

Statistik

02. Deskriptive Statistik

Roland Schäfer

Institut für Germanistische Sprachwissenschaft
Friedrich-Schiller-Universität Jena

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

1 Motivation

2 Skalenniveau

3 Zentraltendenz

4 Empirische Verteilungen und Dispersion

5 Bivariate Statistiken

6 Standardfehler und Konfidenzintervalle

7 Nächste Woche | Überblick

- Deskriptive Statistik als Aggregation von Daten

- Deskriptive Statistik als Aggregation von Daten
- Verteilungen in Stichproben und Grundgesamtheiten:

- Deskriptive Statistik als Aggregation von Daten
- Verteilungen in Stichproben und Grundgesamtheiten:
 - ▶ Zentralmaße

- Deskriptive Statistik als Aggregation von Daten
- Verteilungen in Stichproben und Grundgesamtheiten:
 - ▶ Zentralmaße
 - ▶ Streuung (Varianz)

- Deskriptive Statistik als Aggregation von Daten
- Verteilungen in Stichproben und Grundgesamtheiten:
 - Zentralmaße
 - Streuung (Varianz)
- Theoretische vs. empirische Verteilungen

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

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

Motivation

- Mit unbewaffnetem Auge auf Daten zu blicken, ist meistens zwecklos.

- Mit unbewaffnetem Auge auf Daten zu blicken, ist meistens zwecklos.
- In Zahlen sehen Menschen nur schlecht Tendenzen und Zusammenhänge.

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

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

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

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

- Definition der Grundgesamtheit

Was will man wissen?

- Definition der Grundgesamtheit
- Stichprobengröße (n)

Was will man wissen?

- Definition der Grundgesamtheit
- Stichprobengröße (n)
 - ▶ 200 Sätze aus dem Korpus

Was will man wissen?

- Definition der Grundgesamtheit
- Stichprobengröße (n)
 - ▶ 200 Sätze aus dem Korpus
 - ▶ 1.000 Reaktionen (von 50 Probanden) im Experiment

- Definition der Grundgesamtheit
- Stichprobengröße (n)
 - ▶ 200 Sätze aus dem Korpus
 - ▶ 1.000 Reaktionen (von 50 Probanden) im Experiment
 - ▶ Was sind die elementaren gemessenen Datenpunkte?

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

- 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

- 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 | Stratifizierung und Begründung

Skalenniveau

- dichotom/binär | Menge $\{A, B\}$ | zwei disjunkte Kategorien
männlich, weiblich ; Präteritum, Perfekt

Messvariablen und Skalenniveaus

- dichotom/binär | Menge $\{A, B\}$ | zwei disjunkte Kategorien
männlich, weiblich ; Präteritum, Perfekt
- nominal/kategorial | Menge $\{A, B, ..\}$ | disjunkte Kategorien
Parteilzugehörigkeit ; NP, AP, VP

Messvariablen und Skalenniveaus

- dichotom/binär | Menge $\{A, B\}$ | zwei disjunkte Kategorien
männlich, weiblich ; Präteritum, Perfekt
- nominal/kategorial | Menge $\{A, B, ..\}$ | disjunkte Kategorien
Parteilzugehörigkeit ; 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

Messvariablen und Skalenniveaus

- dichotom/binär | Menge $\{A, B\}$ | zwei disjunkte Kategorien
männlich, weiblich ; Präteritum, Perfekt
- nominal/kategorial | Menge $\{A, B, ..\}$ | disjunkte Kategorien
Parteilzugehörigkeit ; 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

Messvariablen und Skalenniveaus

- dichotom/binär | Menge $\{A, B\}$ | zwei disjunkte Kategorien
männlich, weiblich ; Präteritum, Perfekt
- nominal/kategorial | Menge $\{A, B, ..\}$ | disjunkte Kategorien
Parteilzugehörigkeit ; 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

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
- Zählraten | Keine beobachtbaren Variablen, sondern
Aggregation von dichotomen, nominalen oder ordinalen Variablen

- Verhältnisskala | Größe von Menschen in cm

- Verhältnisskala | Größe von Menschen in cm
 - ▶ $200\text{cm} = 2 \times 100\text{cm}$ usw.

- Verhältnisskala | Größe von Menschen in cm
 - ▶ $200\text{cm} = 2 \times 100\text{cm}$ usw.
 - ▶ Keine Messung unter 0cm

- Verhältnisskala | Größe von Menschen in cm
 - ▶ $200\text{cm} = 2 \times 100\text{cm}$ usw.
 - ▶ Keine Messung unter 0cm
- Intervallskala | Dasselbe als Abweichung vom Mittel

- Verhältnisskala | Größe von Menschen in cm
 - ▶ $200\text{cm} = 2 \times 100\text{cm}$ usw.
 - ▶ Keine Messung unter 0cm
- Intervallskala | Dasselbe als Abweichung vom Mittel
 - ▶ $4\text{cm} = 2 \times 2\text{cm}$ usw.

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

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

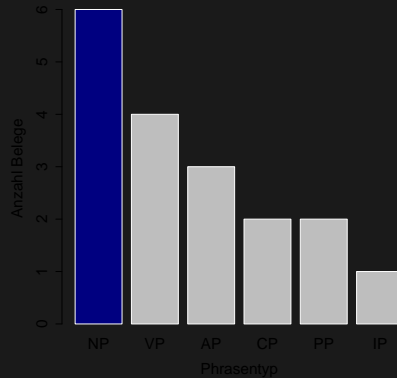
- Bestimmung zulässiger mathematischer Operationen

- Bestimmung zulässiger mathematischer Operationen
- Deskriptive Statistiken je nach Skalenniveau

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

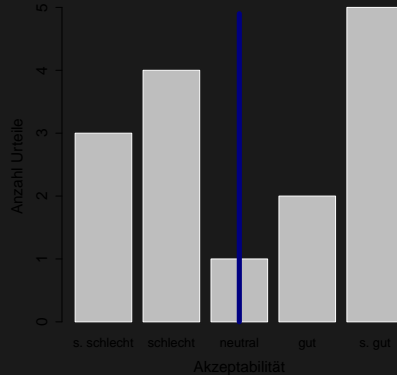
Zentraltendenz

Modus | Der häufigste Wert | **Alle Skalenniveaus**



Zentraltendenz II

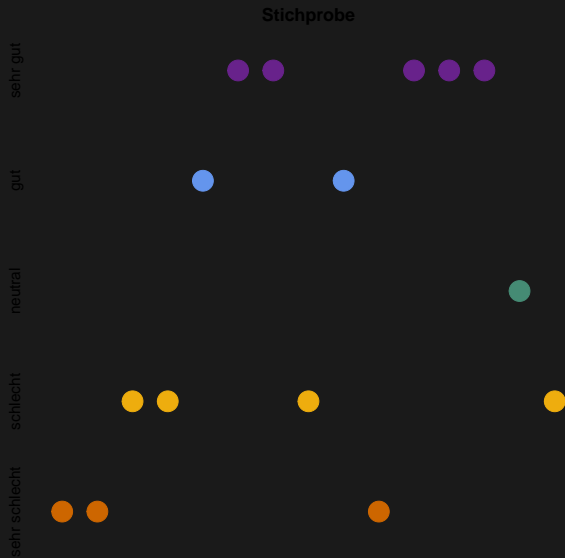
Median | Mitte der sortierten Stichprobe | **ab Ordinalskala**



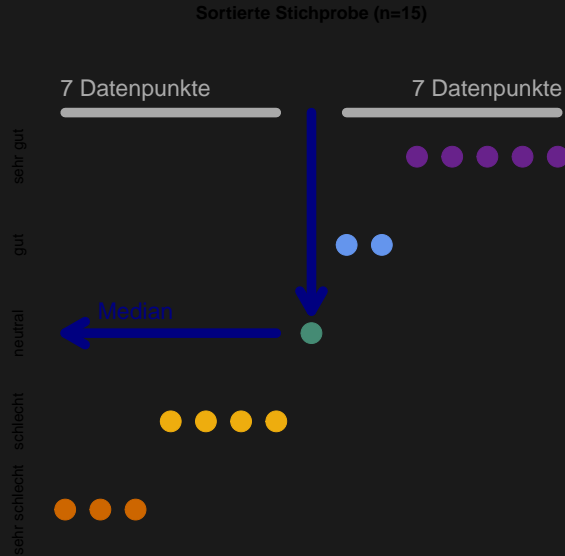
Numerische Messungen | Verschiedene Interpolationsmethoden

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

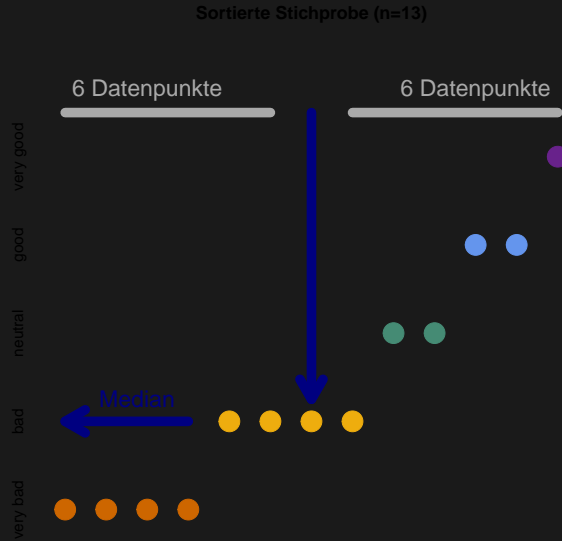
Median bestimmen | Stichprobe



Median bestimmen | Sortierte Stichprobe

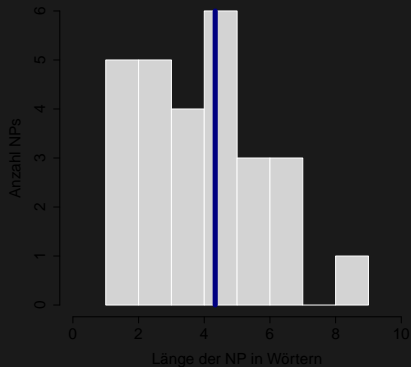


Median bestimmen | Verzerrente sortierte Stichprobe



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

Warum sind Dispersionsmaße wichtig?

Dispersion | Streuung der Daten

- Zentraltendenz | Orientierung über Tendenzen der Stichprobe
- Ein Maß für Zentraltendenz für **beliebig viele Verteilungsformen**

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

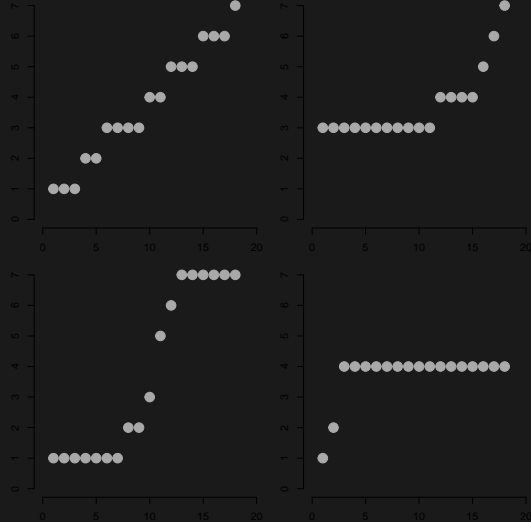
Warum sind Dispersionsmaße wichtig?

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

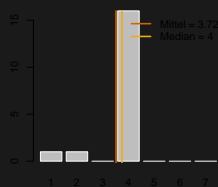
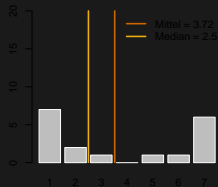
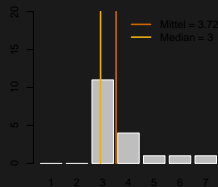
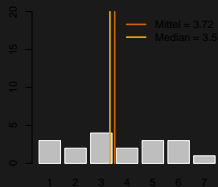
Jeder Punkt entspricht einem Datenpunkt/einer Messung!



Verteilungsformen

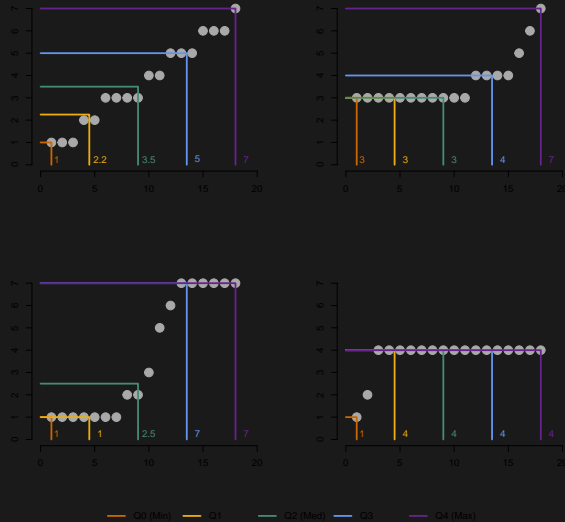
Histogramme | Vier Stichproben mit $\bar{x} = 3.72$ und $n = 18$

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



Quartile

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



- Interquartilbereich $IQR = Q_3 - Q_1$ | Die mittleren 50 %

- Interquartilbereich $IQR = Q_3 - Q_1$ | Die mittleren 50 %
- Boxplots

- Interquartilbereich $IQR = Q_3 - Q_1$ | Die mittleren 50 %
- Boxplots
 - ▶ Median | Linie in der Mitte

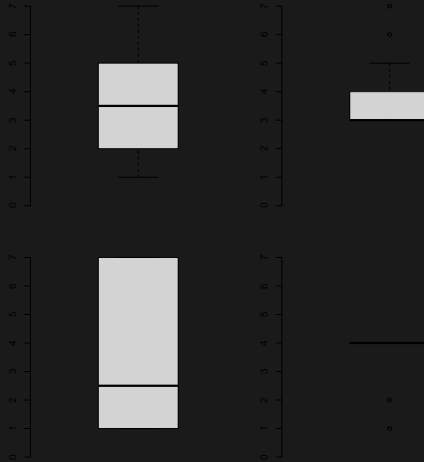
- Interquartilbereich $IQR = Q_3 - Q_1$ | Die mittleren 50 %
- Boxplots
 - ▶ Median | Linie in der Mitte
 - ▶ Oberes und unteres Quartil | Boxen

- 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

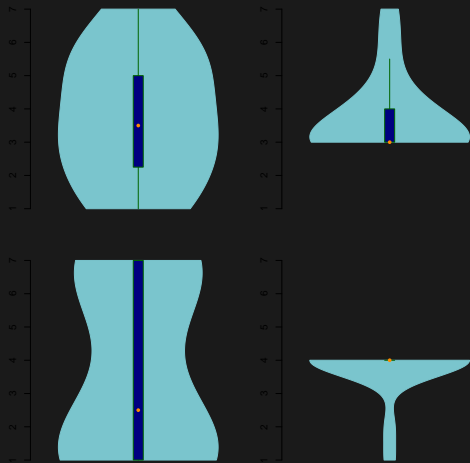
- Interquartilbereich $IQR = Q_3 - Q_1$ | Die mittleren 50 %
- Boxplots
 - ▶ Median | Linie in der Mitte
 - ▶ Oberes und unteres Quartil | Boxen
 - ▶ 1,5-facher Interquartilabstand | gestrichelte Hebel
 - ▶ Ausreißer | Punkte

- 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

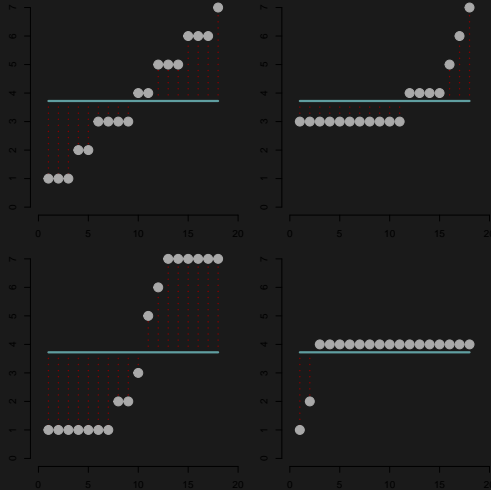


Violinplots | Die noch bessere Zusammenfassung



Was bestimmt die Varianz?

Die Distanzen der Messwerte zum Mittel sind unterschiedlich groß.



Varianz s^2 | Quadrierte mittlere Abweichung vom Mittelwert

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

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

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

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

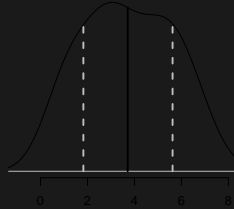
$$s(x) = \sqrt{s^2(x)}$$

Summe der Quadrate | Zählerterm der Varianz

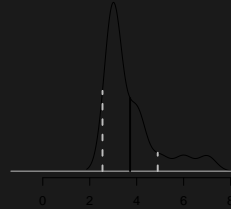
$$SQ(x) = \sum_{i=1}^n (x_i - \bar{x})^2$$

Unterschiedliche Standardabweichungen

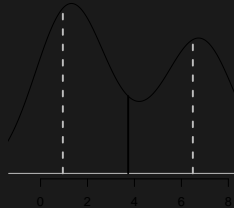
sd = 1.9



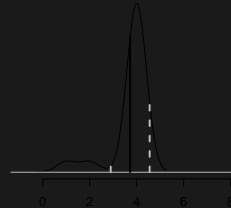
sd = 1.18



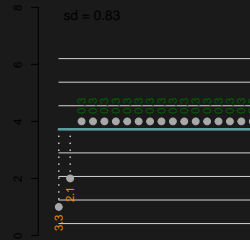
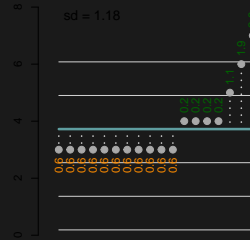
sd = 2.76



sd = 0.83



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



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

- Bsp.: $x = [3.9, 4.3, 7.2, 8.5, 11.1, 12.1, 14.0, 20.7]$
 - ▶ $\bar{x} = 10.225$

- Bsp.: $x = [3.9, 4.3, 7.2, 8.5, 11.1, 12.1, 14.0, 20.7]$
 - ▶ $\bar{x} = 10.225$
 - ▶ $s^2(x) = \frac{(3.9-10.225)^2 + \dots + (20.7-10.225)^2}{8-1} = \frac{215.495}{7} = 30.785$

- Bsp.: $x = [3.9, 4.3, 7.2, 8.5, 11.1, 12.1, 14.0, 20.7]$
 - ▶ $\bar{x} = 10.225$
 - ▶ $s^2(x) = \frac{(3.9-10.225)^2 + \dots + (20.7-10.225)^2}{8-1} = \frac{215.495}{7} = 30.785$
 - ▶ $s(x) = \sqrt{30.785} = 5.548$

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

Bivariate Statistiken

Zähldaten von zwei Variablen

Kreuztabelle | Darstellung der Zähldaten zweier Variablen

	<hr/>	
	Variable 1 Wert 1	Wert2
Variable 2 Wert 1	Anzahl x_{11}	Anzahl x_{12}
Wert 2	Anzahl x_{21}	Anzahl x_{22}
	<hr/>	

Standardfehler und Konfidenzintervalle

- Das Verb *essen* | Manchmal mit, manchmal ohne Akkusativ (direktes Objekt)

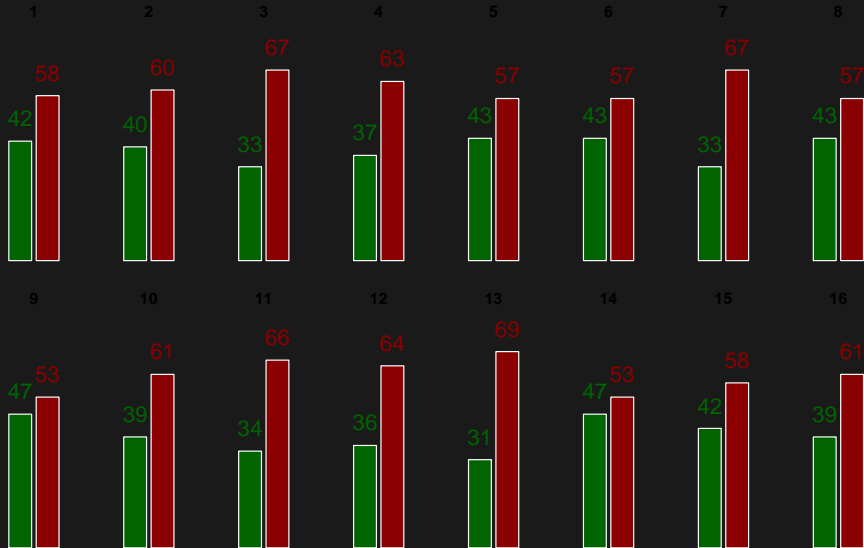
- Das Verb *essen* | Manchmal mit, manchmal ohne Akkusativ (direktes Objekt)
- Angenommenes wahres Verhältnis | Mit Objekt 39 %, ohne Objekt 61 %

- 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

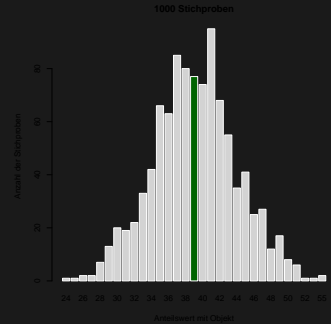
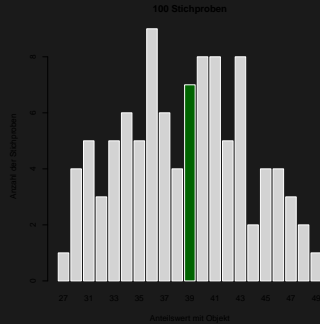
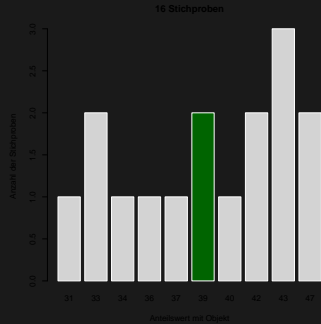
- Das Verb *essen* | Manchmal mit, manchmal ohne Akkusativ (direktes Objekt)
- Angenommenes wahres Verhältnis | Mit Objekt 39 %, ohne Objekt 61 %
- Viele Stichproben mit $n=100$ | Ergebnis **nicht** immer 39 zu 61
- 95%-Konfidenzintervall | **In welchem Bereich liegen 95% aller Messwerte bei $n=100$?**

- 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)



- Die meisten p | Nah am wahren Wert P

- Die meisten p | Nah am wahren Wert P
- Sehr wenige p | Weit von P entfernt

- Die meisten p | Nah am wahren Wert P
- Sehr wenige p | Weit von P entfernt
- Bei unendlich vielen Messungen

- 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

- 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

- 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

- 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
- Standardfehler | Standardabweichung der Messwerte

- 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
- Standardfehler | Standardabweichung der Messwerte
 - Bei gegebener Stichprobengröße n

- 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
- Standardfehler | Standardabweichung der Messwerte
 - ▶ Bei **gegebener Stichprobengröße n**
 - ▶ Bei einem **bekannten Populationsanteil P**

- Für einen wahren Anteilswert P

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

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

- Für einen wahren Anteilswert P
- Bei Stichprobengröße n

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

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

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

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

- Für beliebig viele Stichproben

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

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

- Für beliebig viele Stichproben
- Bei Stichprobengröße $n = 100$

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

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

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

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

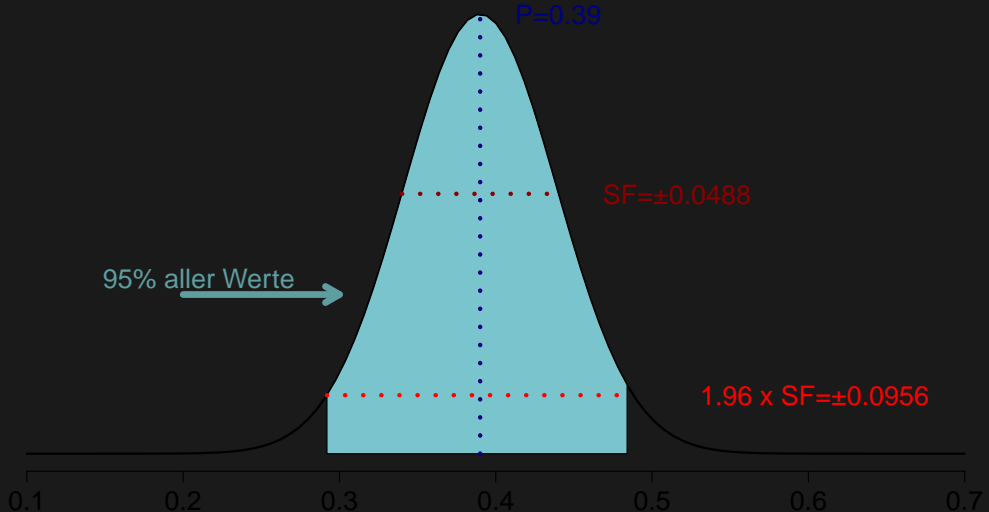
$$\text{Bsp.: } SF(0.39, 100) = \sqrt{\frac{0.39 \cdot (1-0.39)}{100}} = 0.0488$$

- 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 $SF = 0.0488$

Konfidenzintervall | Standardfehler und Normalverteilung

Normal-/Gaussverteilung | Parameter Mittelwert und Standardabweichung

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



- Stichproben normalverteilt

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

- Stichproben normalverteilt
- z-Wert | Wie viele Standardfehler definieren 95% der Fläche unter der Kurve?
- Quantilfunktion der Normalverteilung | In R mit `qnorm()` oder Tabelle

- 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?

- 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)` → $z(0.95) = 1.96$

- Standardfehler | Standardabweichung der Stichprobenwerte

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

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

- Standardfehler | Standardabweichung der Stichprobenwerte
- Konfidenzbreite | **z-Wert multipliziert mit Standardfehler**

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

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

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

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

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

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

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!

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.

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.

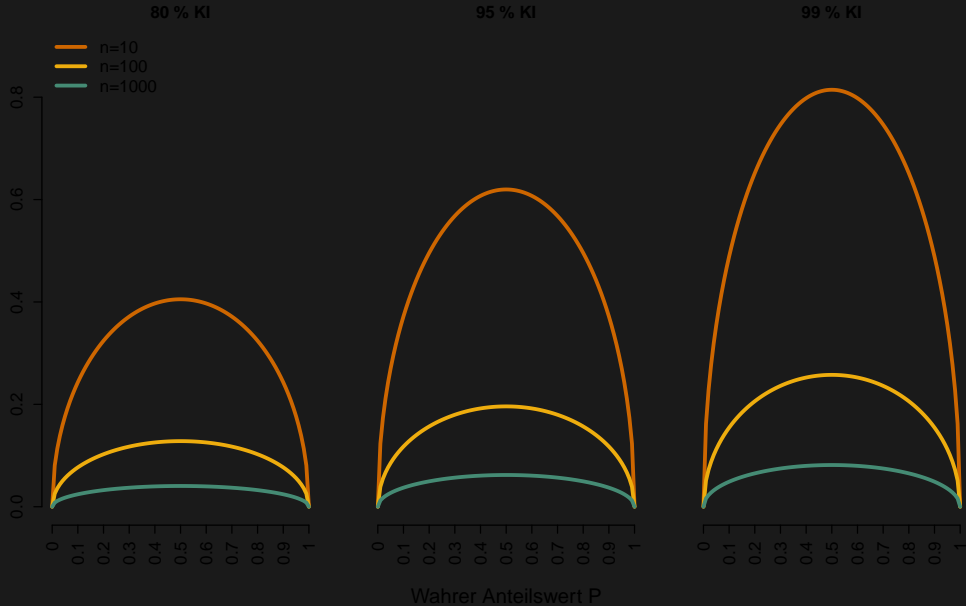
- Praxis | Wahrer Anteil nicht bekannt, daher Schätzung aus Stichprobenanteil p
- Der gemessene Anteil p kann aber eine totale Fehlschätzung sein!
- Die Philosophie bezieht sich auf wiederholte Messungen.
- Entweder liegt der gemessene Wert im Konfidenzintervall, oder ein seltenes Ereignis ist eingetreten.

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

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

- 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.

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

Creative Commons BY-SA-3.0-DE

Dieses Werk ist unter einer Creative Commons Lizenz vom Typ *Namensnennung - Weitergabe unter gleichen Bedingungen 3.0 Deutschland* zugänglich. Um eine Kopie dieser Lizenz einzusehen, konsultieren Sie

<http://creativecommons.org/licenses/by-sa/3.0/de/> oder wenden Sie sich brieflich an Creative Commons, Postfach 1866, Mountain View, California, 94042, USA.