

Mathematik für Biologen

Prof. Dr. Rüdiger W. Braun

Heinrich-Heine-Universität Düsseldorf

20. Oktober 2010

1 empirische Verteilung

2 Lageparameter

- Modalwert
- Arithmetisches Mittel
- Median

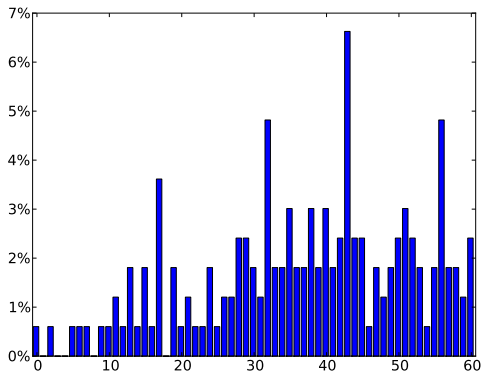
3 Streuungsparameter

- Variationsbreite
- Empirische Varianz und Stichprobenstreuung
- Interquartilabstand

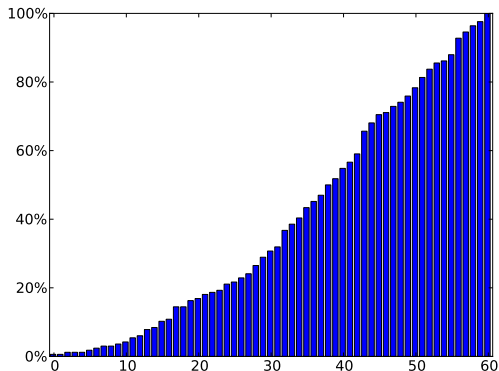
Empirische Häufigkeitsverteilung

- Es liege eine Stichprobe vom Umfang n zu Grunde
- Die *absolute* Häufigkeit einer Merkmalsausprägung ist gleich der Anzahl der Beobachtungen dieser Ausgänge
- Die *relative* Häufigkeit ist gleich der absoluten Häufigkeit geteilt durch den Stichprobenumfang

Beispiel zur empirischen Häufigkeitsverteilung



Kumulierte empirische Häufigkeitsverteilung

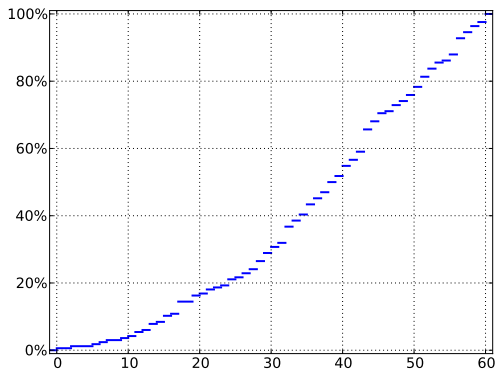


Das bedeutet: der Wert an der Stelle 10 gibt beispielsweise den Prozentsatz aller Ergebnisse mit 10 **oder weniger** Punkten an.

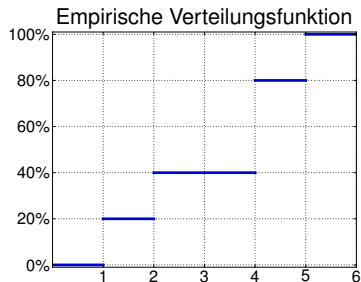
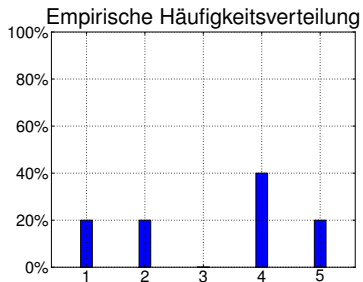
Empirische Verteilungsfunktion

- Durch Kumulation der empirischen Häufigkeitsverteilung erhält man die *empirische Verteilungsfunktion*
- Genauer: Wir untersuchen ein quantitatives Merkmal. Die empirische Verteilungsfunktion $F(t)$ gibt an, für welchen Anteil der Daten sein Wert höchstens gleich t ist

Beispiel für empirische Verteilungsfunktion

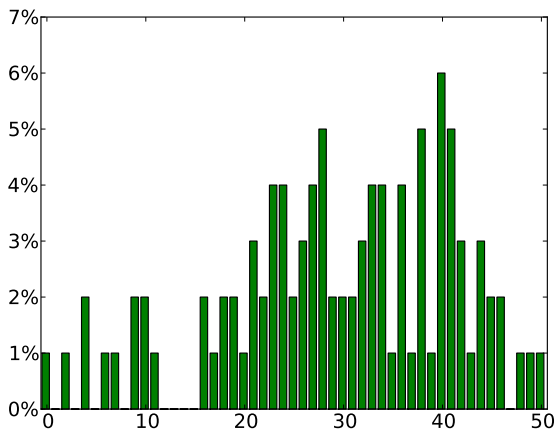


Häufigkeitsverteilung und Verteilungsfunktion

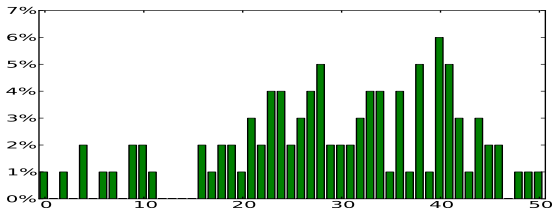
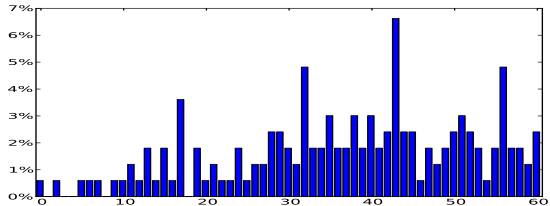


Vergleichbarkeit

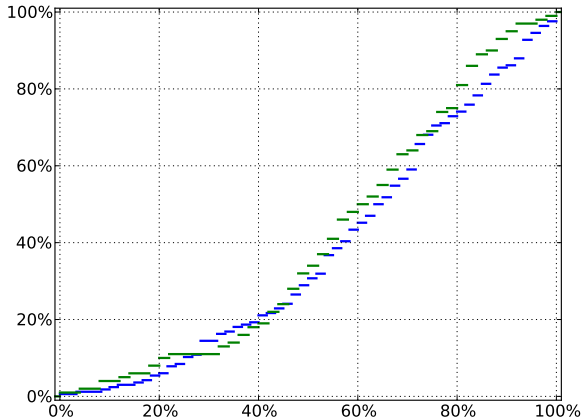
Hypothetisches Ergebnis einer Klausur mit 50 Punkten und
100 Teilnehmern



Vergleich der empirischen Häufigkeitsverteilungen



Vergleich der empirischen Verteilungen



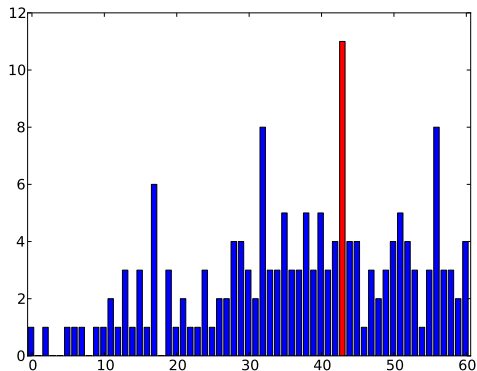
Statistische Maßzahlen

- Lageparameter
 - Modalwert
 - Arithmetisches Mittel
 - Median
- Streuungsparameter
 - Variationsbreite
 - Varianz und Standardabweichung
 - Interquartilabstand

Modalwert

- Der Modalwert einer Menge von Messwerten ist der am häufigsten auftretende Wert.
- Wenn mehrere Werte am häufigsten auftreten, gibt es mehrere Modalwerte.

Modalwert im Beispiel



Für unsere Beispieldaten ist der Modalwert gleich 43.

Arithmetisches Mittel

Das arithmetische Mittel ist der Durchschnitt der Messwerte.

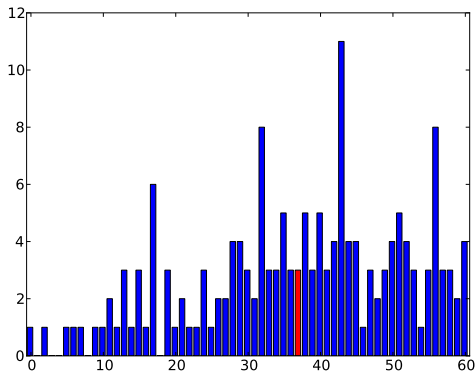
Formel: Beim Stichprobenumfang n seien $x_1, x_2, x_3, \dots, x_n$ die Messwerte, dann ist das arithmetische Mittel gleich

$$\bar{x} = \frac{1}{n} (x_1 + x_2 + x_3 + \dots + x_n)$$

Man schreibt auch

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

Arithmetisches Mittel im Beispiel



Das arithmetische Mittel der Beispieldaten beträgt 36.96.

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 | 5 | 4 | 9 | 10 | 1 | 5

Nach Größe anordnen

1 | 4 | 5 | 5 | 9 | 10 | 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

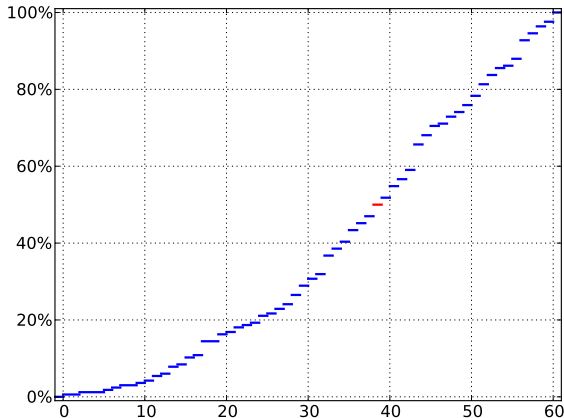
10 | 5 | 4 | 9 | 10 | 1 | 5 | 8

Nach Größe anordnen

1 | 4 | 5 | 5 | 8 | 9 | 10 | 10

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

Median im Beispiel



Den Median kann man aus dem Graphen der Verteilungsfunktion ablesen. Im Beispiel beträgt er 38.5

Robustheit

- Ein *Ausreißer* ist ein Messwert, der weit von fast allen anderen Messwerten entfernt ist. Ausreißer können z.B. von Messfehlern herrühren.
- 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 Robustheit

- Am Altägyptischen Seminar machen 5 Studierende Examen.
Ihre Studiendauern betragen ▶ vor

Lea	Leo	Leon	Lena	Leyla
7	10	8	7	9

- Das arithmetische Mittel ihrer Studiendauern ist 8.2 Semester.
Der Median beträgt 8 Semester
- Außerdem studiert Herbert dort bereits seit 30 Semestern.
Falls er abschließt, ist das arithmetische Mittel gleich 11.8,
und der Median beträgt 8.5

Variationsbreite

- Die Variationsbreite ist die Differenz zwischen dem kleinsten und dem größten Messwert
- In den Beispieldaten hat die Variationsbreite den Wert 60
- Die Variationsbreite ist überhaupt nicht robust und auch sonst kein brauchbares Maß für die Streuung

Empirische Varianz

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

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

Dabei ist $\bar{x}$ 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

Varianz nicht robust

- Die Daten aus dem Beispiel zur Robustheit

Lea	Leo	Leon	Lena	Leyla	Herbert
7	10	8	7	9	30

- Ohne Herbert ist die empirische Varianz gleich 1.7 und die Stichprobenstreuung ist 1.3
Erinnerung: Arithmetisches Mittel gleich 8.2
- Mit Herbert ist die empirische Varianz gleich 81 und die Stichprobenstreuung ist 9
Erinnerung: Arithmetisches Mittel mit Herbert gleich 11.8

Stichprobenstreuung vs. Varianz

- Der Vorteil der Stichprobenstreuung gegenüber der Varianz ist, dass die Stichprobenstreuung richtig skaliert
- Das bedeutet folgendes: Wenn ich alle Daten verdoppele, dann
 - verdoppelt sich das arithmetische Mittel
 - vervierfacht sich die Varianz
 - verdoppelt sich die Stichprobenstreuung

Formeln für die Varianz

- Die Definition ohne Pünktchen

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

► Mit Pünktchen

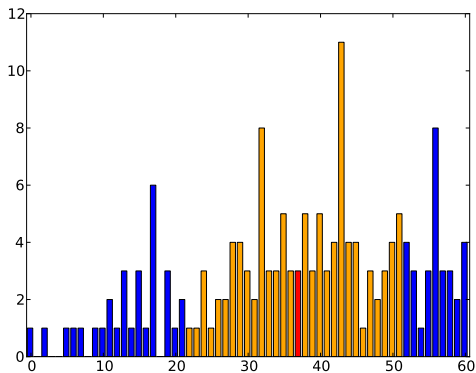
- Eine etwas einfachere Formel, deren Richtigkeit leicht nachgerechnet werden kann

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

Warum $n - 1$ im Nenner?

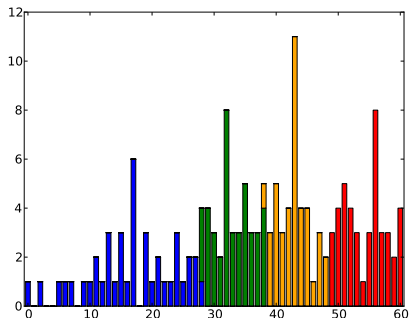
- Das Stichwort ist “erwartungstreuer Schätzer”
- Das bedeutet ungefähr, dass man, wenn man für viele Datensätze die Varianz mit dieser Methode feststellt, im Mittel näher an der “wahren” Varianz ist, als wenn man den Nenner n benutzt
- Die großen Software-Pakete machen das auch so

Stichprobenstreuung im Beispiel



Für die Beispieldaten ist die Streuung gleich 14.6

Quartile im Beispiel



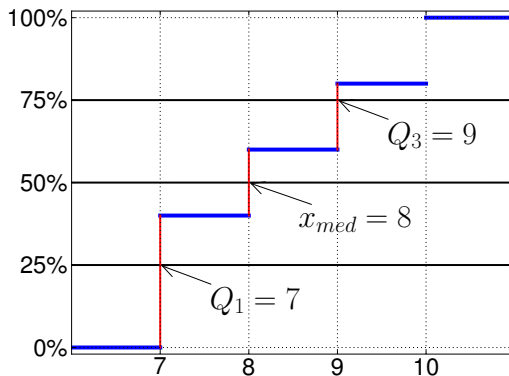
Für die Beispieldaten teilen wir die Beobachtungen in vier gleich große Blöcke. Die Grenzen der beiden äußeren Blöcke sind das erste und das dritte Quartil.

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 $Q_3 = 49$

Quartile und Verteilungsfunktion

Verteilungsfunktion für die Daten aus dem Beispiel [► Seminar](#)



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

Lea	Leo	Leon	Lena	Leyla	Herbert
7	10	8	7	9	30

- Ohne Herbert ist $Q_1 = 7$ und $Q_3 = 9$, also $I_{50} = 2$
- Mit Herbert ist $Q_1 = 7$ und $Q_3 = 10$, also $I_{50} = 3$

Box-Whisker-Plot

Ein Box-Whisker-Plot zeigt von oben nach unten

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

Box-Whisker-Plot der Beispieldaten

