Deutsche Syntax
02. Deskriptive Statistik

Roland Schäfer

Institut für Germanistische Sprachwissenschaft
Friedrich-Schiller-Universität Jena

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

Inhalt

Deskriptiv

Übersicht

- deskriptive Statistik als Datenaggregation
- Verteilungen in Stichproben und Grundgesamtheiten:
  - Zentralmaße
  - Streuung (Varianz)
- Beziehungen zwischen ko-variierenden Messungen
- Genauigkeiten von Schätzungen quantifizieren (Konfidenzintervalle)

Literatur

- Gravetter & Wallnau (2007)
  Achtung! Vermittelt eine falsche Philosophie!
  Nur für die Mathematik benutzen.
- Bortz & Schuster (2010)

Zweck der deskriptiven Statistik

- Mit unbewaffnetem Auge auf Datenmengen zu blicken, ist meistens sinnlos.
- In großen Zahlenkolonnen sehen Menschen
  nur schlecht Tendenzen und Zusammanhänge.
- Um dies zu erleichtern, gruppieren und visualisieren wir die Daten.

Grundlegende Kenndaten, die man berichtet

- Definition und (geschätzte) Größe der Grundgesamtheit.
  (z. B. alle lebenden deutschen Erwachsenen)
- Stichprobengröße (N)
- Stichprobenmethode
  - Zufallsstichprobe (größere Stichprobe)
  - proportional stratifizierte Stichprobe (Quotenstichprobe)

Messvariablen und Skalenniveaus

Variablen sind folgendermaßen skaliert:

- dichotom (binär) = zwei Kategorien:
  männlich, weiblich ; Präteritum, Perfekt
- nominal (kategorial) = disjunkte Kategorien ohne numerische Interpretation:
  Parteizugehörigkeit ; NP, AP, VP
- ordinal = disjunkte Kategorien, nach Rang geordnet:
  Schulnoten ; 5-point oder 7-point scales (Likert scales)
- intervall~ = geordnete Werte mit definierten Abständen,
  aber mit arbiträrem Nullpunkt: Celsius
- verhältnis~ = wie intervall-,
  aber der Nullpunkt ist ein echter Nullpunkt: Kelvin

Intervalle vs. Verhältnisse

- Wir messen die Größe von Menschen in cm auf einer Verhältnisskala.
  - 200cm sind das doppelte von 100cm.
  - Niemand kann kleiner sein als 0cm.

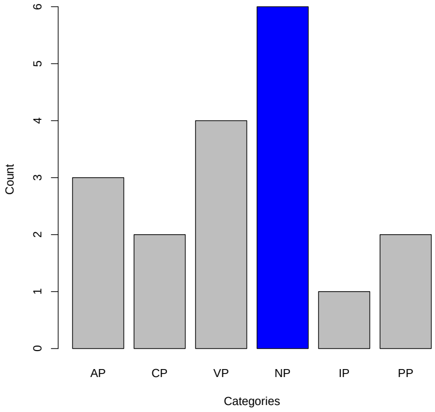
- Dieselbe Messung als Abweichung vom Mittel ergibt eine Intervallskala.
  - Wer 3 cm größer ist als der Durchschnitt ist doppelt soviel größer wie jemand, der 1.5 cm größer ist.
  - Die erste Person ist aber nicht doppelt so groß wie die zweite.
  - Außerdem kann man z.B. -3 cm vom Durchschnitt abweichen.

Relevanz der Skalenniveaus

- Das SN bestimmt die zulässigen mathematischen Operationen (z.B. Rechenarten).
- Also kommen je nach SN nur bestimmte deskriptive Statistiken in Frage.
- Das gleiche gilt für die Zulässigkeit bestimmter inferenzstatistischer Tests je nach Skalenniveau.

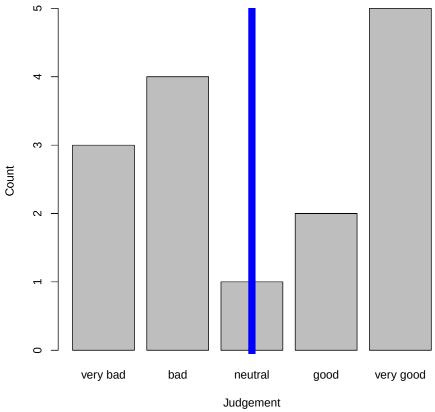
Zentraltendenz I

Der Modus ist der häufigste Wert in einer Grundgesamtheit oder Stichprobe.
Geht bei jedem Skalenniveau.



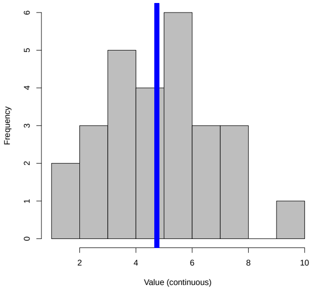
Zentraltendenz II

Der Median ist der Wert über und unter dem gleichviele Werte liegen. Ordinalskala oder höher.



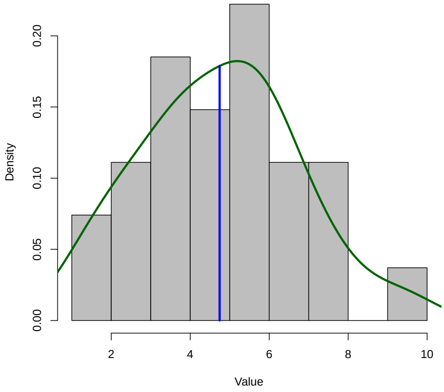
Das arithmetische Mittel $\bar{x}$ ist die Summe aller Werte x dividiert durch Stichprobengröße n.
Intervallskala oder höher.

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



Zentraltendenz IV

Kontinuierliche Variablen und ihr arithmetisches Mittel lassen sich in Dichteplots gut visualisieren (per Software).

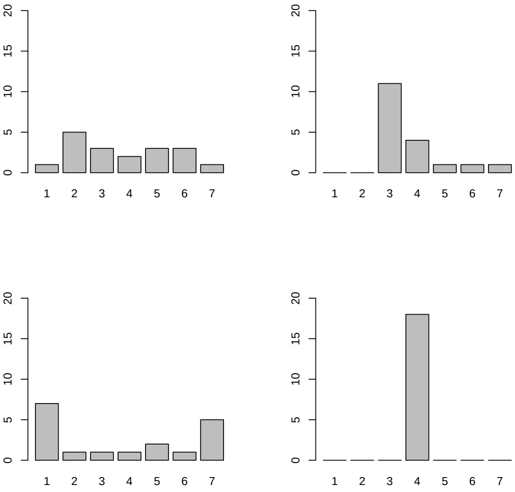


Warum sind Dispersionsmaße wichtig?

1. Das Wissen um die Zentraltendenz ist wichtig
als grobe allgemeine Information über die Population.
2. Aber dieselbe Zentraltendenz kann
das Ergebnis ganz verschiedener Werte sein.
3. Die Verteilung kann flach, chaotisch, glockenförmig usw. sein.

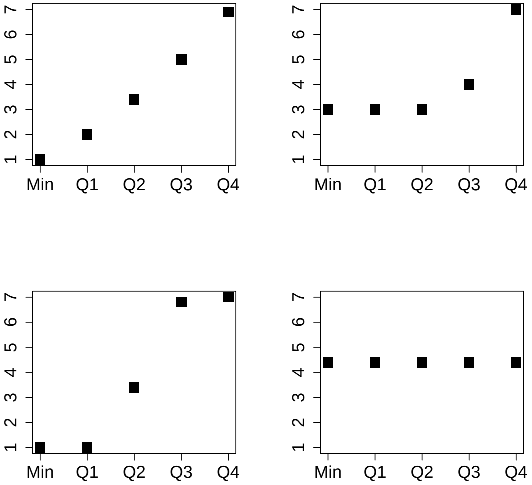
Verteilungsformen

Histogramme von vier Stichproben
mit $\bar{x} = 4.389$ und $n = 18$.



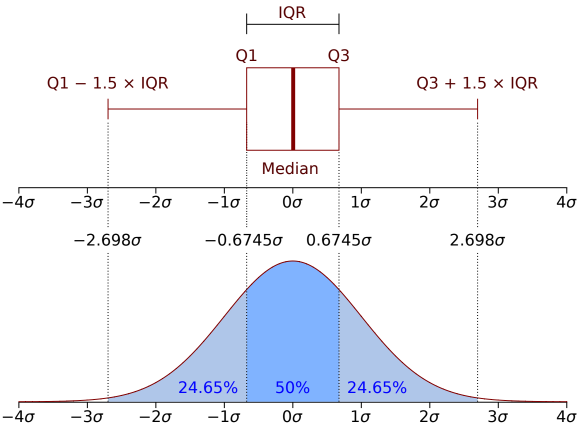
Quartile

Quartile sind die Punkte, unterhalb derer 25%, 50%, 75% und 100% (Maximum) der Werte liegen. Dazu gibt es noch das Minimum (niedrigster Wert).



Quartile und Inter-Quartil-Bereich

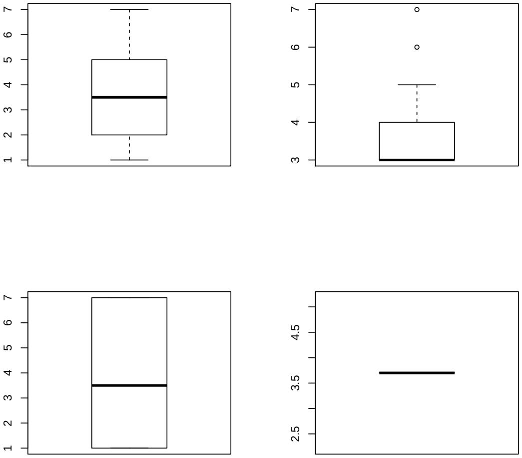
$IQR = Q_3 - Q_1$
oder ganz einfach: die mittleren 50%



Attribution: Jhguch (`http://en.wikipedia.org/wiki/User:Jhguch`) at en.wikipedia

Boxplots als bessere Zusammenfassung

Boxplots zeigen Median (Linie in der Mitte), oberes und unteres Quartil (Boxen), 1,5-fachen Interquartilabstand zu diesen (gestrichelte Hebel) und Ausreißer (Punkte).



Varianz und Standardabweichung

Die Varianz s^2 ist die quadrierte mittlere Abweichung vom Mittel:

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

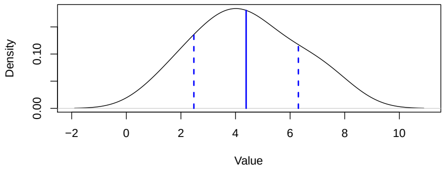
Die Standardabweichung s ist die Quadratwurzel der Varianz:

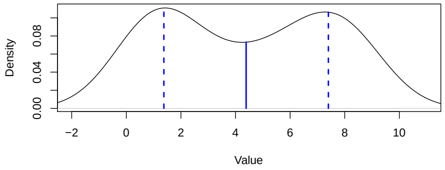
$$s(x) = \sqrt{s^2(x)}$$

Der Zählerterm der Varianz heißt auch Summe der Quadrate:

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

Die erste Stichprobe hat $s = 1.91$,
die zweite $s = 3.01$ (beide $\bar{x} = 4.389$).





z-Wert

Um wie viele Standardabweichungen weicht jeder Datenpunkt vom Mittel ab?
Für jeden Punkt: $z(x_i) = \frac{x_i - \bar{x}}{s(x)}$

Bsp.: $x = [3.9, 4.3, 7.2, 8.5, 11.1, 12.1, 14.0, 20.7]$

$\bar{x} = 10.225$

$s^2(x) = \frac{(3.9-10.255)^2 + ... + (20.7-10.225)^2}{8-1} = \frac{215.495}{7} = 30.785$

$s(x) = \sqrt{30.785} = 5.548$

$z = [\frac{3.9-10.225}{5.548}, ..., \frac{20.7-10.225}{5.548}] =$

$[-1.140, -1.068, -0.545, -0.311, 0.158, 0.338, 0.680, 1.888]$

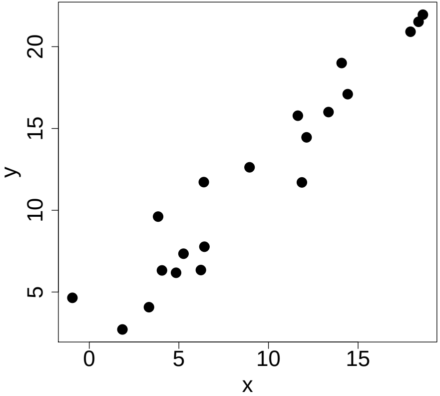
Zähldaten von zwei Variablen

Zähldaten von zwei Variabeln (egal wieviel Ausprägungen) sind ideal als Kreuztabelle darstellbar.

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

Korrelationen

Korrelationskoeffizienten helfen, den Zusammenhang zwischen Variablen, die mindestens ordinalskaliert sind, numerisch zu erfassen.
Z. B. die hier geplotteten *x* und *y*:



Kovarianz

Die Kovarianz kombiniert die Maße, zu denen die zwei Messwerte pro Datenpunkt vom jeweiligen Mittel der Messwertreihen abweichen.

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

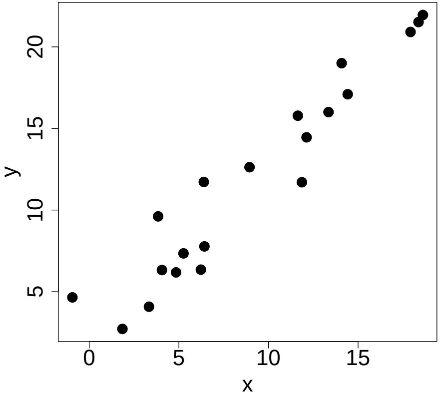
Sind $x_i - \bar{x}$ und $y_i - \bar{y}$ positiv oder negativ, ist der Beitrag ihres Produkts zur Kovarianz positiv, bei ungleichen Vorzeichen negativ.

Der Zählerterm heißt auch Summe der Produkte:

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

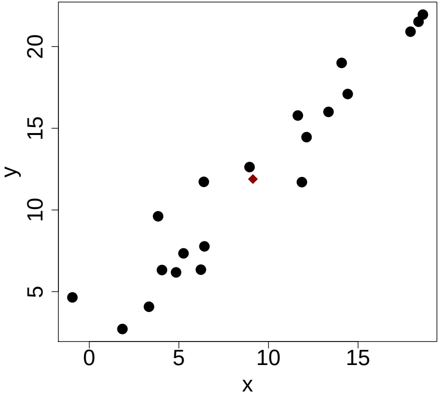
Kovarianz: Illustration 1

Zwei Messvariablen (Vektoren): x und y



Kovarianz: Illustration 2

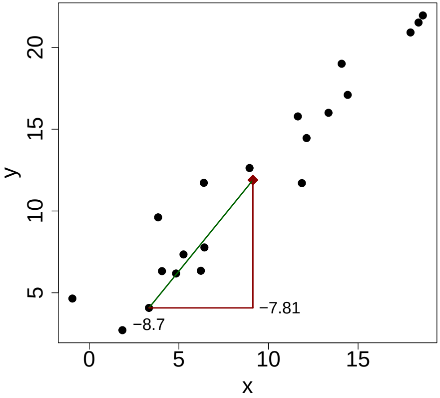
Koordinate von $\langle \bar{x}, \bar{y} \rangle$



Kovarianz: Illustration 3

Punktvarianzen: $x_3 - \bar{x} = -7.81$ und $y_3 - \bar{y} = -5.80$

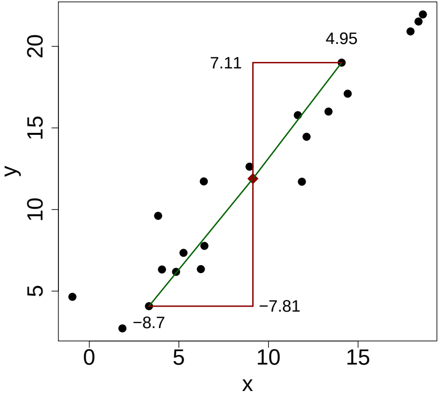
$-7.81 \cdot -5.80 = 45.30$



Kovarianz: Illustration 4

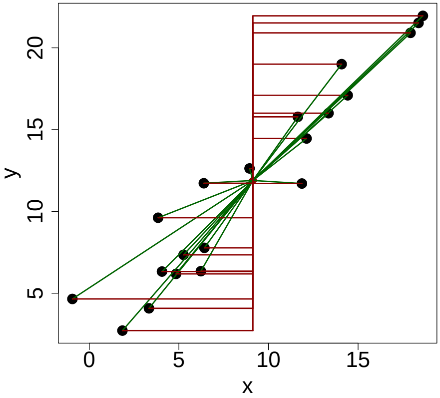
Punktvarianzen: $x_{17} - \bar{x} = 4.95$ und $y_{17} - \bar{y} = 7.11$

$4.95 \cdot 7.11 = 35.19$



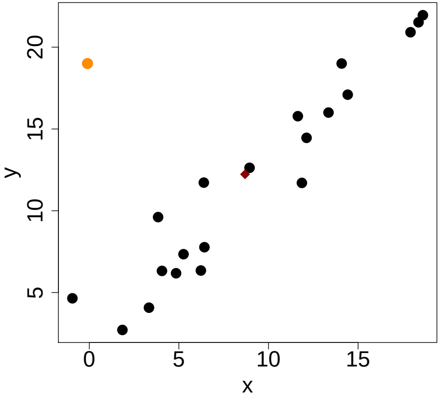
Kovarianz: Illustration 5

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



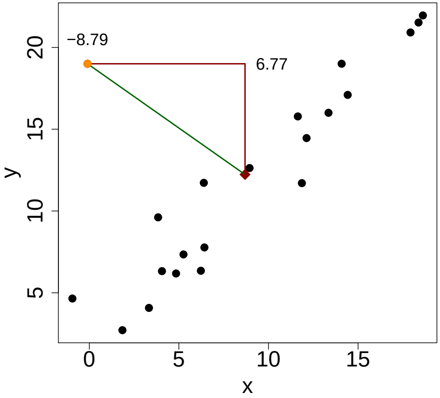
Kovarianz: Illustration 6

„Ausreißer“ bei – im Prinzip – positiver Kovarianz:
Negatives Produkt der Punktvarianzen



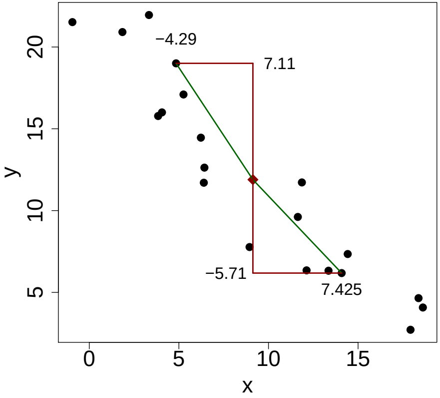
Punktvarianzen: $x_{21} - \bar{x} = 6.77$ und $y_{21} - \bar{y} = -8.79$

$6.77 \cdot -8.79 = -59.51$



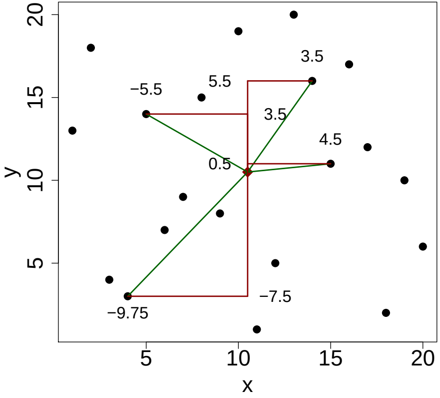
Kovarianz: Negative Kovarianz

Wenn die Abhängigkeit zwischen den Werten tendentiell negativ ist, sind die Produkte der Punktvarianzen überwiegend negativ.
$cov(x, y) = -33.77$



Kovarianz: Null annähernd

Wenn es keine besondere Abhängigkeit gibt,
nähert sich die Kovarianz 0:
$cov(x, y) = -1.74$



Kovarianz zu Korrelationskoeffizient r

Während die Kovarianz von der Größe der Werte abhängt,
macht der Korrelationskoeffizient Kovarianzen vergleichbar:

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

Dies ist die Pearson-Korrelation, später kommen noch andere Korrelationen.

Anteilswerte, Beispiel

- Das Verb *essen* kommt manchmal mit, manchmal ohne Akkusativ (direktes Objekt) vor.
- mit dO 39, ohne dO 61.
- Wenn wir in dieser Situation Stichproben mit n=100 ziehen, werden wir nicht immer genau diese Werte messen, sondern sie zwar häufig gut approximieren, manchmal aber auch stark abweichende Anteilswerte messen.
- In welchem Bereich liegen 95% aller Messwerte bei n=100?
- Diese Frage beantwortet das 95%-Konfidenzintervall.
- Es sagt uns, wie gut Stichproben einer bestimmten Größe bestimmte Anteilswerte approximieren.

Argumentation zum KI

- Annahme: Wahrer Anteilswert in der Grungesamtheit ist P.
- In Stichproben der Größe n misst man einen Stichprobenanteil p.
- Die meisten p liegen nah an P, sehr wenige weit weg davon.
- Wenn man beliebig viele p hat, verteilen sie sich so um P, dass eine Standardabweichung dem Standardfehler entspricht.
- Der Standardfehler ist der Erwartungswert für die Standardabweichung sehr vieler Messwerte (um den wahren Wert).
- Außerdem weiß man, dass die p normalverteilt um P sind. Das folgt für groß genuge Stichproben aus dem Zentralen Grenzwertsatz.
- Bei einer Normalverteilung weiß man, wieviel Prozent der Messwerte in einem Bereich $\pm q \cdot s$ (für beliebige q) vom Mittel liegen.

Erstens: Standardfehler

Wir brauchen also für Stichproben der Größe n den SF für den tatsächlichen Anteilswert P.

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

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

Zum Standardfehler

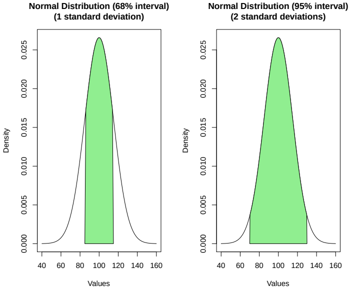
$$SF(p) = \sqrt{\frac{p \cdot (1-p)}{n}}$$

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

- Anders gesagt: Wenn man beliebig viele Stichproben der Größe $n = 100$ aus einer Grundgesamtheit zieht, in der der wahre Anteilswert $P = 0.39$ ist, ist eine Standardabweichung aller p (also der Standardfehler) $SF = 0.0488$.

Normalverteilung und z-Wert für Konfidenzniveau

- Um das KI für die gewünschte Konfidenzniveau zu ermitteln, müssen wir wissen, wie sich Werte um das geschätzte Mittel verteilen.
- Schätzverteilung dank Zentralem Grenzwertsatz: Normalverteilung
- Vorteil: Es ist genau bekannt, wieviel Werte je nach s in einem bestimmten Intervall liegen.



Zweitens: z-Wert

- Wir müssen nun wissen, wieviele Standardabweichungen bei der Normalverteilung 95% der Fläche definieren.
- Wenn es symmetrische 95% werden sollen, müssen oben und unten je 2.5% abgetrennt werden.
- Dazu gibt es Tabellen oder die Quantil-Funktion der Normalverteilung `qnorm()` in R.
- `qnorm(0.025, lower.tail=FALSE)` $\Rightarrow 1.959964$
- Also: $z = 1.96$

Drittens: Konfidenzintervall

- Da der Standardfehler genau einer Standardabweichung entspricht, muss er nun mit dem z-Wert multipliziert werden.

$$KI = p \pm z \cdot SF(p)$$

Bsp.: $KI = 0.39 \pm 1.96 \cdot 0.0488 = 0.39 \pm 0.096 = 0.29, 0.49$

Interpretation

Das Konfidenzinterval ist in unserem Fall also

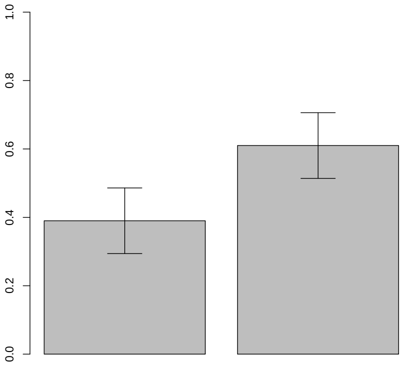
0.29 bis 0.49

- In 95% aller Stichproben mit $n = 100$ läge der Messwert beim wahren Anteil von 0.39 zwischen 0.29 und 0.49.
- Oft wird auf Basis einer Stichprobe mit der Göße n ein Anteilswert p geschätzt und dann für diesen das Konfidenzintervall ausgerechnet.
- Das kann man zwar machen, aber man lernt dadurch nichts über die GG!
- Ggf. kann uns das so errechnete KI einen Eindruck davon geben, wie genau Stichproben der Größe n bei einem Anteil wie dem gemessenen ungefähr sind.

- Der gemessene Anteil p kann aber eine totale Fehlschätzung sein!
- Die Philosophie bezieht sich auf das wiederholte Berechnen von KIs.

Verboten: Balkendiagramm mit Konfidenzintervall

Ein solches Diagramm signalisiert fälschlicherweise,
dass das Konfidenzintervall uns etwas über die GG sagt!



Nächste Woche | Überblick

Einzelthemen

1. Statistik, Inferenz und probabilistische Grammatik
2. Deskriptive Statistik
3. Nichtparametrische Verfahren
4. z-Test und t-Test
5. ANOVA
6. Freiheitsgrade und Effektstärken
7. Power
8. Lineare Modelle
9. Generalisierte Lineare Modelle
10. Gemischte Modelle

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