Statistik
01. Inferenz

Roland Schäfer

Institut für Germanistische Sprachwissenschaft
Friedrich-Schiller-Universität Jena

stets aktuelle Fassungen: https://github.com/rsling/VL-Statistik

Inhalt

Probative Wissenschaft

Empirisch, objektiv, realistisch, exakt(?)

- Beobachtbare Phänomene
- Beobachtungen reproduzierbar
- Messbar = beobachtbar (Sinneswahrnehmung an sich irrelevant)

- Realismus | wirkliche Phänomene und ihre Mechanismen
- Keine postmoderne Realitäts- und Objektivitätsverweigerung

- Kontrolliertes Experiment

Empirie | Gründe für Reproduzierbarkeitsbedingung

- Intrinsische Ungenauigkeiten der Messung (Wirkung plus Störeinflüsse)
- Potentiell inadäquate Messung des theoretischen Konstrukts

→ Vermeidung von Fehlschluss auf unechte Ursachen

→ Relevante Ursachen

→ Insgesamt Stärkung der Validität

Beispiel | Fehlgeleitete Generative Grammatiker

- Gegenstand: interne (mentale) Grammatik (I-Grammatik)
  universeller und individueller Teil
- I-Grammatik bei jedem Sprecher (leicht) verschieden
- I-Grammatik erlaubt immer binäre Grammatikalitätsentscheidung

→ Linguisten können eigene I-Grammatik untersuchen (Introspektion)!

Das Ergebnis ist die aktuelle Krise der Linguistik.

(Logischer) Positivismus

Formale Ableitung von Wissen (= Theorien) aus Beobachtbarem und irgendeiner Logik. Induktion. Keine Metaphysik. Keine Kreativität erwünscht. (Carnap 1928, ...)

Aber suchen wir wirklich nur nach Mustern, z. B. in Korpusdaten?

- Was ist der zugrundeliegende Mechanismus?
- Wie kommen wir zu erklärenden Theorien von Mustern in Daten?
- Datenaufbereitung (z. B. im Korpus) kann dann nicht theoriegeleitet sein.
- Die ART folgt auch nicht einfach so aus Daten!

Inferenz II | Rationalistischer Probativismus (Falsifikationismus)

Rationalistischer Probativismus

Theorien werden aufgestellt von Menschen, die die Welt beobachten. Theorien werden getestet an Daten, aber nicht logisch aus Daten abgeleitet. Wissenschaft lernt aus Fehlern. (Popper 1962, Mayo 1996, ...)

Unter dieser Philosophie werden plötzlich Dinge wichtig ...

- Ist eine Stichprobe repräsentativ für das, was man zeigen will?
- Welche Methode der statistischen Analyse wird verwendet?
- Für eine Korpusstudie muss die Datenaufbereitung damit theoriegeleitet sein!
- Liefert die Studie *a serious Argument from Error*?

  There is evidence an error is absent to the extent that a procedure with a very high capability of signalling the error, if and only if it is present, nevertheless detects no error. (Mayo 2018: 16)

Hypothesen

Die konkrete Hypothesen, die in einem Experiment getestet werden,
sind nie die Primärhypothesen der Theorie.

- Abgeleitete Partikularhypothesen über konkrete Erwartungen im Experiment
- Einfluss zahlreicher Auxiliarhypothesen, z. B. über Messprozeduren
  Duhem (1914), Quine (1951), Laudan (1990)

- „Interessante“ Hypothesen
  - Formulierung relevanter Kausationsbedingung (wenn, dann)
  - Universelle Gültigkeit | ein Sprecher vs. alle Sprecher
  - Also z. B. uninteressant | *Welchen Kasus nimmt wegen?*

Hypothesenprüfung | Probativismus (Falsifikationismus)

Kann die Hypothese weiter angenommen werden,
oder liefert das Experiment starke Evidenz gegen sie?

- Probleme bei Prüfung
  - Falsch abgeleitete Partikularhypothese
  - Falsche Sekundärhypothesen
  - Störeinflüsse, intrinsische Messungenauigkeit
  - Mangelhafte Operationalisierung
  - Zu wenige Daten (oder zu viele Daten?)

Elemente der Empirie

Grundgesamtheit

- Von Interesse | allgemeine Gesetzmäßigkeiten
- Also Untersuchungsgegenstand: alle x (Sprecher, Sätze, ...)
- Untersuchbar | kleine Menge von x

Grundgesamtheit | alle x

Datengenerierender Prozess (DGP) | Prozess, der alle x hervorbringt

Stichprobe | eine kleine Menge x, aus der auf Grundgesamtheit bzw. DGP geschlossen werden soll

Stichprobe

Uniform zufällige Stichprobe

Jedes Element der Grundgesamtheit hat die gleiche Chance beim Ziehen.

Stratifizierte Stichprobe

Die Stichprobe ist so zusammengesetzt, dass wichtige Eigenschaften proportional repräsentiert sind.

- Problem bei Letzterem: haufenweise Auxiliarhypothesen

Operationalisierung und Auxiliarhypothesen

- Operationalisierung | präzise Formulierung der Messmethode für ein theoretisches Konstrukt

- Bsp. Konstrukt „Satzlänge“: Wortanzahl? Phonemanzahl? Phrasenanzahl?
- Bsp. Konstrukt „Satztopik“: Oha!?! (Cook & Bildhauer 2013)

- Alle genannten Beispiele abhängig von Auxiliarhypothesen bzw. anderen theoretischen Konstrukten (Wort, Phonem, Phrase, …)

Variablen I

- Uninteressanter Typ Fragestellung | „Wieviel Prozent X haben Eigenschaft A?“
- Fehlen jeglicher Aussagen über kausale Zusammenhänge
- Bsp. | Wie oft wird *wegen* mit Dat bzw. Gen verwendet?

- Besser | „Wie bedingt Eigenschaft B die Wahrscheinlichkeit von A bei X?“
- Bsp. | Per Hypothese nehmen denominale Präpositionen eher den Gen als den Dat.

Konzeptionell:

	denominale P	andere P
Dat	x_1	x_2
Gen	x_3	x_4

Variablen II

Operationalisierte und gemessene Eigenschaften sind Variablen.

- Im Experiment:
  - Kontrolliere für Theorie irrelevante Variablen (Störvariablen) bzw. verlass dich auf deren Zufallsverteilung (Fisher, s. u.).
  - Variiere „Ursachen-Variablen“ (unabhängige Variablen).
  - Beobachte „Wirkung-Variablen“ (abhängige Variablen).

Experiment und Quasi-Experiment

- Problem in Astronomie, Korpuslinguistik usw. | keine Experimente möglich
- Unabhängige Variablen nicht variierbar
- Daten liegen bereits vor bzw. fallen vom Himmel
- Auswahl von Datensätzen, so dass von den unabhängigen Variablen die zur Theorieprüfung nötigen Permutationen im Datensatz vorkommen
- Dabei Zusatzproblem bei Korpuslinguistik: Korpus meist nicht das eigene, wenig Informationen über mögliche Verzerrungen

- Was ist die Grundgesamtheit bzw. der DGP?

Validität

Statistische Validität

Gefahren für statistische Schlussverfahren

- Falsches Testverfahren für die gegebene Situation
- Mathematische Vorbedingungen für das Testverfahren nicht
- Zu viele Partikulartests einer übergeordneten Hypothese aus denselben Daten
- Zu kleine Stichprobe
- Zu große Stichprobe
- Zu große Variation in der Grundgesamtheit

Interne Validität

- Irrtum beim Herstellen des Kausalzusammenhangs
- Fiktives Bsp.:
  - Korpora | DWDS-Kernkorpus enthält Texte 1900–2000, DECOW12 Texte nach 2000
  - Hypothese | Im DECOW12 kommt öfter das Pronomen *son* vor als im DWDS Kernkorpus, weil es erst nach 2000 zum eigenständigen Pronomen wurde.
  - Die Hypothese wird bestätigt anhand von Stichproben aus den beiden Korpora.
  - Die wirkliche Ursache sind aber Registerunterschiede.

Validität des Konstrukts

- Korrektheit des theoretischen Konstrukts

- Eigentlich aus der Psychologie
- Aber riesiges Problem in der Linguistik

- Echtes Bsp.
  - Beobachtung | Das Deutsche bewahrt genus-typische Pluralflexion am Substantiv.
  - Konstrukt | Nominalklammer/Klammerprinzip (NP-Kongruenzklammer Art – Subst) (Ronneberger-Sibold 2010)
  - Hypothese (post-hoc zur Beobachtung) | Flexionserhalt stärkt Klammerprinzip
  - Das Konstrukt ist hochgradig beliebig und unterdefiniert, damit nicht testbar.
  - Abhilfe: nur Konstrukte/Hypothesen, die starke Vorhersagen generieren

Externe Validität

- Generalisierbarkeit der Ergebnisse (über Raum, Zeit usw.)
- Problem | zu große Homogenität der Stichprobe
  (was für statistische Validität wiederum gut ist)
- Bezug auf Korpora:
  - Zu spezifische Stratifikation (DeReKo)
  - Verzerrte Stichprobe (Webkorpora)

Ronald A. Fisher, Wahrscheinlichkeit, Ereignisraum, Teetassen

Ronald A. Fisher (1890–1962)

- Statistik als Teil der rationalen wissenschaftlichen Argumentation, der Interpretation von Experimenten
- Möglichst kein Mathematik-Jargon, eher intuitiv zugängliche mathematische Konzepte
- Eingeschränkte statistische Inferenz als theoriegeleitete Dateninterpretation
- Kontrolle aller unabhängigen Variablen
- Alle anderen (Stör-)Variablen konzeptuell zufallsgebunden

The Tea-Tasting Lady

Muriel Bristow behauptet, sie könne am Geschmack einer Tasse Tee erkennen, ob die Milch oder der Tee zuerst eingeschenkt wurde. Fisher führt ein Experiment durch (acht Tassen, vier mit dem Tee zuerst) und fragt, wie wir entscheiden können, ob das Ergebnis davon zeugt, dass sie diese Fähigkeit wirklich hat.

- Liegt das Ergebnis deutlich über dem per Zufall erwartbaren Niveau?
- Idee vor Fisher | alle Störvariablen kontrollieren und gleich machen, dann ist induktive Inferenz möglich
- Fisher | Das ist prinzipiell unmöglich, umständlich, teuer und unnötig!
- Wenn alle irrelevanten Störvariablen zufallsverteilt sind, dann gilt:
  - Variiere die relevante unabhängige Variable.
  - Vergleiche das Ergebnis mit zufällig erwartbaren Ergebnissen.

Wahrscheinlichkeit

Bayesische Wahrscheinlichkeit (*inverse probability)*

- Für wie wahrscheinlich hält Individuum I das Ereignis E?
- Subjektiv, berücksichtigt vorherige Überzeugung
- Aktualisierung von Überzeugungen
- Basiert auf Bayes Rule (Tomas Bayes 1763)
- Ereignisraum (s. u.) irrelevant!

Frequentistische Wahrscheinlichkeit

- Wie viele mögliche Ereignisse e_i aus E treten ein?
- Zu jedem Experiment gehört ein Ereignisraum (s. u.)!
- Daher objektiv, unabhängig von Überzeugungen
- Wenn ein Ereignis e_i eingetreten ist, wird seine Wahrscheinlichkeit uninteressant.
- Geeignet für rationalistisch-probativistische Wissenschaftsphilosophie

Frequentismus | Ereignisraum (sample space)

Für ein Experiment gilt:

- Wir beobachten *n* Messungen (Stichprobe),
  jede Messung wird aus einer definierten Menge von möglichen Messungen.
  - Bsp. | 10 Mal einen Würfel werfen {1, 2, 3, 4, 5, 6}.
  - Bsp. | Je 10 Akzeptabilitätsurteile unter 2 Bedingungen von 100 Probanden {Ja, Nein}.
- Wir bekommen ein konkretes Ergebnis.
  - Bsp. | 8 von 10 Würfen mehr als drei Augen.
  - Bsp. | „Mehr“ Ja-Antworten unter Bedingung A (schon deutlich komplexeres Design).
- Wir müssen berücksichtigen, wie viele Ergebnisse (und welche) es insgesamt hätte geben können, um zu bewerten, wie unwahrscheinlich das Ergebnis war.
- Ereignisraum (sample space) | Menge der möglichen Ausgänge des Experiments

Warum „war“?

Wir müssen berücksichtigen, wie viele Ergebnisse (und welche) es insgesamt hätte geben können, um zu bewerten, wie unwahrscheinlich das Ergebnis war.

- Jedes eingetretene Ereignis hat die Wahrscheinlichkeit 1.
  - Die Wahrscheinlichkeit, dass Helmut Kohl 1998 abgewählt wurde, ist 1.
  - Die Wahrscheinlichkeit, dass wir 8 Würfe mit mehr als drei Augen hatten, ist 1.
- Nach dem Experiment | P(konkreter Ausgang des Experiments wurde erzielt) = 1
- Vor dem Experiment | P(konkreter Ausgang des Experiments wird erzielt werden) < 1

Die übelsten Fehler in der Bewertung statistischer Ergebnisse rühren daher, dass Menschen diese Sachverhalte vergessen.

Zurück zur Tea-Tasting Lady

Design des Experiments | Muriel Bristow probiert acht Tassen,
(vier mit Milch zuerst, vier mit Tee zuerst) und wählt die vier mit Tee zuerst aus.

- Mit wie vielen richtigen Treffern wären Sie zufrieden?
- Es muss die frequentistische Wahrscheinlichkeit errechnet werden,
  eine, zwei, drei oder vier Tassen auch per Zufall richtig zu raten.
- Dann können wir beurteilen, ob das Ergebnis
  deutlich über dem erwartbaren Niveau liegt.

Ausgang 1 | Alle vier Tassen korrekt

Allgemein:

$$P(\text{konkreter Ausgang}) = \frac{\text{Anzahl richtiger Zuweisungen}}{\text{Anzahl aller potentiellen Zuweisungen}} \tag{1}$$

Für diesen Ausgang:

$$P(\text{vier Tassen korrekt}) = \frac{1}{\text{Anzahl aller potentiellen Zuweisungen}} \tag{2}$$

Wie viele Möglichkeiten gibt es?

Wir wählen vier *Tee zuerst*-Tassen (TZ) aus acht Tassen aus:

- erste TZ-Tasse: eine von 8 (bleiben 7)
- zweite TZ-Tasse: eine von 7 (bleiben 6)
- dritte TZ-Tasse: eine von 6 (bleiben 5)
- vierte TZ-Tasse: eine von 5 (bleiben 4)

→ STOPP | alle anderen 4 Tassen automatisch MZ

Also naiv gedacht | $8 \cdot 7 \cdot 6 \cdot 5 = 1680$

Die Reihenfolge

1680 ist zu hoch, denn je nachdem, welche Tasse aus den verbleibenden wir wählen, ergeben sich andere Permutationen (Reihenfolgen) desselben Ergebnisses.

- Bsp. | Auswahl von Tasse 7, 3, 6, 1 identisch zu 3, 1, 6, 7 usw.
- Es gibt von jeder möglichen Auswahl gleich viele Permutationen.
- Und zwar die Anzahl der Möglichkeiten, vier Tassen zu ordnen: $4 \cdot 3 \cdot 2 \cdot 1$

$$\text{Anzahl aller potentiellen Zuweisungen} = \frac{8 \cdot 7 \cdot 6 \cdot 5}{4 \cdot 3 \cdot 2 \cdot 1} = \frac{1680}{24} = 70 \tag{3}$$

Wenn sie also genau richtig liegt ...

Wahrscheinlichkeit, per Zufall genau richtig zu liegen:

$$P(\text{vier Tassen korrekt}) = \frac{1}{70} = 0.014 \tag{4}$$

Interpretieren Sie das Ergebnis.

Für die Generalisierung der Berechnung

- Eigentlich haben wir es mit Binomialkoeffizienten zu tun.
- „Lotto-Kombinationen“ | k aus n
  ohne Zurücklegen und ohne Beachtung der Reihenfolge

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{5}$$

Ausgang 2 | Drei Tassen korrekt

Berechnung mit dem Binomialkoeffizienten

- Die drei richtigen aus vier TZ| $\binom{4}{3}$
- Die eine falsche aus vier TZ | $\binom{4}{1}$

$$P(\textit{drei richtig per Zufall}) = \frac{\binom{4}{3} \cdot \binom{4}{1}}{70} = \frac{16}{70} = 0.229 \tag{6}$$

Interpretieren Sie das Ergebnis.

Darstellung als Kreuztabelle

Ausgang 1

		Realität	
		Tee zuerst	**Milch zuerst**
Lady	**Tee zuerst**	4	0
	Milch zuerst	0	4

Ausgang 2

		Realität	
		Tee zuerst	**Milch zuerst**
Lady	**Tee zuerst**	3	1
	Milch zuerst	1	3

Stichprobengröße und Effektstärke

- Unbefriedigendes Ergebnis bei 3 von 4 richtigen Tassen
- Sehr kleine Stichprobe | nur perfektes Ergebnis zufriedenstellend
- Effektstärke | Vielleicht kann MB ca. 75 % aller Tassen richtig erkennen.

Bei größerer Stichprobe | Was ist mit 30 von 40 richtigen Tassen, also insgesamt 80 Tassen?

Das wäre die gleiche Effektstärke, aber eine größere Stichprobe.

Letzte Warnung

Was zeigt man mit so einem Experiment? Und was nicht?

- Der Ausgang war ziemlich unwahrscheinlich,
  bevor das Experiment durchgeführt wurde.
- Daher gehen wir bis auf Weiteres davon aus, dass ein Effekt vorliegt ...
- ... oder zufällig ein seltenes Ereignis eingetreten ist!!!

- Wenn Sie mit den Geburtsdaten Ihrer Familie im Lotto gewinnen,
  ist ein seltenes Ereignis eingetreten, Sie haben aber nicht gezeigt,
  dass Ihre Geburtsdaten die Lottokugeln beeinflussen!

- Ein solches Ergebnis beweist also nichts!
- Die Logik basiert auf der Annahme einer wiederholten Testung.

→ Wenn wir das Experiment sehr oft machen, und es gibt keinen Effekt,
dann nähert sich die Verteilung der Ergebnisse der Zufallsverteilung an.

Nächste Woche | Überblick

Einzelthemen

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

Literatur I

Carnap, Rudolf. 1928. *Der logische Aufbau der Welt*. Berlin: Weltkreis Verlag.

Cook, Philippa & Felix Bildhauer. 2013. Identifying “aboutness topics”: two annotation experiments. *Dialogue and Discourse* 4(2), 118–141.

Duhem, Pierre. 1914. *La Théorie Physique: Son Objet et sa Structure*. Marcel Riviera & Cie.

Laudan, Larry. 1990. Demystifying Underdetermination. In C. Wade Savage (Hrsg.), *Scientific Theories*, 267–297. Minneapolis: University of Minnesota Press.

Mayo, Deborah G. 1996. *Error and the growth of experimental knowledge*. Chicago: University of Chicago Press.

Mayo, Deborah G. 2018. *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge: Cambridge University Press.

Popper, Karl Raimund. 1962. *Conjections and Refutations: The Growth of Scientific Knowledge*. New York: Basic Books.

Quine, Willard Van Orman. 1951. From a Logical Point of View. In 2. Aufl. Cambridge: Harvard University Press. Kap. Two Dogmas of Empiricism, 20–46.

Ronneberger-Sibold, Elke. 2010. Der Numerus – das Genus – die Klammer : die Entstehung der deutschen Nominalklammer im innergermanischen Vergleich. In Antje Dammel, Sebastian Kürschner & Damaris Nübling (Hrsg.), *Kontrastive Germanistische Linguistik. Teilband 2*, Bd. 206/209 (Germanistische Linguistik), 719–748. Hildesheim: Olms.

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