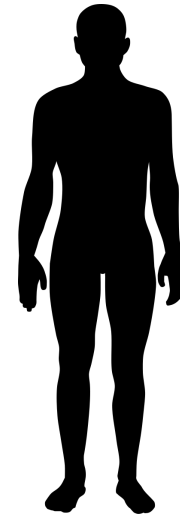
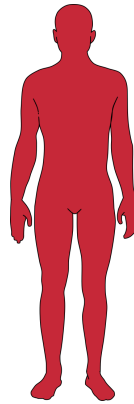
Definition: Median

Definition: Median

Die Merkmalsausprägung, die bei aufsteigend sortierten Beobachtungen in der Mitte liegt, nennt man Median.

Bei geradem n wird das arithmetische Mittel der beiden mittleren Werte gebildet.

Median: der mittlere Wert einer sortierten Reihe



Der Median (1,79 m, rot) als Wert des “mittleren” Objekts bei aufsteigend sortierten Objekten: gleich viele Personen sind kleiner wie größer.

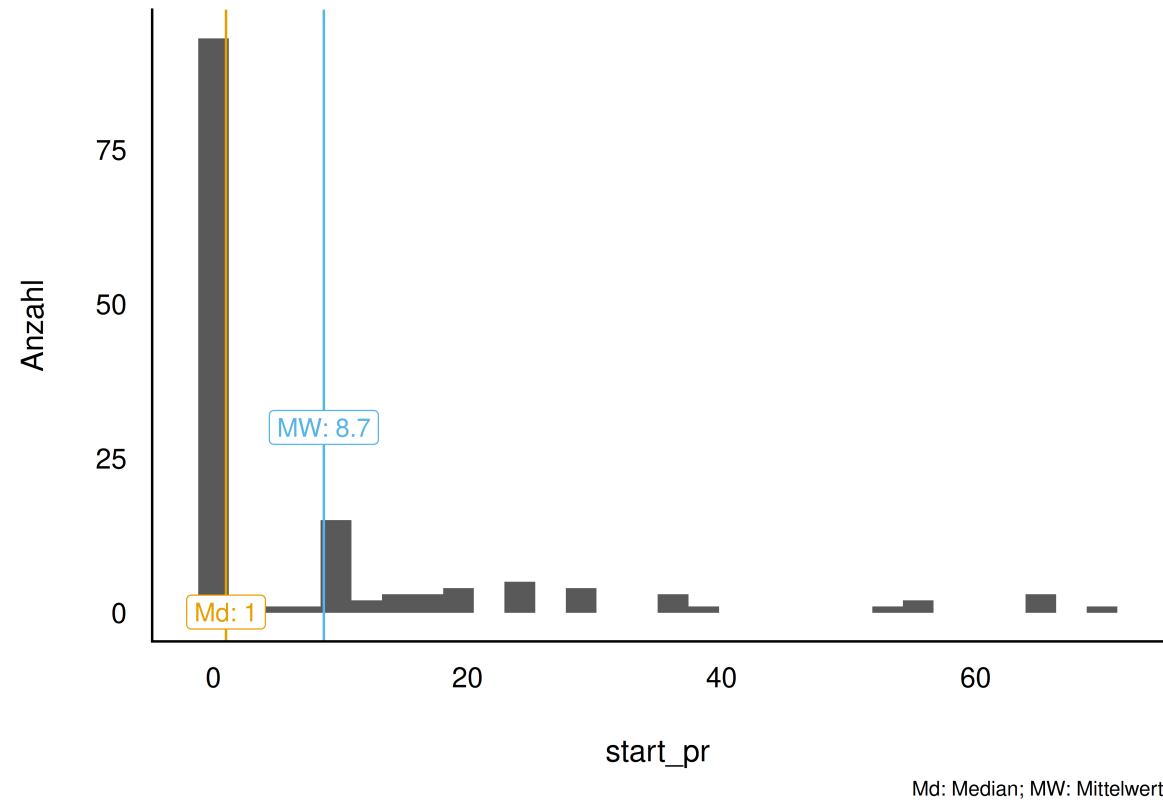
Der Median ist robust

Erhöht sich das Einkommen einer Person im Hörsaal um das Hundertfache (z. B. durch ein Patent), bleibt der Median nahezu unverändert.

! Wichtig

Bei (sehr) schiefen Verteilungen ist der Mittelwert wenig aussagekräftig, da er nicht mehr den “typischen” Wert widerspiegelt.

Beispiel: Mariokart — Median vs. Mittelwert



Ist der Mittelwert größer als der Median, nennt man die Verteilung **rechtsschief**

Der Median als Modell

Quantile

Quantile als Verallgemeinerung des Medians

Der Median teilt eine Verteilung in eine untere und eine obere Hälfte — eine “50-Prozent-Marke”.

- Median der billigeren Hälfte → trennt das billigste Viertel vom Rest ab.
- Median der teureren Hälfte → trennt das teuerste Viertel vom Rest ab.

Warum nur in 25-%-Schritten aufteilen? Man kann genauso gut in 10-%- oder 1-%-Schritten aufteilen.

Definition: Quartile

Definition: Quartile

Sortiert man die Daten aufsteigend, trennt das *erste Quartil* (Q1, 25 %) das Viertel der kleinsten Werte vom Rest ab. Den Median nennt man das *zweite Quartil* (Q2, 50 %). Das *dritte Quartil* (Q3, 75 %) trennt die drei Viertel kleinsten Werte vom obersten Viertel ab.

Definition: Dezile & Quantile

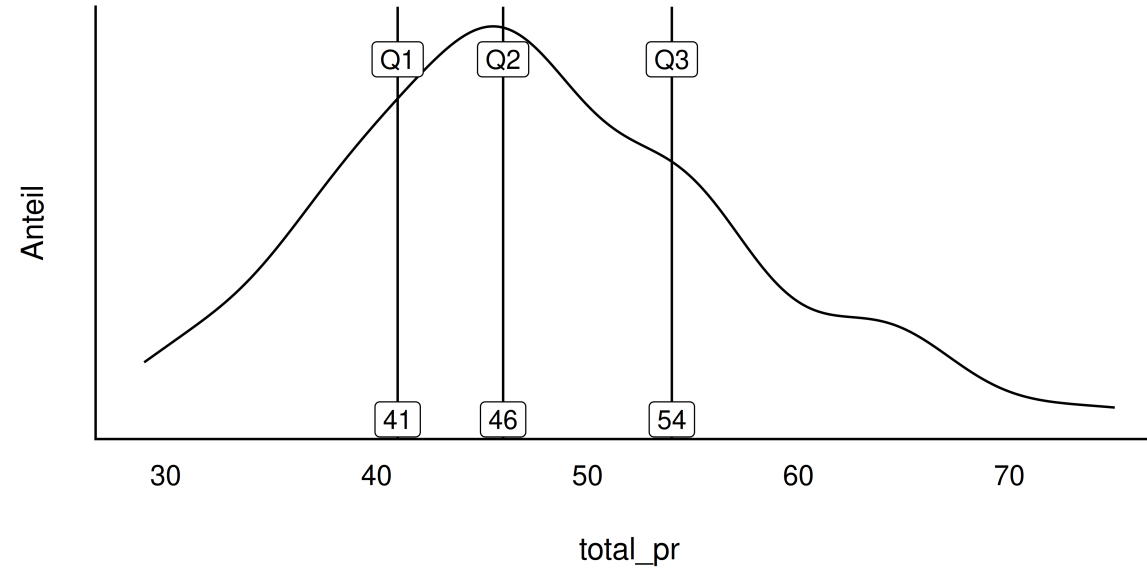
Definition: Dezile

Die neun Quantile $p = 0.1, 0.2, \dots, 1$, die eine Verteilung in 10 gleich große Teile unterteilen, nennt man Dezile.

Definition: Quantile

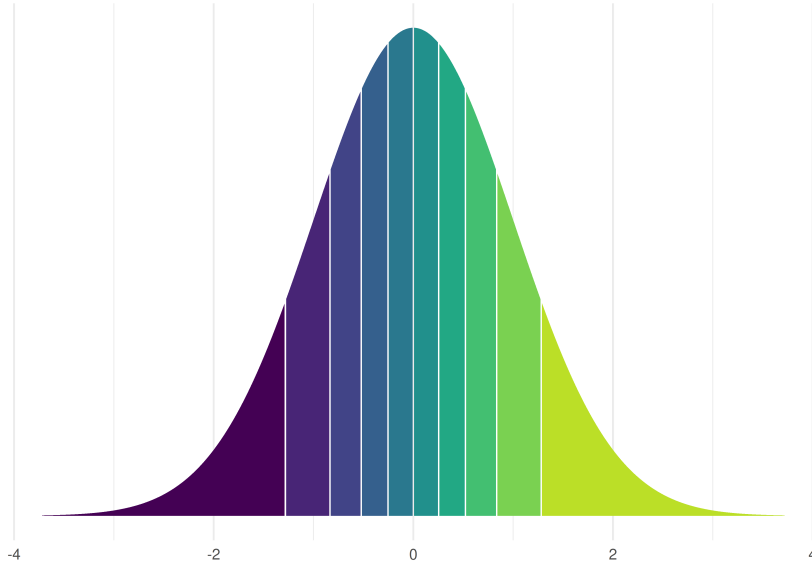
Ein p -Quantil ist der Wert, der von p Prozent der Werte nicht überschritten wird. Quantil ist der Oberbegriff für Quartile, Dezile etc.

Quartile für Mariokart-Verkaufsgebote

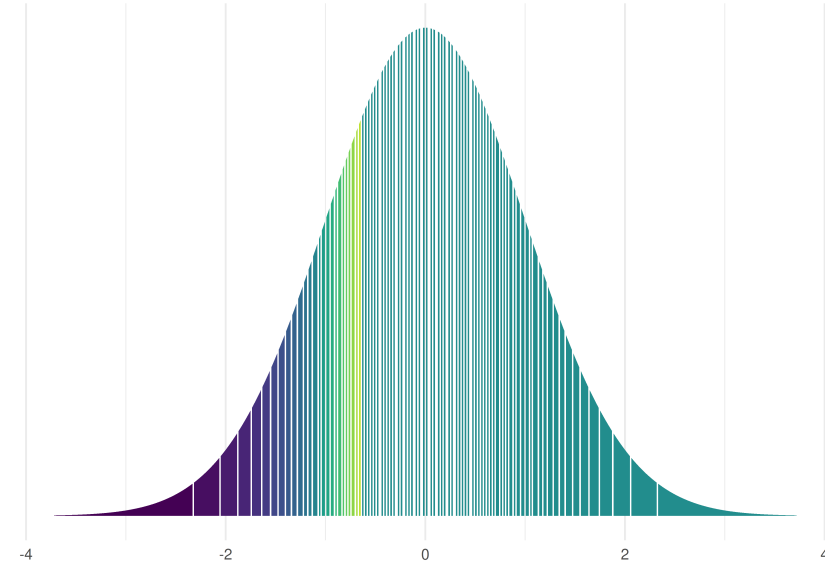


Quantile

Quantile visualisiert



10%-Schritte: Dezile



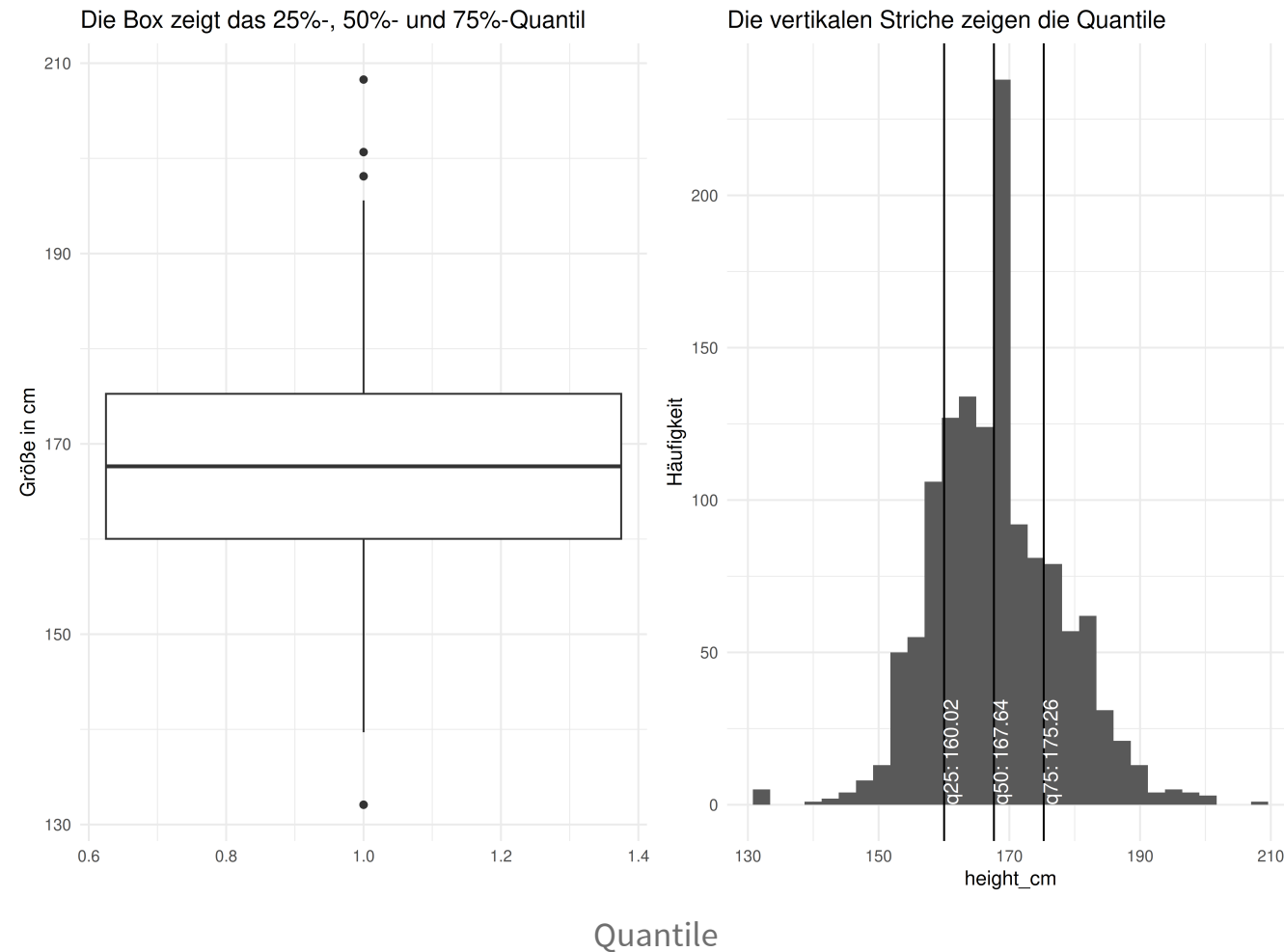
1%-Schritte: Perzentile

Alle Flächen sind gleich groß.

Quantile berechnen mit R

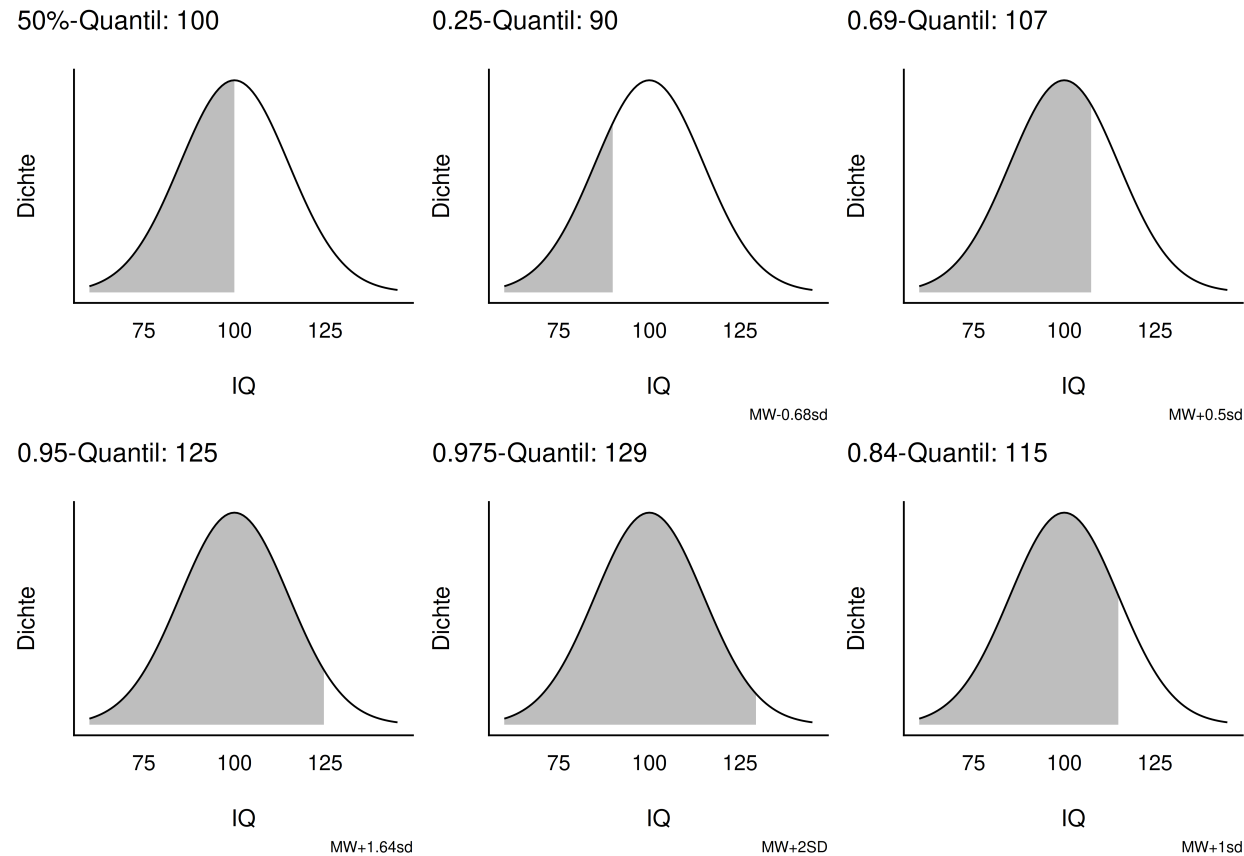
```
1 mariokart %>%
2   filter(total_pr < 100) %>%
3   summarise(
4     q25 = quantile(total_pr, .25), # 1. Quartil
5     q50 = quantile(total_pr, .50), # 2. Quartil
6     q75 = quantile(total_pr, .75)) # 3. Quartil
```

Beispiel: Körpergröße von Studierenden



Beispiel: Quantile der IQ-Verteilung

Der IQ wird üblicherweise als normalverteilt mit Mittelwert 100 und Streuung 15 angenommen.



Quantile

Lagemaße

Definition: Lagemaß

Definition: Lagemaß

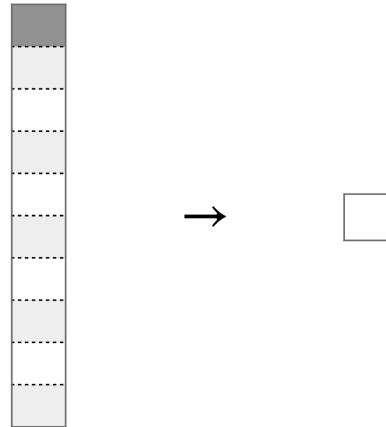
Ein *Lagemaß* (synonym: Maß der zentralen Tendenz) gibt einen Vorschlag, welchen Wert einer Verteilung wir als typisch, normal, erwartbar oder “mittel” ansehen sollten.

Gebräuchliche Lagemaße: Mittelwert, Median, Quantile (z. B. Quartile), Minimum, Maximum, Modus.

Definition: Punktmodell

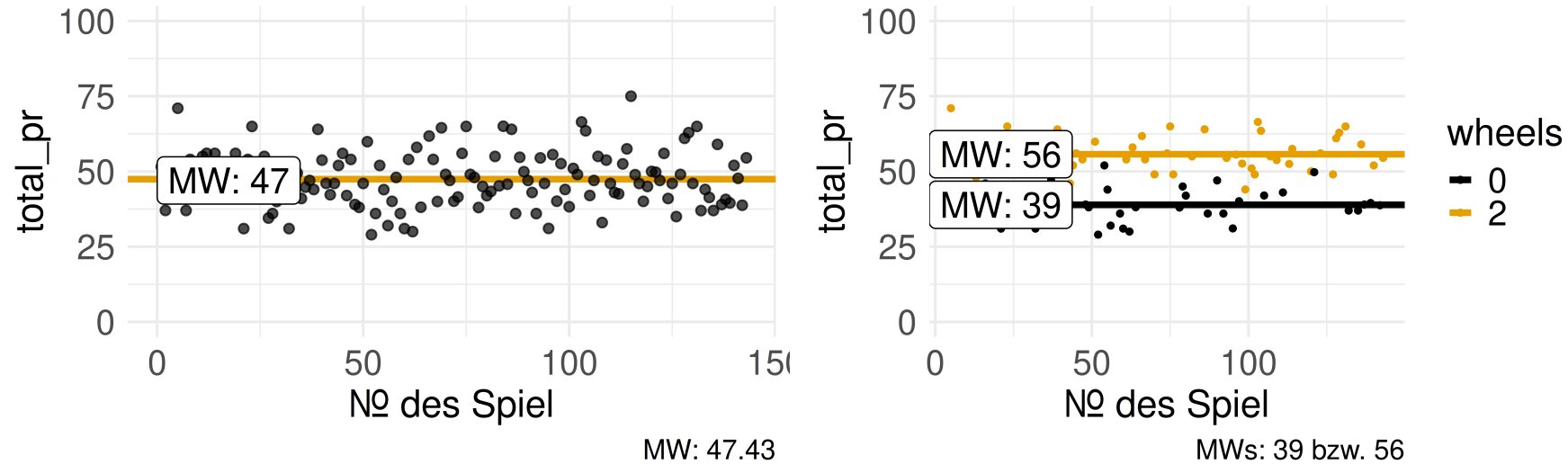
Definition: Punktmodell

Ein Modell, das für alle Beobachtungen ein und denselben Wert vorhersagt, heißt *Punktmodell*.



Mittelwert, Median und Quartile sind Beispiele für Punktmodelle.

Gruppierte Lagemaße



Innerhalb der Gruppen sind die Abweichungen zum Gruppen-Mittelwert oft geringer als zum ungruppierten Mittelwert — das Residuum wird kleiner.

Wie man mit Statistik lügt

Checkliste: Transparenz einer Analyse

1. Wurde die Art und Zeitdauer der Datenerhebung vorab festgelegt und berichtet?
2. Wurden ausreichend Daten gesammelt (z. B. mind. 20 Beobachtungen pro Gruppe)?
3. Wurden alle untersuchten Variablen berichtet?
4. Wurden alle durchgeführten Interventionen berichtet?
5. Wurden Daten aus der Analyse entfernt? Wenn ja, gibt es eine stichhaltige Begründung?

! Wichtig

Stellen Sie hohe Anforderungen an die Transparenz einer statistischen Analyse.

Zusammenfassung

- **Mittelwert:** Summe der Werte geteilt durch deren Anzahl — nicht robust gegenüber Extremwerten
- **Median:** mittlerer Wert bei sortierten Daten — robust gegenüber Extremwerten
- **Quantile:** verallgemeinern den Median (Quartile, Dezile, Perzentile)
- **Lagemaße** sind Punktmodelle: Sie fassen eine Verteilung zu einem “typischen” Wert zusammen
- Gruppierte Lagemaße können Residuen verkleinern
- Transparenz ist der Schlüssel gegen Manipulation statistischer Analysen

Literaturhinweise

- Mittag & Schüller (2020) und Oestreich & Romberg (2014) — Einführungen in die Statistik
- Bortz & Schuster (2010) — ein Klassiker
- Sauer (2019) — mit Fokus auf R
- Çetinkaya-Runde & Hardin (2021), Poldrack (2023) — englischsprachig, online verfügbar

Referenzen

Bortz, J., & Schuster, C. (2010). *Statistik Für Human- und Sozialwissenschaftler*. Springer. <https://doi.org/10.1007/978-3-642-12770-0>

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

Mittag, H.-J., & Schüller, K. (2020). *Statistik: Eine Einführung mit interaktiven Elementen*. Springer. <https://doi.org/10.1007/978-3-662-61912-4>

Oestreich, M., & Romberg, O. (2014). *Keine Panik vor Statistik!: Erfolg und Spaß im Horrorfach nichttechnischer Studiengänge*. Springer. <https://doi.org/10.1007/978-3-658-04605-7>

Poldrack, R. A. (2023). *Statistical Thinking: Analyzing Data in an Uncertain World*. Princeton University Press. <https://statstheinking21.github.io/statstheinking21-core-site/>

Sauer, S. (2019). *Moderne Datenanalyse mit R: Daten einlesen, aufbereiten, visualisieren und modellieren*. Springer. <https://www.springer.com/de/book/9783658215866>