Statistik
02. Deskriptive Statistik

Roland Schäfer

Institut für Germanistische Sprachwissenschaft
Friedrich-Schiller-Universität Jena

stets aktuelle Fassungen: `https://github.com/rsling/VL-Deutsche-Syntax`

Inhalt

Übersicht

- Deskriptive Statistik als Aggregation von Daten
- Verteilungen in Stichproben und Grundgesamtheiten:
  - Zentralmaße
  - Streuung (Varianz)
- Theoretische vs. empirische Verteilungen
- Kovarianz | Miteinander variierende Variablen
- Konfidenzintervalle | Genauigkeiten von Schätzungen?

Literatur

- Google, Stackoverflow usw.
- Gravetter & Wallnau (2007)
  Achtung! Vermittelt eine falsche Philosophie bei Anwendung der Tests!
- Bortz & Schuster (2010)

Motivation

Zweck der deskriptiven Statistik

- Mit unbewaffnetem Auge auf Daten zu blicken, ist meistens zwecklos.
- In Zahlen sehen Menschen nur schlecht Tendenzen und Zusammanhänge.
- Deskriptive Statistik
  - Zusammenfassen
  - Gruppieren
  - Visualisieren

Was will man wissen?

- Definition der Grundgesamtheit
- Stichprobengröße (n)
  - 200 Sätze aus dem Korpus
  - 1.000 Reaktionen (von 50 Probanden) im Experiment
  - Was sind die elementaren gemessenen Datenpunkte?
- Stichprobenmethode
  - Zufallsstichprobe | Nachweis der uniformen Zufälligkeit
  - Quotenstichprobe | Stratifzierung und Begründung

Skalenniveau

Messvariablen und Skalenniveaus

- dichotom/binär | Menge $\{A, B\}$ | zwei disjunkte Kategorien
  männlich, weiblich ; Präteritum, Perfekt
- nominal/kategorial | Menge $\{A, B, ..\}$ | disjunkte Kategorien
  Parteizugehörigkeit ; NP, AP, VP
- ordinal | Tupel $\langle A, B, ..\rangle$, nicht $\mathbb{N}$ oder $\mathbb{Z}$ | disjunkte Kategorien mit Rang
  Schulnoten ; 5- oder 7-Punkt-Skalen für Akzeptabilität
- Verhältnis | $+\mathbb{Q}_0$ | geordnete Werte mit Nullpunkt
  Temperatur in Kelvin ; Lesezeiten
- Intervall | $\mathbb{Q}$ | Wie Verhältnis, aber ohne Nullpunkt
  Temperatur in Celsius

- Zähldaten | Keine beobachtbaren Variablen, sondern
  Aggregation von dichtotomen, nominalen oder ordinalen Variablen

Intervalle vs. Verhältnisse

- Verhältnisskala | Größe von Menschen in cm
  - $200cm = 2 \times 100cm$ usw.
  - Keine Messung unter $0cm$
- Intervallskala | Dasselbe als Abweichung vom Mittel
  - $4cm = 2 \times 2cm$ usw.
  - $184cm \neq 2 \times 182cm$
  - Negative Messungen möglich

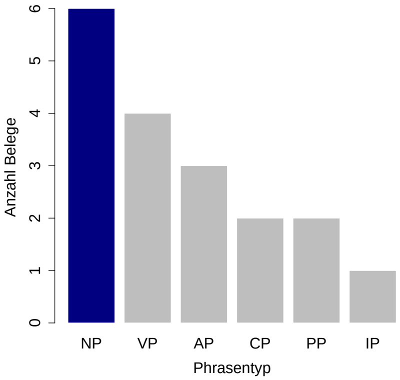
Relevanz der Skalenniveaus

- Bestimmung zulässiger mathematischer Operationen
- Deskriptive Statistiken je nach Skalenniveau
- Zulässigkeit von inferenzstatistischen Tests je nach Skalenniveau

Zentraltendenz

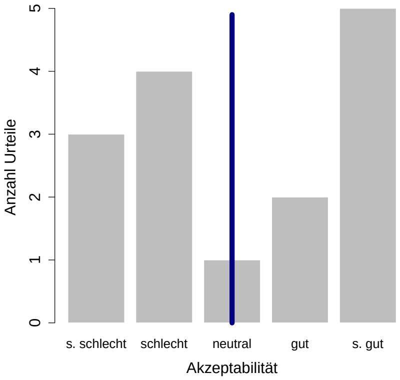
Zentraltendenz I

Modus | Der häufigste Wert | Alle Skalenniveaus



Zentraltendenz II

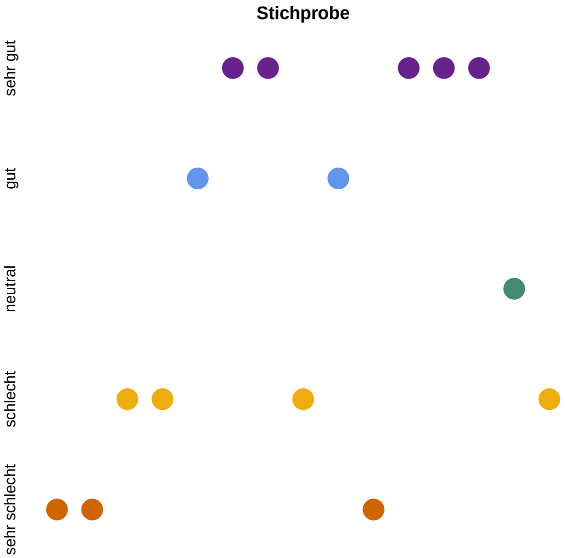
Median | Mitte der sortierten Stichprobe | ab Ordinalskala



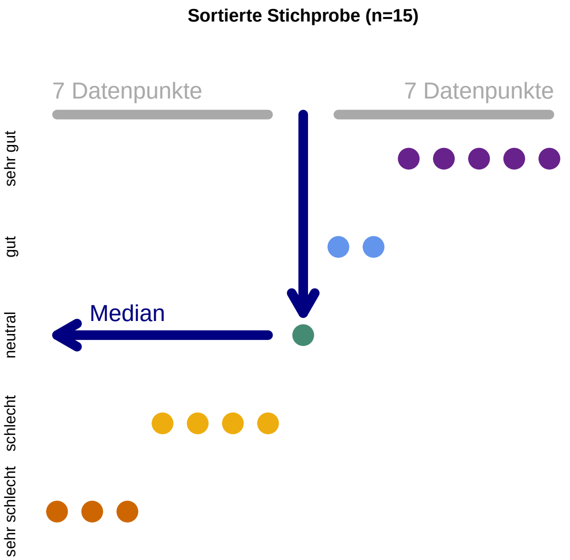
Numerische Messungen | Verschiedene Interpolationsmethoden

https://en.wikipedia.org/wiki/Quantile#Estimating_quantiles_from_a_sample

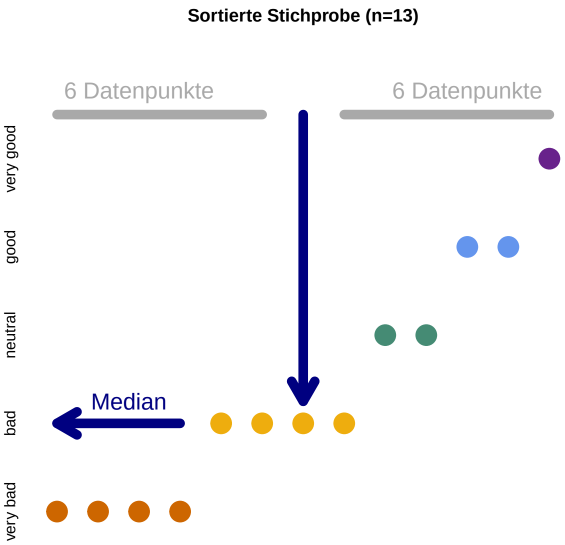
Median bestimmen | Stichprobe



Median bestimmen | Sortierte Stichprobe



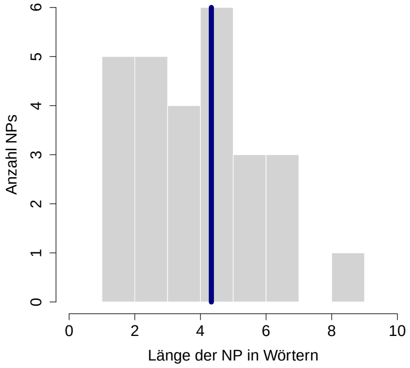
Median bestimmen | Verzerrtere sortierte Stichprobe



Zentraltendenz III

Arithmetisches Mittel $\bar{x}$ | Summe aller Werte geteilt durch n | ab Intervallskala

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



Empirische Verteilungen und Dispersion

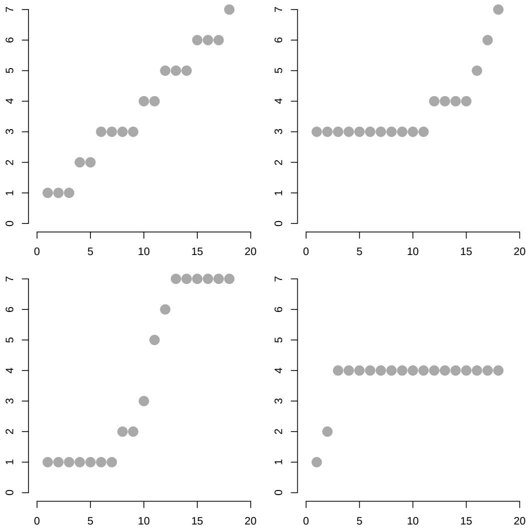
Warum sind Dispersionsmaße wichtig?

Dispersion | Streuung der Daten

- Zentraltendenz | Orientierung über Tendenzen der Stichprobe
- Ein Maß für Zentraltendenz für beliebig viele Verteilungsformen
- Arithmetisches Mittel | deskriptiv oft unbrauchbar ohne Betrachtung der Verteilung
- Median | auch nur bedingt besser

Vier sortierte Stichproben

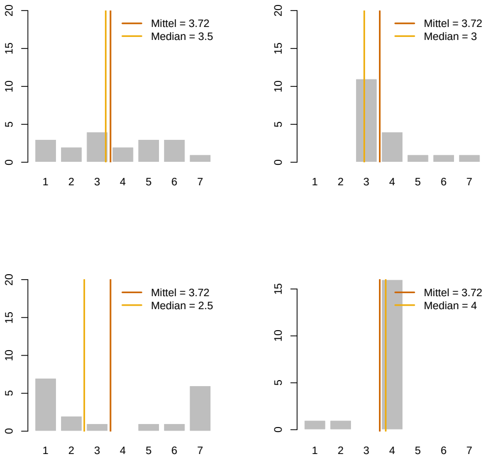
Jeder Punkt entspricht einem Datenpunkt/einer Messung!



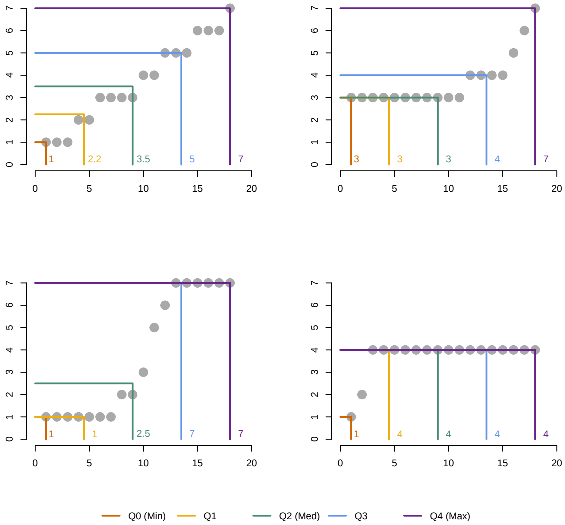
Verteilungsformen

Histogramme | Vier Stichproben mit $\bar{x} = 3.72$ und $n = 18$

Zum Beispiel 18 Bewertungen eines Probanden auf einer 7-Punkt-Skala



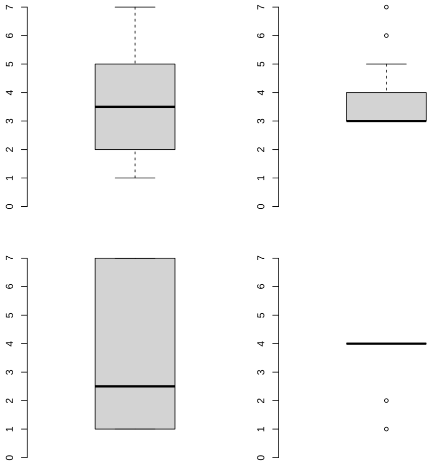
Quartile | Generalisierung des Medians (bei 25 %, 50 %, 75 %)



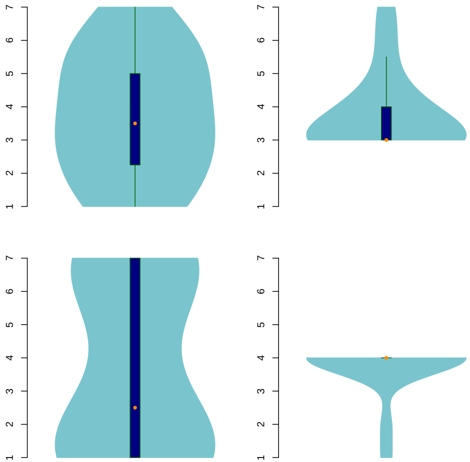
Interquartilbereich, Boxplots und Violinplots

- Interquartilbereich $IQR = Q_3 - Q_1$ | Die mittleren 50 %
- Boxplots
  - Median | Linie in der Mitte
  - Oberes und unteres Quartil | Boxen
  - 1,5-facher Interquartilabstand | gestrichelte Hebel
  - Ausreißer | Punkte
- Violinplots | Zusätzlich Plot der Verteilungsdichte (statt Box)

Boxplots | Die bessere Zusammenfassung

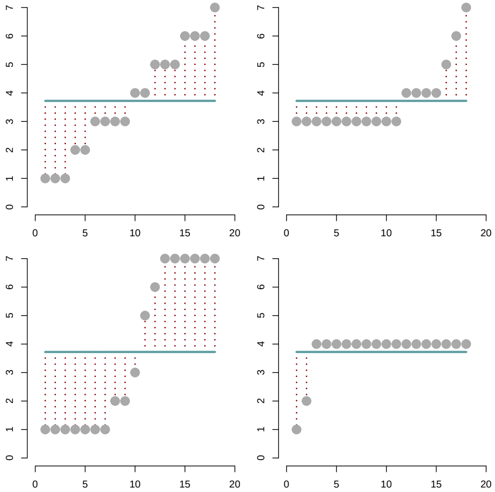


Violinplots | Die noch bessere Zusammenfassung



Was bestimmt die Varianz?

Die Distanzen der Messwerte zum Mittel sind unterschiedlich groß.



Varianz und Standardabweichung

Varianz s^2 | Quadrierte mittlere Abweichung vom Mittelwert

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

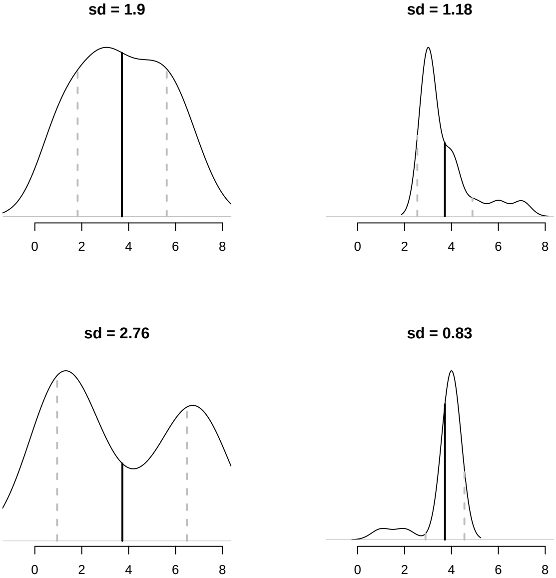
Standardabweichung s | Quadratwurzel der Varianz

$$s(x) = \sqrt{s^2(x)}$$

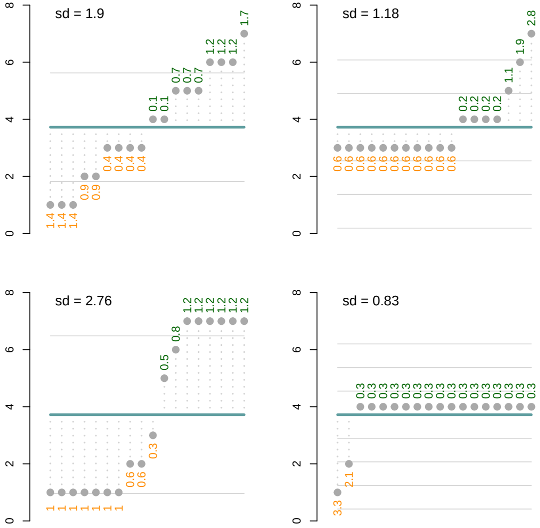
Summe der Quadrate | Zählerterm der Varianz

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

Unterschiedliche Standardabweichungen



Für jeden Messpunkt x_i | $z_i = \frac{x_i - \bar{x}}{s(x)}$



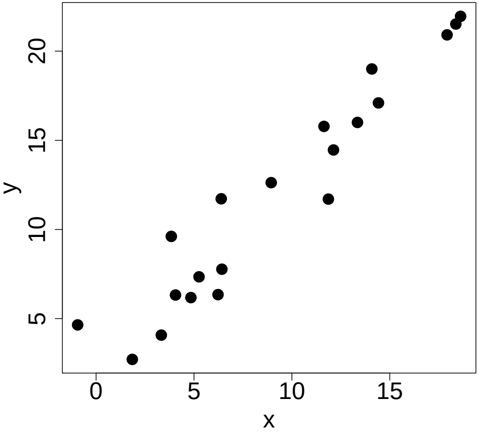
- Bsp.: $x = [3.9, 4.3, 7.2, 8.5, 11.1, 12.1, 14.0, 20.7]$
  - $\bar{x} = 10.225$
  - $s^2(x) = \frac{(3.9-10.255)^2+...+(20.7-10.225)^2}{8-1} = \frac{215.495}{7} = 30.785$
  - $s(x) = \sqrt{30.785} = 5.548$
  - $z = [\frac{3.9-10.225}{5.548}, ..., \frac{20.7-10.225}{5.548}] = [-1.140, -1.068, -0.545, -0.311, 0.158, 0.338, 0.680, 1.888]$

Bivariate Statistiken

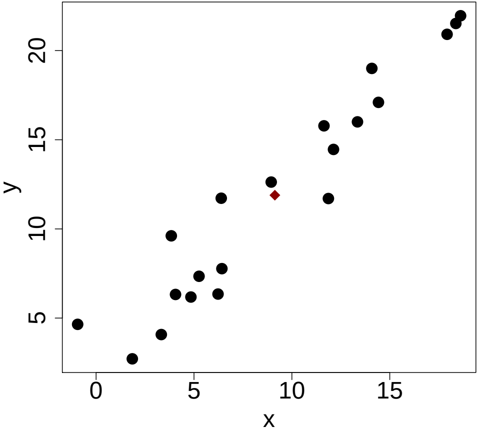
Kreuztabelle | Darstellung der Zähldaten zweier Variablen

	Variable 1 | Wert 1	**Wert2**
Variable 2 | Wert 1	Anzahl x_{11}	Anzahl x_{12}
Wert 2	Anzahl x_{21}	Anzahl x_{22}

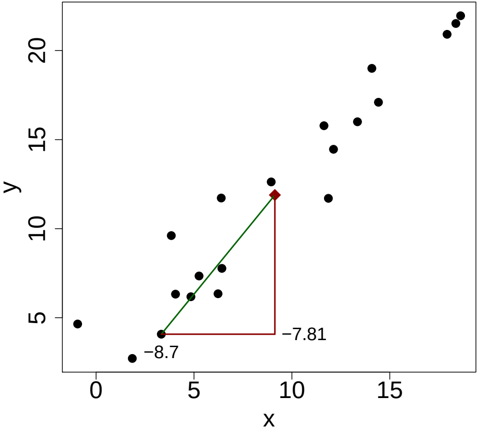
Bivariate Korrelationskoeffizienten | ab Ordinalskala



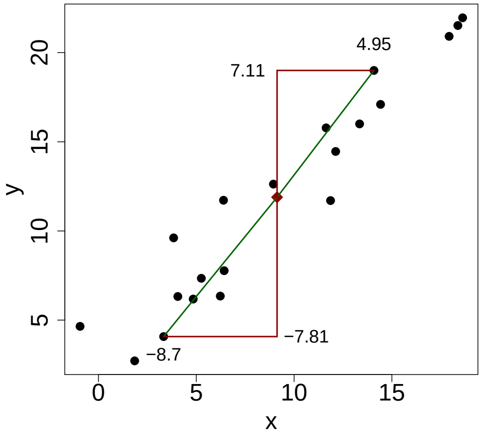
Koordinate von $\langle \bar{x}, \bar{y} \rangle$ | Mittel der beiden gemessenen Variablen



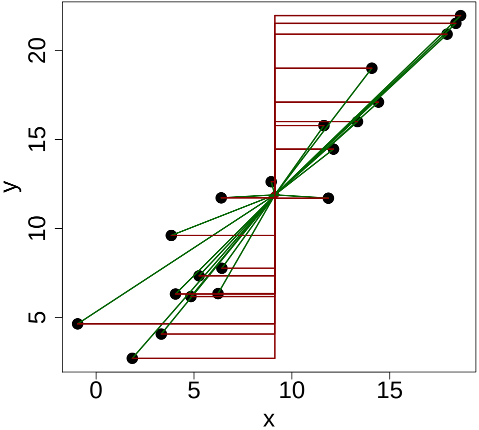
Punktvarianzen | $x_3 - \bar{x} = -7.81$ und $y_3 - \bar{y} = -5.80$ | $-7.81 \cdot -5.80 = 45.30$



Punktvarianzen | $x_{17} - \bar{x} = 4.95$ und $y_{17} - \bar{y} = 7.11$ | $4.95 \cdot 7.11 = 35.19$

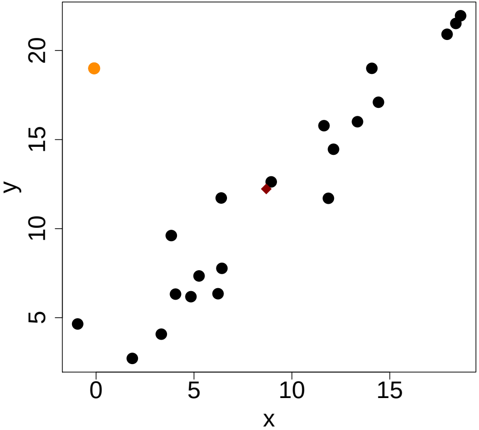


Puntvarianzen für alle $\langle x_i, y_i \rangle$ $cov(x, y) = 34.52$

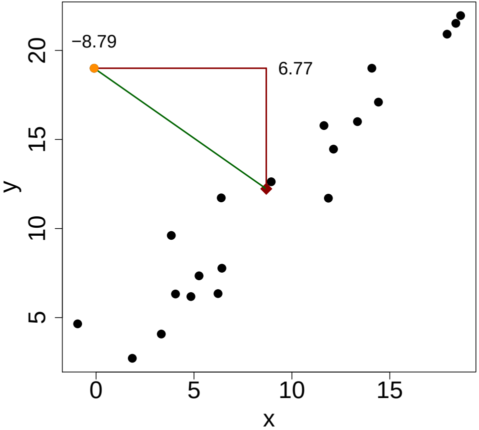


Kovarianz | Illustration 5

Ausreißer bei ansonsten positiver Kovarianz | Negatives Produkt der Punktvarianzen

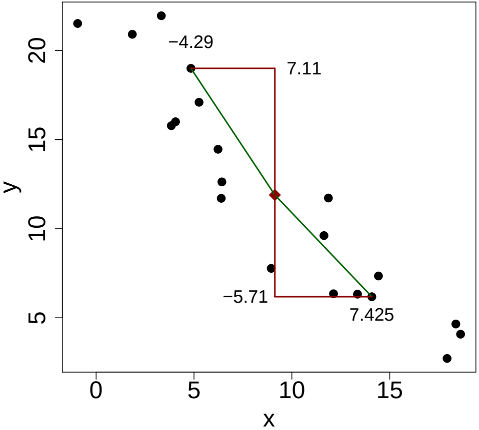


Punktvarianzen | $x_{21} - \bar{x} = 6.77$ und $y_{21} - \bar{y} = -8.79$ | $6.77 \cdot -8.79 = -59.51$



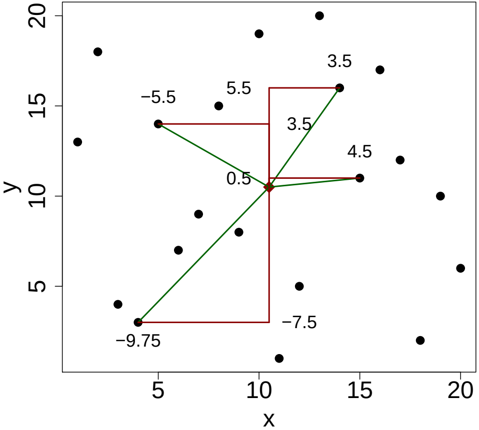
Negative Kovarianz

Tendenziell negative Abhängigkeit | Punktvarianzen überwiegend | $cov(x, y) = -33.77$



Kovarianz nahe Null

Ohne Abhängigkeit | Kovarianz nahe 0 | $cov(x, y) = -1.74$



Kovarianz

Kovarianz | Kombination der Abweichung der Messpunkte vom jeweiligen Mittel

$$cov(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{n-1}$$

Summe der Produkte | Der Zählerterm | $SP(x,y) = \sum_{i=1}^{n}(x_i - \bar{x})\cdot(y_i - \bar{y})$

- $x_i - \bar{x} > 0$ und $y_i - \bar{y} > 0$ | Beitrag zur Kovarianz positiv
- $x_i - \bar{x} < 0$ und $y_i - \bar{y} < 0$ | Beitrag zur Kovarianz positiv

- $x_i - \bar{x} > 0$ und $y_i - \bar{y} < 0$ | Beitrag zur Kovarianz negativ
- $x_i - \bar{x} < 0$ und $y_i - \bar{y} > 0$ | Beitrag zur Kovarianz negativ

Korrelationskoeffizient

Korrelationskoeffizient | Im Gegensatz zur Kovarianz skalenunabhängig

$$r(x,y) = \frac{cov(x,y)}{s(x) \cdot s(y)}$$

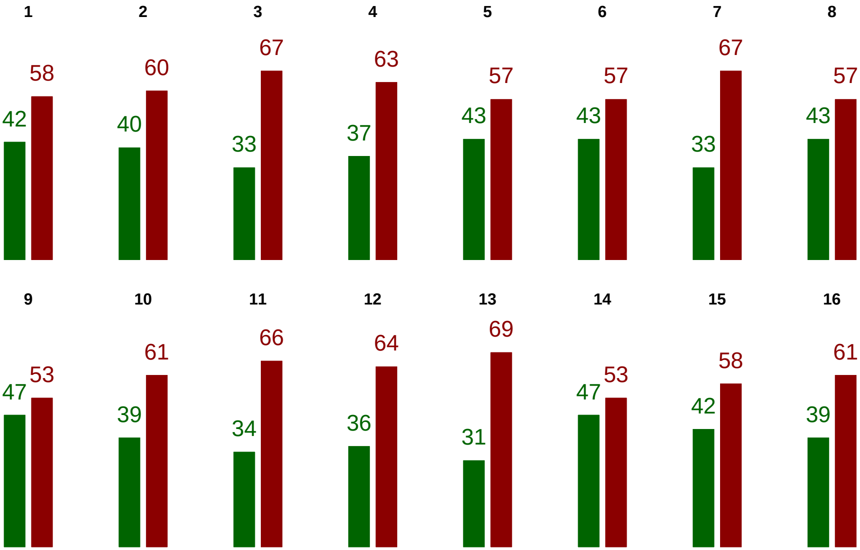
Pearson-Korrelation

Standardfehler und Konfidenzintervalle

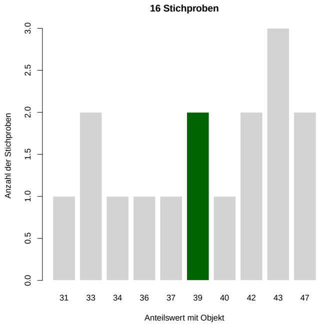
Anteilswerte und Stichproben

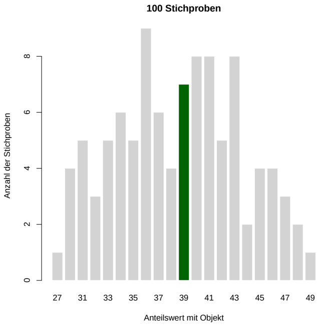
- Das Verb *essen* | Manchmal mit, manchmal ohne Akkusativ (direktes Objekt)
- Angenommenes wahres Verhältnis | Mit Objekt 39 %, ohne Objekt 61 %
- Viele Stichproben mit n=100 | Ergebnis nicht immer 39 zu 61
- 95%-Konfidenzintervall | In welchem Bereich liegen 95% aller Messwerte bei n=100?
- Güte von Stichproben einer bestimmten Größe angesichts gegebener Proportionen

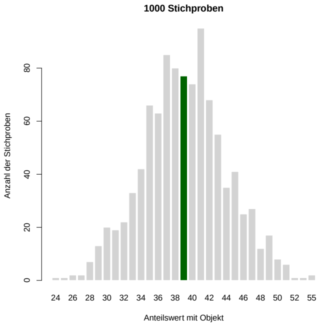
Sechzehn simulierte Stichprobenentnahmen (n=100)



Wiederholte Stichprobenentnahmen (n=100)







Standardfehler

- Die meisten p | Nah am wahren Wert P
- Sehr wenige p | Weit von P entfernt
- Bei unendlich vielen Messungen
  - Mittelwert der gemessenen Anteilswerte gleich P
  - Gemessene Anteilswerte normalverteilt um P
  - Standardabweichung der Messwerte um P bekannt → Standardfehler
- Standarfehler | Standardabweichung der Messwerte
  - Bei gegebener Stichprobengröße n
  - Bei einem bekannten Populationsanteil P

- Für einen wahren Anteilswert P
- Bei Stichprobengröße n

$$SF(P,n) = \sqrt{\frac{P \cdot (1-P)}{n}}$$

Bsp. für $P = 0.39$ und $n = 100$ | $SF(p) = \sqrt{\frac{0.39 \cdot (1-0.39)}{100}} = 0.0488$

Standardfehler | Interpretation

$$SF(P,n) = \sqrt{\frac{P \cdot (1-P)}{n}}$$

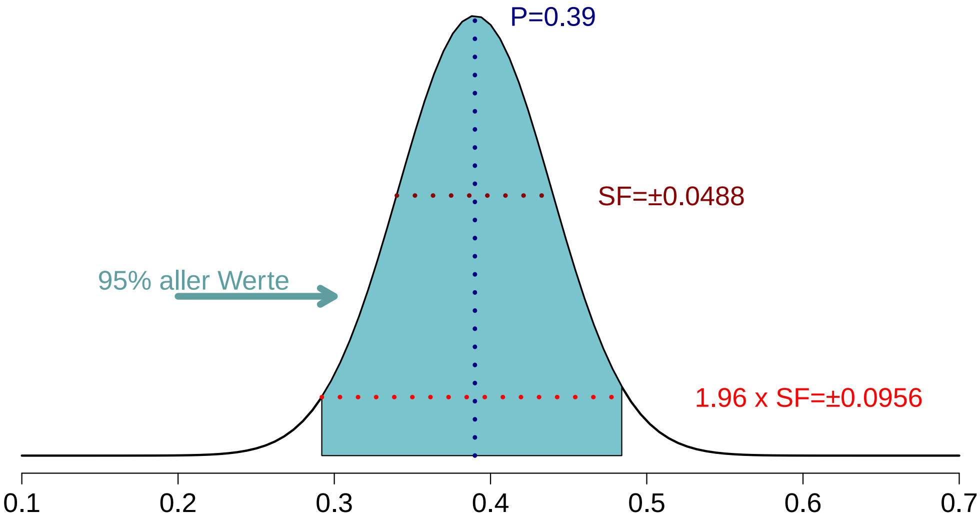
$$\text{Bsp.: } SF(0.39, 100) = \sqrt{\frac{0.39 \cdot (1-0.39)}{100}} = 0.0488$$

- Für beliebig viele Stichproben
- Bei Stichprobengröße $n = 100$
- Aus einer Grundgesamtheit mit wahrem Anteilswert $P = 0.39$

- Abweichung der gemessenen Anteile von $P = 0.39$ mit einem $SF = 0.0488$

Normal-/Gaussverteilung | Parameter Mittelwert und Standardabweichung

→ Mathematisch exhaustiv bekannt, Flächen unter der Kurve usw. berechenbar



Konfidenzintervall | z-Werte

- Stichproben normalverteilt
- z-Wert | Wie viele Standardfehler definieren 95% der Fläche unter der Kurve?

- Quantilfunktion der Normalverteilung | In R mit `qnorm()` oder Tabelle
- Quantilfunktion | Wie viele Standardabweichungen trennen auf jeder Seite 2.5% ab?
- `qnorm(0.025, lower.tail=FALSE)` → $z(0.95) = 1.96$

Konfidenzintervall | Standardfehler um wahren Anteilswert

- Standardfehler | Standardabweichung der Stichprobenwerte
- Konfidenzbreite | z-Wert multipliziert mit Standardfehler
- 95% der Werte | Intervall Wahrer Anteilswert ± Konfidenzbreite

$$KI(P, n, s) = P \pm z(s) \cdot SF(P, n)$$

Bsp.: $KI(0.39, 100, 0.95) = 0.39 \pm 1.96 \cdot 0.0488 = 0.39 \pm 0.096 = [0.29, 0.49]$

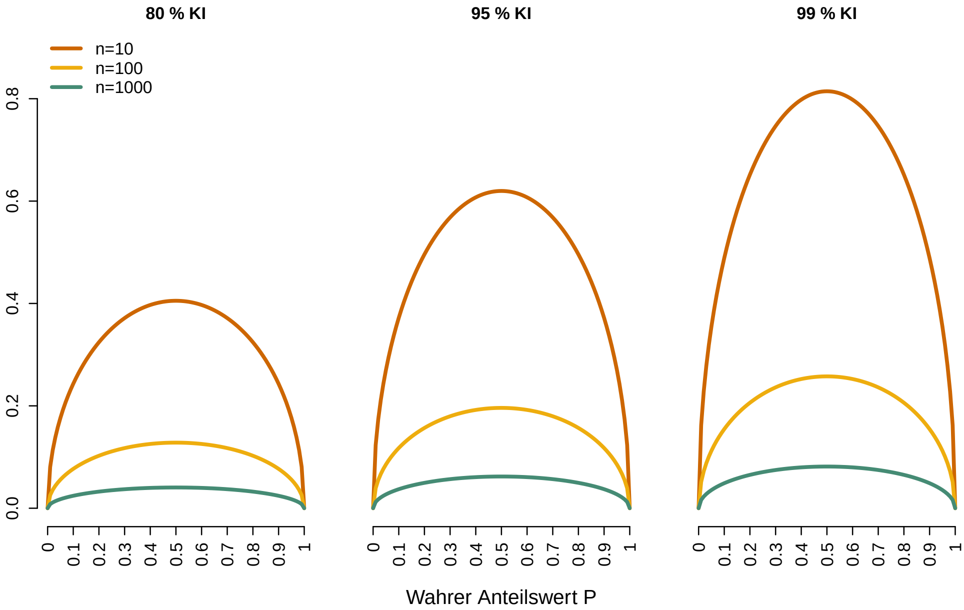
Interpretation

Konfidenzintervall im Beispiel | 0.29 bis 0.49

In 95% aller Stichproben mit $n = 100$ liegt der Messwert zwischen 0.29 und 0.49 bei einem wahren Anteil von 0.39.

- Praxis | Wahrer Anteil nicht bekannt, daher Schätzung aus Stichprobenanteil p
- Der gemessene Anteil p kann aber eine totale Fehlschätzung sein!
- Die Philosophie bezieht sich auf wiederholte Messungen.
- Entweder liegt der gemessene Wert im Konfidenzintervall, oder ein seltenes Ereignis ist eingetreten.
- Wir sind **nicht** zu 95% sicher, dass der wahre Wert zwischen 0.29 und 0.49 liegt!

Konfidenzintervall | Breite bei verschiedenen P, n und Niveaus



Nächste Woche | Überblick

Einzelthemen

Literatur I

Bortz, Jürgen & Christof Schuster. 2010. *Statistik für Human- und Sozialwissenschaftler*. 7. Aufl. Berlin: Springer.

Gravetter, Frederick J. & Larry B. Wallnau. 2007. *Statistics for the Behavioral Sciences*. 7. Aufl. Belmont: Thomson.

Autor

Kontakt

Prof. Dr. Roland Schäfer
Institut für Germanistische Sprachwissenschaft
Friedrich-Schiller-Universität Jena
Fürstengraben 30
07743 Jena

https://rolandschaefer.net
roland.schaefer@uni-jena.de

Lizenz