Willkommen zur Vorlesung Statistik (Master)

Thema dieser Vorlesung:
Letzte Worte zur Inferenzstatistik, v. a. zu Signifikanztests

Prof. Dr. Wolfgang Ludwig-Mayerhofer

Universität Siegen – Philosophische Fakultät, Seminar für Sozialwissenschaften

Immenente und exmanente Kritik

Immanente Kritik

- Einfache Zufallsstichproben – komplexe Stichproben
- Was tun, wenn Verteilung eines Schätzers nicht bekannt ist?
- Daten aus Vollerhebungen

Exmanente Kritik

- Signifikanztests als Ritual
- Falsches/unzulängliches Verständnis von Signifikanztests

Komplexere Stichprobendesigns: Einführung

In der Forschungspraxis ist die Annahme einer einfachen Zufallsstichprobe selten gegeben (gerade bei großen Bevölkerungsstichproben); vor allem geklumpte Stichproben sind häufig. Alles, was wir bislang gelernt haben, bezieht sich auf einfache Zufallsstichproben (i. i. d.-Annahme – „independent and identically distributed“ = unabhängig und identisch verteilt). Hinzu kommen Gewichtungen, mit denen beispielsweise Ausfallwahrscheinlichkeiten korrigiert werden sollen.

Ist das Stichprobendesign bekannt, kann man es (häufig allerdings nur annähernd) durch geeignete Schätzung der Standardfehler berücksichtigen. Entsprechende Möglichkeiten wurde in professioneller Statistiksoftware in den letzten Jahren zunehmend implementiert.

Komplexere Stichprobendesigns: Ein einfaches Beispiel

Bei Ziehung von Klumpenstichproben wird der Standardfehler für Mittelwerte (im einfachsten Fall) berechnet als

$$S.E. = \sqrt{\frac{MQS_{zwischen}}{n}}$$

mit den Klumpen als Gruppierungsvariable. $MQS_{zwischen}$ (die „Varianz“ der Mittelwerte der Klumpen) tritt also an die Stelle von $\widehat{\sigma}^2$. (Entsprechend werden die Freiheitsgrade für die t-Statistik für das Konfidenzintervall aus der Zahl Klumpen berechnet.)

Wenn die Mittelwerte der Klumpen sich wenig unterscheiden (und $MQS_{zwischen}$ entsprechend klein ist), kann das Konfidenzintervall sogar präziser sein als bei einer einfachen Zufallsstichprobe. Allerdings ist dieser Fall eher selten.

Komplexere Stichprobendesigns: Der Designeffekt

Designeffekt = Unterschied zwischen Varianz der Stichprobenkennwerte einer Größe θ, also $V(\theta)$, in einer komplexen Stichproben und der Varianz einer vergleichbaren einfachen Zufallsstichprobe.

Kishs Designeffekt: $DEFF = \frac{V(\widehat{\theta})}{V(\widehat{\theta}_{srs})}$

mit $V(\widehat{\theta}_{srs})$ als der Varianz einer vergleichbaren einfachen Zufallsstichprobe. Manchmal wird statt dem (auf die Varianz der Schätzwerte bezogenen) $DEFF$ auch $DEFT \approx \sqrt{DEFF}$ berichtet, also das Ausmaß, in dem sich der Standardfehler ändert.

Eine andere Größe zur Beurteilung der Auswirkungen von Klumpenstichproben setzt $V(\widehat{\theta})$ in Beziehung zu der „fehlspezifierten“ Varianz $V(\widehat{\theta}_{msp})$, die sich aus der Klumpenstichprobe ergibt, wenn diese wie eine einfache Stichprobe behandelt wird.

Fehlspezifikationseffekt: $MEFF = \frac{V(\widehat{\theta})}{V(\widehat{\theta}_{msp})}$

Unbekannte Verteilung der Stichprobenkennwerte

Für zahlreiche weniger gängige (bzw. StatistikerInnen weniger interessierende) Größen ist die Verteilung der Stichprobenkennwerte nicht bekannt. Für diese oder andere Fälle (z. B. kleine Fallzahlen bei Statistiken, für die nur die large sample-Verteilung bekannt ist) wurden Resampling-Verfahren entwickelt:

Jackknife (gilt heute als weitgehend überholt): Aus der Stichprobe werden nacheinander die einzelnen Fälle (oder: kleine Gruppen von Fällen) entfernt und anhand der so gezogenen Teilstichproben die Standardfehler geschätzt.

Bootstrap: Aus der Stichprobe werden viele Stichproben (vom Umfang n) *mit Zurücklegen* gezogen und aus ihnen die Standardfehler geschätzt (verschiedene Verfahren der Schätzung existieren).

NB! Resampling-Verfahren können oft auch genutzt werden, um komplexe Stichproben auszuwerten.

Daten aus Vollerhebungen

Inferenzstatistik soll (vor allem) Zufallskomponenten berücksichtigen, die aus Stichprobenziehung resultieren → Signifikanztests bei Vollerhebungen angemessen?

Mögliche Antworten:

- Nein.
- Ja, wenn Messfehler vorliegen.
- Ja, wenn stochastische Komponenten vorliegen (z.B. Fehlerterm (=Residuen) im linearen Regressionsmodell) (Broscheid & Gschwend 2005).

Signifikanztests als Ritual bzw. Religion

- „Signifikant“ = Ergebnis; „nicht signifikant“ = kein Ergebnis.
- „ $<$ 5 Prozent“: Hurra, ein Ergebnis! „6 Prozent“: kein Ergebnis (Willkür der 5-Prozent-Grenze).
- „Signifikant“ zählt; selten Frage nach Größe des Effekts.
- Fehlende Berücksichtigung der „Power“ von Tests (Wahrscheinlichkeit, mit der ein in der Grundgesamtheit tatsächlich vorhandener Effekt mit einem auf eine Stichprobe angewandten statistischen Test auch entdeckt werden kann; entspricht $1 - \beta$ [Fehler 2. Art]).

Was sagt ein Signifikanztest?

„Unfortunately, literature suggests that after a statistics course the average students cannot describe the underlying idea of NHST (scil. Null Hypothesis Significance Testing). What is mastered is the mere calculation of a significance test.“ (Haller & Krauss, 2002).

Die Zustimmungsraten zu sechs Aussagen in einer Untersuchung durch diese Autoren bestätigen das:

Aussage	Methoden-Dozenten	Andere Wissenschaftler	Studierende
1)	10	15	34
2)	17	16	32
3)	10	13	20
4)	33	33	59
5)	73	67	68
6)	37	49	41

Was sagen Signifikanztests (Fortsetzung)?

Alles, was ein Signifikanztest sagt, ist, dass die Daten unter der Annahme, dass die Nullhypothese gilt, unwahrscheinlich sind (mit gegebenem Signifikanzniveau).

Die Annahme, dass die Nullhypothese gilt, ist aber äußerst unwahrscheinlich (vor allem im typischen Fall der Nullhypothese, dass kein Unterschied besteht). Warum also eine Hypothese widerlegen, von der man schon weiß, dass sie nicht stimmt?

Sagt uns der Signifikanztest, was wir wissen wollen? Wir wollen doch wissen, wie (un)wahrscheinlich die Nullhypothese ist, gegeben unsere Daten. Wir testen aber, wie unwahrscheinlich die Daten sind, gegeben die Nullhypothese.

Was sagen Signifikanztests (Fortsetzung)?

Eine bayesianische Antwort (siehe Cohen, Dubben/Beck-Bornholdt u.a.)

Unsere Frage ähnelt stark dem bayesianischen Herangehen.

Um zu beurteilen, was unsere Daten sagen, müssen wir etwas über das (Nicht-)Zutreffen von H0 in der Grundgesamtheit wissen (bzw. annehmen) – genau wie wir beim klassischen Diagnoseproblem wissen müssen, wie viel „Gesunde“ und wie viel „Kranke“ es gibt. Außerdem müssen wir etwas über die Power des Tests wissen (entspricht Sensitivität beim Diagnoseproblem).

Was sagen Signifikanztests (Fortsetzung)?

Irrtumswahrscheinlichkeit

nach Dubben/Beck-Bornholdt 2006 bei Power von 80 Prozent

	Anzahl Studien	Signifikant	Nicht signifikant
A besser als B	100	80 % richtig signifikant: 80 Studien	Falsch nicht signifikant: 20 Studien
A gleich gut wie B	900	5 % falsch signifikant: 45 Studien	Richtig nicht signifikant: 855 Studien
Summe	1000	125	

„Positiver prädiktiver Wert“: $80/125 = 64\,\%$

Irrtumswahrscheinlichkeit $100 - 64 = 36\,\%$

Statistische Ergebnisse von Studien I

Ein Beispiel (aus Cumming 2012, S. 1 ff.):

Zwei Studien untersuchen Wirkung eines neuen Medikaments (im Vergleich zu Standardtherapie):

- Lucky (2008): Je Gruppe 22 Vp., Mittelwertunterschied: 3,61, $t = 2{,}43$, $p = 0{,}02$
- Noluck (2008): Je Gruppe 18 Vp., Mittelwertunterschied 2,23, $t = 1{,}25$, $p = 0{,}22$

Sind die Studien (a) einander widersprechend (inkonsistent), (b) uneindeutig, oder (c) konsistent?

Statistische Ergebnisse von Studien II

Die gleichen Ergebnisse, hier als Konfidenzintervalle dargestellt.

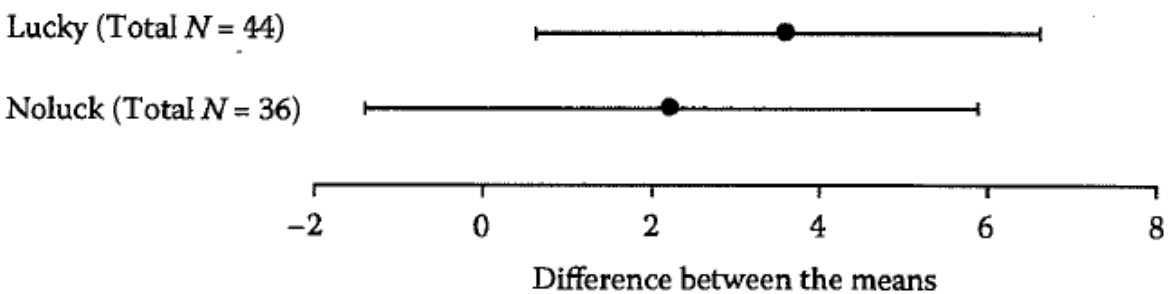

FIGURE 1.1
Difference between the means (mean for new treatment minus mean for current treatment) for treatments for insomnia in the Lucky (2008) and Noluck (2008) studies, with 95% confidence intervals. A positive difference indicates an advantage for the new treatment.

Statistische Ergebnisse von Studien II

Die gleichen Ergebnisse, hier als Metaanalyse

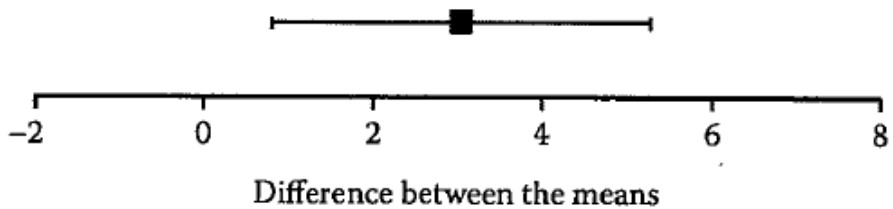

FIGURE 1.2
Difference between the means (mean for new treatment minus mean for current treatment) for treatments for insomnia, with its 95% confidence interval, from a meta-analysis of two studies that compared a new treatment with the current treatment. Total $N = 80$. A positive difference indicates an advantage for the new treatment. The null hypothesis of no difference was rejected, $p = .008$.

Was tun?

- Mir (und den Kritikern von Signifikanztests) nicht glauben – derzeit, und wohl bis auf weiteres, die beste Strategie.
- Jacob Cohen und viele andere: Konfidenzintervalle berichten (WLM: ersatzweise, wenn nicht besser: Standardfehler; außerdem neue Visualisierungsmöglichkeiten für die multivariate Modelle).
- Kein „One Size Fits All“-Denken. Will sagen: Nicht routinemäßig Standardverfahren anwenden, sondern überlegen (und sich gegebenenfalls erkundigen): Was habe ich für Daten? Wie sind sie entstanden? Welche Eigenschaften (z. B. Messfehler) haben sie? In Abhängigkeit von den Antworten gegebenenfalls Non-Standard-Verfahren anwenden.

Literatur

- Broscheid, Andreas/Gschwend, Thomas: Zur statistischen Analyse von Vollerhebungen, in: Politische Vierteljahresschrift 46, 2005, S. O16-O26.
- Cohen, Jacob: The earth is round ($p < .05$), in: Harlow, L. L./Mulaik, S. A./Steiger, J. H. (Hrsg.): What if there were no significance tests? Mahwah, New Jersey; London: Lawrence Erlbaum Associates, 1997, S. 21-35.
- Cumming, Geoff (2012): Understanding the New Statistics. Effect Sizes, Confidence Intervals, and Meta-Analysis, London: Routledge
- Dubben, Hans-Hermann/Beck-Bornholdt, Hans-Peter: Die Bedeutung der statistischen Signifikanz, in: Diekmann, A. (Hrsg.): Methoden der Sozialforschung. Sonderheft 44/2004 der Kölner Zeitschrift für Soziologie und Sozialpsychologie. Wiesbaden: VS Verlag für Sozialwissenschaften, 2006, S. 61-74.
- Haller, Heiko/Krauss, Stefan: Misinterpretations of Significance: A Problem Students Share with Their Teachers? in: Methods of Psychological Research Online 7, 2002, S. 1-20.
- Kohler, Ulrich: Schätzer für komplexe Stichproben, in: Behnke, J./Gschwend, T./Schindler, D./Schnapp, K.-U. (Hrsg.): Methoden der Politikwissenschaft. Neuere qualitative und quantitative Analyseverfahren. Baden-Baden: Nomos, 2006, S. 309-320.
- Shikano, Susumu: Bootstrap and Jackknife, in: Behnke, J./Gschwend, T./Schindler, D./Schnapp, K.-U. (Hrsg.): Methoden der Politikwissenschaft. Neuere qualitative und quantitative Analyseverfahren. Baden-Baden: Nomos, 2006, S. 68-79.