Einführung in die Statistik

Praktische Übung – Jürgen Hermes – IDH – SoSe 2023

Programm heute

- Wiederholung der Einführung in die Deskriptive Statistik
- Mittelwerte
- Dispersionsmaße
- Zusammenfassung / Hausaufgabe

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

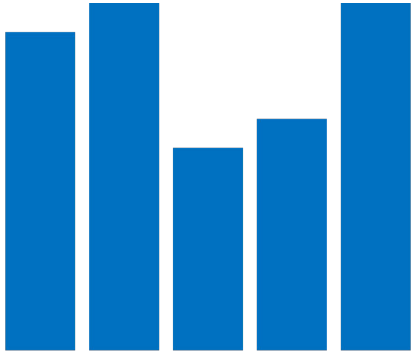
Deskriptive Statistik

- 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")

Zusammenfassung Skalen / Mittelwerte / Dispersionsmaße

Niveau	**Häufigkeit**	**Rangfolge**	**Abstand**	**Nullpunkt**	**Zentrale Tendenz**	**Dispersions-maß**
Nominal	messbar				Modus	
Odinal	messbar	messbar			Mod + Median	IQA / Spannweite
Intervall	messbar	messbar	messbar		Mod + Med + arithmetisches Mittel	IQA / Spannw. + Varianz / SA
Verhältnis	messbar	messbar	messbar	absolut	Mod + Med + arithmetisches Mittel	IQA / Spannw. + Varianz / SA

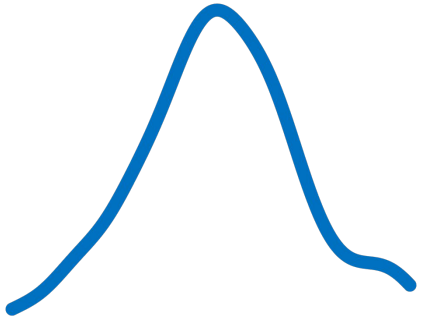
Verteilungen



Häufigkeitsverteilung

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

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



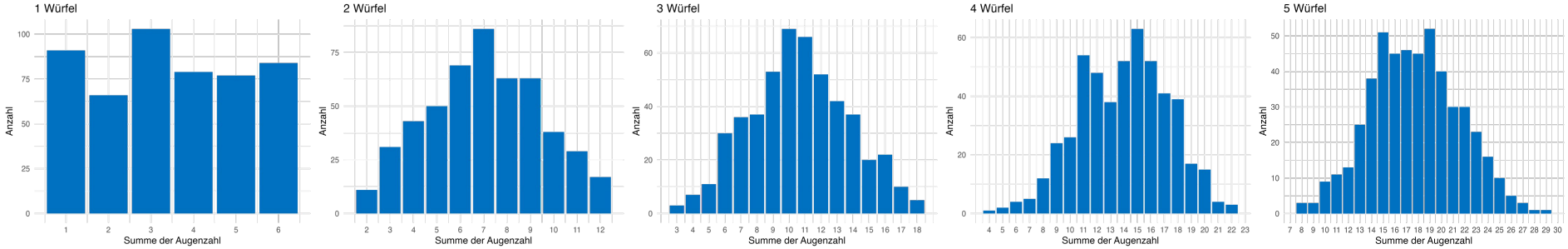
Dichteverteilung

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

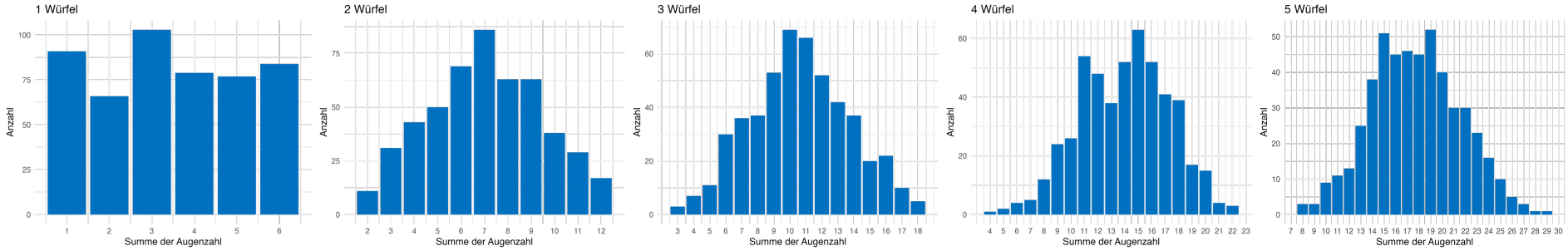
Beispiel: Anzahl von Reaktionszeiten zwischen 200 und 300 msec.

Nebenbemerkung

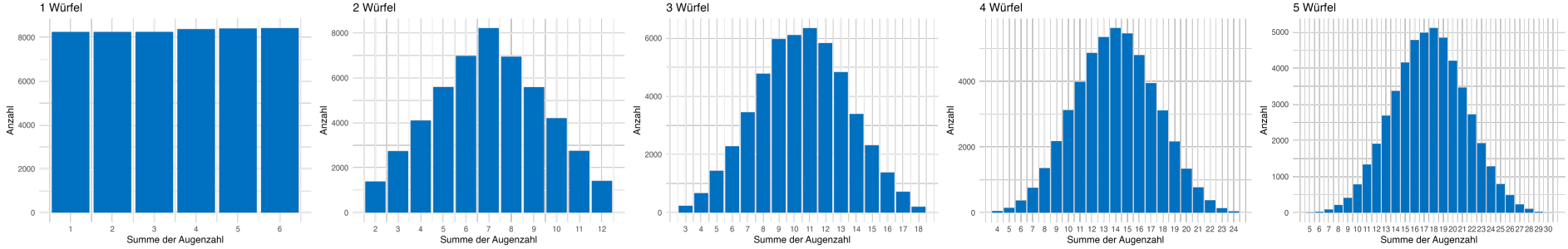
- So groß sind die Unterschiede gar nicht.
  - Eine Häufigkeitsverteilung kann in eine Dichteverteilung übergehen.
- Versuch: Wir würfeln 500x mit einer steigenden Anzahl von Würfeln (1 bis 5) und notieren uns für jeden Wurf die Summe der Augenzahl.

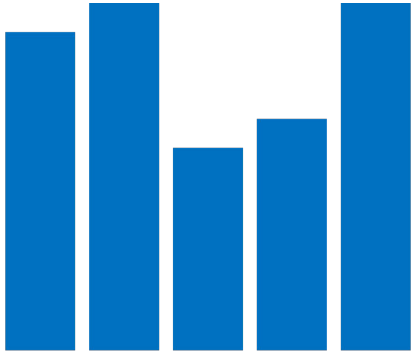


500 Mal

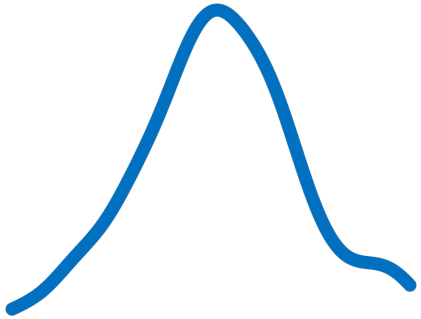


5000 Mal



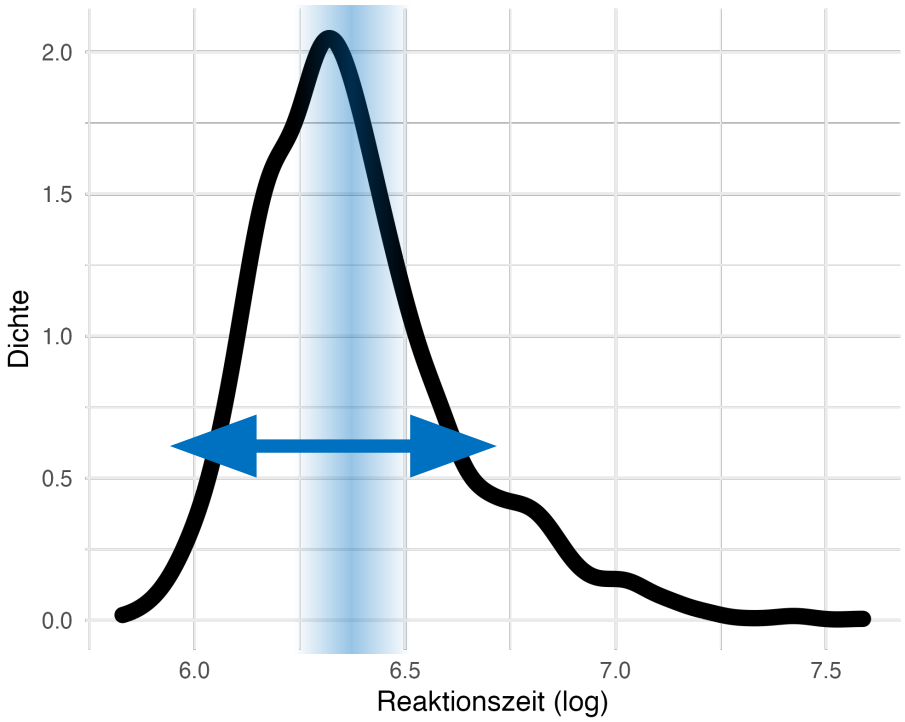


Häufigkeitsverteilung



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	Median	Arithmetisches Mittel
Der am häufigsten vorkommende Wert	Der Wert, der die Datenreihe in zwei Hälften teilt.	Der Mittelwert: Summe aller Werte geteilt durch die Anzahl

Aufsummieren aller Werte von 1 bis *n*

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

***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.

12	8	312	-2	15
-2	8	12	15	312

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

Mittelwert

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

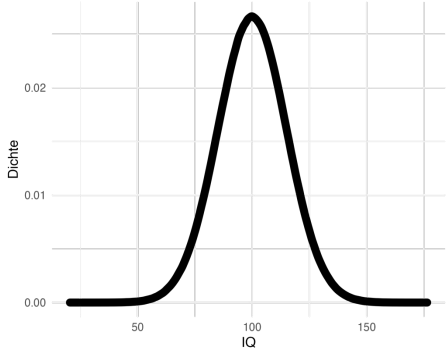
- Summe aller Werte geteilt durch die Anzahl.
- Kann nur für metrisch skalierte Variablen berechnet werden.
  - Beispiel: Was sollte der Mittelwert aus den Erstsprachen Französisch, Englisch und Italienisch sein?

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

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

Fehlende Werte

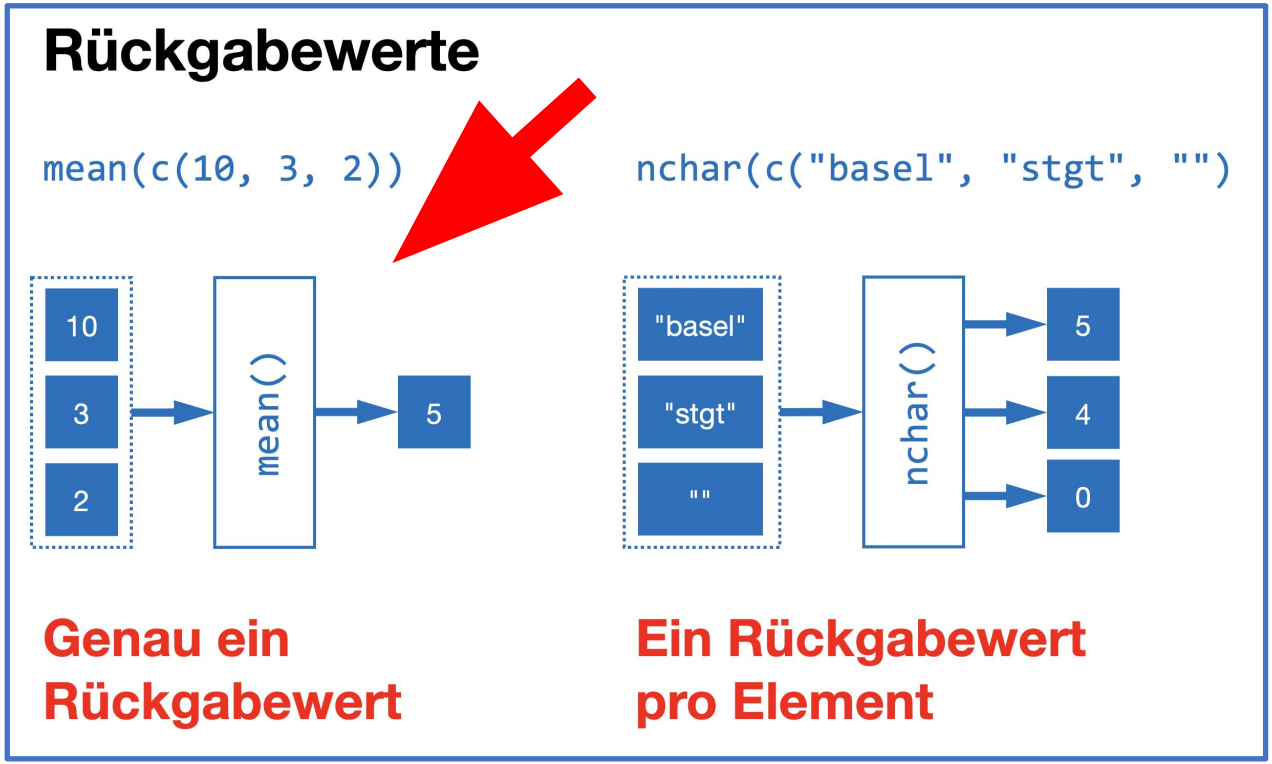
- Enthält eine Datenreihe einen fehlenden Wert, gibt R für `median()` und `mean()` ebenfalls NA zurück.
- Argument `na.rm = T` entfernt erst die fehlenden Werte und berechnet dann den Median/Mittelwert (`na.rm` = *NA remove*).
- Das gilt auch für die Funktionen für Standardabweichung und Varianz (s.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

Streuungsmaße

Interquartilabstand

Der Abstand, der die mittleren 50% aller Werte umfasst

Varianz

Die aufsummierten quadrierten Abweichungen vom Mittelwert, geteilt durch die Anzahl der Werte

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

Standardabweichung

Die Quadratwurzel aus der Varianz

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

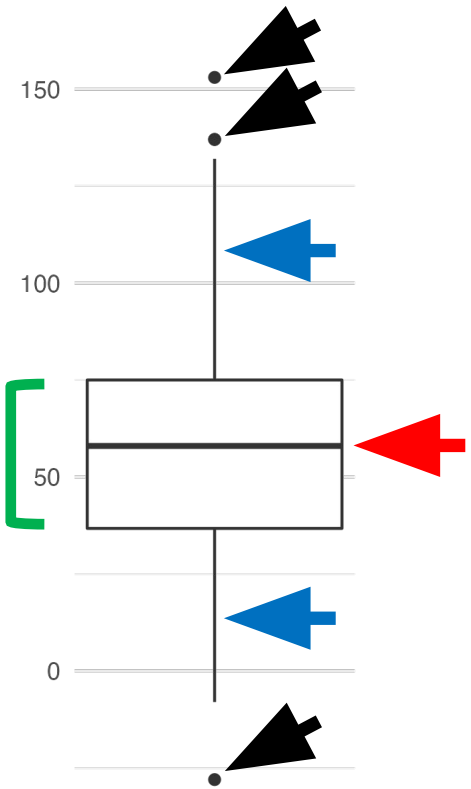
Interquartilabstand (IQR)

- Der Median befindet sich in der Mitte aller beobachteten Werte. Unter und über dem Median sind somit 50% aller Werte.
- Der Interquartilabstand umfasst die **mittleren 50%** aller Werte.
- Je weiter die Werte streuen, desto grösser muss dieser Abstand sein, um die mittleren 50% zu erfassen.

Interquartilabstand (IQR) und Spannweite

- In R: `IQR(<Vektor>)`
- IQR = interquartile range
- Man kann zusätzlich die **komplette** Spannweite / Range der Daten angeben, also Maximalwert minus Minimalwert.
  - `range()` gibt einen Vektor mit zwei Werten aus: Minimal- und Maximalwert.
  - Mit `diff(range(<Vektor>))` bekommen Sie den **Abstand** zwischen Minimal- und Maximalwert.

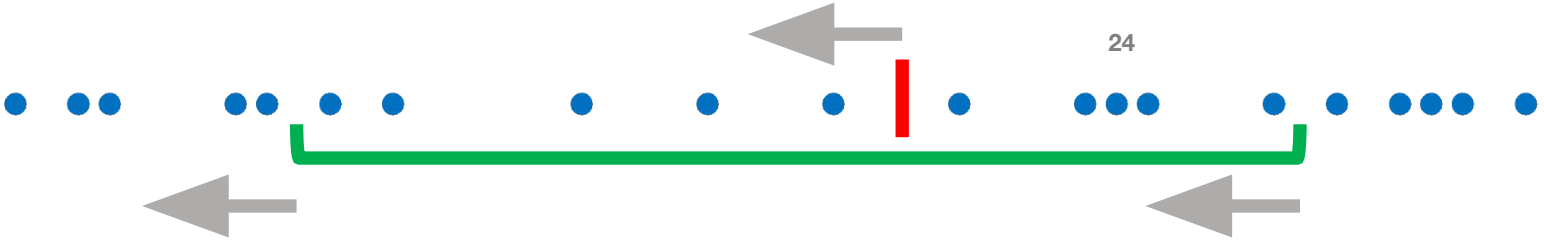
Boxplot



- Median und Interquartilabstand sind Bestandteil des Boxplots.
  - Median: Horizontale Linie in der Box
  - IQR: Höhe der Box
- "Hinges": 1,5*IQR, bis zum letzten beobachteten Wert in diesem Abstand
- Outlier: Alle Punkte, die nicht innerhalb IQR +/- 1,5*IQR liegen.
- Boxplot zeigt viele Informationen auf einmal und bleibt dabei einigermaßen übersichtlich.
- In R: `boxplot(x)`

Quartile und Perzentile

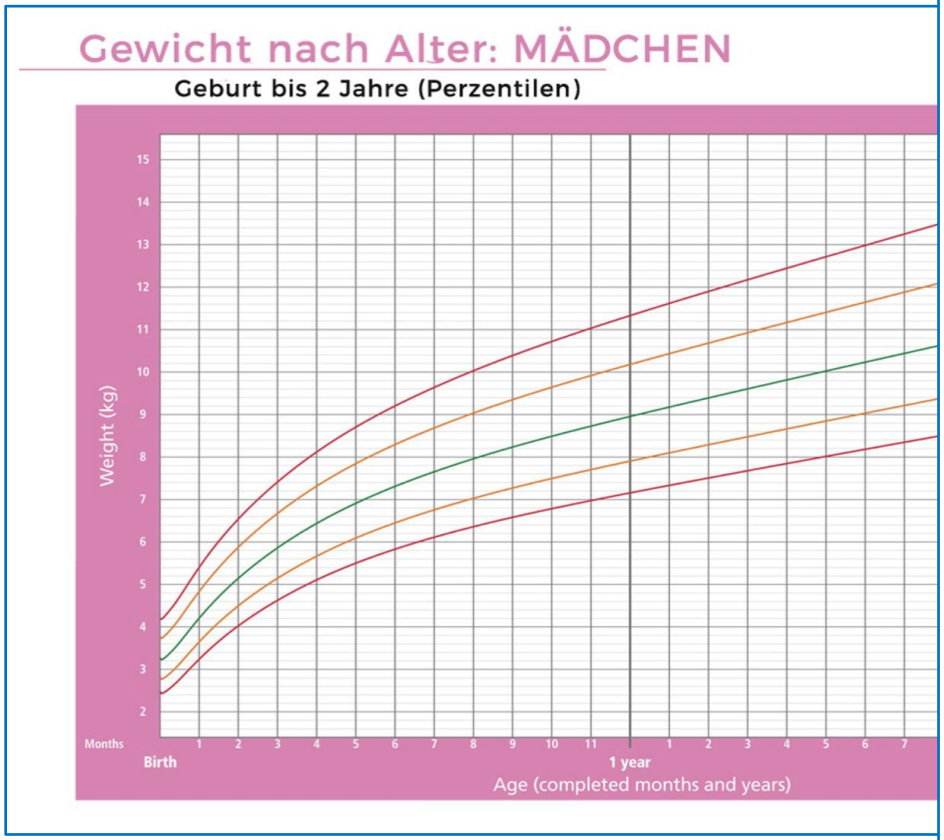
- Nochmal: **Median** teilt die Datenpunkte in zwei Hälften. Unter dem Median sind also 50% aller Daten.
- Unter der **unteren Grenze** des IQR sind 25% aller Daten.
- Unter der **oberen Grenze** sind 75% aller Daten.
- Man nennt die Grenzen deshalb auch erstes und drittes **Quartil**.
- Haben Sie eine Idee, wie man den Median noch nennt?

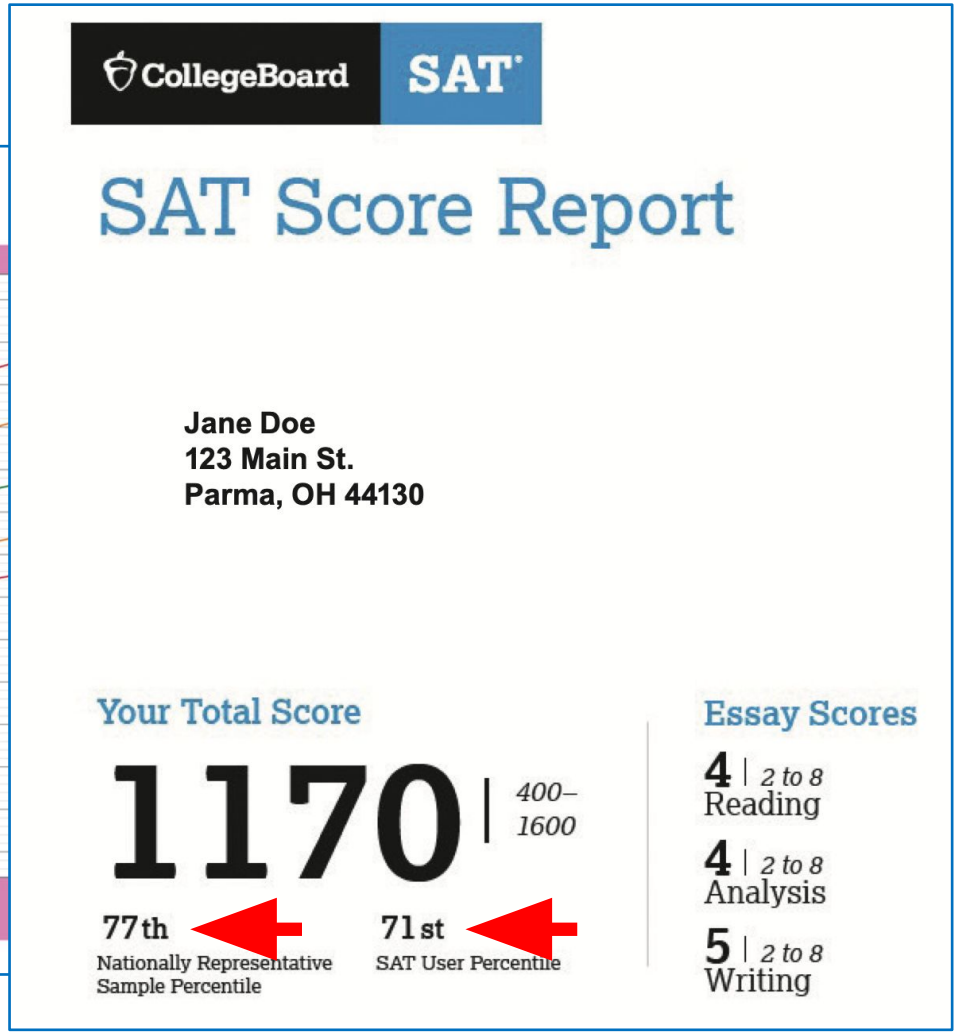


Quartile und Perzentile

- Verallgemeinerbar auf alle "Abschnitte" außer Quartilen.
  - Z. B.: Punkt, unter dem 42% aller Daten liegen.
- Name: **Perzentile** (auch: **Quantile**)
- Definition: Das *x*-te Perzentil ist jener Punkt, unter dem *x* Prozent aller Daten liegen.
- `quantile(X=Vektor, probs = c(quartile))`

Perzentile in der Welt





Varianz

- Idee: Daten streuen mehr, wenn die einzelnen Werte weiter von ihrem Mittelwert entfernt sind.
- Wir können also …
  1. die Abweichungen von jedem Wert zum Mittelwert berechnen,
  2. diese aufsummieren
  3. und dann durch die Anzahl der Werte teilen.
- Ergebnis: Durchschnittliche Abweichung zum Mittelwert
- Weil wir große Abweichungen stärker gewichten möchten, quadrieren wir zusätzlich die Abweichungen in 1.

Varianz: Beispielrechnung

8,2

MW

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

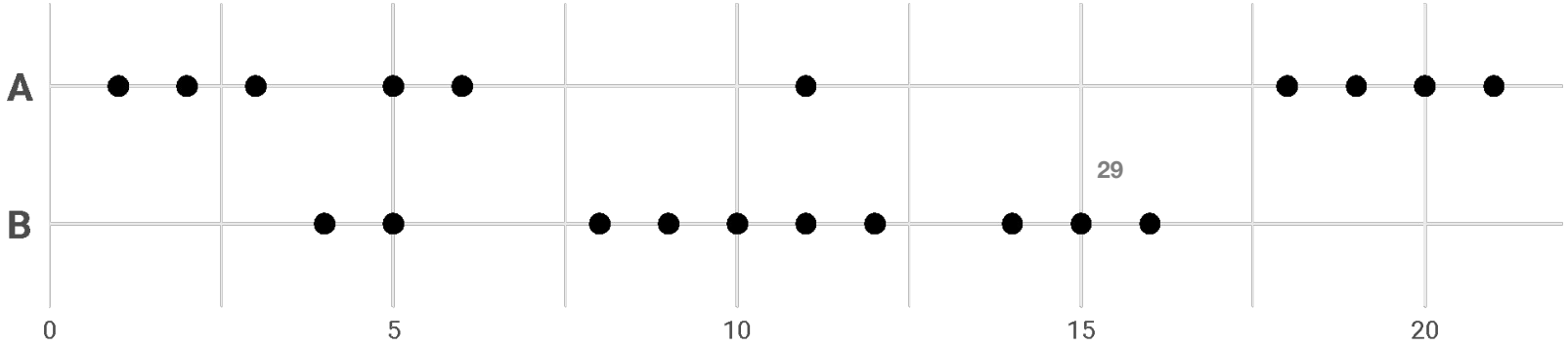
x	Abw. zum MW	quadriert
5	-3,2	10,24
10	1,8	3,24
8	-0,2	0,04
7	-1,2	1,44
11	2,8	7,84

Summe: 22,8 / 5 = s^2: 4,56

Varianz

Varianz für Datenreihe A: 66,5
Varianz für Datenreihe B: 16,3

Fällt Ihnen etwas auf, wenn Sie die Größenordnung der Varianz mit den Messwerten vergleichen (insb. in Datenreihe A)?



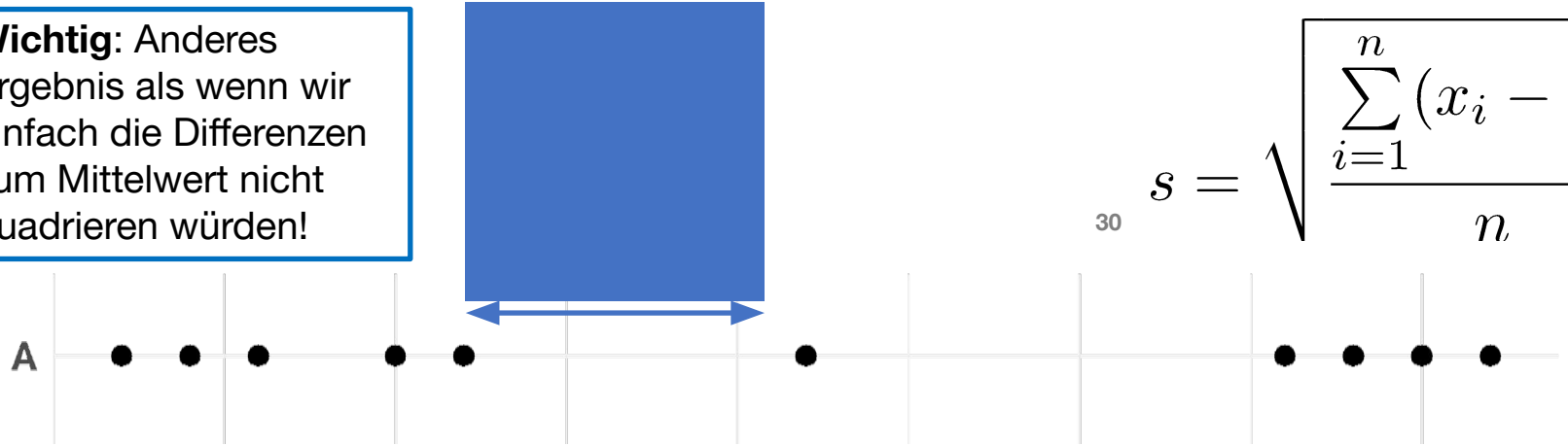
Varianz und Standardabweichung

$$s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

- Problem: Durch die Quadrierung der Differenzen zum Mittelwert hat die Varianz eine andere Skalierung als die Datenpunkte.
- Lösung: Ziehen der Quadratwurzel → Standardabweichung *s*

Wichtig: Anderes Ergebnis als wenn wir einfach die Differenzen zum Mittelwert nicht quadrieren würden!

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$



Dispersionsmaße in R

- Interquartilabstand: `IQR(x)`, Range: `range(x)`
- Perzentile: `quantile(x, <Wert>)` – Prozentwert zwischen 0 und 1
- `summary(x)` gibt 6 Werte aus:
  - Minimum & Maximum
  - 1. Quartil, 2. Quartil (Median), 3. Quartil
  - Mittelwert
- Varianz: `var(x)`
- Standardabweichung: `sd(x)` – *standard deviation*
- Achtung: `var(x)` und `sd(x)` werden mit $n - 1$ im Nenner berechnet (Schätzung von Populationsparametern).

Zusammenfassung

- In der deskriptiven Statistik beschreiben wir die **Stichprobe**, die wir gesammelt haben.
- Wir können unterscheiden zwischen Häufigkeitsverteilungen (**diskrete** Variablen) und Dichteverteilungen (**kontinuierliche** Variablen).
- Bei kontinuierlichen Variablen wollen wir wissen:
  - Wo ist die Mitte der Verteilung? → **Maße der zentralen Tendenz**
  - Wie stark streut die Variable? → **Streuungs-/Dispersionsmaße**

Zusammenfassung

- Maße der zentralen Tendenz:
  - **Modus / Modalwert**
  - **Median**
  - **Arithmetischer Mittelwert**
- Modalwert ist nur für diskrete Variablen sinnvoll.
- Mittelwert ist anfälliger für **Ausreisserwerte** als der Median.

Zusammenfassung

- Streuungsmaße:
  - **Interquartilabstand (IQR) / Spannweite**
  - **Varianz**
  - **Standardabweichung**
- **Quartile** teilen Daten in Viertel.
- Unter dem *x*ten **Perzentil** liegen *x* Prozent aller Datenpunkte.
  - Auch "**Quantil**" genannt.

Zusammenfassung Skalen / Mittelwerte / Dispersionsmaße

Niveau	**Häufigkeit**	**Rangfolge**	**Abstand**	**Nullpunkt**	**Zentrale Tendenz**	**Dispersions-maß**
Nominal	messbar				Modus	
Odinal	messbar	messbar			Mod + Median	IQA / Spannweite
Intervall	messbar	messbar	messbar		Mod + Med + arithmetisches Mittel	IQA / Spannw. + Varianz / SA
Verhältnis	messbar	messbar	messbar	absolut	Mod + Med + arithmetisches Mittel	IQA / Spannw. + Varianz / SA

Begriffe

Deskriptive Statistik

Dispersionsmaße

Spannweite / Range

Stichprobe

Modus / Modalwert

Quartil

Population

Median

Perzentil

Verteilungen

Mittelwert

Boxplot

Maße der zentralen Tendenz

Interquartilabstand (IQR)

Varianz

Standardabweich.

Hausaufgabe 2

- Laden Sie die Datei `Exp.csv` (wenn noch nicht geschehen) aus ILIAS in R
- Berechnen Sie IQR, Spannweite, Varianz und Standardabweichung für die Spalte `RT`.
- Lassen Sie sich eine Summary der Spalte geben.
- Erstellen Sie einen Boxplot der Spalte.
  - Versuchen Sie einen Boxplot unter Ausschluss des Ausreissers zu plotten.
- Bonusaufgabe: Berechnen Sie die Standardabweichung mit *n* statt mit *n* - 1 im Nenner (Lösung: `1257.265`).