Mathematik für Biologen

Prof. Dr. Rüdiger W. Braun
http://blog.ruediger-braun.net

Heinrich-Heine-Universität Düsseldorf

05. Dezember 2014

Termine

- Mittwoch, 10.12.: Doppelstunde Vorlesung, Ausgabe von Übungsblatt 8
- Freitag, 12.12.: Doppelstunde Vorlesung
- Mittwoch, 17.12.: Vorlesung & Übung, Rückgabe von Blatt 6
- Freitag, 19.12: Doppelstunde Vorlesung
- *Weihnachtspause*

Teil I

Deskriptive Statistik

Literaturempfehlungen

- Rudolf, Kuhlisch: *Biostatistik*
- McKillup: *Statistics Explained*
- Whitlock, Schluter: *The Analysis of Biological Data*
- Köhler, Schachtel, Voleske: *Biostatistik* (gibt es auch elektronisch unter `http://dx.doi.org/10.1007/978-3-540-37712-2`)
- Henze: *Stochastik für Einsteiger* (mathematischer als die anderen Titel)

Alle diese Werke enthalten weit mehr Stoff als die Vorlesung.

Grundbegriffe

Grundgesamtheit	(Population)
Merkmal	(Variable)
Ausprägung	(Realisierung)

- die Elemente der Grundgesamtheit sind Träger von Merkmalen
- die Merkmale haben verschiedene Ausprägungen
- jedes Element der Grundgesamtheit besitzt für jedes Merkmal nur eine Ausprägung

Typen von Merkmalen

- *Quantitatives Merkmal:*
  zahlenmäßig erfassbar; Zahlenwerte besitzen Bedeutung
  - *stetiges Merkmal:*
    Zahlenwerte variieren kontinuierlich (z.B. Gewicht)
  - *diskretes Merkmal:*
    Skala ohne Zwischenwerte (z.B. Anzahl)
- *Qualitatives Merkmal:*
  alle anderen

Beispiele zu den Grundbegriffen

Grundgesamtheit: alle Bäume einer Baumschule

- Merkmal: Art (qualitatives Merkmal)
  Ausprägung: Fichte
- Merkmal: Größe (quantitativ stetiges Merkmal)
  Ausprägung: 3.38 m
- Merkmal: Pflanzdatum (quantitativ diskretes Merkmal)
  Ausprägung: 9.10.2003

Beispiel: Matrikelnummer

Die Matrikelnummer ist ein qualitatives Merkmal, da sie keine eigenständige Bedeutung hat

Rundung von Merkmalswerten

- kontinuierliche quantitative Merkmalswerte werden auf Messgenauigkeit gerundet
- diskrete quantitative Merkmalswerte sind auf beliebig viele Stellen genau
- "Messwerte wurden im Abstand von 2 Stunden erhoben" bedeutet: Die Experimentatoren haben sich bemüht, den Abstand von zwei Stunden so genau wie möglich zu realisieren
- aggregierte Werte (z. B. das arithmetische Mittel) werden mit zwei zusätzlichen Stellen Genauigkeit angegeben
- wir werden im Abschnitt über Konfidenzintervalle sehen, wie man die Genauigkeit aggregierter Daten feststellt und sauber beschreibt
- Zwischenrechnungen immer so genau wie (mit vernünftigem Aufwand) möglich

Stichproben

- Eine *Stichprobe* ist eine **zufällig** gewonnene Teilmenge aus der zu untersuchenden Grundgesamtheit
- Der *Stichprobenumfang* ist die Anzahl der Elemente in der Stichprobe
- Die *Daten* sind die beobachteten Ausprägungen des Merkmals bzw. der Merkmale
- Die Erfassung der Daten geschieht in der *Urliste*, auch Protokoll genannt

Zufall

- Zufall bedeutet hier: Kein erkennbares Muster
- Zufällige Auswahl ist nicht einfach. Verwende
  - Würfel
  - Zufallsgenerator
  - Zufallstafeln

Versuchsplanung

Folgendes Experiment:

- 25 Fische zufällig ausgewählt
- Fische lernen, in einem Labyrinth Futter zu suchen; Zeit wird gemessen
- die Fische werden an 25 Artgenossen verfüttert
- die Artgenossen sollen im selben Labyrinth Futter suchen; Zeit wird gemessen
- die neuen Fische sind schneller
- Nobelpreis?

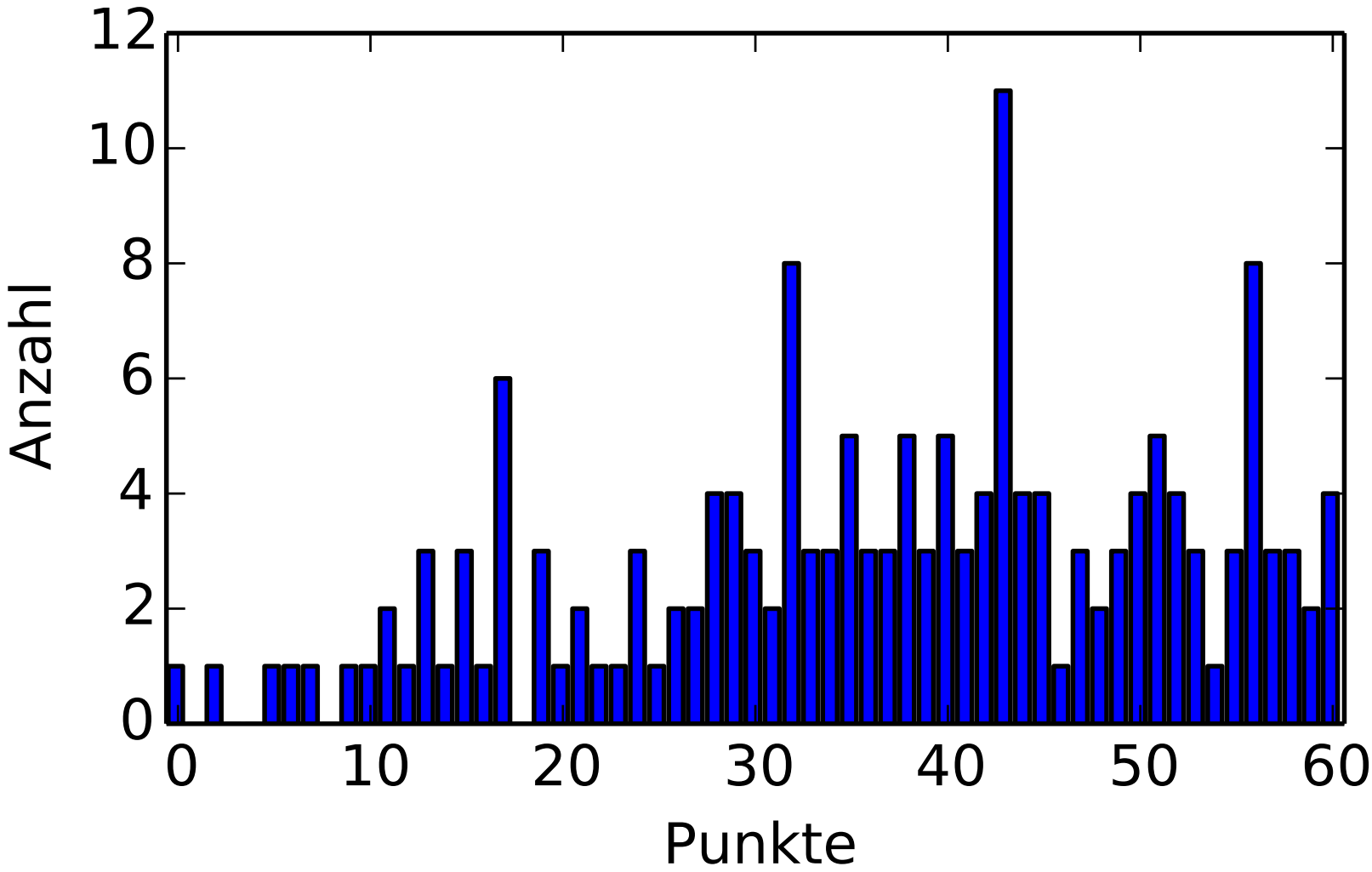
grafische Darstellung

- Balkendiagramm: Für jeden möglichen Wert ein Balken, der die Anzahl anzeigt
- Histogramm: Wie Balkendiagramm, aber Werte werden vorher in Klassen zusammengefasst
  Bei Stichprobenumfang n Anzahl der Klassen ungefähr $\sqrt{n}$
- Tortendiagramm: Anteile an der Gesamtpopulation werden grafisch dargestellt

Beispieldatensatz

- Datensatz: Ein altes Klausurergebnis
- Anzahl der Messwerte: $n = 166$
- Diskretes quantitatives Merkmal mit 61 Ausprägungen: Punktezahl zwischen 0 und 60

Balkendiagramm im Beispiel



Balkendiagramm der Häufigkeitsverteilung des Klausurergebnisses

Zusammenfassung in Klassen

- Bei stetigen Merkmalen kommt häufig jede Ausprägung nur einmal vor
- Dann fasst man benachbarte Ausprägungen zu Klassen zusammen und zeigt die Häufigkeitsverteilung der Klassen
- Histogramm

Beispiel: Reaktionszeiten

Die folgenden 30 Reaktionszeiten sind gemessen worden (in [s])

0.31	0.53	0.51	0.36	0.42	0.61	0.46	1.37	0.54	0.32
0.57	0.19	1.12	0.31	0.21	0.43	0.52	0.66	0.89	1.20
0.51	0.66	0.61	0.82	0.75	0.57	0.59	0.48	0.11	0.47

Der kleinste Wert ist $0.11s$, der größte $1.37s$.

Wie viele Klassen

- Anzahl der Klassen ungefähr die Wurzel aus dem Stichprobenumfang
- Klassen müssen gleich breit sein
- Klassengrenzen dürfen nicht krumm sein

Im Beispiel führt eine Klassenbreite

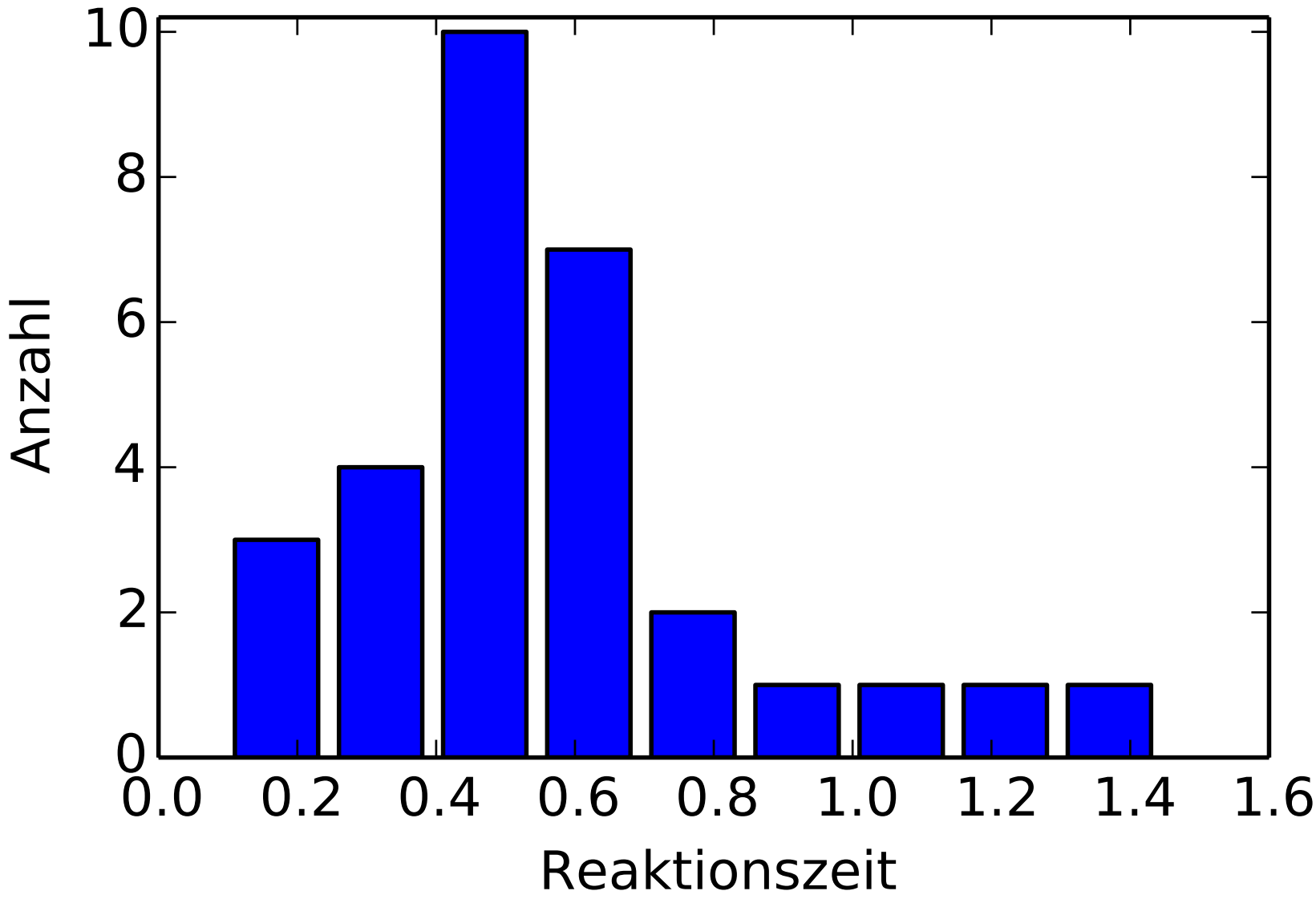
von $0.1s$ zu 13 Klassen

von $0.15s$ zu 9 Klassen

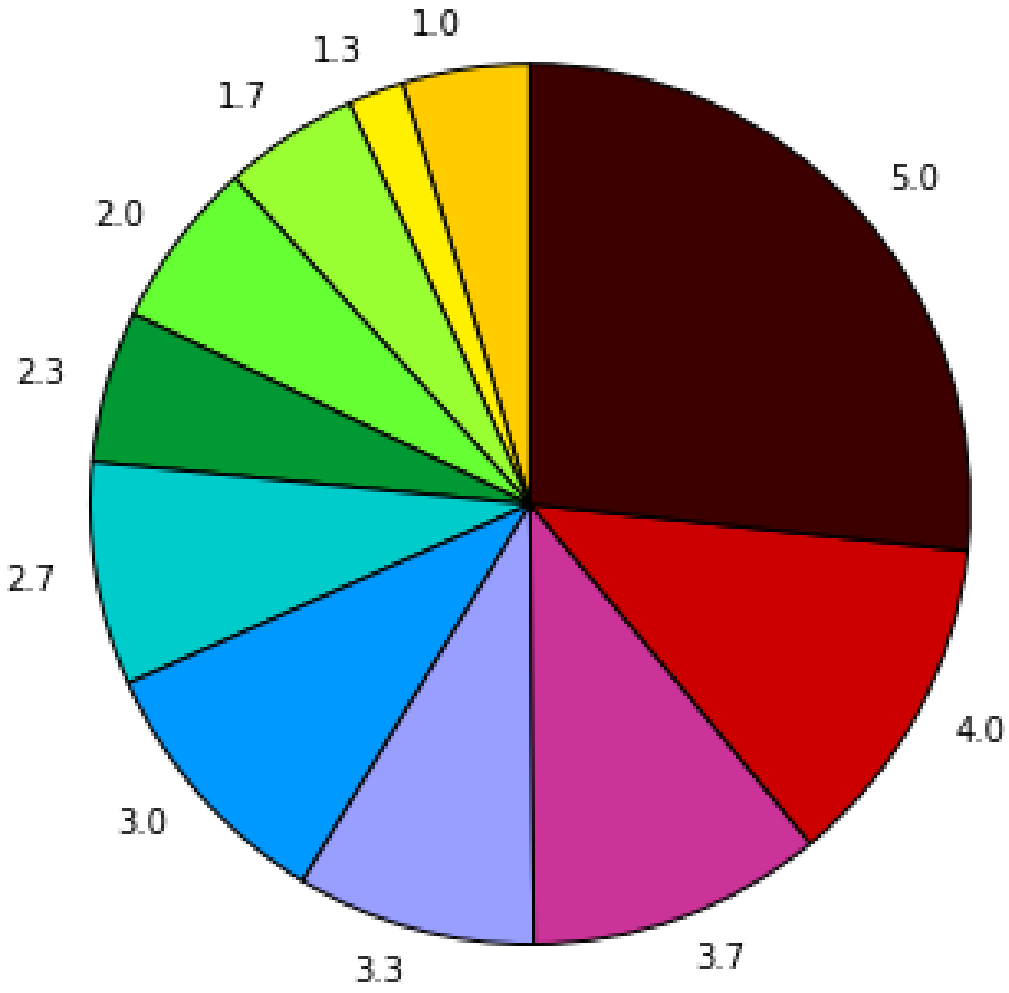
von $0.2s$ zu 7 Klassen

Wir wählen $0.15s$.

Histogramm im Beispiel



Tortendiagramm im Beispiel



Arithmetisches Mittel

Das arithmetische Mittel ist der Durchschnitt der Messwerte.

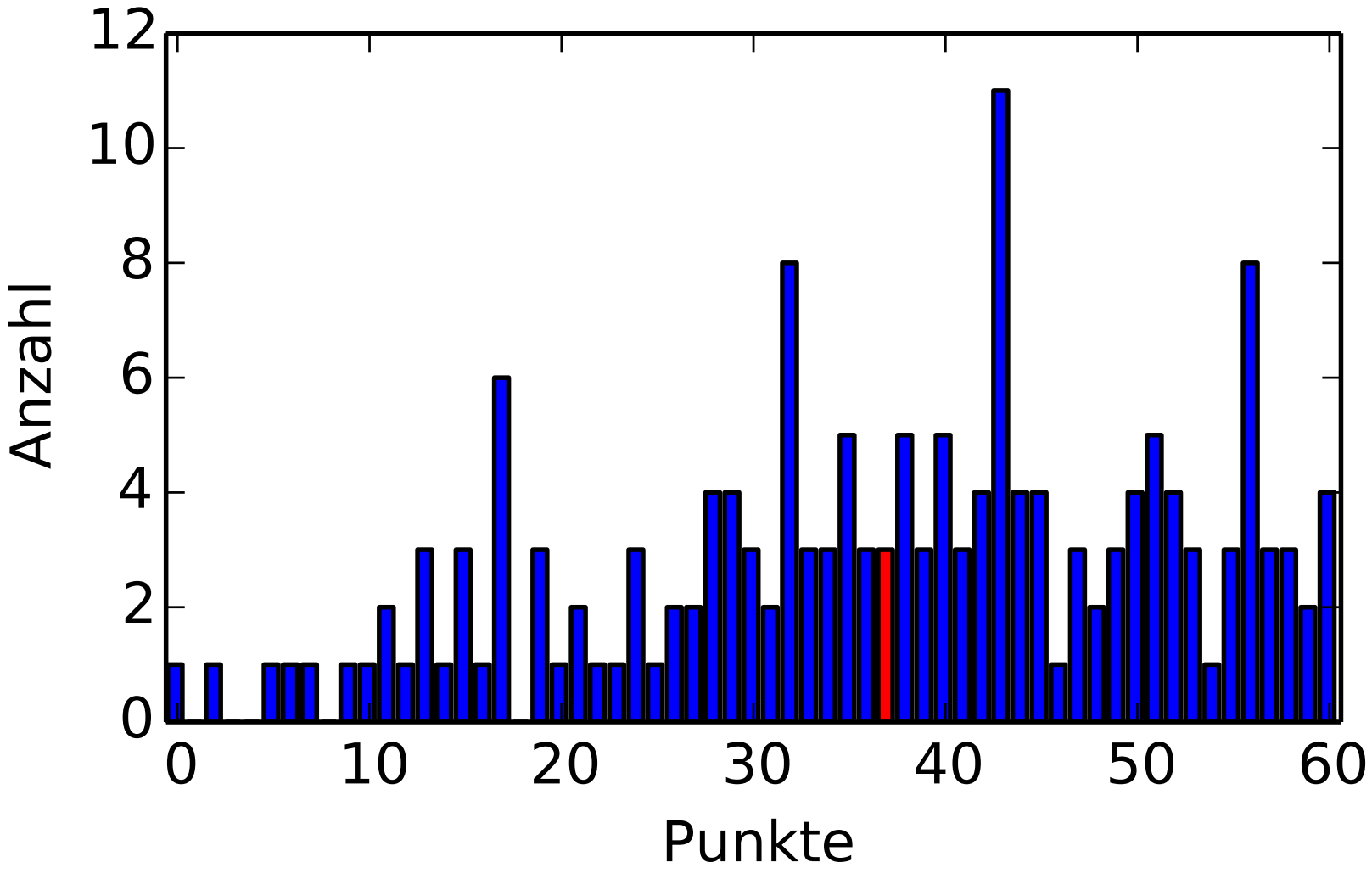
Formel: Beim Stichprobenumfang n seien $x_1, x_2, x_3, \ldots, x_n$ die Messwerte, dann ist das arithmetische Mittel gleich

$$\overline{x} = \frac{1}{n}\left(x_1 + x_2 + x_3 + \cdots + x_n\right)$$

Man schreibt auch

$$\overline{x} = \frac{1}{n}\sum_{j=1}^{n} x_j$$

Arithmetisches Mittel im Beispiel



Das arithmetische Mittel der Beispieldaten beträgt 36.96 Punkte.

Rundung aggregierter Daten

- Durch Versuchswiederholung sinkt der relative Fehler des Gesamtergebnisses
- Deshalb werden aggregierte Daten wie z. B. das arithmetische Mittel mit zwei gültigen Stellen mehr als die Einzelmesswerte angegeben
- Die Bestimmung der tatsächlichen Genauigkeit aggregierter Daten ist eine zentrale Frage der Statistik

Median

Der Median ist ein Wert mit der Eigenschaft, dass in der Menge der nach Größe geordneten Messwerte gleich viele Daten unterhalb und oberhalb des Medians liegen.

Beispiel: 7 Messwerte

$$10 \mid 5 \mid 4 \mid 9 \mid 10 \mid 1 \mid 5$$

Nach Größe anordnen

$$1 \mid 4 \mid 5 \mid 5 \mid 9 \mid 10 \mid 10$$

Median für geraden Stichprobenumfang

Falls die Anzahl der Daten gerade ist, stehen in der Menge der nach Größe geordneten Messwerte zwei Zahlen in der Mitte. Der Median x_{med} ist dann deren arithmetisches Mittel.

Beispiel

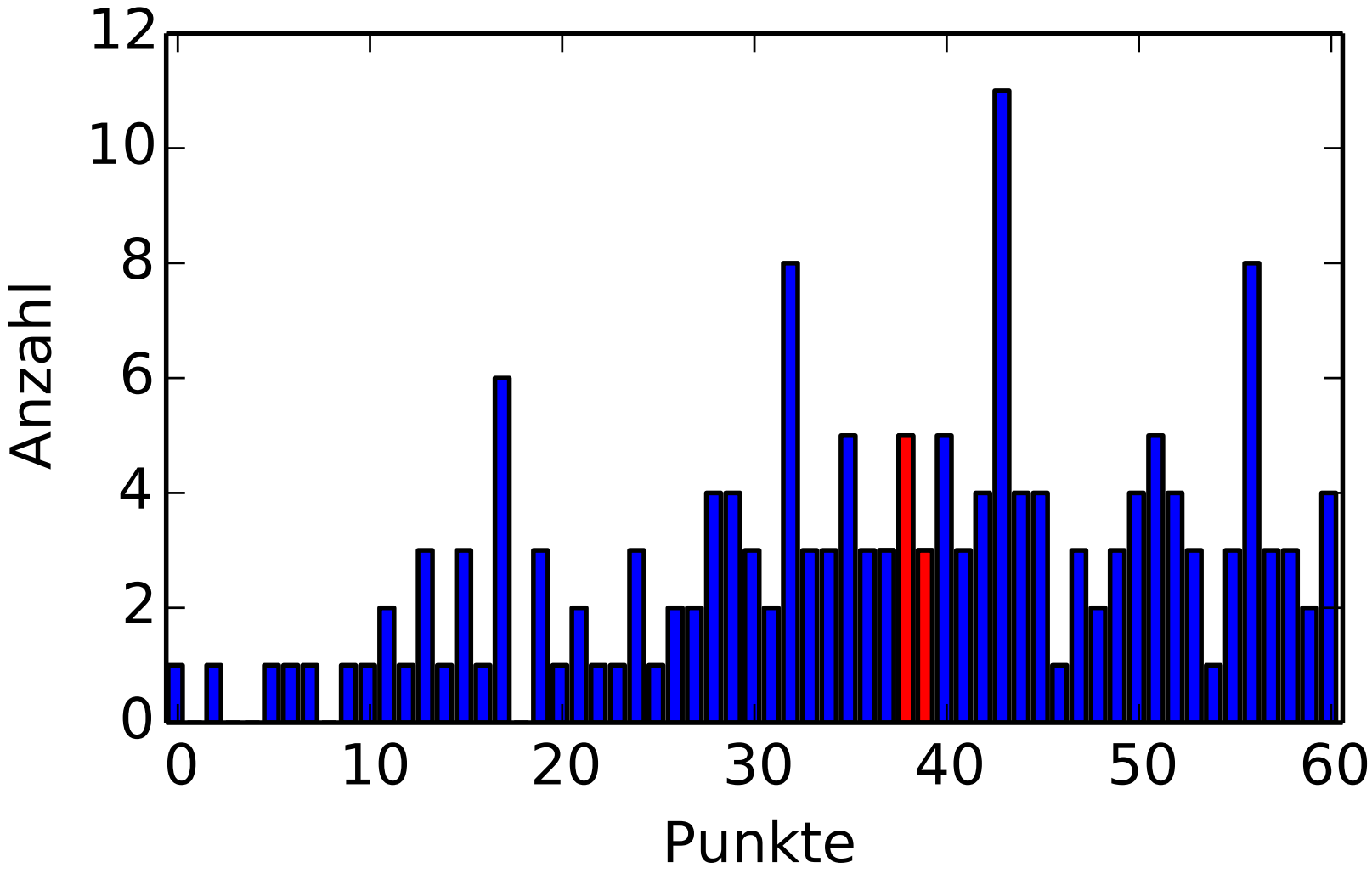
$$10 \mid 5 \mid 4 \mid 9 \mid 10 \mid 1 \mid 5 \mid 8$$

Nach Größe anordnen

$$1 \mid 4 \mid 5 \mid 5 \mid 8 \mid 9 \mid 10 \mid 10$$

Der Median ist $x_{\text{med}} = 6.5$

Median im Beispiel



Im Beispiel der Klausurdaten beträgt der Median 38.5 Punkte

Robustheit

- Ein *Ausreißer* ist ein Messwert, der weit von fast allen anderen Messwerten entfernt ist. Ausreißer können z.B. von Messfehlern herrühren.
- Ausreißer dürfen nicht einfach aus dem Datensatz entfernt werden!
- Eine statistische Größe ist *robust*, wenn sie unempfindlich gegen Ausreißer ist.
- Das arithmetische Mittel ist nicht robust. Ausreißer gehen genauso ein wie alle anderen.
- Der Median ist robust.

Beispiel zur Robustheit

- Beim Eintippen der Klausurergebnisse wurden für die beste Arbeit versehentlich 600 Punkte notiert.
- Das arithmetische Mittel steigt um 3.3 Punkte auf 40.2 Punkte.
- Der Median ändert sich gar nicht.

Empirische Varianz

- Beim Stichprobenumfang n seien $x_1, x_2, x_3, \ldots, x_n$ die Messwerte, dann ist das *empirische Varianz* gleich

$$s^2 = \frac{(x_1 - \overline{x})^2 + (x_2 - \overline{x})^2 + (x_3 - \overline{x})^2 + \cdots + (x_n - \overline{x})^2}{n - 1}$$

  Dabei ist $\overline{x}$ das arithmetische Mittel ▸ Ohne Pünktchen
- Die empirische Varianz wird mit s^2 bezeichnet. Die Zahl s heißt *empirische Standardabweichung* oder *Stichprobenstreuung*
- Die Stichprobenstreuung ist also die Quadratwurzel der empirischen Varianz

Konkrete Rechnung

- Bei fünf Proben wurden die folgenden Massen gewogen: $1.1g$, $1.3g$, $1.6g$, $1.3g$, $2.0g$
- Man erhält $\overline{x} = \frac{7.3g}{5} = 1.46g$
- Rechnung:

j	$x_j - \overline{x}$	$(x_j - \overline{x})^2$
1	$-0.36g$	$0.1296g^2$
2	$-0.16g$	$0.0256g^2$
3	$0.14g$	$0.0196g^2$
4	$-0.16g$	$0.0256g^2$
5	$0.54g$	$0.2916g^2$
Summe	$0g$	$0.4920g^2$

- Also

$$s^2 = \frac{0.4920g^2}{4} = 0.1230g^2 \qquad s = \sqrt{0.1230g^2} = 0.3507g$$

Stichprobenstreuung vs. Varianz

- Der Vorteil der Stichprobenstreuung gegenüber der Varianz ist, dass die Stichprobenstreuung richtig skaliert
- Das bedeutet folgendes: Wenn ich alle Daten mit 1 000 multipliziere, dann
  - multipliziert sich das arithmetische Mittel mit 1 000
  - multipliziert sich die Varianz mit 1 000 000
  - multipliziert sich die Stichprobenstreuung mit 1 000
- Das bedeutet auch, dass die Stichprobenstreuung in derselben Einheit angegegeben wird wie die Daten

Formeln für die Varianz

- Die Definition ohne Pünktchen

$$s^2 = \frac{1}{n-1} \sum_{j=1}^{n} (x_j - \overline{x})^2$$

▸ Mit Pünktchen

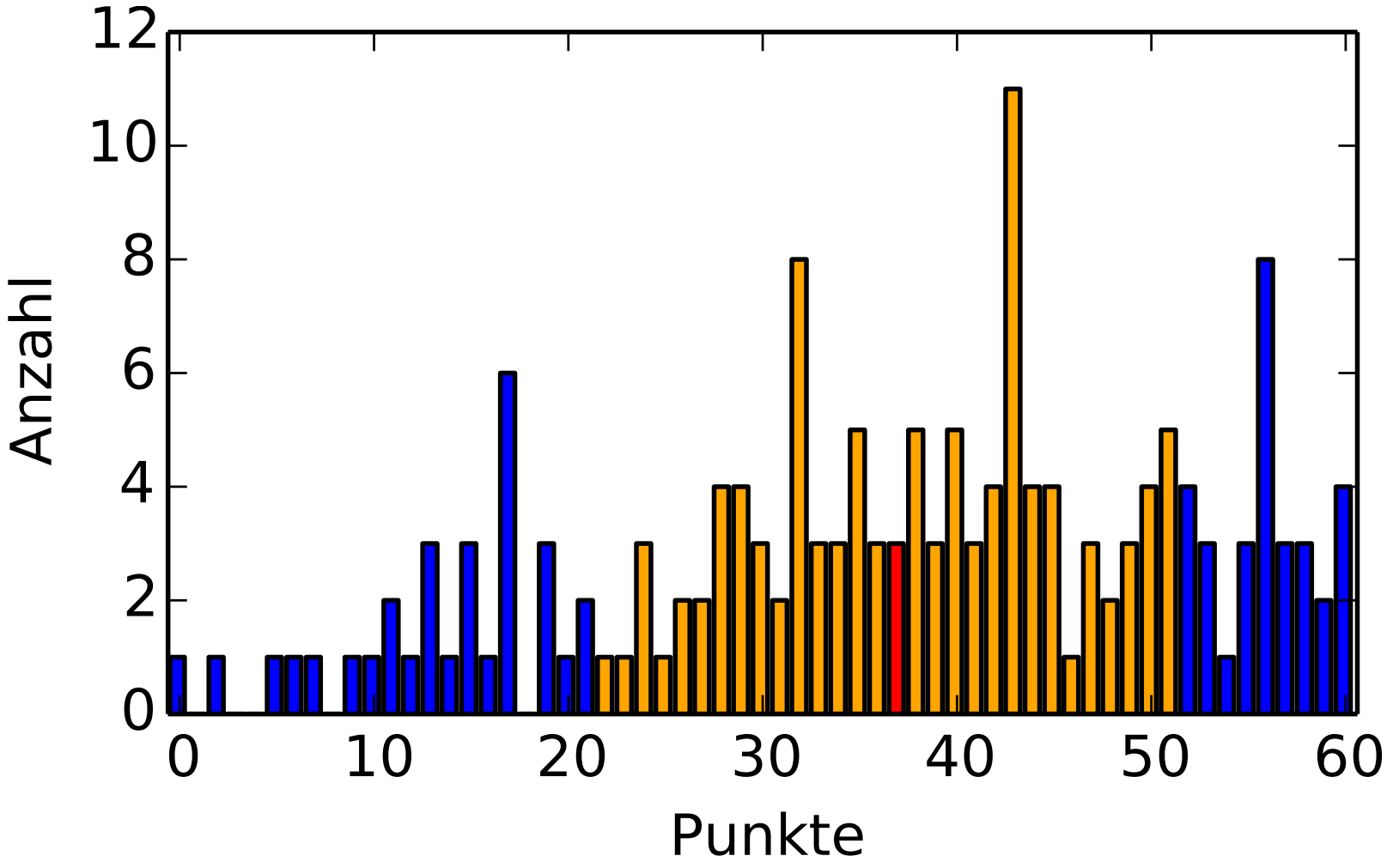
- Eine etwas einfachere Formel, deren Richtigkeit leicht nachgerechnet werden kann

$$s^2 = \frac{1}{n-1} \left(\left(\sum_{j=1}^{n} x_j^2 \right) - n\overline{x}^2 \right)$$

Warum $n-1$ im Nenner?

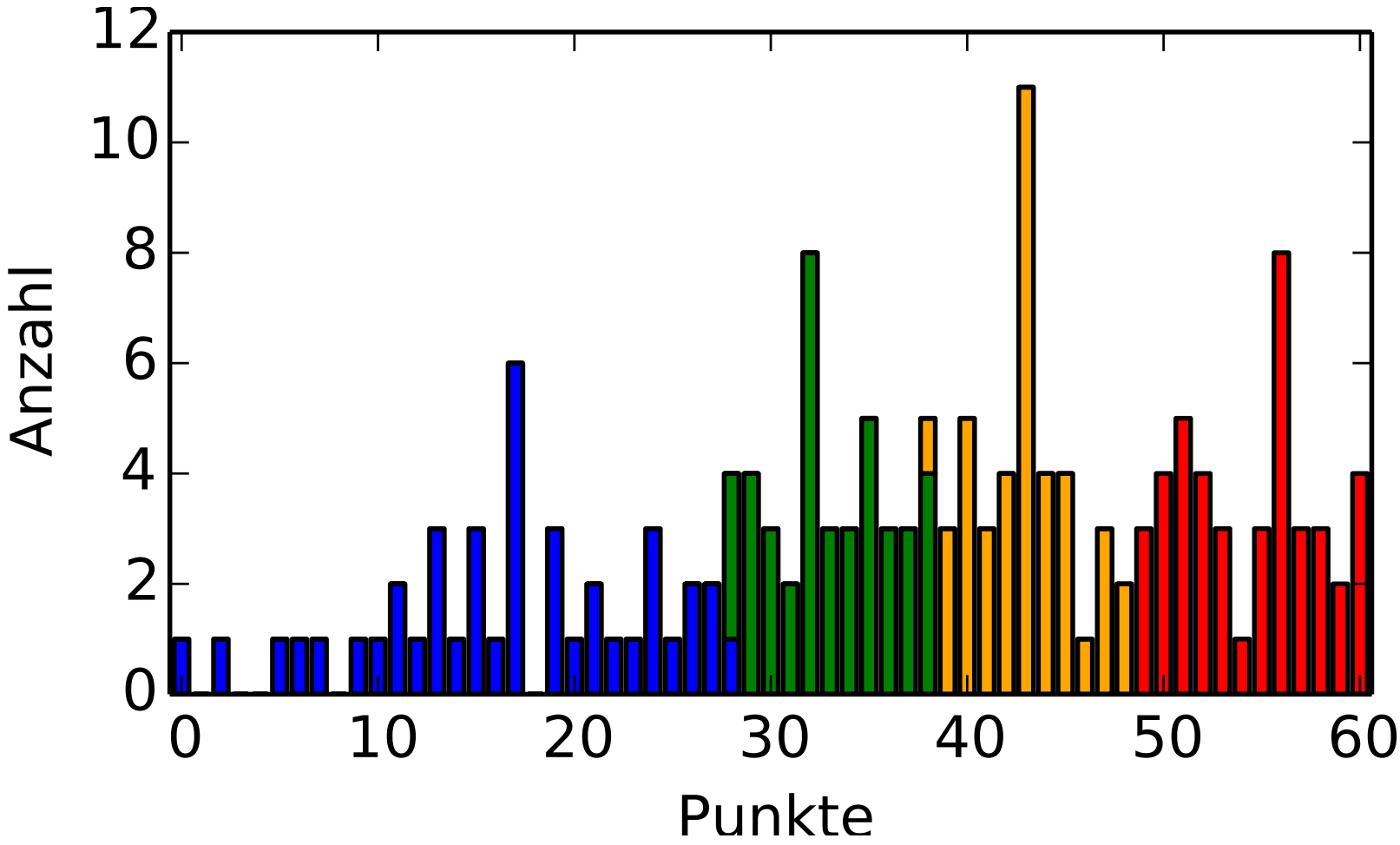
- Das Stichwort ist “erwartungstreuer Schätzer”
- Das bedeutet ungefähr, dass man, wenn man für viele Datensätze die Varianz mit dieser Methode feststellt, im Mittel näher an der “wahren” Varianz ist, als wenn man den Nenner n benutzt
- Die großen Software-Pakete machen das auch so

Stichprobenstreuung im Beispiel



Für die Beispieldaten ist die Streuung gleich 14.6 Punkte.
Gelb ist der Bereich von $\overline{x} - s$ bis $\overline{x} + s$

Quartile im Beispiel



Für die Beispieldaten teilen wir die Beobachtungen in vier gleich große Blöcke. Die Grenzen der beiden äußeren Blöcke sind das erste und das dritte Quartil.

Quartile

- Das Quartile Q_1 ist als derjenige Wert definiert, unter dem 25% und über dem 75% der Messwerte liegen
- Das Quartile Q_3 ist als derjenige Wert definiert, über dem 25% und unter dem 75% der Messwerte liegen
- Im Beispiel: $Q_1 = 28$ und $Q_3 = 49$

Konkrete Berechnung der Quartile

- Daten der Größe nach ordnen
- Bei ungeradem Stichprobenumfang das mittlere Element entfernen
- Der Median der unteren Hälfte ist das erste Quartil, der Median der oberen Hälfte das dritte
- Beispieldaten (bereits geordnet)

  1 3 5 7 7 9 10 11 11

  Der Median ist 7, das erste Quartil beträgt 4, das dritte Quartil beträgt 10.5

Interquartilabstand

- Der Interquartilabstand ist definiert als

$$I_{50} = Q_3 - Q_1$$

- Der Interquartilabstand ist ein robustes Maß für die Streuung
- Für die Beispieldaten aus der Klausur ist $I_{50} = 21$

Robustheit des Interquartilabstands

- Tippfehler in den Beispieldaten: Einmal 600 statt 60 Punkte
- Dadurch steigt die Streuung von 14.6 auf 46.0
- Der Interquartilabstand ändert sich überhaupt nicht

Box-Whisker-Plot

Ein Box-Whisker-Plot zeigt von oben nach unten

- den größten Datenwert
- Q_3
- den Median
- Q_1
- den kleinsten Datenwert

Dabei bezeichnet

- Q_1 das erste *Quartil*. Das ist derjenige Wert, der von 25% der Daten unter- und von 75% der Daten überschritten wird
- Q_3 das dritte Quartil. Das ist derjenige Wert, der von 75% der Daten unter- und von 25% der Daten überschritten wird

Box-Whisker-Plot von Klausurergebnissen