Statistik für Digital Humanities

Empirische Forschung

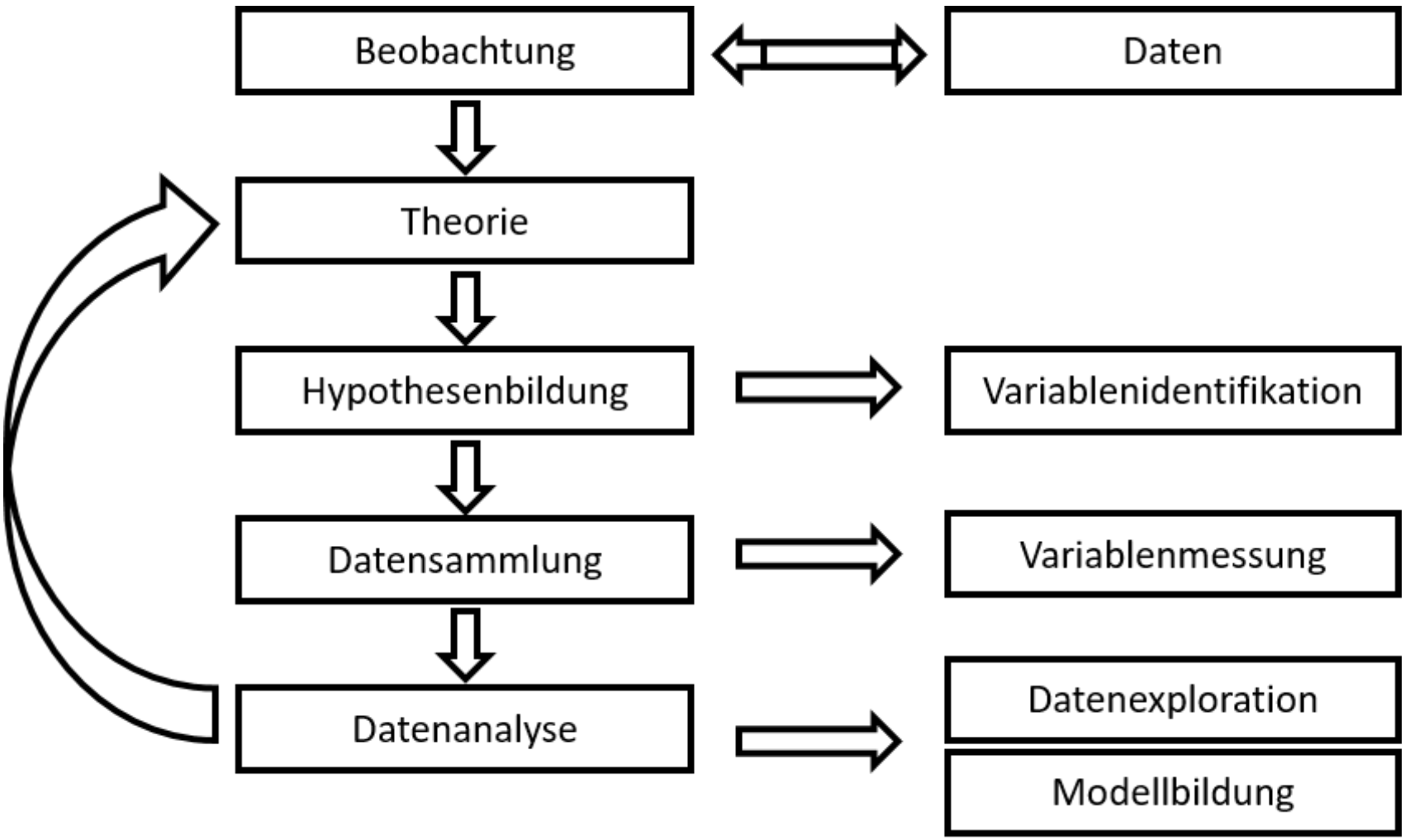
Dr. Jochen Tiepmar

Institut für Informatik
Computational Humanities
Universität Leipzig

19. April 2021

[Letzte Aktualisierung: 18/05/2021, 09:03]

Empirische Forschung



Beobachtung

- Subjektive Wahrnehmung eines Phänomens in der Welt
- Anhand der Daten verifizierbares Statement über *etwas*
  - $\rightarrow$ Big Brother Teilnehmer haben narzisstische Persönlichkeitsstörungen
  - $\rightarrow$ Coca Cola ist ein Spermizit
- Auf Basis von Beobachtungen werden Theorien gebildet

Theorien

- Versuch einer Erklärung einer Beobachtung
- Dieselbe Beobachtung kann mehrere Theorien hervorrufen
  - → Personen mit narzisstischer Persönlichkeitsstörung bewerben sich mit erhöhter Wahrscheinlichkeit bei Big Brother.
  - → Die Produzenten von Big Brother wählen mit erhöhter Wahrscheinlichkeit Personen mit narzisstischer Persönlichkeitsstörung.
- Wissenschaftliche Theorien sollten sich anhand empirisch widerlegbarer Hypothesen bestätigen, falsifizieren oder abweisen lassen.

Hypothesenbildung

- Vorhersage auf Basis einer Theorie
- (Empirisch) widerlegbar
  - → Personen mit narzisstischer Persönlichkeitsstörung bewerben sich mit erhöhter Wahrscheinlichkeit bei Big Brother.
    → Der Anteil derer mit narzisstischer Persönlichkeitsstörung ist höher bei den BewerberInnen bei Big Brother als bei der generellen Öffentlichkeit (1%).
  - → Die Produzenten von Big Brother wählen mit erhöhter Wahrscheinlichkeit Personen mit narzisstischer Persönlichkeitsstörung.
    → Der Anteil derer mit narzisstischer Persönlichkeitsstörung ist bei den FinalistInnen höher als bei den anfänglichen Bewerbungen.
- Hypothesen können **gerichtet** (gut, besser, schlechter, dicker) oder **ungerichtet** (abweichend, anders, verändert) sein.

Beispiel für einen Hypothesentest

	keine NPS	NPS	Summe
Angenommen	3	9	12
Abgewiesen	6805	845	7650
Summe	6808	854	7662

- Personen mit narzisstischer Persönlichkeitsstörung bewerben sich mit erhöhter Wahrscheinlichkeit bei Big Brother.
  $\rightarrow$ Der Anteil der Personen mit narzisstischer Persönlichkeitsstörung ist höher bei den BewerberInnen bei Big Brother als bei der generellen Öffentlichkeit (1%).
- Die Produzenten von Big Brother wählen mit erhöhter Wahrscheinlichkeit Personen mit narzisstischer Persönlichkeitsstörung.
  $\rightarrow$ Der Anteil derer mit narzisstischer Persönlichkeitsstörung ist bei den FinalistInnen höher als bei den anfänglichen Bewerbungen.
- Die Daten bestätigen beide Hypothesen und damit die jeweiligen Theorien.

Daten und Variablen

- Daten können konstant oder variabel (veränderlich) sein
- Statistik arbeitet idR mit Variablen
- Variablen enthalten konkrete Messwerte
- Hypothesen bestehen oft aus 2 Variablen
  - → Ursache → Wirkung
  - → Unabhängige Variable → Abhängige Variable
  - → Coca Cola → Tote Spermien
- Falls Kausalzusammenhang nicht gegeben (und korrekter)
  - → Prädiktor → Ergebnis (outcome)

Daten und Variablen

- Variablen können stetig und diskret sein
  - $\rightarrow$ stetige Variablen können jeden beliebigen Wert im Intervall annehmen
  - $\rightarrow$ diskrete Variablen können nur vorher konkret definierte Werte annehmen
- (Messungen erzeugen eigentlich immer diskrete Variablen)
- Der Unterschied zwischen *Integer* und *Float* ist ähnlich aber *Float* kann ebenfalls diskret sein
- Bsp.: *Alter* ist stetig aber jede Altersangabe ist diskret
- Für uns gilt: stetig wenn es weiter zerlegbar ist als angegeben, nicht ganzzahlig

Feste vs. Zufällige Variablen

Feste Variablen fixed

- Ändern sich nicht
- Beispiel: Geschlecht

Zufällige Variablen

- Ändern sich
- Beispiel: Alter

Feste vs. Zufällige Effekte

Feste Effekte fixed

- Alle Werte einer Variable enthalten
- Beispiel Würfeln: wir zählen alle ungeraden Zahlen 1,3,5

Zufällige Effekte

- Zufälliges Subset aller Werte enthalten
- Beispiel Würfeln: wir zählen alle Zahlen 1,2,3,4,5,6

Datenskalierung

- Kategorische Skalierung
- Numerische Skalierung

Datenskalierung

- Kategorische Skalierung
  - → Binär
  - → Nominal
  - → Ordinal
- Numerische Skalierung

Binär & Nominal

- Eigenschaften wie "krank" – "gesund", "Raucher", "Nichtraucher", Geschlecht, Farben, Berufsgruppe, Tierart, Apfelsorte
- jede Beobachtung einer Merkmalsausprägung wird genau einer bestimmten Klasse (Kategorie) zugeordnet
- Klassen können nicht geordnet sondern nur unterschieden werden
- Klassen auch z.B. durch natürliche Zahlen oder Buchstaben charakterisiert
- Binär: 2 Kategorien (Biologisches Geburtsgeschlecht)

Ordinal

- Bewertung durch Noten 1 - 5, Antwortmuster: "stark ablehnend" - "ablehnend" - "unentschieden" - "zustimmend" - "stark zustimmend"
- Platzierungen, Güteklassen, Ratingskalen
- Präferenzstruktur
- Unterschiede zwischen den Werten bzw. Klassen nicht vergleichbar (keine analysierbaren Abstände)
- sinnvolles Ordnen der Beobachtungen möglich
- Abstände der Werte ohne Informationswert / nicht analysierbar
- wenn Klassen, dann üblicherweise durch natürliche Zahlen repräsentiert

Datenskalierung

- Kategorische Skalierung
- Numerische Skalierung
  - → Intervallskalierung
  - → Ratio / Absolute Skalierung

Intervall

- physikalische Größen wie Temperatur in Grad Celsius
- Abstände zwischen den Werten der Skala besitzen eine Bedeutung; Berechnung von Differenzen sinnvoll
- kein absoluter Nullpunkt, deshalb z.B. Aussage: "20 Grad Celsius sind doppelt so warm wie 10 Grad Celsius" unsinnig

Absolut / Verhältnis

wie Intervallskala, aber mit absolutem Nullpunkt

- Grad Kelvin, Körpergrösse, Häufigkeit eines Wortes

Datensammlung

- Korrelation
  - → Ohne Manipulation der Variablen
  - → Passive Beobachtung
  - → mit Variablenwerten beschriebener Ausschnitt aus Raum und Zeit
- Experiment
  - → Gezielte Manipulation von Prädiktoren
  - → Messung des Einflusses auf Ergebnis
  - → Bspw. Positive Reinforcement oder Bestrafung von Probanden
  - → Systematische Variation: Beabsichtigte Änderungen und daraus folgende Effekte
  - → Unsystematische Variation: Alles andere (Meta-Lerneffekte, Langeweile)
  - → Randomisierung hilft, unbeabsichtigte Effekte zu vermeiden
- Vorsicht vor Drittvariablen
  - → Augenscheinliche Korrelationen können durch dritte Variablen ausgelöst werden
  - → Korrelation zwischen Brustvergößerung und Selbstmord
    → vermutete Drittvariable geringes Selbstwertgefühl (Koot et. al, 2003)

Ausflug Korrelation und Kausalität

Kausalität nach David Hume: (1) Ursache und Wirkung müssen zeitlich nah sein (2) Ursache muss vor der Wirkung passieren (3) Wirkung sollte nicht ohne Ursache auftreten.

- Korrelation
  - → Kausalität kann nicht festgestelllt werden
  - → Phänomene und Effekte korrelieren (koexistieren) nur
  - → Unbekannt, was Prädiktor und Ergebnis ist
- Experiment
  - → Kausalität wird untersucht
  - → Prädiktor wird anhand Hypothese bestimmt und variiert
  - → Änderung des Ergebnis nach Variation des Prädiktors kann Kausalzusammenhang bedeuten
- Kausalität wird durch Datensammlung bestimmt, nicht durch mathematische Testverfahren
  - → Kausalität nicht mathematisch berechenbar

Variablenmessung

- Variablenmessung ist eine handwerkliche Fehlerquelle
- Qualität der Messung folgt aus der Qualität des richtig ausgewählten Messgerätes
- Korrektheit

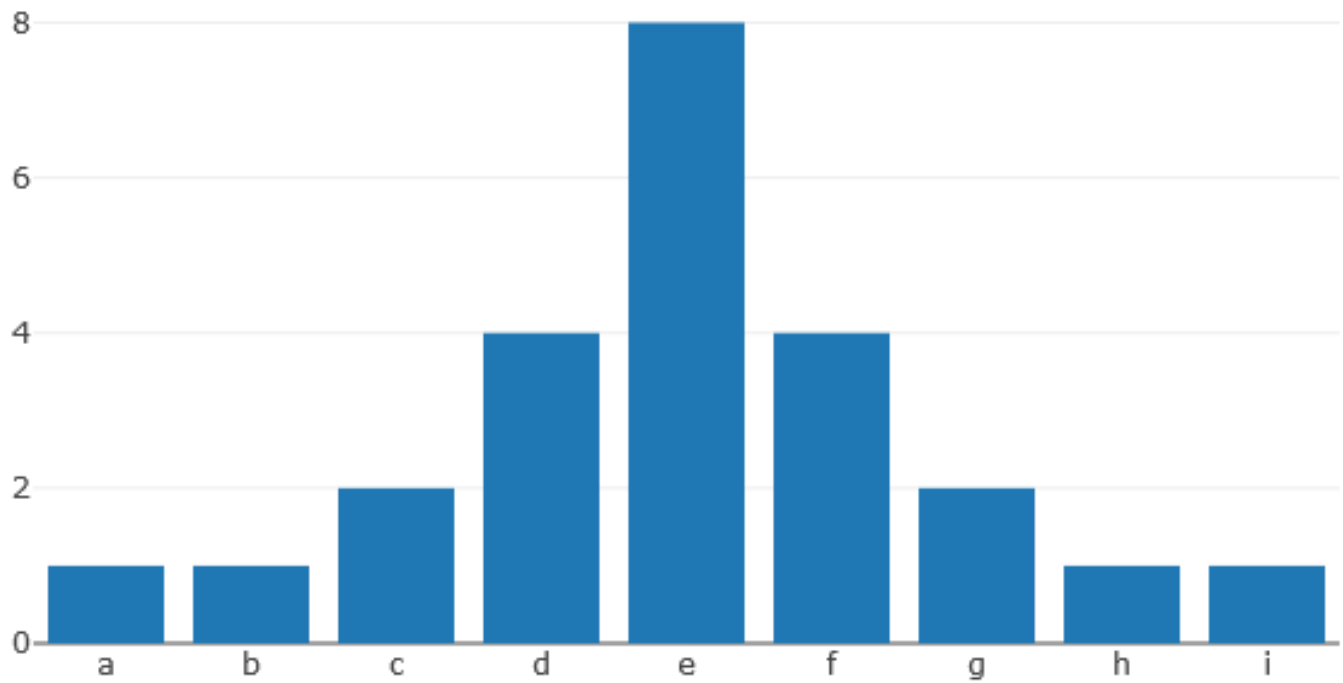
  $\rightarrow$ Wird der Wert korrekt gemessen?
- Verlässlichkeit

  $\rightarrow$ Erzeugt das Messgerät dieselben Messwerte in denselben Situationen?

Häufigkeitsverteilung

Standardfall Normalverteilung

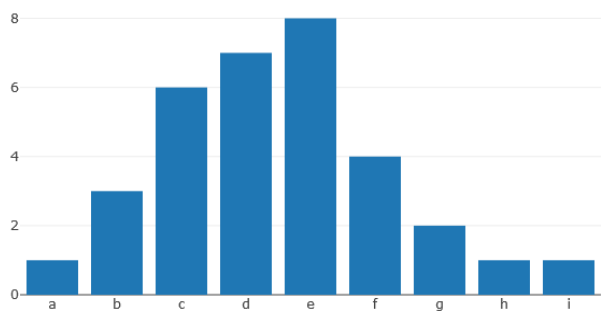
- Masse im Zentrum, viele Einzelwerte im Rand (Head, Tail)
- "Glockenkurve"
- erst sinnvoll ab ordinalskalierten Daten

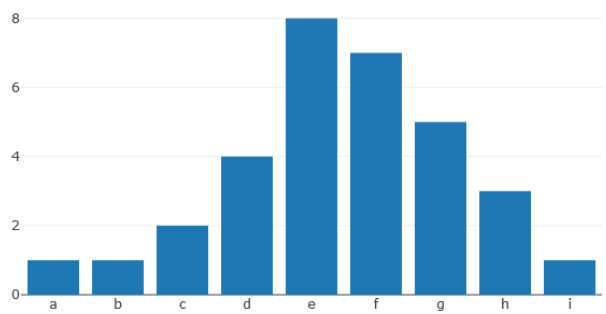


Häufigkeitsverteilung

(Nicht ganz) Standardfall Normalverteilung

- Verschiebung in Richtung links / head (positiv) oder rechts / tail (negativ)

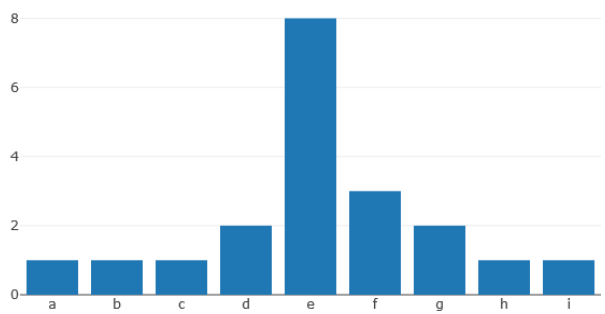


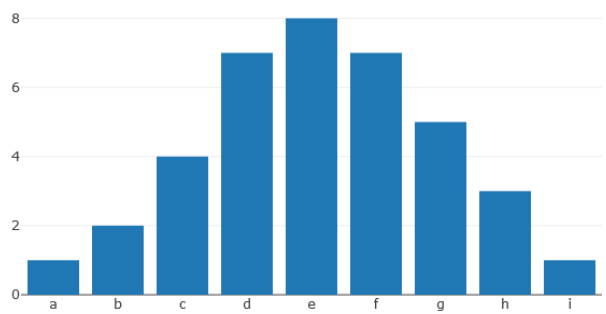


Häufigkeitsverteilung

(Nicht ganz) Standardfall Normalverteilung

- Wölbung (Kurtosis) spitz (positiv, leptokurtisch) oder flach (negativ, platykurtisch)





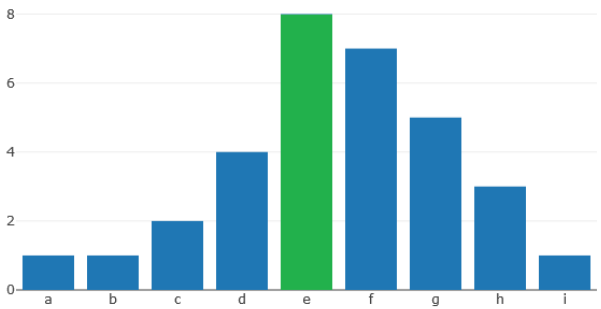
Grundlegende statistische Maße

3 übliche Werte

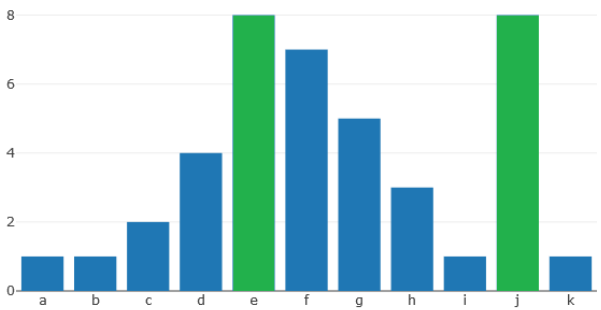
- Modalwert
- Mittelwert
- Median

Modalwert

Modalwert = Häufigste Variablenausprägung / Maximum der Häufigkeitsverteilung



Bimodal / Multimodal: Mehrere Modalwerte möglich



Zentrumsmaße

- Arithmetisches Mittel $\overline{x} = \frac{\sum_{i=1}^{n}(x_0, x_1, \ldots, x_n)}{n}$
- Median $\tilde{x} = \frac{x_{\lceil \frac{n+1}{2} \rceil} + x_{\lfloor \frac{n+1}{2} \rfloor}}{2}$
  - $\rightarrow$ Obermedian $\tilde{x}_o = x_{\lceil \frac{n+1}{2} \rceil}$
  - $\rightarrow$ Untermedian $\tilde{x}_u = x_{\lfloor \frac{n+1}{2} \rfloor}$
- Median robust gegenüber Extremwerten
- Modalwert und Median ignorieren Großteil der in den Daten enthaltenen Information

Median und Mittelwert Beispiele

Anzahl der MySpace Freunde: $X = \{22, 40, 53, 57, 93, 98, 103, 108, 116, 121, 252\}$

- $\overline{x} = \frac{22+40+53+57+93+98+103+108+116+121+252}{11} = \underline{96.64}$
- $\tilde{x} = 22, 40, 53, 57, 93, \overline{98}, 103, 108, 116, 121, 252 = \underline{98}$

$X = \{22, 40, 53, 57, 93, 98, 103, 108, 116, 121, \cancel{252}\}$

- $\overline{x} = \frac{22+40+53+57+93+98+103+108+116+121}{10} = \underline{81.10}$
- $\tilde{x} = 22, 40, 53, 57, \overline{93, 98}, 103, 108, 116, 121 = \underline{95.5}$

Ausflug Mittelwert und Erwartungswert

- Arithmetisches Mittel $\overline{x} = \frac{\sum_{i=1}^{n}(x_0, x_1, \ldots, x_n)}{n}$
- Erwartungswert E = Mittel aller möglichen Ausprägungen

Beispiel: 6 seitiger Würfel 5 mal geworfen:$X = \{1, 2, 3, 5, 6\}$

- $\overline{x} = \frac{1+2+3+5+6}{5} = 3.4$
- $E = \frac{1+2+3+4+5+6}{6} = 3.5$

Streuung

- Range $R = x_{max} - x_{min}$
- Sehr anfällig für Extremwerte, deshalb
- $R_m = m$ Extremwerte an beiden Seiten ignorieren
- Interquartilsabstand / Interquartile Range $R_{25\%} = 25\%$ bei x_{max} und x_{min} ignorieren

  $\rightarrow$ Lower Quartile $Q1$: 25 %
  $\rightarrow$ Median $\tilde{x}$: 50 %
  $\rightarrow$ Upper Quartile $Q3$: 75 %
- Achtung: Informationsverlust

Range Beispiele

Anzahl der StudiVZ Freunde:

$X = \{22, 40, 53_{Q1}, 57, 93, 98_{\tilde{x}}, 103, 108, 116_{Q3}, 121, 252\}$

- $R = x_{max} - x_{min} = 252 - 22 = \underline{230}$
- $R_1 = x_{max_{-1}} - x_{min_{+1}} = 121 - 40 = \underline{81}$
- $R_{25\%} = x_{max_{25\%}} - x_{min_{25\%}} = 116 - 53 = \underline{63}$

Modellbildung

- Experimentelle (alternative) Hypothese H_1 = ursprüngliche Hypothese
  - $\rightarrow$ Der Anteil der Personen mit narzisstischer Persönlichkeitsstörung ist höher bei den BewerberInnen bei Big Brother als bei der generellen Öffentlichkeit (1%).
- Nullhypothese H_0 = Verneinung von H_1
  - $\rightarrow$ Der Anteil der Personen mit narzisstischer Persönlichkeitsstörung bei den BewerberInnen für Big Brother ist kleiner/gleich dem Anteil in der generellen Öffentlichkeit (1%).
- $\rightarrow$ Binärentscheidung möglich zwischen H_0 mit Wahrscheinlichkeit p oder H_1 mit Gegenwahrscheinlichkeit $1 - p$

  Eins der beiden ist in der Regel wahrscheinlicher als das andere
- "Unsere Stichprobe wäre unwahrscheinlich, wenn H_0 wahr wäre, daher ist H_1 wahrscheinlicher."