Statistik für Digital Humanities

Statistische Modelle

Dr. Jochen Tiepmar

Institut für Informatik
Abteilung Computational Humanities
Universität Leipzig

28. Oktober 2019

[Letzte Aktualisierung: 28/10/2019, 13:07]

Statistische Modelle

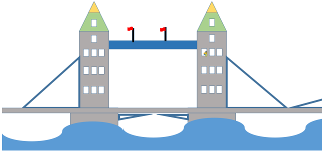
Agenda

Modellbildung

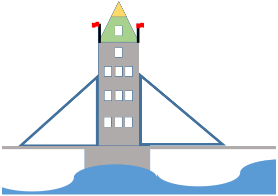
- Phänomene meistens nicht anhand der Realität erforschbar
  - Aufwand
  - Störfaktoren / Variablenisolierung
  - Wiederholbarkeit
- meist Forschung anhand von Auszügen der Realität (Modelle)
- Fitting eines Modells = Übertragbarkeit auf Realität (gut, moderat, schlecht)
- Schlechtes Fitting führt zu fehlerhaften und gefährlichen Schlüssen
- Zu genaues Fitting (Overfitting) führt zu Fehlschlüssen und mangelnder Wiederholbarkeit

Modellfitting

Gut



Moderat



Schlecht

Stichprobenbildung / Sampling

- Population = Alle Betroffenen / Grundgesamtheit
- Stichprobe (Sample) = Ausschnitt aus einer Population
- Stichprobenverteilung (Sampling Distribution) = Verteilung über alle Stichproben (Schätzfunktion auf (unbekannte) Population)
- n = Stichprobengröße, N = Populationsgröße, n_{group} =Gruppengröße
  - **Achtung:** Andy Field verwendet N für Stichprobengröße, aber sonst scheinbar kaum jemand
- Forschungsarbeit anhand Sample, anschließend (meist induktiver) Schluss auf gesamte Gruppe
- Je größer das Sample, desto wahrscheinlicher ist ein guter Fit
  - ($\rightarrow$ Gesetz der großen Zahlen)
- Zusammensetzung des Samples von Experiment abhängig
  - für die Hypothese identifizierte Variablen sollten vorhanden sein
  - weitere Varianten sollten random-ish auftreten

Mittelwert als Modell

Die folgenden Berechnungen können analog für andere Modelle angewendet werden, sind hier aber beispielhaft auf den Mittelwert bezogen

Wiederholung Mittelwert

Artithmetisches Mittel $\overline{x} = \frac{\sum(x_0, x_1, \ldots, x_n)}{n}$

Beispiel Anzahl der Twitter Follower: $X = \{22, 40, 53, 57\}$

- $\overline{x} = \frac{22+40+53+57}{4} = \underline{43}$

Fitness des Mittelwerts

- Abweichung (deviance) $= x_i - \overline{x}$
- Naiv: Abweichungen addieren $= \sum(x_i - \overline{x})$

Fitness des Mittelwerts

- Abweichung (deviance) $= x_i - \overline{x}$
- Naiv: Abweichungen addieren $= \sum(x_i - \overline{x})$
    - $X = \{22, 40, 53, 57\}$
    - $\overline{x} = 43$
    - Totaler Fehler $= -21 + -3 + 10 + 14 = 0$ ☹
- Halbgut: Quadratabweichungen addieren SS $= \sum(x_i - \overline{x})^2$

Fitness des Mittelwerts

- Abweichung (deviance) $= x_i - \overline{x}$
- Naiv: Abweichungen addieren $= \sum(x_i - \overline{x})$
  - $X = \{22, 40, 53, 57\}$
  - $\overline{x} = 43$
  - Totaler Fehler $= -21 + -3 + 10 + 14 = 0$ ☹
- Halbgut: Quadratabweichungen addieren SS $= \sum(x_i - \overline{x})^2$
  - Sum of Squares steigt mit Stichprobengröße☹
- Gut: SS mit Stichprobengröße normalisieren
  Varianz $s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1}$
  Standardabweichung $s = \sqrt{s^2}$
  - $n - 1$ gleicht stichprobenbezogenen statistischen Fehler bei $\overline{x}$ (etwas) aus (Für genauere Informationen Siehe Freiheitsgrade bezogen auf Grundgesamtheit und Stichproben)

Fitness des Mittelwerts

Beispiel Anzahl der Instagram Follower

- $X = \{22, 40, 53, 57\}$
- $\overline{x} = 43$

Varianz $s^2 =$

Fitness des Mittelwerts

Beispiel Anzahl der Instagram Follower

- $X = \{22, 40, 53, 57\}$
- $\overline{x} = 43$

Varianz $s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1} = \frac{(-21)^2+(-3)^2+10^2+14^2}{3} = \frac{746}{3} = 248.67$

Standardabweichung $s =$

Fitness des Mittelwerts

Beispiel Anzahl der Instagram Follower

- $X = \{22, 40, 53, 57\}$
- $\overline{x} = 43$

Varianz $s^2 = \frac{\sum(x_i - \overline{x})^2}{n-1} = \frac{(-21)^2+(-3)^2+10^2+14^2}{3} = \frac{746}{3} = 248.67$
Standardabweichung $s = \sqrt{s^2} = \sqrt{248.67} = 15.77$

Fitness eines Modells

- $Ergebnis = Modell + Fehler$
- $Abweichung = \sum (Beobachtung - Modell)^2$
- s und s^2 beschreiben statistischen Fehler des Modells, also das Ausmaß, in dem beispielsweise das Modell *mean* von den Daten der Stichprobe abweicht.

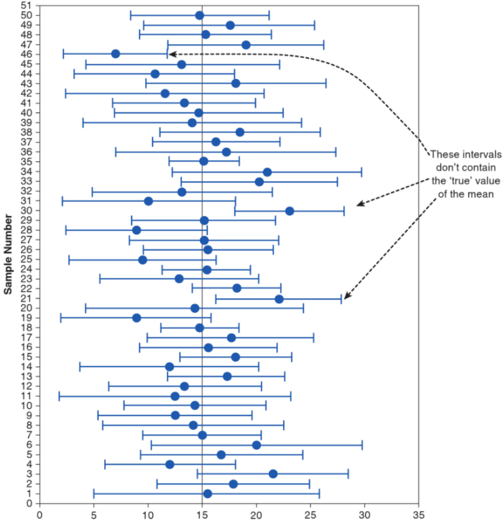
Standardfehler

- Stichprobenvarianz: Gleiche Modelle ergeben verschiedene Ergebnisse bei verschiedenen Stichproben
- Ergebnisse verschiedener Samples unterliegen also einer Häufigkeitsverteilung, auch beim Modell *mean*
- für $n > 30$ folgt *mean* einer Normalverteilung
- **Standardfehler** $\sigma =$ Standardabweichungen aller möglichen Stichproben
  - Praktisch idR nicht berechenbar
- **Central Limit Theorem**
  if $n > 30$: $\sigma \approx \frac{s}{\sqrt{n}}$
- σ beschreibt den statistischen Fehler bezogen auf die Stichprobenverteilung

Ausflug Central Limit Theorem

- Stichprobenverteilung oft nicht vollständig erfassbar
- $\rightarrow$ Abschätzung von Stichprobe auf Stichprobenverteilung
- **Central Limit Theorem**
- $\rightarrow$ Wenn Stichprobe tendenziell normalverteilt dann Stichprobenverteilung ebenfalls
  if $n > 30$:
  - $\sigma \approx \frac{s}{\sqrt{n}}$
  - $\overline{x}_{Stichprobenverteilung} \approx \overline{x}_{population}$
  - Stichprobenverteilung tendenziell normalverteilt

Konfidenzintervall



Ausflug z-Score

- Normalverteilung erlaubt Abschätzen der Wahrscheinlichkeit des Auftretens von Werten
  - $\overline{x} = 0, s = 1$
- z-Score "transformiert" Werte zu entsprechender Normalverteilung
- $z = \frac{x - \overline{x}}{s}$
- Wahrscheinlichkeiten für Auftreten von x aus z-Score Tabelle ablesbar
- $z = 1.96$ entspricht 2.5% der höchsten Werte, $z = -1.96$ 2.5% der niedrigsten Werte
- 95% der Werte haben z-score zwischen -1.96 und 1.96

Ausflug z-Score

Beispiel Anzahl der StudiVZ Freunde

- $X = \{22, 40, 53, 57\}$
  $\overline{x} = 43$
  $s = \sqrt{s^2} = \sqrt{248.67} = 15.77$
- Wie wahrscheinlich ist es, dass der nächste Wert mindestens 30 ist?
- $z = \frac{x-\overline{x}}{s} = \frac{30-43}{15.77} = -0.82$
  z-Tabelle sagt .79389
  $P(x >= 30) = 79,38\%$

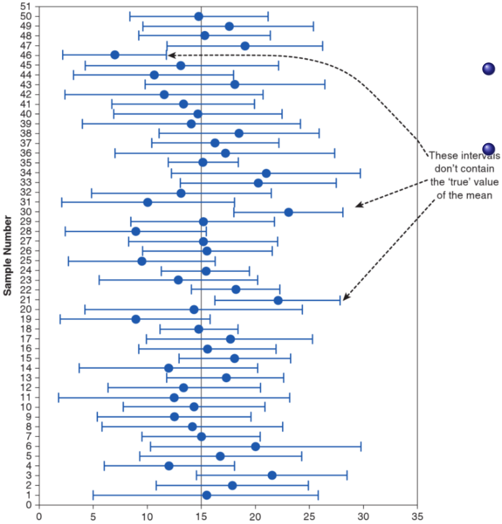
Konfidenzintervall

- Jedes Sample hat ein Konfidenzintervall bezogen auf ein Modell
- Höhe vorher festgelegt
- Meist 95% (manchmal 99%)
- Konfidenzintervalle von 95% der Samples enthalten den wahren Wert der Population
- Berechnung für 95%:
  Untergrenze $= \overline{x} - (1.96 * \sigma)$
  Obergrenze $= \overline{x} + (1.96 * \sigma)$

Konfidenzintervall

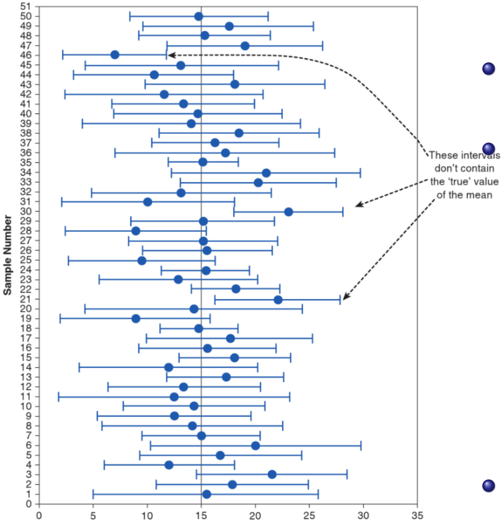
- Jedes Sample hat ein Konfidenzintervall bezogen auf ein Modell
- Höhe vorher festgelegt
- Meist 95% (manchmal 99%)
- Konfidenzintervalle von 95% der Samples enthalten den wahren Wert der Population
- Berechnung für 95%:
  Untergrenze $= \overline{x} - (1.96 * \sigma)$
  Obergrenze $= \overline{x} + (1.96 * \sigma)$
  $\sigma =$ Standardfehler
- z-Score für 99%: ± 2.58
- Bei kleinen Stichproben (< 30) t-Score (two-tailed) statt z-Score verwenden mit df = n-1

Konfidenzintervall



- Visualisierung mittels Fehlerbalken (Error Bar)
- Überschneidungsfreiheit zweier Samples bedeutet:

Konfidenzintervall



- Visualisierung mittels Fehlerbalken (Error Bar)
- Überschneidungsfreiheit zweier Samples bedeutet:
  - Ein Sample enthält nicht den "wahren" Populationswert (5% wahrscheinlich)
  - Samples stammen aus verschiedenen Populationen bspw. vor und nach experimenteller Manipulation (95% wahrscheinlich)
- Ein präziseres Modell hat ein kleineres Konfidenzintervall

Modellbildung

- Experimentelle (alternative) Hypothese H_1 = ursprüngliche Hypothese
- Nullhypothese H_0 = Verneinung von H_1
- Beweis von H_1 mit Statistik nicht möglich, aber Ablehnung von H_0 für unser Experiment
- "Unsere Stichprobe wäre unwahrscheinlich, wenn H_0 wahr wäre, daher ist H_1 wahrscheinlicher."
- "Unsere Stichprobe wäre 5% wahrscheinlich, wenn H_0 wahr wäre, daher ist H_1 wahrscheinlicher."
- $\rightarrow$ Jedes 20. mal liegt man damit daneben, da die Zahlen zufällig auftraten

Teststatistik

Teststatistik (grob) $= \frac{Varianzerklartdurch Modell}{Varianznichterklartdurch Modell}$

- auch Prüfgröße, Testgröße oder Prüffunktion
- Teststatistiken messen, wie gut das Modell zu den Daten passt
- Verschiedene Teststatistiken existieren (t, F, X^2)
- Gegeben ein zur Hypothese passendes Modell, sagt eine signifikante Teststatistik dass es unwahrscheinlich wäre, dass das Modell so gut zu den Daten passen würde, wenn die Nullhypothese wahr wäre.
  - Man testet also eigentlich die statistische Wahrscheinlichkeit von H_0
- **One Tailed Tests**: gerichtete Hypothesen (5% Wahrscheinlichkeit)
- **Two Tailed Tests**: ungerichtete Hypothesen (je 2.5% Wahrscheinlichkeit)

Typ 1 und Typ 2 Fehler

- Fehler Erster Art:
  - Effekt fälschlicherweise bestätigt
  - α-level
  - bspw. 5% akzeptabel
- Fehler Zweiter Art:
  - Effekt fälschlicherweise übersehen
  - β-level
  - bis 20% akzeptabel (Cohen 1992)
- indirekt proportionaler Zusammenhang vorhanden aber nicht genau bestimmbar

Typ 1 und Typ 2 Fehler

		Wirklichkeit	
		H_0 ist wahr	H_1 ist wahr
Entscheidung des Tests	für H_0	Spezifität True Positive Wahrscheinlichkeit: $1 - \alpha$	**Fehler 2. Art** False Negative Wahrscheinlichkeit: β
	für H_1	**Fehler 1. Art** False Positive Wahrscheinlichkeit: α	Sensitivität, Trennschärfe True Negative Wahrscheinlichkeit: $1\text{-}\beta$

Effektstärke (Effect Size)

- Standardisierte Maße für Einfluss einzelner Variablen auf andere
- Pearson's r Korrelationskoeffizient
  - 0.1 : Schwach (1% der Variation)
  - 0.3 : Mittel (9% der Variation)
  - 0.5 : Stark (25% der Variation)
- Cohen's d
- Quotenverhältnis (Odds Ratio)
- ...dazu später mehr

Statistische Modelle

Zusammenfassung

- Hypothese
- Stichprobe
- Passendes (fitting) Modell finden, welches Effekt der Hypothese beschreibt
- Mit Konfidenzintervall Vorhersagepräzision des Modells berechnen
- Teststatistik/Prüfzahl des Modells berechnen
- Fehler erster und zweiter Art der Teststatistik untersuchen
- Teststatistik signifikant (Effekt mathematisch unwahrscheinlich)
  → Effekt/Zusammenhang trat wahrscheinlich auf
- Teststatistik nicht signifikant (Effekt mathematisch wahrscheinlich)
  → Effekt/Zusammenhang zu klein um gemessen zu werden
- Effektstärke berechnen