Einführung in die Statistik

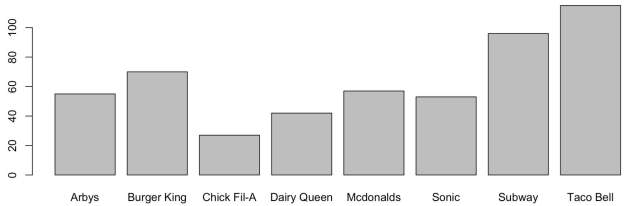
Praktische Übung – Jürgen Hermes – IDH – SoSe 2023

Programm heute

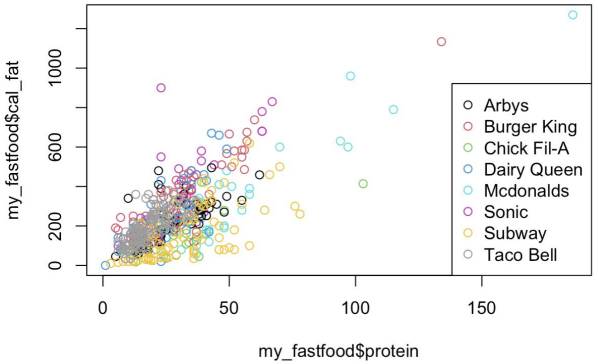
- Besprechung der Klausur
- Visualisierungen (Exkurs)
- Einführung in die deskriptive Statistik

Visualisierungen (Exkurs)

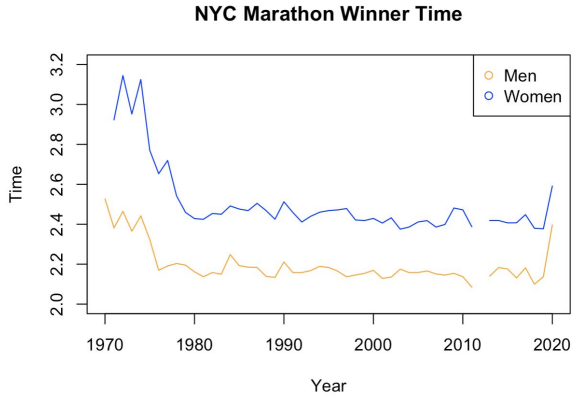
Barplots



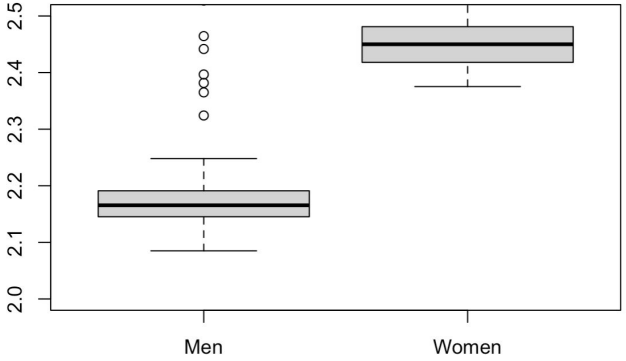
Punktplots



Linienplots

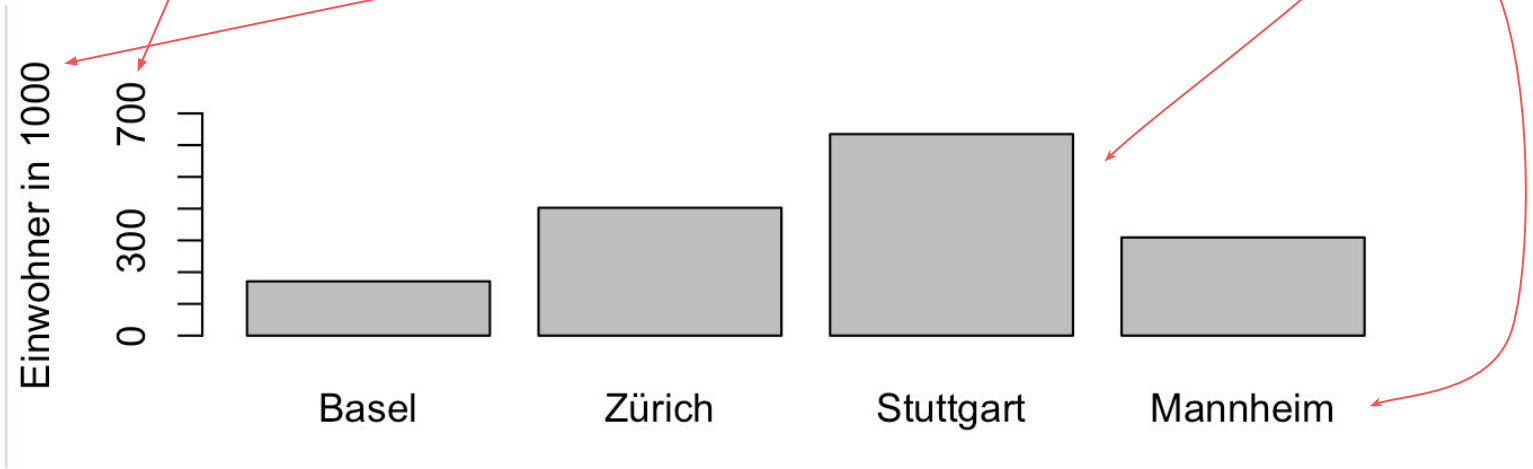


Boxplots



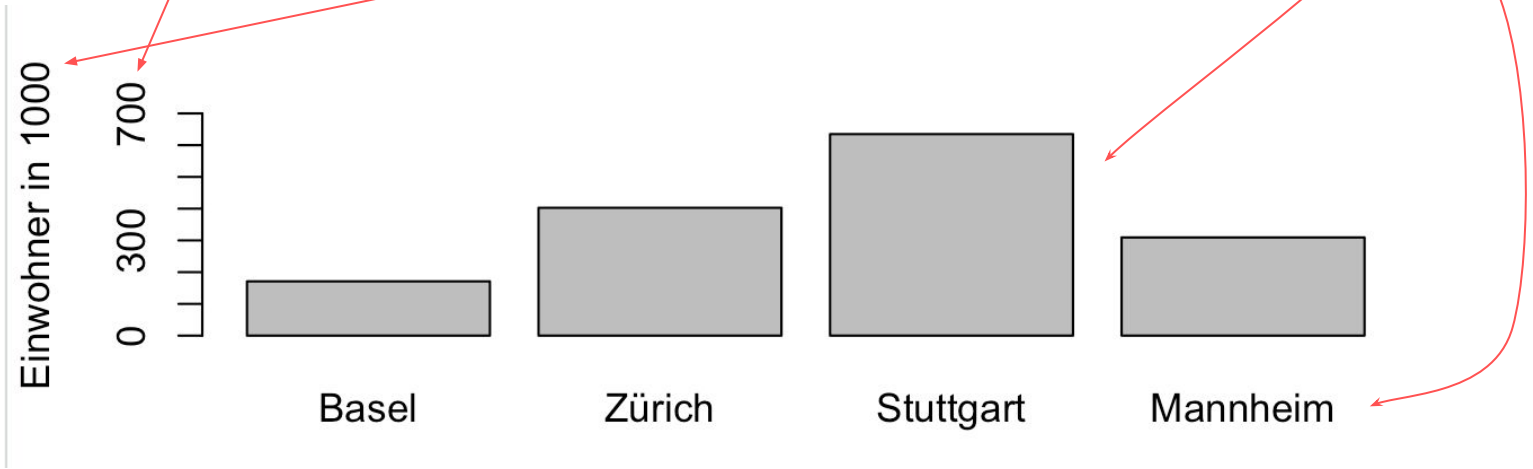
Visualisierung: Barplots

- `barplot(height = cit$einw/1000, names.arg = cit.name, ylim = c(0, 700), ylab= "Einwohner in 1000")`



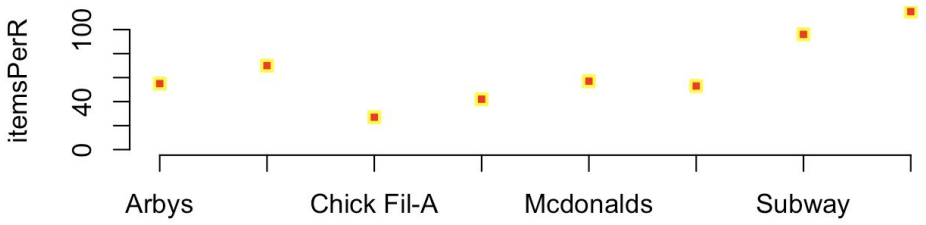
Visualisierung: Barplots

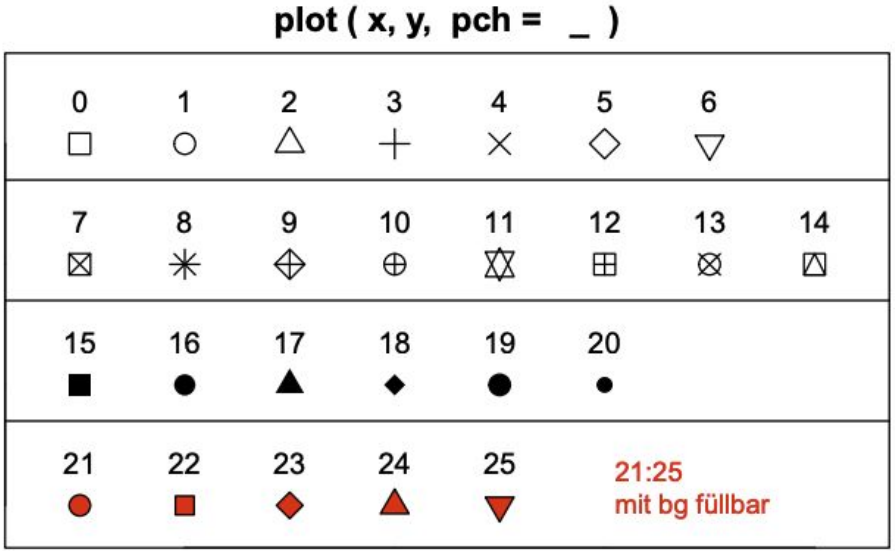
- `barplot(height = cit$einw/1000, names.arg = cit.name, ylim = c(0, 700), ylab= "Einwohner in 1000")`

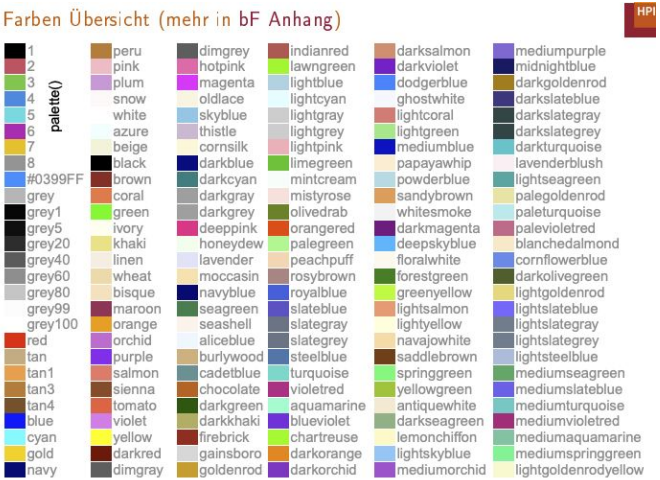


Visualisierung: Punktdiagramme

```
plot(itemsPerR, pch = 22,
     bg = "red", color = "yellow")
```

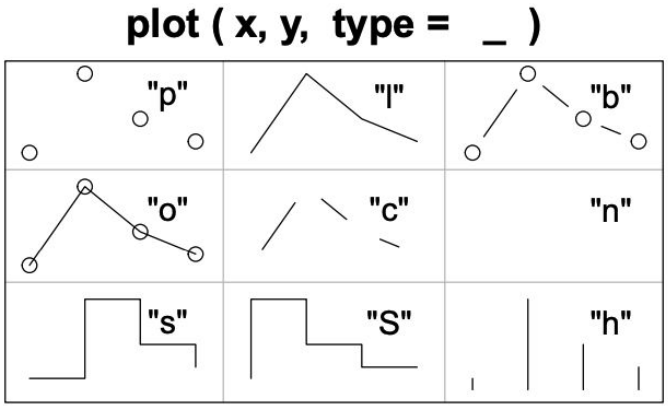


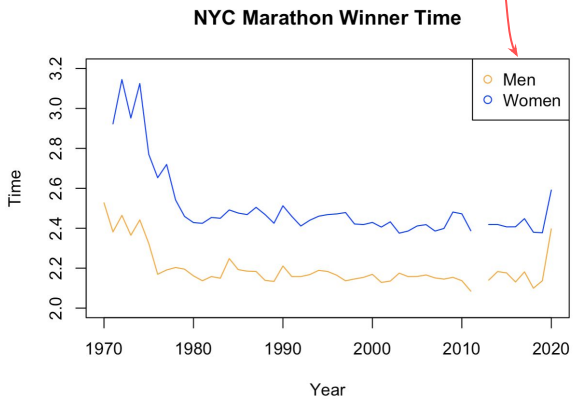




Visualisierung: Liniendiagramme

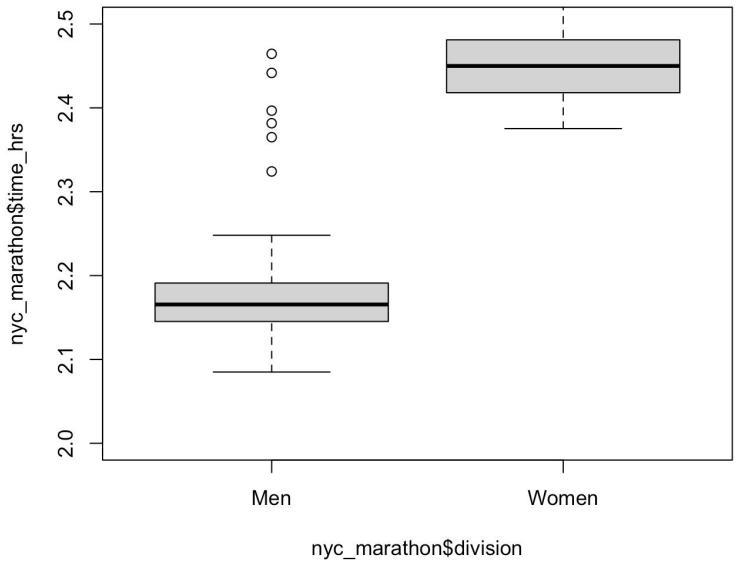
```
plot(y=nycm_women$time_hrs, x=nycm_women$year, col="blue", type = "l", ylim=c(2,3.2),
     xlab = "Year", ylab="Time", main = "NYC Marathon Winner Time")

lines(y=nycm_men$time_hrs, x=nycm_men$year, col="orange")

legend("topright", levels(as.factor(nyc_marathon$division)),
     col =c("orange","blue"), pch = 1)
```





Visualisierung: Boxplots

```
boxplot(nyc_marathon$time_hrs ~ nyc_marathon$division, ylim=c(2,3.2))
```



Deskriptive Statistik

Diese und die folgenden Folien sind erstellt worden von Sascha Wolfer für seinen Kurs “Statistik mit R” an der Uni Basel. Ich nutze sie mit seiner freundlichen Genehmigung. DOI für die Materialien ist 10.5281/zenodo.7431504

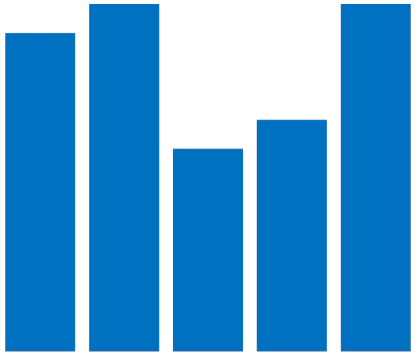
Deskriptive Statistik

- Ziel: **Beschreiben** einer Stichprobe
- Wir tun das für uns und andere!
  - Eindruck von den gesammelten Daten **bekommen**
    - Fehler entdecken
    - Muster entdecken
  - Überblick über die Daten **geben**
    - Offene Darstellung der eigenen Daten im wissenschaftlichen Prozess
- Bei der deskriptiven Statistik bleiben wir eng an der **Stichprobe**.
  - Keine Verallgemeinerung auf die Grundgesamtheit ("Population")

Deskriptive Statistik

- **Visualisierungen** sind ein Mittel der deskriptiven Statistik.
- Visualisierungen brauchen aber numerische Werte.
- Kennzahlen zu Charakteristika von **Verteilungen**.
  - Nominal-/Ordinalskalierte ("**diskrete**") Variablen: Häufigkeitsverteilung
  - Metrisch skalierte ("**kontinuierliche**") Variablen: Dichteverteilung

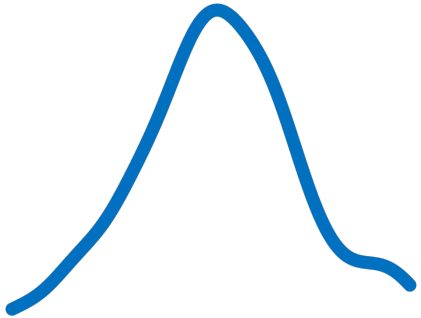
Verteilungen



Häufigkeitsverteilung

Gibt für jeden Wert einer Variable an, wie oft dieser Wert vorkommt.

Beispiel: Häufigkeit von Augenzahlen bei 100 Würfen mit einem Würfel.



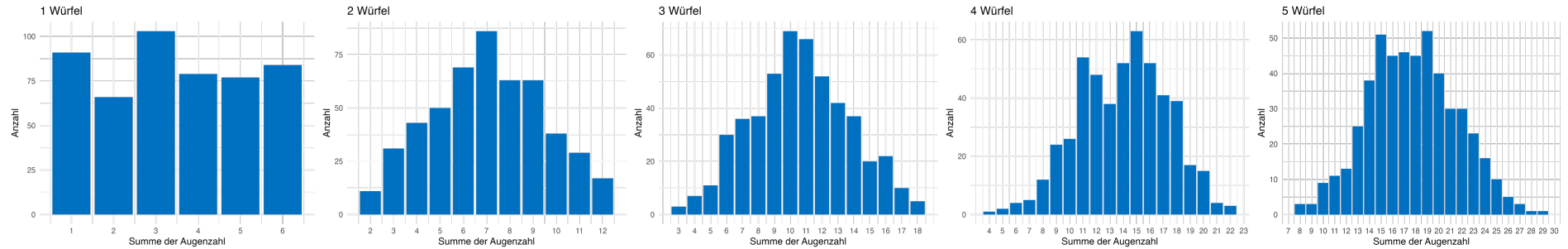
Dichteverteilung

Gibt an, wie viele Werte in einem bestimmten *Bereich* vorkommen.

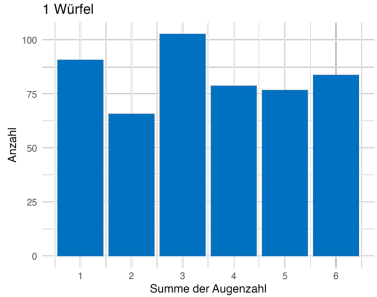
Beispiel: Anzahl von Reaktionszeiten zwischen 200 und 300 msec.

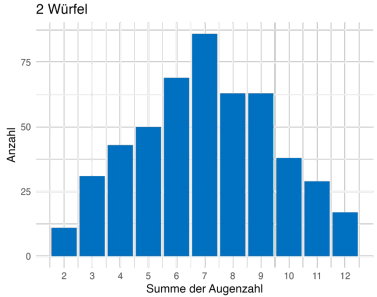
Nebenbemerkung

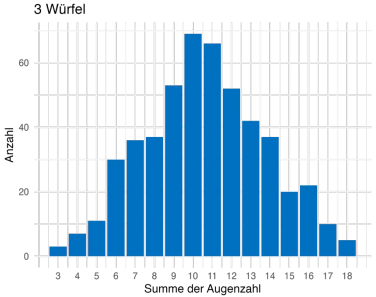
- So groß sind die Unterschiede gar nicht.
  - Eine Häufigkeitsverteilung kann in eine Dichteverteilung übergehen.
- Versuch: Wir würfeln 500x mit einer steigenden Anzahl von Würfeln (1 bis 5) und notieren uns für jeden Wurf die Summe der Augenzahl.

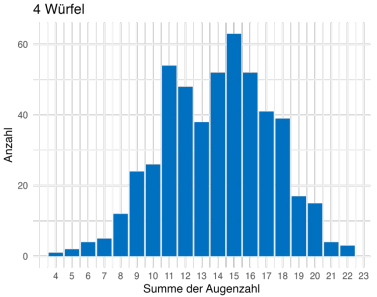


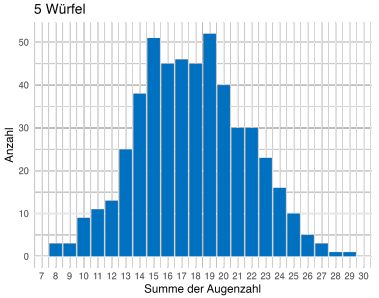
500 Mal



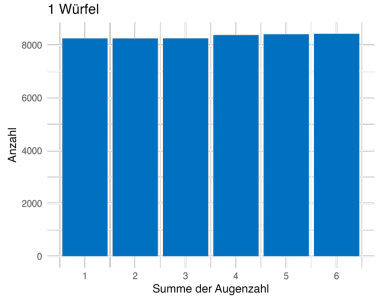


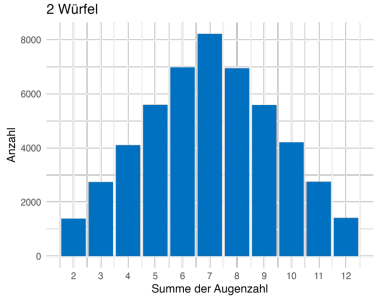


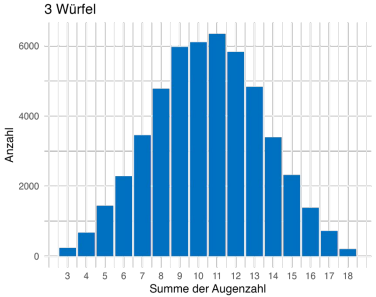


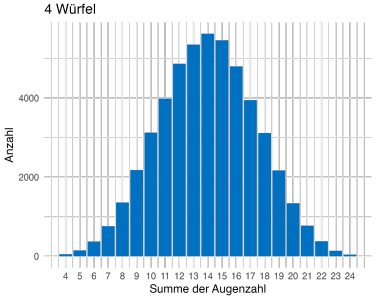


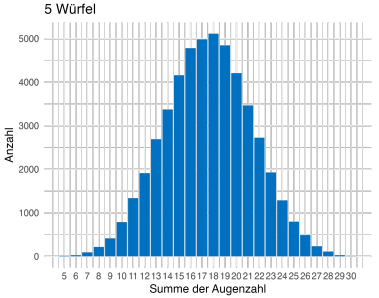
5000 Mal

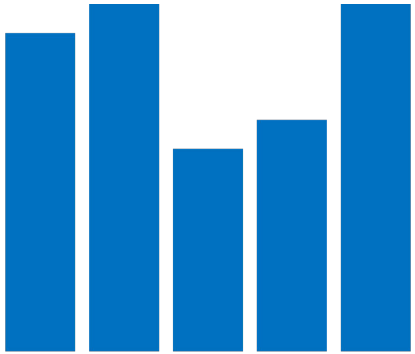








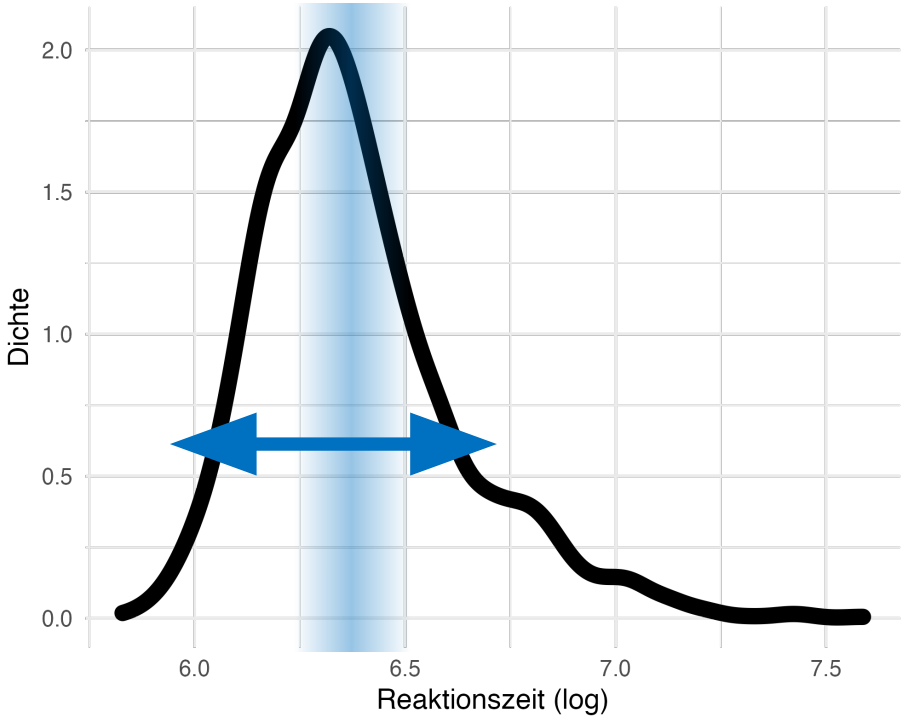




Häufigkeitsverteilung

Dichteverteilung

Kontinuierliche Variablen



- Wie können wir die Verteilung beschreiben?
- Wir wollen wissen:
  - Wo ist die "**Mitte**" der Verteilung?
  - Wie stark **streut** die Variable?

Masse der zentralen Tendenz

Streuungs-/Dispersionsmasse

Maße der zentralen Tendenz

Modus / Modalwert

Der am häufigsten vorkommende Wert

Median

Der Wert, der die Datenreihe in zwei Hälften teilt.

Arithmetisches Mittel

Der Mittelwert: Summe aller Werte geteilt durch die Anzahl

Aufsummieren aller Werte von 1 bis *n*

"x quer" $\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$

***n*: Anzahl Werte**

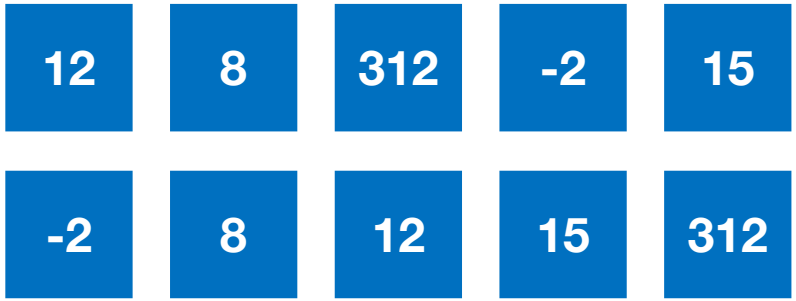
Modus / Modalwert

- Bei diskreten Variablen (Häufigkeitsverteilungen) kann nur der Modus sinnvoll berechnet werden.
  - Welcher Wert kommt am häufigsten vor?
- In R: `which.max(table(<Vektor>))`
- Bei kontinuierlichen Variablen ist der Modalwert hingegen meist sinnlos.
  - Hier wäre es angebrachter, ein bestimmtes **Intervall** anzugeben, in dem die meisten Werte vorkommen.

Median

Bei einer geraden Anzahl an Werten wird die Mitte der mittleren beiden Werte als Median angenommen.

- Der Median liegt in der Mitte aller Werte.
- Oder: Der Median ist der Wert, der alle vorkommenden Werte in zwei Hälften teilt.
  - Über und unter dem Median befinden sich 50% aller Werte.
- Kann ab der Ordinalskala berechnet werden.




```
median(c(12, 8, 312, -2, 15))
```

Mittelwert

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

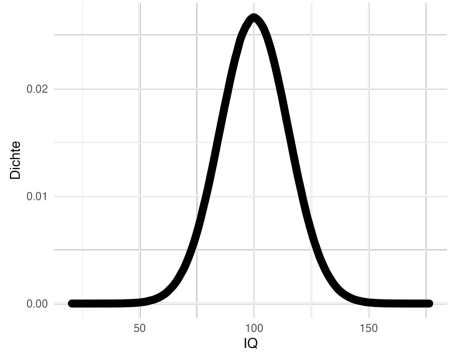
- Summe aller Werte geteilt durch die Anzahl.
- Kann nur für metrisch skalierte Variablen berechnet werden.
  - Beispiel: Was sollte der Mittelwert aus den Erstsprachen Französisch, Englisch und Italienisch sein?

12 + **8** + **312** + **-2** + **15** = **345**

345 / **5** = **69**

```
mean(c(12, 8, 312, -2, 15))
```

Median und Mittelwert




```
median(c(12, 8, 312, -2, 15))  → 12
mean(c(12, 8, 312, -2, 15))    → 69
```

- Beides sind Maße der zentralen Tendenz, können aber zu unterschiedlichen Ergebnissen führen.
- Der Mittelwert ist deutlich anfälliger gegenüber **Ausreisserwerten**!
  - Oben: `312`

Beachten Sie aber auch das Argument `trim`. Siehe `?mean`

- Identisch sind Median und Mittelwert nur bei exakt symmetrischen Verteilungen (z. B. Normalverteilung).

Fehlende Werte

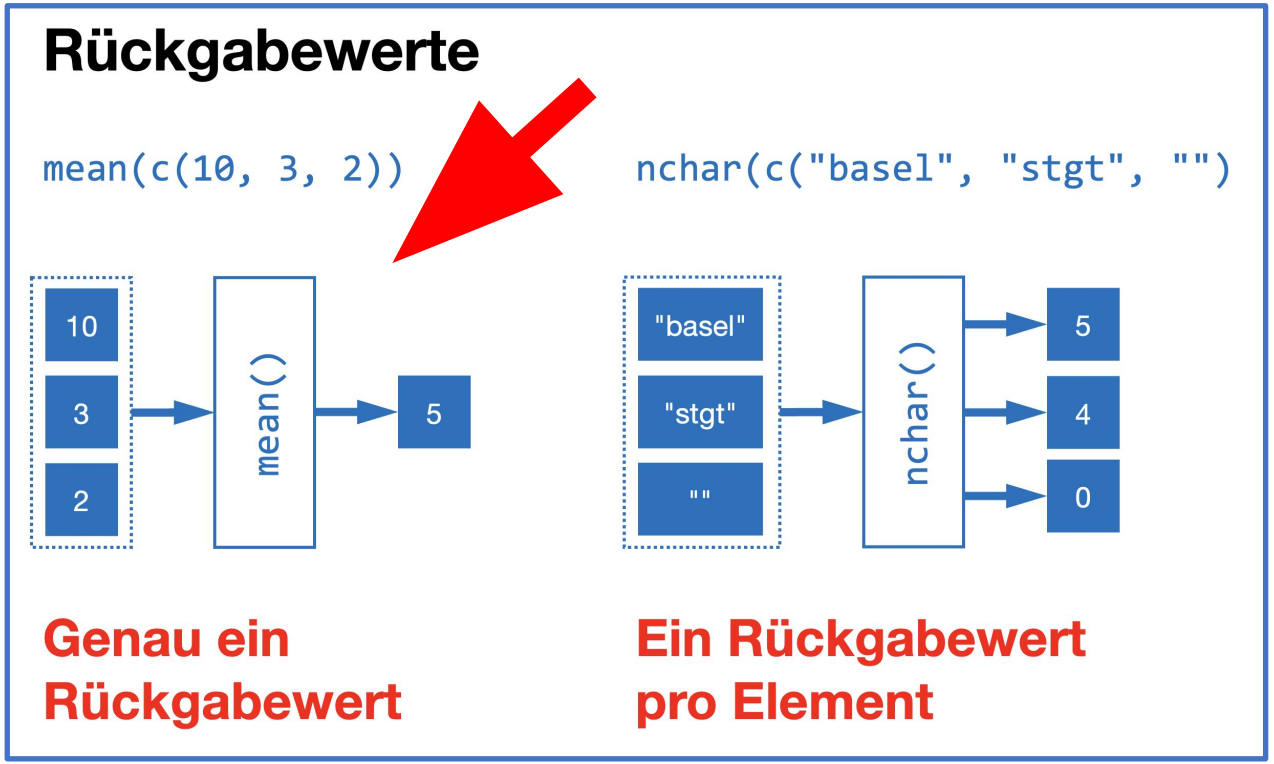
- Enthält eine Datenreihe einen fehlenden Wert, gibt R für `median()` und `mean()` ebenfalls NA zurück.
- Argument `na.rm = T` entfernt erst die fehlenden Werte und berechnet dann den Median/Mittelwert (`na.rm` = *NA remove*).
- Das gilt auch für die Funktionen für Standardabweichung und Varianz (s. später)!

Berechnung nach Gruppen

- Mit `tapply()` kann man Berechnungen gruppieren.
  - `tapply(<Vektor>, <Gruppierungsvektor>, <Funktion>)`
- Zum Beispiel: Mittelwert von Spalte `WLen` gruppiert nach `POS`
  - `tapply(data$WLen, data$POS, mean)`
- Nach `<Funktion>` können Argumente folgen, die `<Funktion>` übernimmt (bspw. `na.rm = T`).

Berechnung nach Gruppen

`tapply()` wird typischerweise nur mit Funktionen angewendet, die genau **einen** Rückgabewert haben!



Hausaufgabe

- Laden Sie die Datei `Exp.tsv` (finden Sie in ILIAS) in R.
- Berechnen Sie Mittelwert und Median für die Spalte `RT`.
  - `NA` ist keine gültige Lösung!
- Identifizieren Sie den Grund dafür, dass Median und Mittelwert so weit auseinanderliegen (schreiben Sie diesen in einen Kommentar).
- Erstellen Sie ein einfaches Diagramm der Daten in der Spalte RT: `plot(<Spalte>)`
- Berechnen Sie den Mittelwert für jede Bedingung (Spalte `Bedingung`)