

Statistik für Digital Humanities

Statistische Modelle

Dr. Jochen Tiepmar

Institut für Informatik
Computational Humanities
Universität Leipzig

20. April 2020

[Letzte Aktualisierung: 19/04/2020, 22:43]

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

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
 - (→ 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

Arithmetisches Mittel $\bar{x} = \frac{\sum(x_0, x_1, \dots, x_n)}{n}$

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

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

Fitness des Mittelwerts

- Abweichung (deviance) = $x_i - \bar{x}$
- Naiv: Abweichungen addieren = $\sum(x_i - \bar{x})$
 - $X = \{22, 40, 53, 57\}$
 - $\bar{x} = 43$
 - Totaler Fehler = $-21 + -3 + 10 + 14 = 0$
- Halbgut: Quadratabweichungen addieren $SS = \sum(x_i - \bar{x})^2$
 - Sum of Squares steigt mit Stichprobengröße
- Gut: SS mit Stichprobengröße normalisieren

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

Standardabweichung $s = \sqrt{s^2}$

- $n - 1$ gleicht stichprobenbezogenen statistischen Fehler bei \bar{x} (etwas) aus
(Für genauere Informationen Siehe Freiheitsgrade bezogen auf Grundgesamtheit und Stichproben)

Fitness des Mittelwerts

Beispiel Anzahl der Instagram Follower

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

$$- \bar{x} = 43$$

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

$$\text{Standardabweichung } s = \sqrt{s^2} = \sqrt{248.67} = 15.77$$

Fitness eines Modells

- *Ergebnis* = *Modell* + *Fehler*
- *Abweichung* = $\sum (\text{Beobachtung} - \text{Modell})^2$
- 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** σ = 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
- → Abschätzung von Stichprobe auf Stichprobenverteilung
- **Central Limit Theorem**
- → Wenn Stichprobe tendenziell normalverteilt dann Stichprobenverteilung ebenfalls
if $n > 30$:
 - $\sigma \approx \frac{s}{\sqrt{n}}$
 - $\bar{X}_{\text{Stichprobenverteilung}} \approx \bar{X}_{\text{population}}$
 - Stichprobenverteilung tendenziell normalverteilt

Ausflug z-Score

- Normalverteilung erlaubt Abschätzen der Wahrscheinlichkeit des Auftretens von Werten
 - $\bar{x} = 0, s = 1$
- z-Score "transformiert" Werte zu entsprechender Normalverteilung
- $z = \frac{x - \bar{x}}{s}$
- 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

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

$$\bar{x} = 43$$

$$s = \sqrt{s^2} = \sqrt{248.67} = 15.77$$

- Wie wahrscheinlich ist es, dass der nächste Wert mindestens 30 ist?

$$- z = \frac{x - \bar{x}}{s} = \frac{30 - 43}{15.77} = -0.82$$

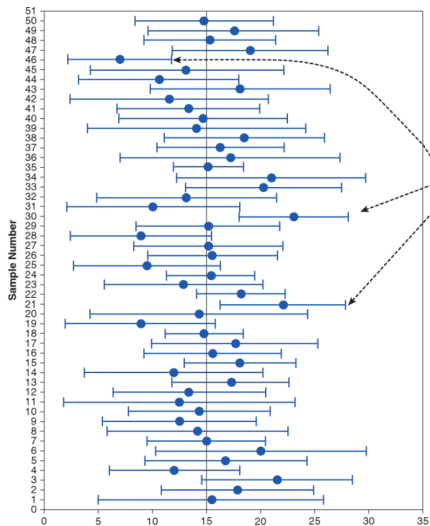
z-Tabelle sagt .79389

$$P(x \geq 30) = 79,38\%$$

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 = $\bar{x} - (1.96 * \sigma)$
Obergrenze = $\bar{x} + (1.96 * \sigma)$
 σ = 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:

→ 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 kleinere Konfidenzintervalle

Modellbildung

- Experimentelle (alternative) Hypothese H_1 = ursprüngliche Hypothese
- Nullhypothese H_0 = Verneinung von H_1
- → 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."
- "Unsere Stichprobe wäre 5% wahrscheinlich, wenn H_0 wahr wäre, daher ist H_1 wahrscheinlicher."
- → Jedes 20. mal liegt man damit daneben, da die Zahlen zufällig auftraten

Teststatistik

$$\text{Teststatistik (grob)} = \frac{\text{Varianz erklart durch Modell}}{\text{Varianz nicht erklart durch Modell}}$$

- auch Prüfgröße, Testgröße oder Prüffunktion
- Teststatistiken messen, wie gut das Modell zu den Daten passt
- Verschiedene Teststatistiken existieren (t , F , χ^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 mathematische 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, J. (1992). A power primer. *Psychological Bulletin*.)

- 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: $1 - \beta$

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

Zusammenfassung

- Hypothese
- Stichprobe
- Passendes (fitting) Modell finden, welches Zusammenhang 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