

Statistik1

Daten verbildlichen

Sebastian Sauer

Einstieg

Lernziele

- Sie können erläutern, wann und wozu das Visualisieren statistischer Inhalte sinnvoll ist.
- Sie kennen typische Arten von Datendiagrammen.
- Sie können typische Datendiagramme mit R visualisieren.
- Sie können zentrale Ergebnisse aus Datendiagrammen herauslesen.

Benötigte R-Pakete und Daten

Neben `tidyverse` und `easystats` benötigen wir in diesem Kapitel:

```
1 library(ggpubr)      # Datenvisualisierung
2 library(ggstatsplot) # optional, Datenvisualisierung
```

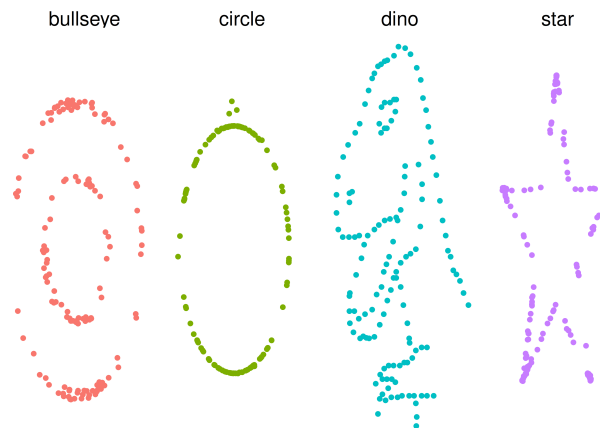
Wir arbeiten wieder mit dem Datensatz `mariokart`.

Wozu das alles?

Ein Bild sagt mehr als 1000 Worte — Datendiagramme helfen, Erkenntnisse aus Daten zu ziehen, ohne mühsam einzelne Zahlen zu vergleichen.

**Ein Dino sagt mehr als
1000 Worte**

Gleiche Kennwerte, ganz andere Bilder



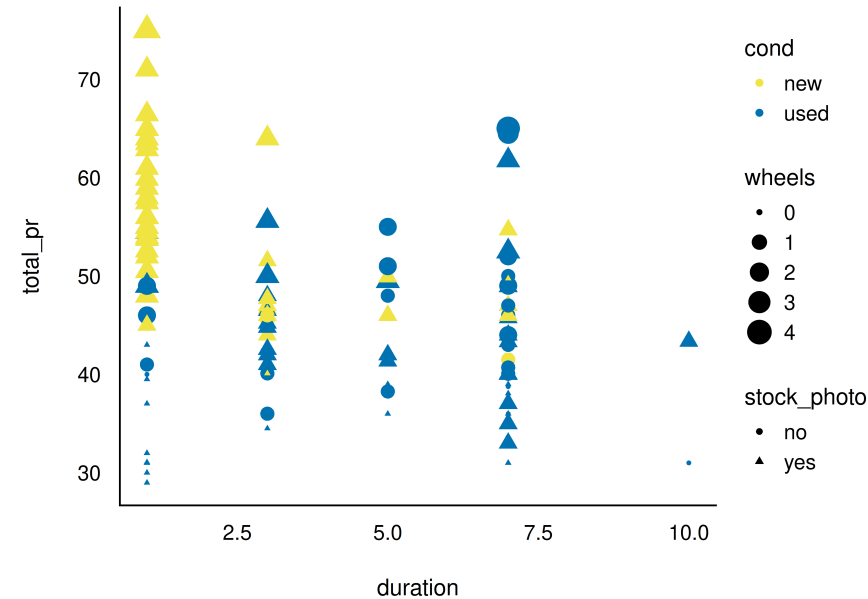
- Verschiedene Datensätze können (nahezu) identische Kennwerte (Mittelwert, Streuung, Korrelation) haben — und trotzdem völlig verschieden aussehen.
- Dasselbe zeigt “Anscombes Quartett”: vier grundverschiedene Streudiagramme mit fast identischen Statistiken.
- Fazit: Eine Visualisierung enthüllt oft, was eine Kennzahl verhüllt.

Definition: Datendiagramm

Definition: Datendiagramm

Ein *Datendiagramm* (kurz: Diagramm) ist ein Diagramm, das Daten und Statistiken zeigt, mit dem Zweck, Erkenntnisse daraus zu ziehen.

Grenzen der Dimensionalität



Ein Bild kann nur eine begrenzte Anzahl an Variablen sinnvoll zeigen — vielleicht vier bis sechs. Bei mehr Variablen brauchen wir Statistik statt (nur) Bilder.

Nomenklatur von Datendiagrammen

Eine kleine Nomenklatur

Erkenntnisziel	qualitativ	quantitativ
Verteilung	Balkendiagramm	Histogramm / Dichtediagramm
Zusammenhang	gefülltes Balkendiagramm	Streudiagramm
Unterschied	gefülltes Balkendiagramm	Boxplot

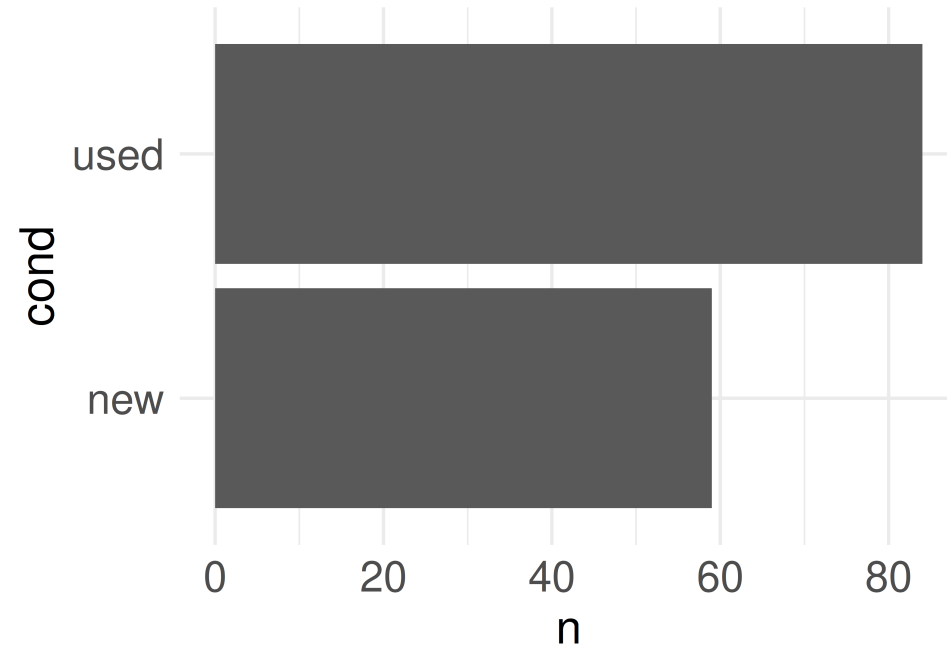
Verteilungen verbildlichen

Definition: Verteilung

Definition: Verteilung

Eine (Häufigkeits-)Verteilung einer Variablen X schlüsselt auf, wie häufig jede Ausprägung von X ist.

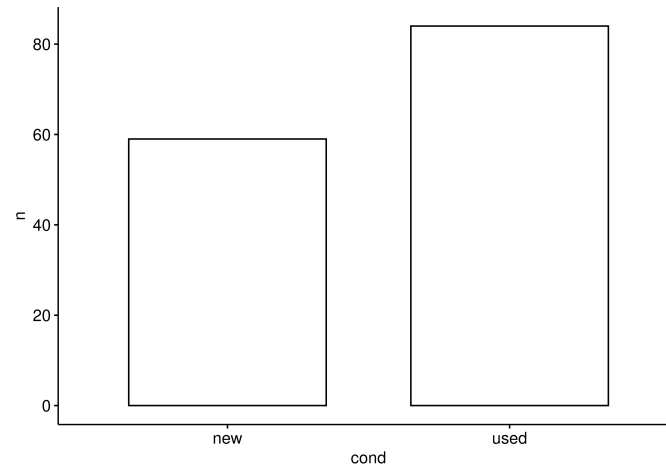
Verteilung einer nominalen Variable: Balkendiagramm



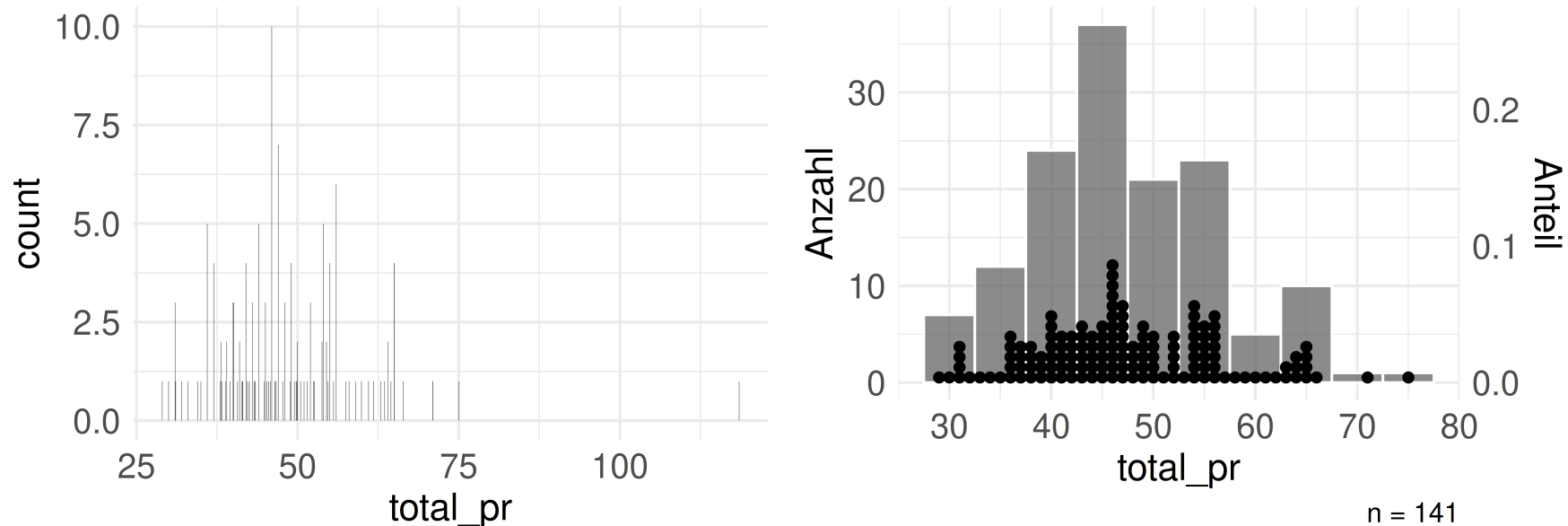
Ein Balkendiagramm zeigt Häufigkeiten von Kategorien als Balken — die Länge ist proportional zur Häufigkeit. Die Ausrichtung (horizontal/vertikal) ist Geschmackssache.

Syntax: Balkendiagramm mit **ggpubr**

```
1 mariokart %>%  
2   count(cond) %>%  
3   ggbarplot(x = "cond", y = "n")
```



Verteilung einer quantitativen Variable



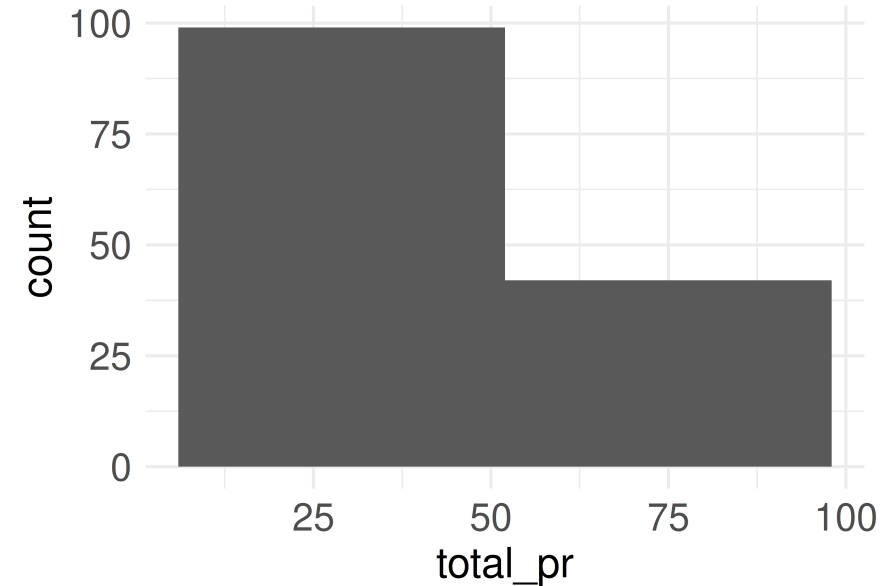
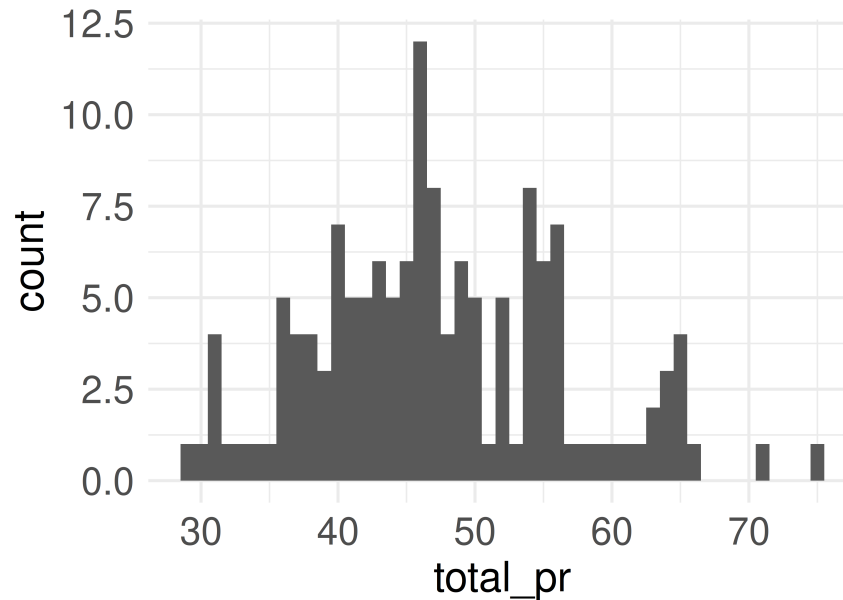
Bei vielen Ausprägungen ist ein Balkendiagramm unübersichtlich. Lösung: Werte in Gruppen (Klassen) zusammenfassen — das ergibt ein **Histogramm**.

Definition: Histogramm

Definition: Histogramm

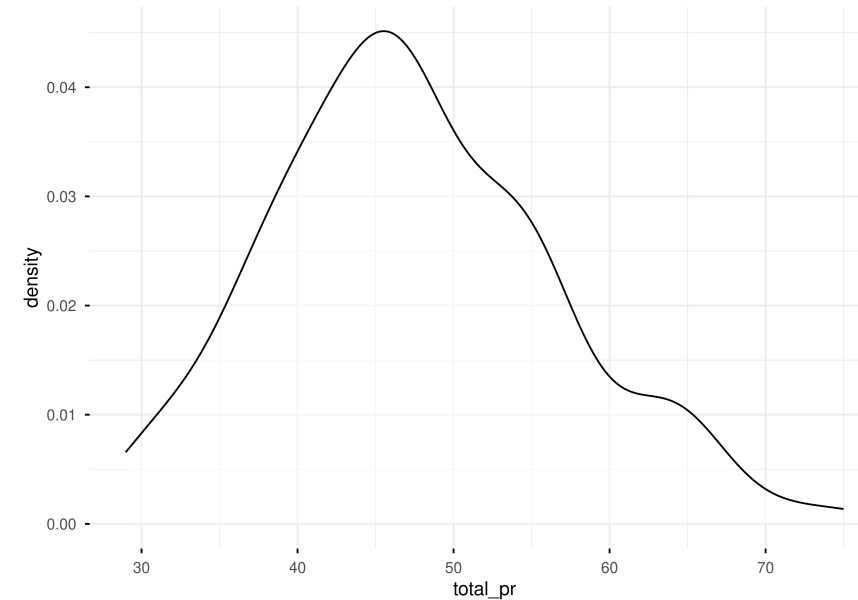
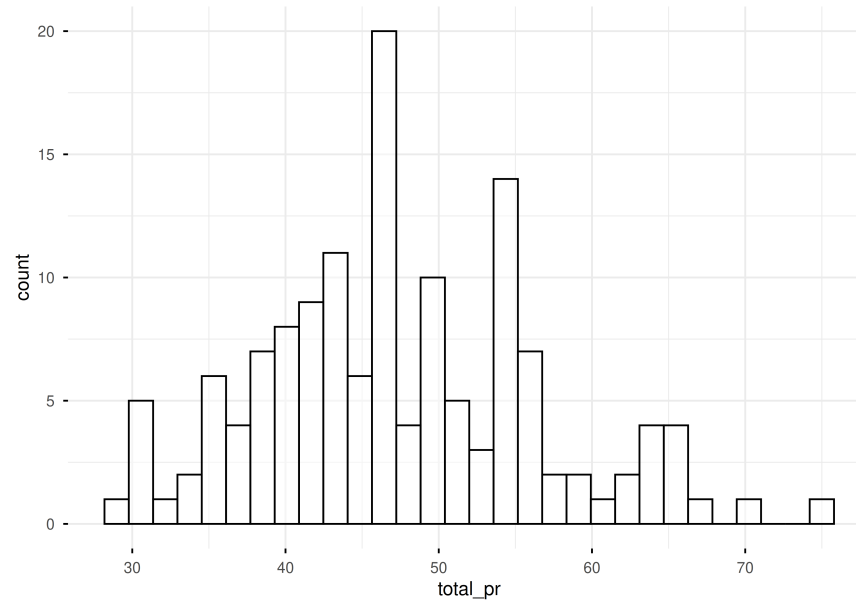
Ein Histogramm ist ein Diagramm zur Darstellung der Häufigkeitsverteilung einer quantitativen Variablen. Die Daten werden in Klassen eingeteilt, die durch Balken dargestellt werden; die Balkenhöhe zeigt die Häufigkeit pro Klasse.

Wie viele Balken? Nicht zu viele, nicht zu wenige



Es gibt keine feste Regel für die Balkenanzahl — nur: nicht zu grob, nicht zu fein.

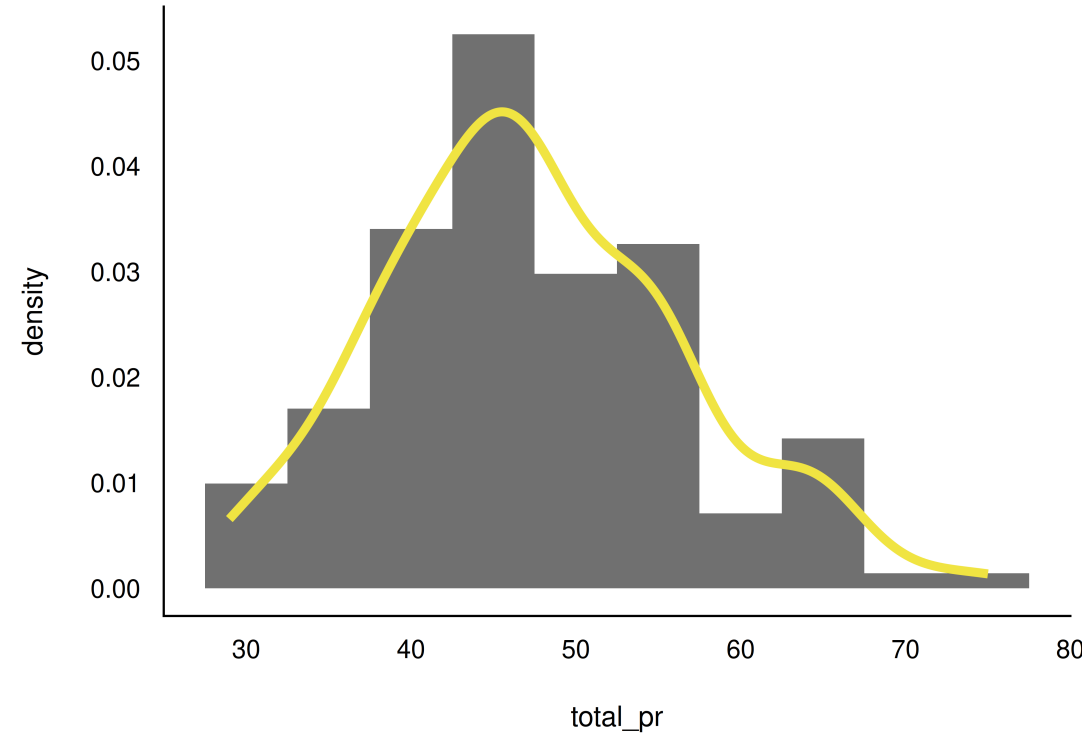
Histogramm vs. Dichtediagramm



Definition: Dichtediagramm

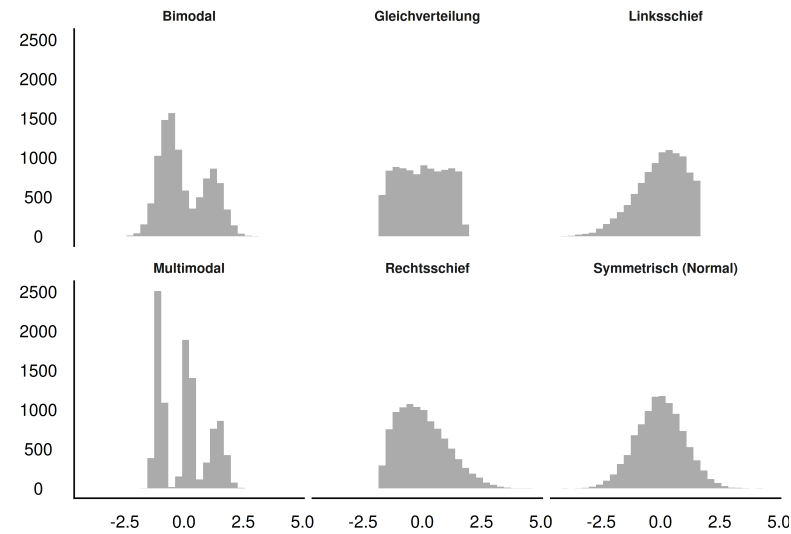
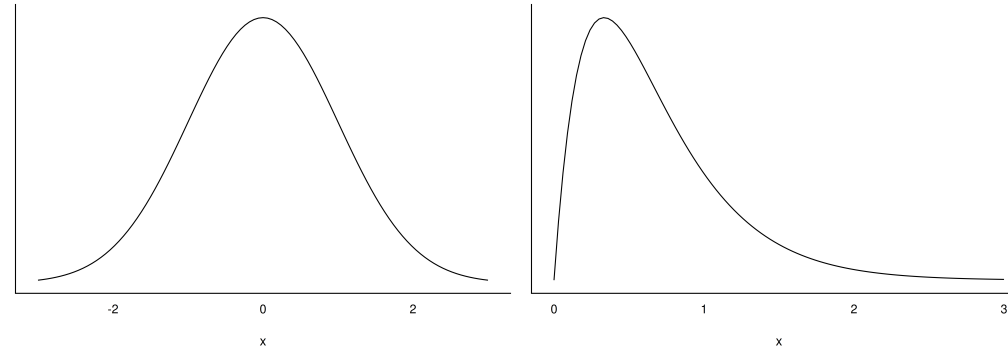
Ein Dichtediagramm visualisiert die Verteilung einer stetigen Variablen — geglättet, sodass Rauschen weniger stört als beim Histogramm.

Histogramm + Dichtelinie kombiniert



Verteilungen verbildlichen

Verteilungsformen



Verteilungen verbildlichen

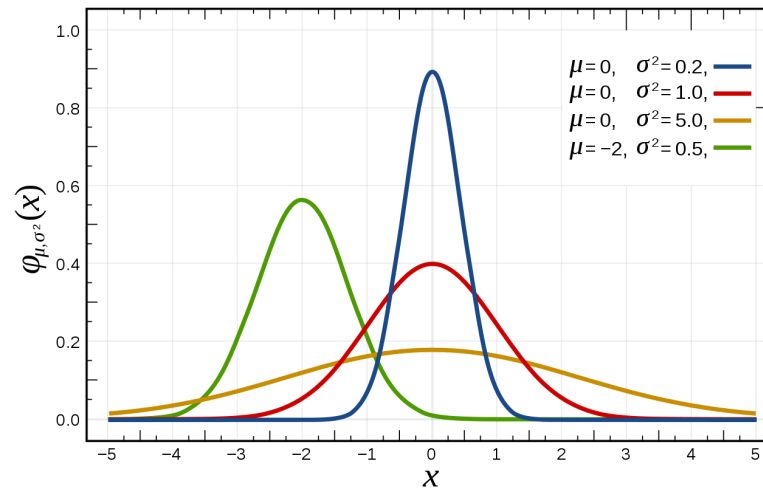
Spezialfall Normalverteilung

Definition: Normalverteilung

Definition: Normalverteilung

Normalverteilungen haben eine charakteristische, symmetrische Glockenform. Sie unterscheiden sich in Mittelwert μ und Streuung σ — diese beiden Parameter bestimmen den gesamten Kurvenverlauf.

Kurzschreibweise: $X \sim \mathcal{N}(\mu, \sigma)$

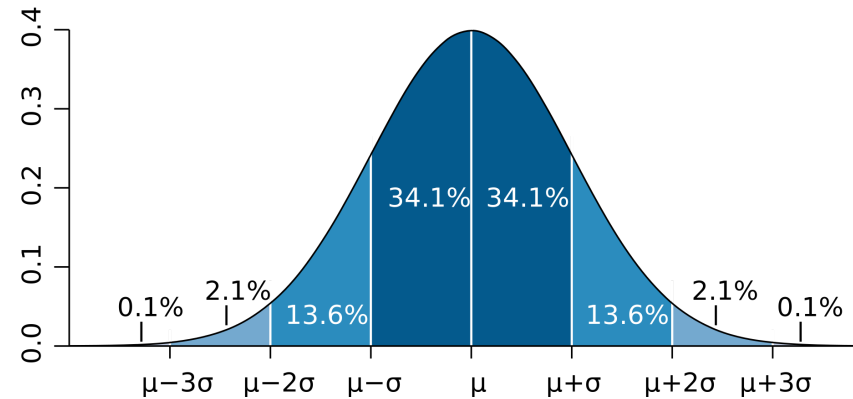


Normalverteilungen mit verschiedenen Mittelwerten und Streuungen ([InductiveLoad, 2008](#))

Beispiele normalverteilter Variablen

Körpergröße, IQ-Werte, Prüfungsergebnisse, Messfehler,
Lebensdauer von Glühbirnen, Gewicht von Brotlaiben,
Milchproduktion von Kühen ...

Fläche unter der Kurve



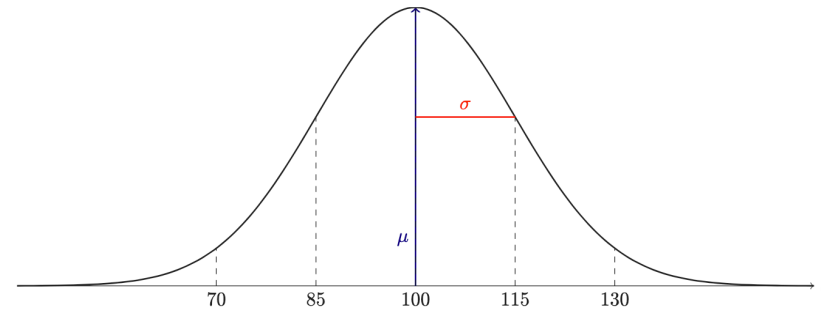
Flächeninhalte (Wahrscheinlichkeitsmasse) einer Normalverteilung ([Ainali, 2007](#))

Die 68-95-99,7-Prozentregel:

- 68 % im Bereich $\mu \pm 1 \cdot \sigma$
- 95 % im Bereich $\mu \pm 2 \cdot \sigma$
- 99.7 % im Bereich $\mu \pm 3 \cdot \sigma$

Beispiel: IQ-Verteilung

$$IQ \sim \mathcal{N}(100, 15)$$



Entstehung einer Normalverteilung

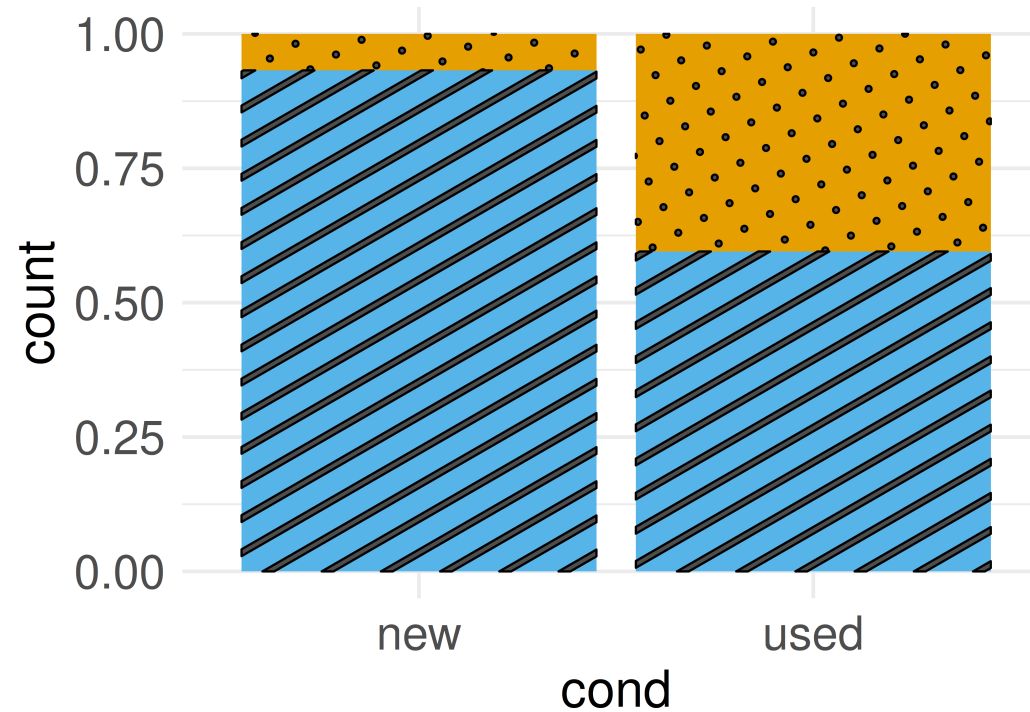
Definition: Entstehung einer Normalverteilung

Wenn sich eine Variable X als Summe mehrerer, unabhängiger, etwa gleich starker Summanden ergibt, tendiert X dazu, normalverteilt zu sein.

Anschaulich zeigt das Galton-Brett, wie aus vielen kleinen Zufallsereignissen eine Glockenkurve entsteht.

Zusammenhänge verbildlichen

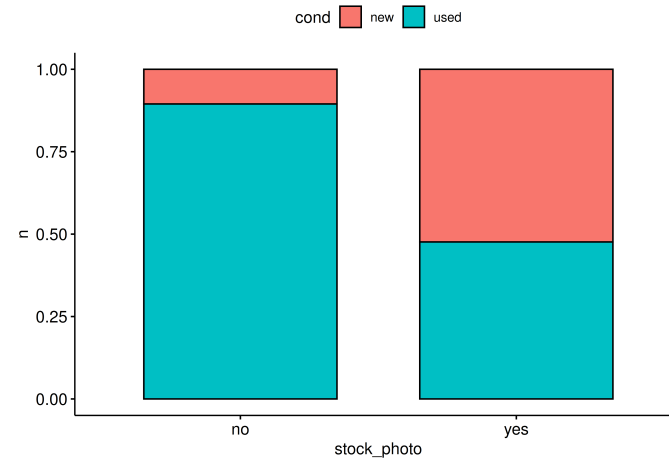
Zusammenhang nominaler Variablen



Bei neuen Spielen war fast immer ein Foto dabei; bei gebrauchten nur bei gut der Hälfte — ein Zusammenhang zwischen `cond` und `stock_photo`.

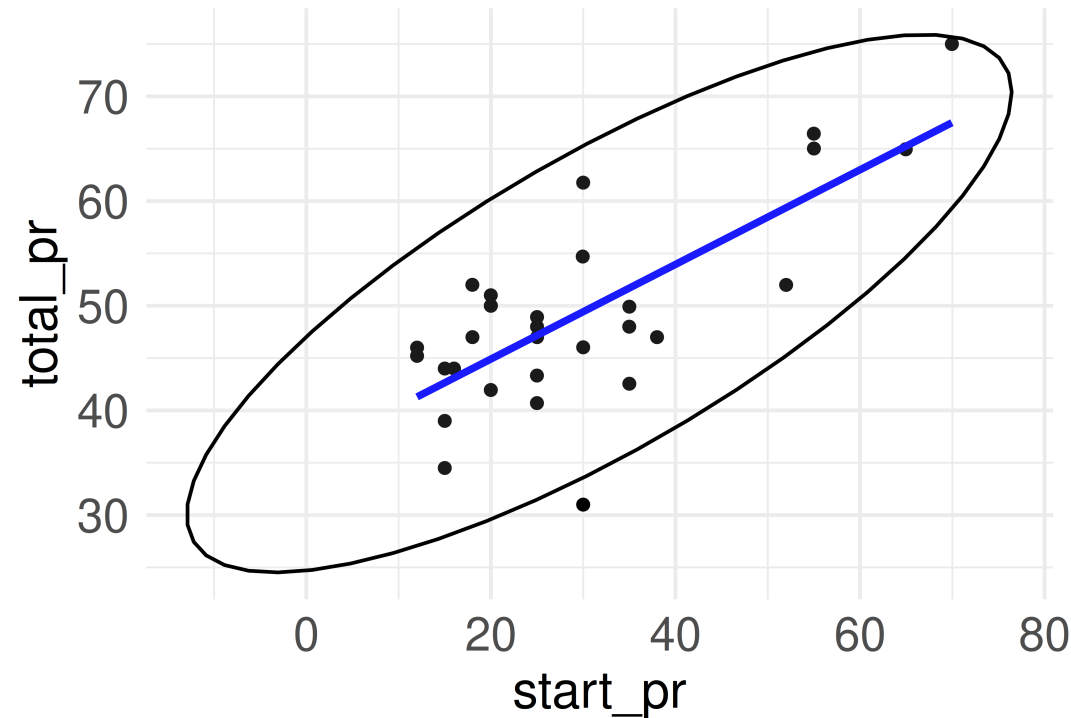
Gefüllte Balkendiagramme

```
1 mariokart %>%  
2   count(stock_photo, cond) %>%  
3   ggbarplot(x = "stock_photo", y = "n", fill = "cond",  
4             position = position_fill()) # aus dem Paket ggpubr
```



Geeignet, wenn (mindestens) eine der beiden Variablen nur zwei Ausprägungen hat.

Zusammenhang bei metrischen Variablen: Streudiagramm



Die Trendgerade (Regressionsgerade) zeigt Richtung des Zusammenhangs; die Ellipse zeigt, wie eng die Daten um sie streuen.

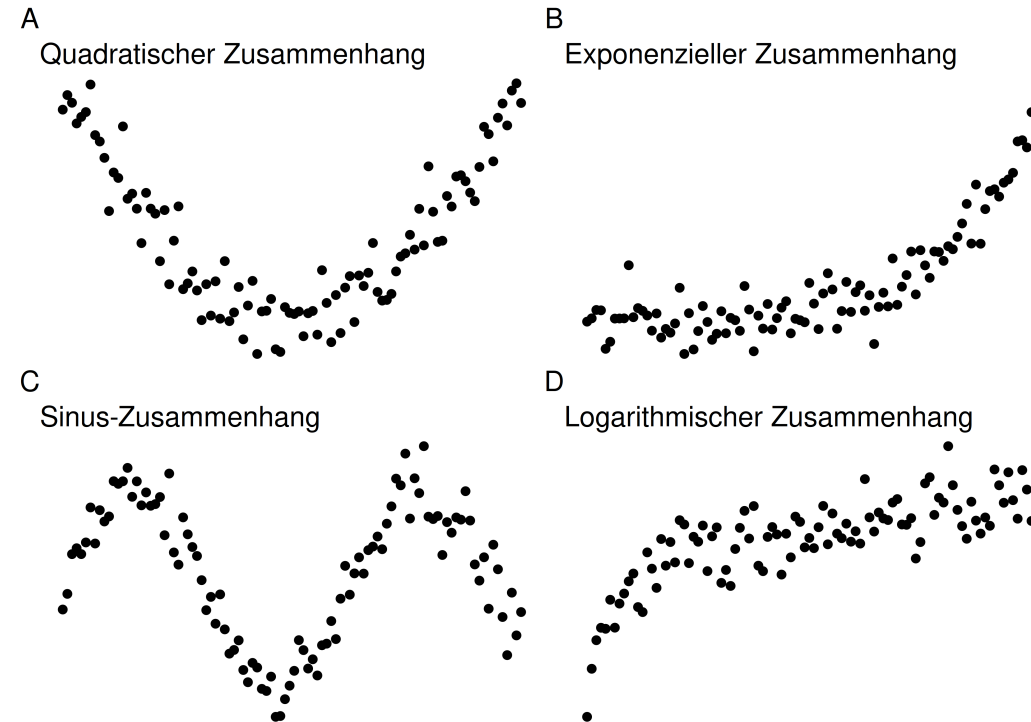
Zusammenhänge verbildlichen

Definition: Linearer Zusammenhang

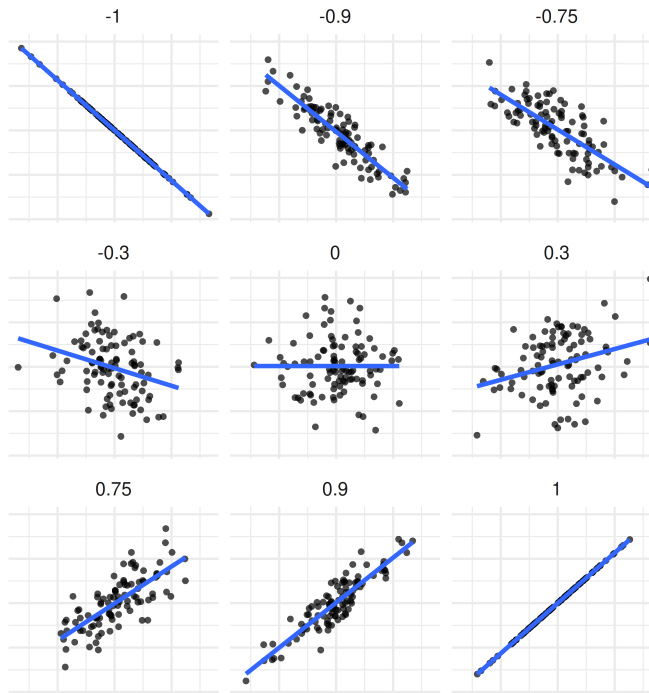
Definition: Linearer Zusammenhang

Lässt sich die Beziehung zweier Variablen gut mit einer Geraden beschreiben, spricht man von einem linearen Zusammenhang. *Gleichsinnige* (positive) Zusammenhänge zeigen aufsteigende Trendgeraden ↗; *gegensinnige* (negative) zeigen absteigende Trendgeraden ↘.

Auch nicht-lineare Zusammenhänge gibt es



Stärke und Richtung von Zusammenhängen

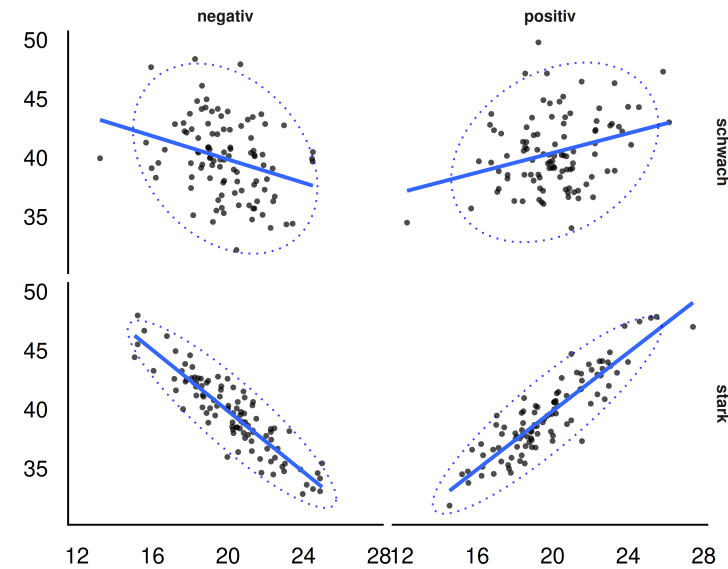


Schmale Ellipsen (“Baguette”) = starker Zusammenhang; breite Ellipsen (“Torte”) = schwacher Zusammenhang.

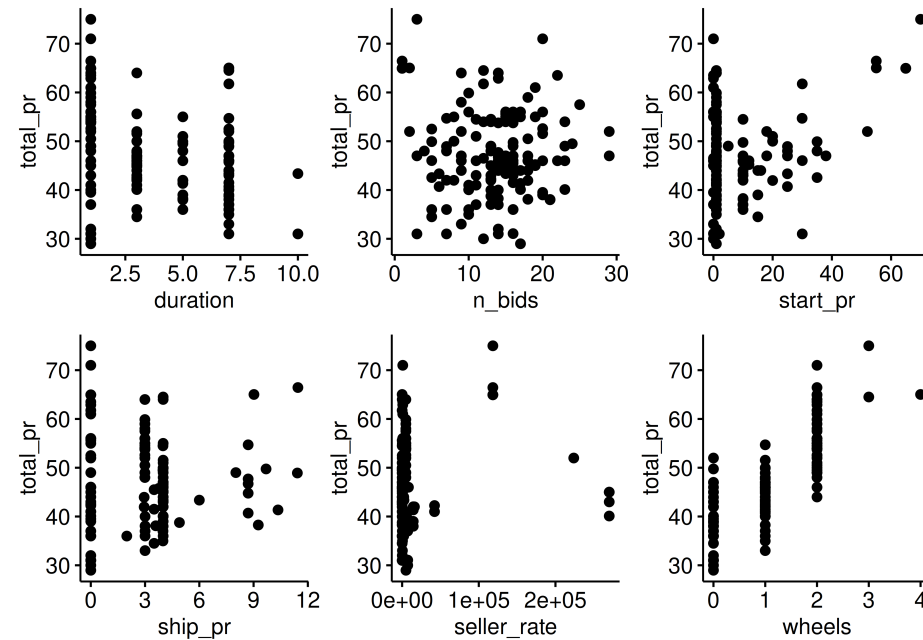
Zusammenhänge verbildlichen

Faustregeln für den Korrelationskoeffizienten

- $r \approx 0$: kein Zusammenhang
- $r \pm .1$: schwacher Zusammenhang
- $r \pm .3$: mittlerer Zusammenhang
- $r \pm .5$: starker Zusammenhang
- $r = 1$: perfekter Zusammenhang



Mehrere Zusammenhänge auf einen Blick



Mit `map()` bauen wir für jede Variable ein `ggscatter()`-Diagramm gegen die Zielvariable `total_pr` und fügen alle mit `ggarrange()` zu einem Panel zusammen — `start_pr` und `wheels` hängen am stärksten mit `total_pr` zusammen.

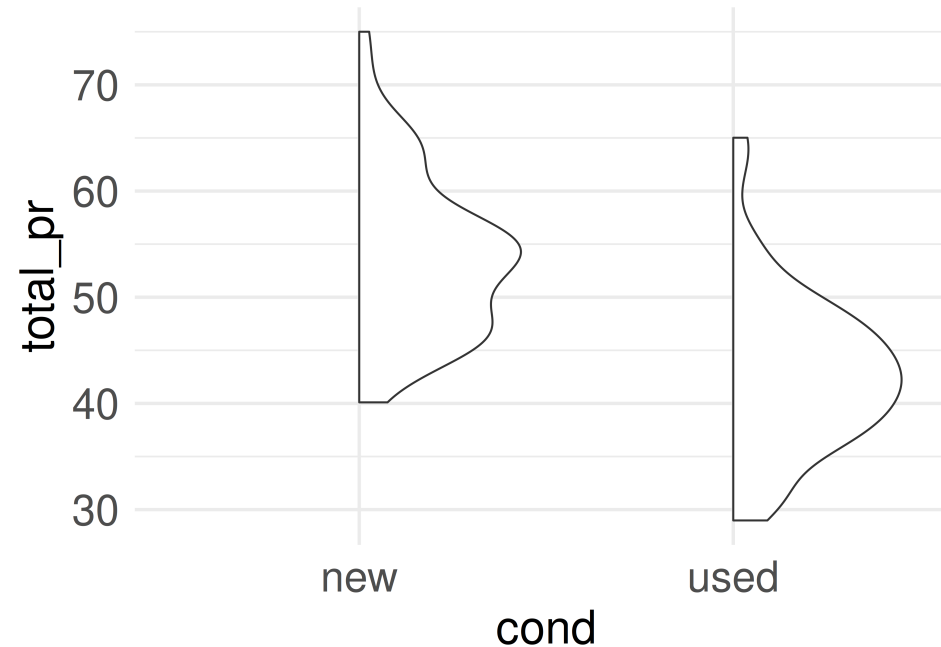
Unterschiede verbildlichen

Unterschiede bei nominalen Variablen

Für nominale Variablen eignen sich Balkendiagramme sowohl für Zusammenhänge als auch für Unterschiede — beides ist letztlich dieselbe Fragestellung, nur anders formuliert.

Tortendiagramme sind dagegen ungeeignet: Flächenvergleiche sind für das Auge schwer.

Unterschiede bei quantitativen Variablen: Boxplot

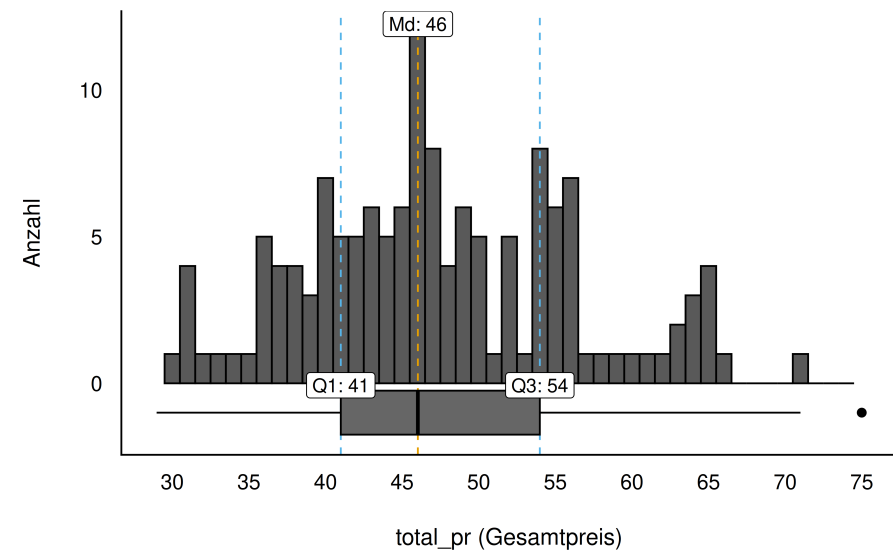


Der Boxplot vereinfacht die Verteilungsdarstellung — kompakter als ein Histogramm, aber mit weniger Detail.

Definition: Boxplot

Definition: Boxplot

Der Boxplot ist eine Vereinfachung bzw. Zusammenfassung eines Histogramms. Damit stellt er ebenfalls eine Verteilung (einer metrischen Variablen) dar.



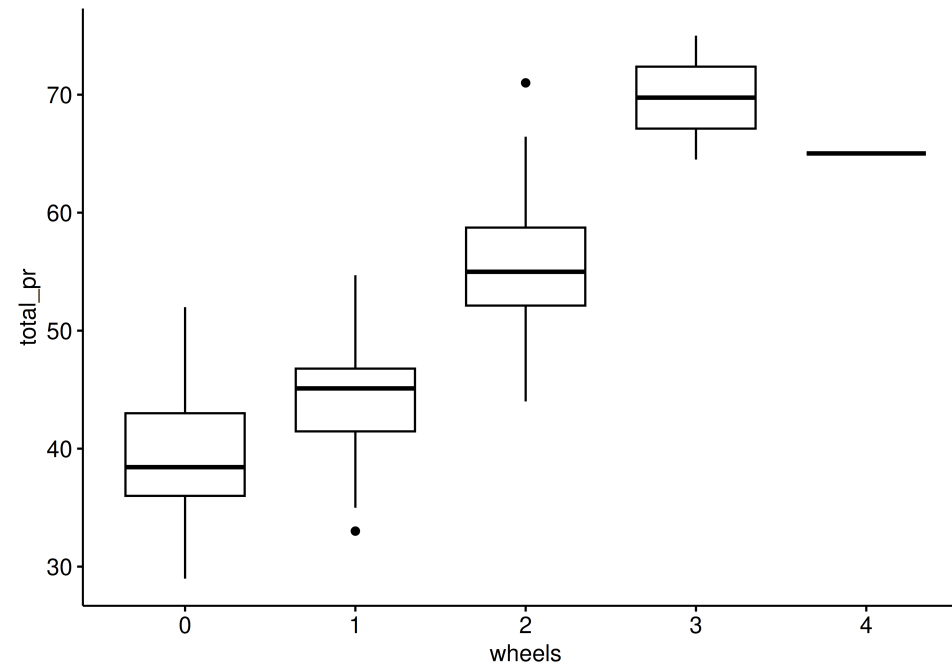
Md: Median; Q1/3: 1./3. Quartil

Unterschiede verbildlichen

Anatomie des Boxplots

1. **Dicker Strich:** Median der Verteilung.
2. **Enden der Box:** 1. und 3. Quartil — die Breite ist der *Interquartilsabstand* (IQR).
3. **Antennen:** Streuung der äußeren 25 % (links/rechts), begrenzt auf das 1,5-fache der Boxlänge.
4. **Einzelne Punkte:** Extremwerte außerhalb der Antennen.
5. Median mittig in der Box + gleich lange Antennen → Verteilung symmetrisch; sonst schief.

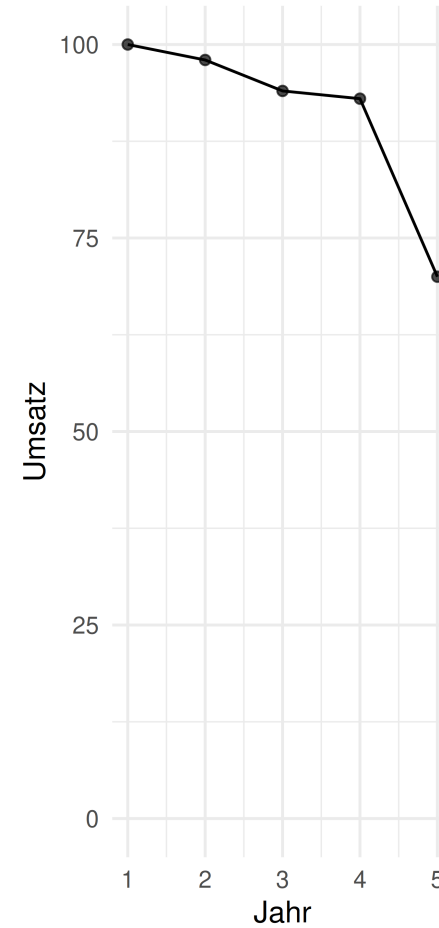
Boxplot: metrisch vs. nominal gruppiert



`factor(wheels)` erzwingt, dass jeder Wert (0, 1, 2, 3, 4) als eigene Gruppe behandelt wird — sonst gruppiert R metrische Variablen automatisch (und nicht immer sinnvoll).

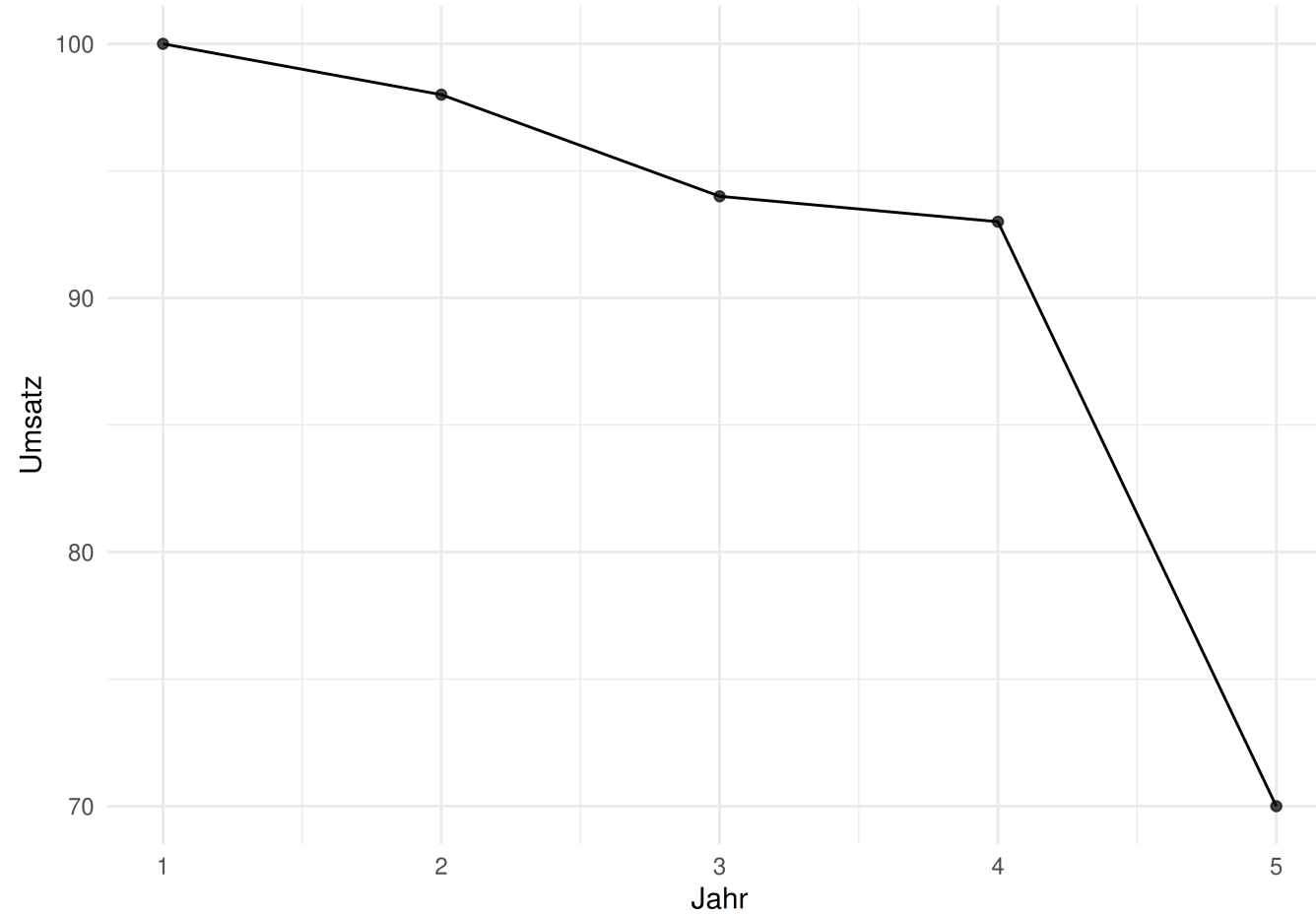
So lügt man mit Statistik

Achsen stauchen und strecken



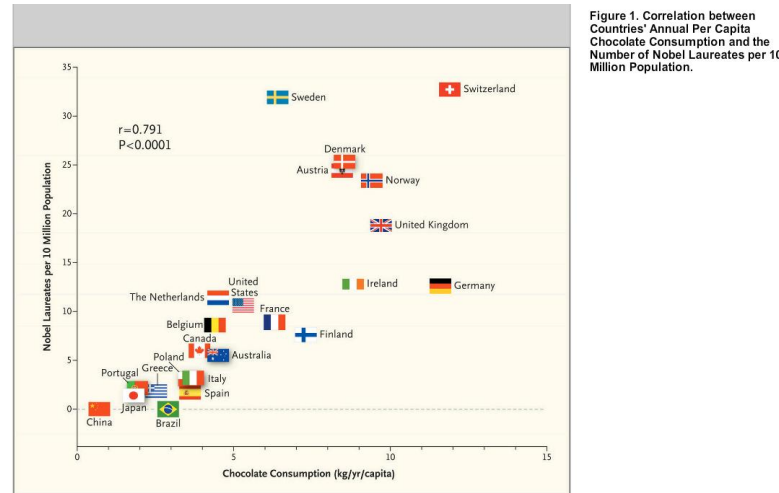
So lügt man mit Statistik

Die Y-Achse abschneiden



So lügt man mit Statistik

Scheinkorrelationen als Kausalität verkaufen

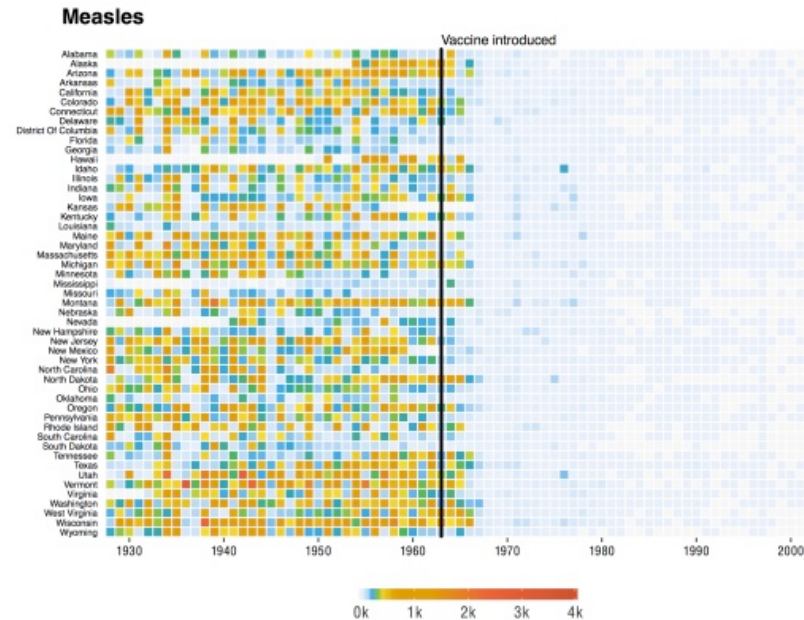


Schokoladenkonsum und Nobelpreise ([Messerli, 2012](#))

Schokoladenkonsum hängt mit der Anzahl der Nobelpreise (pro Land) zusammen — aber nicht ursächlich. Eine Drittvariable (z. B. allgemeiner Entwicklungsstand) erklärt vermutlich beides gleichzeitig.

Praxisbezug

Die Stärke von Datendiagrammen



Häufigkeit von Masern vor und nach Einführung der Impfung ([Moore, 2015](#))

Ein Diagramm kann eine Botschaft (hier: Wirksamkeit der Impfung) auf einen Blick vermitteln — deutlich eindrücklicher als eine Tabelle mit Zahlen.

Tortendiagramme bleiben aber meist eine schlechte Wahl.

Zusammenfassung

- Datendiagramme decken auf, was reine Kennzahlen verbergen (Dino-Datensatz, Anscombes Quartett)
- **Verteilung:** Balkendiagramm (nominal), Histogramm/Dichtediagramm (quantitativ) — bei Normalverteilung: μ und σ bestimmen alles
- **Zusammenhang:** gefülltes Balkendiagramm (nominal), Streudiagramm (metrisch) — Stärke via Ellipsenform/Korrelation
- **Unterschied:** Boxplot fasst eine Verteilung kompakt zusammen (Median, Quartile, Antennen, Extremwerte)
- Vorsicht: Diagramme lassen sich manipulieren (gestauchte Achsen, abgeschnittene Y-Achse, Scheinkorrelationen)

Literaturhinweise

- Wickham (2016) — das Standardwerk zu `ggplot2`, kostenfrei online
- Wilke (2019) — hervorragender, praxisnaher Überblick über Datenvisualisierung mit R
- Fisher & Meyer (2018) — verwandte Perspektive auf gutes Visualisieren

Referenzen

Ainali. (2007). *Standard Deviation Diagram Micro* [Artwork]. <https://commons.wikimedia.org/w/index.php?curid=3141713>

Fisher, D., & Meyer, M. (2018). *Making Data Visual: A Practical Guide to Using Visualization for Insight*. O'Reilly.

Inductiveload. (2008). *Normal Distribution PDF* [Artwork]. https://commons.wikimedia.org/wiki/File:Normal_Distribution_PDF.svg

Messerli, F. H. (2012). Chocolate Consumption, Cognitive Function, and Nobel Laureates. *New England Journal of Medicine*, 367(16), 1562–1564. <https://doi.org/10.1056/NEJMon1211064>

Moore, B. (2015, April 9). *Recreating the Vaccination Heatmaps in R*. Benomics. <https://benjaminlmoore.wordpress.com/2015/04/09/recreating-the-vaccination-heatmaps-in-r/>

Wickham, H. (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. <https://ggplot2.tidyverse.org>

Wilke, C. (2019). *Fundamentals of Data Visualization: A Primer on Making Informative and Compelling Figures*. O'Reilly. <https://clauswilke.com/dataviz/>