Willkommen zur Vorlesung Statistik (Master)

Thema dieser Vorlesung:
Statistisches Testen

Prof. Dr. Wolfgang Ludwig-Mayerhofer

Universität Siegen – Philosophische Fakultät, Seminar für Sozialwissenschaften

Statistische Hypothesentests (Signifikanztests)

- Einführung
- Signifikanztests nach R. A. Fisher
- Statistisches Testen nach Neyman & Pearson
- Die Praxis statistischer Tests in den Sozialwissenschaften
- Signifikanztests und Konfidenzintervalle
- Abschließendes und Literatur

Statistische Tests: Zwei typische Probleme

Ein-Stichprobenfall / Annahme über einen Populationswert:

In einer Stichprobe ergibt sich ein Kennwert, der von einem „ex negativo“ angenommenen Parameter in der Grundgesamtheit abweicht.
$\rightarrow$ Ist diese Abweichung nur durch die Zufälligkeiten der Stichprobenziehung zustande gekommen, oder ist sie „real“?

Zwei-Stichprobenfall / Annahme über zwei Populationswerte:

In Stichprobendaten zeigt sich ein Unterschied zwischen zwei Gruppen (etwa hinsichtlich eines Mittel- oder Anteilswerts).
$\rightarrow$ Ist dieser Unterschied nur durch die Zufälligkeiten der Stichprobenziehung zustande gekommen, oder ist er „real“?

Über einige Grundzüge des Vorgehens bei der Beantwortung solcher Fragen besteht heute weitgehend Einigkeit. Dennoch gibt es noch heute Debatten über wichtige Details.

Statistische Tests: Das Maß aller Dinge?

Es gibt allerdings auch intensive Kritik an der aktuellen Praxis statistischer Tests. Diese richten sich allerdings teilweise mehr gegen häufige Fehler (und Fehlinterpretationen) als gegen die grundlegende Logik des Testens. Dennoch sind statistische Tests umstritten, sei es, weil man nicht glaubt, dass die Praxis verbessert werden kann, sei es, weil man die grundlegende Logik der Tests für falsch hält.

Beispielsweise wurde in der American Psychological Association Mitte der 1990-er Jahre intensiv diskutiert, ob man statistische Tests nicht abschaffen sollte (allerdings wurde letztlich nur eine Empfehlung ausgesprochen, sie zu verbessern bzw. zu ergänzen, siehe Wilkinson 1999).

Aber schon weil die Praxis der Sozialforschung wenig Anzeichen macht, die Kritik ernst zu nehmen, müssen wir uns mit der gängigen Testpraxis auseinandersetzen.

Zum Sprachgebrauch

Die Unterschiede in den Empfehlungen für das Vorgehen bei statistischen Tests schlagen sich auch im Sprachgebrauch nieder.

Manche AutorInnen (hauptsächlich aber: Autoren!) sprechen bei statistischem Testen generell von Signifikanztests. Andere reservieren diesen Begriff für eine spezifische Teilmenge statistischer Tests oder für eine spezifische Vorgehensweise (z. B. K & K, neueste Auflage [2012]: Signifikanztest = Testen nach R. A. Fisher) und gebrauchen als Oberbegriff allgemeine Begriffe wie statistisches Testen oder Hypothesentests. (Im angelsächsischen Sprachraum verwendet man dagegen sogar einen noch spezielleren Begriff, den des „Null Hypothesis Significance Testing“ (NHST)).

Da wir in allen (gängigen) Testprozeduren am Ende zu Aussagen kommen wollen, ob ein Stichprobenergebnis „signifikant“ ist (ob es auf ein „reales“ Ergebnis hindeutet, also eines, das auch in der GG gültig ist), gebrauche ich den Begriff des Signifikanztests im weiteren Sinne.

Zum weiteren Vorgehen

In den letzten Jahren wird man sich dessen bewusst, dass die aktuelle Praxis der Signifikanztests ein (nicht unbedingt geglücktes) Amalgam aus unterschiedlichen „Testlogiken“ ist. Da es die historische Entwicklung (und Diskussion) m. E. auch leichter macht, statistisches Testen zu verstehen,

1. beschreibe ich zunächst die Logik des Testens nach R. A. Fisher,
2. gehe dann auf die Weiterentwicklung des Testens durch Jerzy Neyman und Egon S. Pearson ein, und
3. skizziere vor diesem Hintergrund die heute gängige Praxis (die in den Sozialwissenschaften kaum über Fisher hinaus ist).

Die Grundidee von Signifikanztests I

Der moderne Signifikanztest geht auf R. A. Fisher zurück. Am Beispiel des Zwei-Stichproben-Falles formuliert, ist die Grundüberlegung:

Es wird ausgegangen von Annahmen darüber, welche Stichprobenergebnisse mehr bzw. weniger wahrscheinlich wären, wenn unsere Vermutung über den beobachteten Unterschied in der Grundgesamtheit nicht zuträfe.

Wenn unser Stichprobenergebnis im Lichte dieser Annahme (dass in der Grundgesamtheit der beobachtete Unterschied nicht vorzufinden ist) unwahrscheinlich ist, entscheiden wir, dass die Annahme über die Grundgesamtheit (wahrscheinlich) falsch ist und erachten die Annahme, dass der Unterschied ‚wirklich' (d. h. in der Grundgesamtheit) besteht, als vorläufig bestätigt (natürlich mit einem Risiko, dass wir uns irren).

Ein in diesem Sinne unwahrscheinliches Ergebnis heißt (statistisch) signifikant.

Die Grundidee von Signifikanztests II

Um Aussagen darüber zu machen, ob ein Stichprobenergebnis unwahrscheinlich ist (bezogen auf die Annahme, dass kein Unterschied besteht!), muss das Ergebnis als Wert einer Zufallsvariablen ausgedrückt werden können, deren Verteilung bekannt ist.

Dabei greifen wir wieder auf die Ergebnisse der theoretischen Statistik zurück, die begründet, welches die Standardfehler (= Streuungen von Stichprobenergebnissen) für spezifische Testgrößen sind und welcher Verteilung diese Standardfehler folgen.

Beispiel: Signifikanztest für Mittelwertunterschiede I

Der Standardfehler für einen Mittelwertunterschied zwischen zwei Gruppen beträgt unter bestimmten Umständen

$$\sqrt{\left(\frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_2^2}{n_2}\right)}$$

wobei $\widehat{\sigma}_1^2$ und $\widehat{\sigma}_2^2$ die (aus der Stichprobe geschätzten) Varianzen und n_1 und n_2 Stichprobenumfänge von Gruppe 1 bzw. Gruppe 2 sind. Die mit Hilfe des Standardfehlers gleichsam „standardisierte“ Differenz der Mittelwerte zweier Gruppen, die wir als $\bar{x}_1$ und $\bar{x}_2$ bezeichnen wollen, lautet dann:

$$T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\left(\frac{\widehat{\sigma}_1^2}{n_1} + \frac{\widehat{\sigma}_2^2}{n_2}\right)}}$$

Diese Statistik folgt einer t-Verteilung oder – bei großen Fallzahlen – einer Standardnormalverteilung.

Beispiel: Signifikanztest für Mittelwertunterschiede II

Unsere Annahme laute: $\mu_1 > \mu_2$ (eine Annahme über die Mittelwerte der Grundgesamtheit!). Sie wird kontrastiert mit der hypothetischen Annahme $\mu_1 \leq \mu_2$. Die mit unser Forschungshypothese kontrastierende hypothetische Annahme heißt Nullhypothese, meist als H_0 abgekürzt.

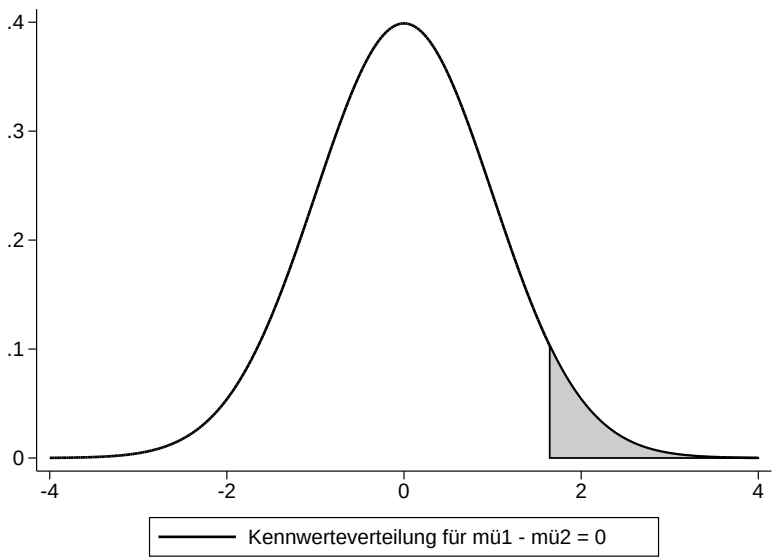
Nach Ansicht der meisten Autoren (siehe etwa Cohen 1990) heißt die Nullhypothese so, weil sie „nullifiziert“, zunichte gemacht werden soll, nicht, weil sie einen „Null-Unterschied“ postuliert, auch wenn dies die gängigste Nullhypothese ist. Eine solche H_0 heißt im Englischen manchmal *nil null hypothesis*.

Die mit der H_0 gerade noch kompatible Mittelwertdifferenz ist $\mu_1 - \mu_2 = 0$. Ist im Lichte dieser Mittelwertdifferenz das Stichprobenergebnis unwahrscheinlich, nehmen wir dies als Indikator, dass die H_0 nicht zutrifft.

Beispiel: Signifikanztest für Mittelwertunterschiede III

Um ein Stichprobenergebnis als „unwahrscheinlich“ zu qualifizieren, müssen wir noch den entsprechenden Grad der (Un-)Wahrscheinlichkeit festlegen (meist wird auch hier 5 Prozent gewählt, also ein Ergebnis, das nur in 5 Prozent der Fälle auftreten kann). Diese Wahrscheinlichkeit nennt man Signifikanzniveau (manchmal auch Irrtumswahrscheinlichkeit); sie wird mit α bezeichnet.

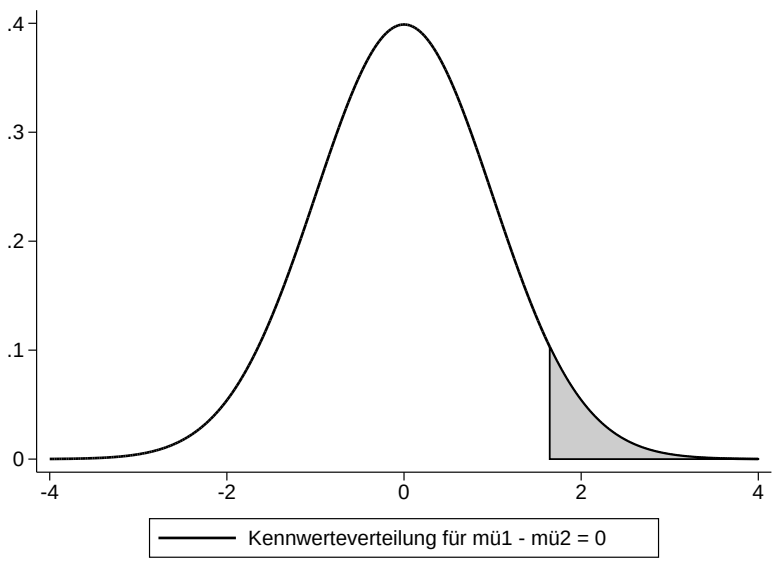
In unserem Beispiel würden wir bei $\alpha = 0{,}05$ alle Stichprobenergebnisse $> 1{,}645$ in diesem Sinne als unwahrscheinlich bezeichnen (bei Verwendung der SNV).



Beispiel: Signifikanztest für Mittelwertunterschiede IV

Der Bereich $> 1{,}645$ wird auch als Ablehnungsbereich bezeichnet, der Wert 1,645 selbst als „kritischer Wert“.

(Zur Erinnerung – dies gilt nur für die H_0, dass $\mu_1 \leq \mu_2$ und für ein Signifikanzniveau von 5 Prozent).



Beispiel: Signifikanztest für Mittelwertunterschiede V

Beispiel: Gegeben sei eine Stichprobenerhebung mit folgenden Größen:

$\bar{x}_1$:	2100	$\bar{x}_2$:	1800
$\widehat{\sigma}_1^2$:	1000000	$\widehat{\sigma}_2^2$:	800000
n_1:	100	n_2:	80

Eingesetzt in die Formel für die T-Statistik erhalten wir:

$$T = \frac{2100 - 1800}{\sqrt{\left(\frac{1000000}{100} + \frac{800000}{80}\right)}} = \frac{300}{141{,}421356} = 2{,}12132$$

der Wert liegt also im Ablehnungsbereich (bei $\alpha = 0{,}05$).

Der Wert entspricht dem 0,983-Quantil der SNV (bzw. 0,982-Quantil einer t-Verteilung mit 178 Freiheitsgraden). Statistik-Software gibt in der Regel das Quantil 1-T als „p-Wert“ aus (hier also 0,0177 bzw. 0,0178). Dieser Wert wird auch als „empirisches Signifikanzniveau“ bezeichnet.

Interpretation eines Signifikanztests nach Fisher

Haben wir ein signifikantes Ergebnis erhalten, so können wir schließen:

> *„Either an exceptionally rare chance has occurred, or the theory . . . is not true“ (Fisher 1973, S. 43)*

– entweder ist ein außergewöhnlich seltener Zufall aufgetreten, oder die Theorie (gemeint ist hier die geprüfte Nullhypothese) ist nicht wahr.

Ist das Ergebnis des Tests nicht signifikant, so können wir jedoch *nicht* schließen, dass die H_0 zutrifft. Vielmehr enthalten wir uns dann einer Aussage.

Probleme des Signifikanztests nach Fisher

Der Signifikanztest nach Fisher befasst sich ausschließlich mit der Wahrscheinlichkeit einer Fehlentscheidung unter der Annahme, dass die H_0 zutrifft (nämlich α). Man kann auf dieser Grundlage aber

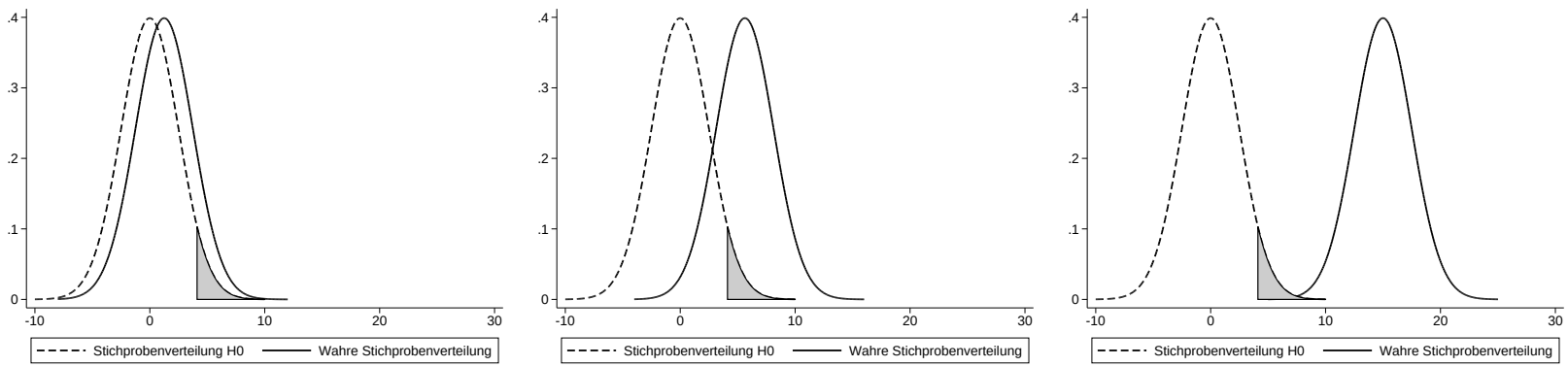
1. keine Aussage machen, wie groß die Chancen sind, einen vorhandenen Effekt überhaupt zu entdecken,
2. (wie schon gesagt) ein nicht-signifikantes Ergebnis nicht interpretieren.

Diese beiden (miteinander verwandten) Probleme lassen sich nur lösen, wenn man der H_0 eine möglichst spezifische Alternativhypothese (H_A oder H_1) gegenüberstellt.

Alternativhypothesen und Testverteilungen

Auf der Grundlage von Annahmen (!) über den „wahren“ (oder: über einen „lohnenswerten“) Unterschied sowie anhand des Standardfehlers (bekanntlich zusammengesetzt aus Streuung des Merkmals und Stichprobenumfang) kann man Aussagen darüber machen, wie wahrscheinlich es ist, einen vorhandenen Unterschied tatsächlich zu entdecken. Man spricht hier von der Power (auch: Teststärke, Trennschärfe) eines Tests.

Die folgenden Abbildungen zeigen (v. l. n. r.) Testsituationen mit geringer, mittlerer und sehr hoher Power (gestrichelt: Stichprobenverteilung für H_0, durchgezogen: Verteilung für H_1).



Power – rechnerisch verdeutlicht

Die Abbildungen der vorherigen Folie beruhen auf Verteilungen mit einer S.D. (hier S.E.) von 2,28. Der kritische Wert (5-Prozent-Signifikanzniveau) entspricht also $1{,}645 \cdot 2{,}28 = 3{,}75$ (der Mittelwert unter H_0 beträgt 0).

- In der linken Graphik (geringe Power) beträgt der Mittelwert unter der H_A 2. Der Wert 3,75 entspricht standardisiert dem Wert $(3{,}75 - 2)/2{,}28 = 0{,}7675$.
  Wir sehen also bei z = 0,7675 in der SNV-Tabelle nach und finden, dass dieser Wert ungefähr dem Quantil 0,7775 entsprechen dürfte (in der Tabelle sind nur die Werte 0,755 [Quantil 0,775] und 0,772 [Quantil 0,780] ausgewiesen). Mit einer Wahrscheinlichkeit von ca. 0,7775 liegt also ein Testergebnis nicht im Ablehnungsbereich.
- In der rechten Graphik (Mittelwert H_A: 15) entspricht der Wert 3,75 standardisiert dem Wert $-4{,}934211$. Dieser Wert liegt extrem weit links in der SNV (ausgewiesen ist nur das 0,5-Prozent Quantil mit dem Wert $-2{,}576$). Die Wahrscheinlichkeit, einen Wert zu erhalten, der nicht im Ablehnungsbereich liegt, ist also außerordentlich gering.

Situationen mit niedriger oder sehr hoher Power

In einer Testsituation mit sehr niedriger Power ist die Durchführung des geplanten Tests Ressourcenvergeudung (die Wahrscheinlichkeit, die H_0 abzulehnen, ist sehr gering; der Test ist nicht informativ). Man muss überlegen,

- ob es sich lohnt, sich überhaupt auf die Suche nach einem Effekt zu machen, und
- falls ja, die Power durch Verringerung des Standardfehlers (=Vergrößerung der Stichprobe) zu erhöhen.

In einer Testsituation mit sehr hoher Power

- gibt man wahrscheinlich zu viel Geld aus (geringere Stichprobe wäre ausreichend),
- läuft man (bei großen Stichproben) Gefahr, triviale (=sehr geringe) Effekte als „Ergebnis“ zu bejubeln.

Fehler 1. und 2. Art

Wir müssen also tatsächlich mit zwei möglichen Fehlern rechnen.

1. Der Fehler 1. Art, auch α-Fehler genannt, bezeichnet die Wahrscheinlichkeit, die Nullhypothese zu verwerfen, obwohl sie tatsächlich zutrifft (also in der Grundgesamtheit gilt).
2. Der Fehler 2. Art, auch β-Fehler genannt, bezeichnet die Wahrscheinlichkeit, die Nullhypothese beizubehalten, obwohl sie tatsächlich *nicht* zutrifft (also in der Grundgesamtheit nicht gilt). Das Risiko des β-Fehlers kann (in Situation mit geringer Power) bis zu $1 - \alpha$ betragen!

Die Power oder Trennschärfe eines Tests ist $1 - \beta$. Für einige Testsituationen kann man die erforderlichen Stichprobenumfänge nachlesen (Cohen 1988) oder berechnen lassen (z. B. mit Stata oder mit Freeware G*Power, siehe `http://www.psycho.uni-duesseldorf.de/aap/projects/gpower/`).

Power, Fehler 1. und 2. Art

Grundsätzlich gilt: Je kleiner man den Fehler 1. Art festlegt, desto größer wird der Fehler 2. Art.

Nach Neyman & Pearson sollte man Verhältnis von Fehler 1. Art und Fehler 2. Art jeweils in Abhängigkeit vom Erkenntnisziel (und von der Power) festlegen. Prüft man beispielsweise ein neues Medikament, das weniger unerwünschte Arzneimittelwirkungen (UAW) hat als ein etabliertes, so könnte man den Fehler 1. Art größer halten als gewöhnlich, da die irrtümliche Annahme, dass das Medikament besser wirkt (mit Blick auf den Therapieerfolg), immer noch mit einem Nutzen für die Patienten verbunden ist (weniger UAW).

Soweit es um wissenschaftliche Erkennnisziele geht, besteht aber weitgehend Einigkeit, dass der Fehler 1. Art gering gehalten werden sollte. Allerdings sollte man auch den Fehler 2. Art nicht zu groß werden lassen (Ressourcenvergeudung durch Untersuchungen mit wenig Aussicht, ein Ergebnis zu finden). Das gilt erst recht, wenn die H_1 lautet, dass *kein* Unterschied besteht.

„Der Hybrid“: Die Praxis statistischer Tests I

In der aktuellen Praxis gerade von Soziologie, Politikwissenschaft oder Ökonometrie spielen Fragen der Test-Power keine Rolle (zur Ökonometrie siehe aber Diskussion im Journal of Socio-Economics Bd. 33, 2004).

Man könnte aber argumentieren, dass angesichts der üblicherweise großen Stichproben in den Sozialwissenschaften diese Frage wenig relevant ist. Eher wäre zu diskutieren, ob nicht die Gefahr besteht, auch schwache Effekte als „wissenschaftlich relevant“ einzustufen. Allgemein ist zu konstatieren, dass die Frage der Größe der Effekte selten diskutiert wird.

Tatsächlich wird in den meisten Untersuchungen nur geprüft, ob überhaupt ein Effekt besteht (z. B., ob die Regressionskoeffizienten signifikant von Null verschieden sind). Es werden zwar häufig H_0 und H_1 formuliert (wenn auch teilweise nur implizit), aber die H_1 ist in der Regel trivial, eben: „Es besteht ein Effekt“ (u. U.: „Es besteht ein positiver (oder negativer] Effekt“).

„Der Hybrid“: Die Praxis statistischer Tests

Die sozialwissenschaftliche Praxis ist also näher am Vorgehen nach Fisher; von Neyman & Pearson übernimmt sie zwar die Unterscheidung zwischen H_0 und H_1, nimmt sie aber nicht ernst. Von Fisher übernimmt sie übrigens nicht die Überlegung, dass ein einzelner Signifikanztest nur ein Indiz ist (siehe Gigerenzer et al. 1989, von denen auch der Ausdruck „Hybrid“ stammt).

Der Unterschied zwischen dem Vorgehen nach Fisher und dem Neyman & Pearson ist im übrigen gering, was das Ziel des statistischen Testens angeht; in beiden Fällen wird vorrangig geprüft, ob der Wert der Teststatistik im Ablehnungsbereich der H_0 liegt oder nicht. Der Unterschied besteht vor allem in der Sorgfalt der Überlegungen hinsichtlich der Test-Power und der damit verknüpften Möglichkeit einer besseren Interpretation nicht signifikanter Ergebnisse.

Überblick: Das Vorgehen bei Signifikanztests

Die aktuelle Praxis von Signifikanztests in den Sozialwissenschaften sieht etwa so aus (siehe auch Lehrbücher):

1. Formulierung von Null- und Alternativhypothese
2. Auswahl der statistischen Prüfgröße (Teststatistik)
3. Festlegung des Signifikanzniveaus und (damit) des Ablehnungsbereiches
4. Berechnung der Teststatistik und Entscheidung über Akzeptanz oder Ablehnung der Nullhypothese

Kühnel & Krebs (Ausgabe 2012) schlagen als Schritt 5 vor: Überprüfung der Anwendungsvoraussetzungen des Tests. Die geschieht in den Sozialwissenschaften mal mehr, mal weniger.

Signifikanztests I: Null- und Alternativhypothese

Beispiel (Zwei-Stichproben-Fall): Es soll (anhand von Stichprobendaten) geprüft werden, ob Frauen weniger Lohn erhalten als Männer.

Diese unsere Vermutung (Hypothese) heißt Alternativhypothese (Kürzel: H_1).

Die Nullhypothese (H_0) formuliert das Gegenteil: Frauen erhalten den gleichen Lohn wie Männer, oder sogar mehr Lohn als diese.

Signifikanztests I: Arten von Hypothesen

Gerichtete (oder einseitige) Hypothesen: Männer verdienen mehr als Frauen (oder: Frauen verdienen weniger als Männer) (allgemein: die Richtung eines Unterschieds oder eines Zusammenhangs wird angegeben).

$$H_0: \mu_1 \leqslant \mu_2$$

$$H_1: \mu_1 > \mu_2$$

Ungerichtete (oder zweiseitige) Hypothesen: Frauen verdienen nicht das Gleiche wie Männer (allgemein: es wird behauptet, dass ein Unterschied oder ein Zusammenhang besteht; über dessen Richtung wird nichts ausgesagt).

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

Signifikanztests I: Arten von Hypothesen

Die wohl am häufigsten getesteten Hypothesen (neben Hypothesen über Unterschiede zwischen Gruppen in Mittel- oder Anteilswerten, allgemein: in der Verteilung der Werte) lauten:

- Es besteht ein Zusammenhang zwischen zwei Merkmalen (ungerichtete Hypothese), oder: es besteht ein positiver bzw. negativer Zusammenhang (gerichtete Hypothese).
- Ein Regressionskoeffizient ist von 0 verschieden (ungerichtete Hypothese), oder: ein Regressionskoeffizient ist größer bzw. kleiner als 0.

Zur Beachtung: Es gibt nicht wenige Tests, die nur die undifferenzierte H_1 prüfen, dass „irgendein Zusammenhang“ (oder „irgendein Unterschied“) besteht.

Umgekehrt ließen sich in vielen Fällen noch wesentlich differenziertere Hypothesen testen (z. B.: der Unterschied [Zusammenhang, Regressionskoeffizient] hat mindestens den Betrag X); dies geschieht in den Sozialwissenschaften sehr selten.

Signifikanztests II: Auswahl der Prüfgröße/Teststatistik

Die Prüfgröße oder Teststatistik: Eine aus den Stichprobendaten zu berechnende Kennzahl, die Aufschluss darüber gibt, ob die Daten mit der Nullhypothese vereinbar sind oder nicht. Die Auswahl und Anwendung der Teststatistik wird primär von der Art der Daten (z. B. Skalenniveau) bzw. der Fragestellung (der zu prüfenden Hypothese) motiviert.

Die Teststatik ist die Überführung der in den Stichprobendaten vorhandenen Information über einen Unterschied (Zusammenhang , Regressionskoeffizienten etc.) in eine Größe, deren Verteilung bekannt ist.

Teststatistiken für häufige Anwendungsfälle zu vermitteln ist das wichtigste Ziel der nächsten Vorlesungsstunden.

Für viele Tests gelten bestimmte Anwendungsvoraussetzungen. Sind diese verletzt, so sind gegebenenfalls alternative Testverfahren zu verwenden.

Signifikanztests III: Signifikanzniveau, Ablehnungsbereich

Ähnlich wie beim Berechnen von Konfidenzintervallen ist klar, dass wir eine gewisse Unsicherheit in Kauf nehmen – genauer: ein Risiko, zu einer falschen Entscheidung über die Hypothesen zu kommen. Bezogen auf die H_0 ist dieses Risiko das Signifikanzniveau.

Signifikanzniveau heißt: Wahrscheinlichkeit, die H_0 abzulehnen, obwohl sie tatsächlich zutrifft.

Diese Wahrscheinlichkeit sollte möglichst niedrig sein, z. B. 0,05 oder noch geringer.

Wir wissen aber nicht, ob H_0 zutrifft oder nicht (sonst würden wir ja nicht testen); noch mehr: im strengen Sinn dürfte die H_0 nie zutreffen, denn irgendein Unterschied, und sei er noch so klein, besteht immer zwischen zwei Gruppen. Das Signikanzniveau ist also eine rein hypothetische Konstruktion, die unsere Entscheidungen anleitet, keine Wahrscheinlichkeit dafür, dass etwas bestimmtes der Fall ist (oder nicht).

Signifikanztests III: Signifikanzniveau, Ablehnungsbereich

Charakteristischerweise hat die Teststatistik den Wert 0, wenn (in der Stichprobe) kein Zusammenhang / Unterschied besteht (entspricht der H_0). Werte, die im Lichte der H_0 sehr unwahrscheinlich sind, liegen am Rand der Verteilung, im Ablehnungsbereich. Dessen Größe ergibt sich aus dem Signifikanzniveau ($\alpha = 0{,}05 \rightarrow$ die fünf Prozent der extremsten Werte werden als Indiz gegen die H_0 gesehen).

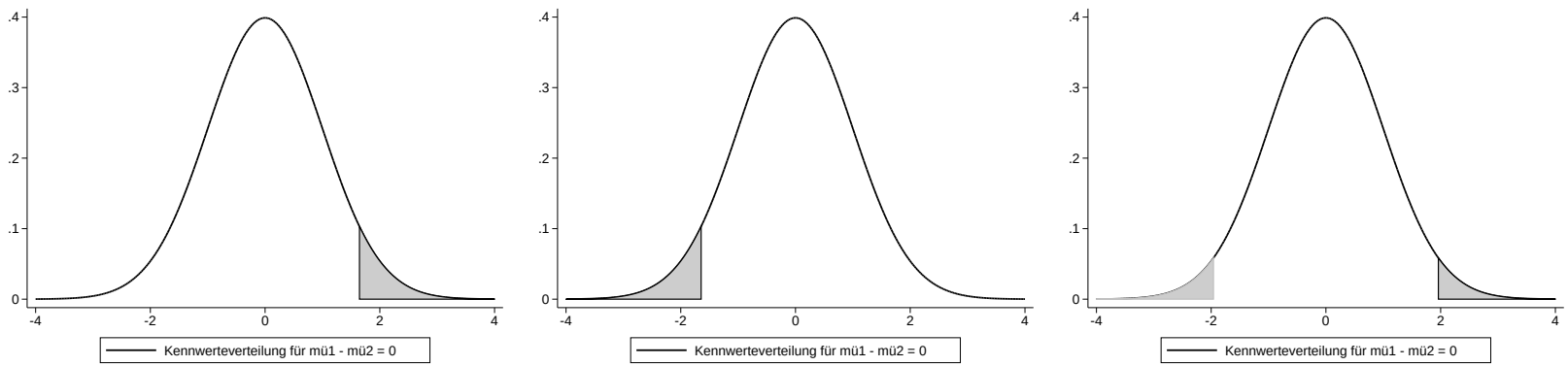
Im Ablehnungsbereich liegen all jene (möglichen) Werte der Teststatistik, die bei gegebenem α nicht mehr mit der Nullhypothese vereinbar sind.

Angenommen, die Teststatistik folgt einer Standardnormalverteilung (SNV). Bei einer ungerichteten Hypothese und einem Signifikanzniveau von 5 % beschreiben jeweils die „äußersten“ 2,5 Prozent der Verteilung den Ablehnungsbereich, also die Werte $< -1{,}96$ und $> +1{,}96$.

Bei einer gerichteten (einseitigen) Hypothese und gleichem Signifikanzniveau beschreiben die obersten (oder untersten) 5 % der SNV den Ablehnungsbereich, also die Werte $> +1{,}645$ (wenn die Hypothese einen positiven Unterschied behauptet) oder $< -1{,}645$ (bei negativem Unterschied).

Signifikanztests III: Ablehnungsbereiche allgemein

Je nach Art und Richtung der Forschungshypothese liegt der Ablehnungsbereich im oberen Bereich ($\mu_1 > \mu_2$, linke Graphik), im unteren Bereich ($\mu_1 < \mu_2$, mittlere Graphik) oder im oberen und unteren Bereich ($\mu_1 \neq \mu_2$, rechte Graphik). In letzterem Fall entspricht der Ablehnungsbereich auf jeder Seite $\alpha/2$.



Die Graphiken zeigen das Beispiel der vorherigen Seite: SNV und $\alpha = 0{,}05$.

Signifikanztests IV: Berechnung der Teststatistik und Entscheidung

Liegt die Teststatistik im Ablehnungsbereich, wird die Nullhypothese verworfen. Dabei geht man aber ein Risiko in der Höhe von α ein, eine Nullhypothese zu verwerfen, obwohl in Wahrheit die Nullhypothese gilt. Dies ist der Fehler 1. Art. Wie gesagt, ist dies ein rein hypothetisches Risiko, da die H_0 im strengen Sinn nie gilt.

Den Fehler 2. Art kann man, wie gesagt, ebenso nur hypothetisch bestimmen, nämlich unter Annahmen über den wahren Wert des Unterschieds in der Grundgesamtheit (und bei gegebenem S. E.).

Empirische Signifikanz: der p-Wert

Statistikprogramme gegen im allgemeinen nicht an, ob ein bestimmtes Signifikanzniveau überschritten wird; sondern sie geben exakt das Quantil (fast immer als 1-p) an, das von der jeweiligen Prüfgröße abgeschnitten wird (siehe Beispiel).

Objektive Armut	west-/ostdeutschland		
50% Grenze	west	ost inkl.	Total
trifft nicht zu	572	304	876
	91.37	85.15	89.11
trifft zu	54	53	107
	8.63	14.85	10.89
Total	626	357	983
	100.00	100.00	100.00

Pearson chi2(1) = 9.0668 Pr = 0.003

Wichtig: Ein p von „0,000“ ist ein gerundeter Wert und bedeutet: das empirische Signifikanzniveau ist $< 0{,}0005$. Ein p von exakt 0 gibt es nicht.

Grade der Signifikanz

In den Sozialwissenschaften hat sich eingebürgert, ausgehend vom p-Wert verschiedene „Grade der (empirischen) Signifikanz“ zu unterscheiden und durch Sterne zu markieren:

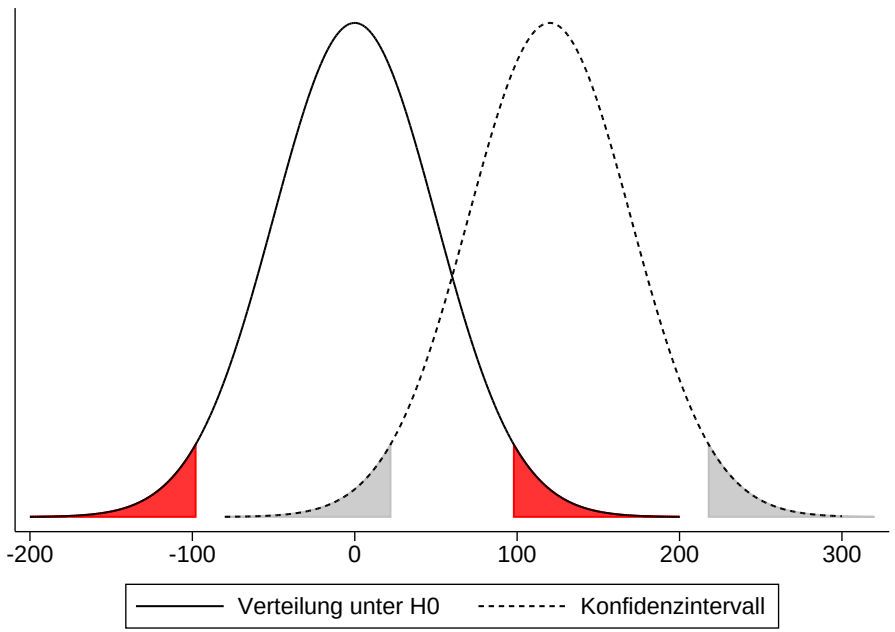
$0{,}01 \leq p < 0{,}05$:	*	„signifikant“
$0{,}001 \leq p < 0{,}01$:	**	„sehr signifikant“
$p < 0{,}001$:	***	„höchst signifikant“

Gelegentlich werden Werte $0{,}05 \leq p < 0{,}10$ ebenfalls gekennzeichnet (etwa durch +).

Signifikanztest und Konfidenzintervalle

Konfidenzintervalle können in einer Reihe von Fällen an die Stelle von Signifikanztests treten.

Liegt ein Stichprobenmittelwert im Ablehnungsbereich einer H_0, so ist das (grosso modo) äquivalent mit der Aussage, dass das entsprechende Konfidenzintervall um den Stichprobenmittelwert den der H_0 entsprechenden Wert nicht einschließt.



Signifikanztest und KI: formal

Beispiel Z-Test, ungerichtete Hypothese. Die Nullhypothese wird beibehalten, wenn gilt:

$$|z| = \left| \frac{\bar{x}-\mu}{\frac{\sigma_X}{\sqrt{n}}} \right| < z_{1-\alpha/2}, \quad \text{d. h. wenn gilt:}$$

$$\bar{x} - \mu > -z_{1-\alpha/2} \cdot \frac{\sigma_X}{\sqrt{n}} \quad \text{und} \quad \bar{x} - \mu < z_{1-\alpha/2} \cdot \frac{\sigma_X}{\sqrt{n}}.$$

Umformen:

$$-\mu > -\bar{x} - z_{1-\alpha/2} \cdot \frac{\sigma_X}{\sqrt{n}} \quad \text{bzw.} \quad -\mu < -\bar{x} + z_{1-\alpha/2} \cdot \frac{\sigma_X}{\sqrt{n}}$$

$$\mu < \bar{x} + z_{1-\alpha/2} \cdot \frac{\sigma_X}{\sqrt{n}} \quad \text{bzw.} \quad \mu > \bar{x} - z_{1-\alpha/2} \cdot \frac{\sigma_X}{\sqrt{n}}.$$

Das sind aber die Grenzen des Konfidenzintervalls für α. Die H_0 wird also beibehalten, wenn der ihr entsprechende Wert $\bar{x}$ (oder allgemein: $\widehat{\theta}$) innerhalb des Konfidenzintervalls, und verworfen, wenn er außerhalb desselben liegt.

Zum Abschluss: Korrekter Umgang mit Signifikanztests

- Das Ergebnis eines Signifikanztests sagt nur: Die Daten sind (bei einem gewissen α) mit der H_0 inkompatibel, daher sollten wir die H_0 verwerfen. Es sagt nichts darüber, wie wahrscheinlich oder unwahrscheinlich irgendetwas (z. B. die H_0 oder die H_1) „tatsächlich ist“.
- Das Ergebnis eines Signifikanztests sagt nichts darüber, ob ein Unterschied/Zusammenhang/Regressionskoeffizient groß, wichtig, bedeutend oder stark ist. Bei kleinem Standardfehler (geringe Streuung und/oder große Stichprobe) fällt fast jeder statistische Test signifikant aus.
- Ein *nicht* signifikantes Ergebnis ist ohne Überlegungen zur Power uninterpretierbar.

Zum Abschluss: Korrekter Umgang mit Signifikanztests

Ein α von 0,05 besagt, dass man durchschnittlich in 5 Prozent der Fälle, in denen die H_0 zutrifft, gleichwohl ein signifikantes Ergebnis erhält. Das heißt: Auch wenn man Daten hätte, in denen durchgängig nur die H_0 zutrifft, würden bei Durchführung vieler Tests im Durchschnitt 5 Prozent derselben signifikante Ergebnisse liefern.

Fährt man also viele Tests durch, gelangt man mehr oder zwangsläufig zu signifikanten Ergebnissen (genau so, wie auch gelegentlich Menschen im Lotto gewinnen, obwohl die Chance außerordentlich gering ist – es gibt eben sehr viele Menschen, die Lotto spielen).

$\rightarrow$ Man sollte nur ex ante (im vorhinein) formulierte Hypothesen prüfen, nicht „in den Daten nach Zusammenhängen suchen“.

In der Psychologie sind auch Regeln dafür verbreitet, wie man α bei Durchführung mehrerer Tests (anhand der gleichen Stichprobe) anpasst.

Statistische und substanzielle Hypothesen

Anfang des 18. Jahrhundert untersuchte John Arbuthnot anhand von Londoner Geburtsstatistiken die Hypothese: Es gibt eine göttliche Vorsehung.

Da Männer infolge von Krieg etc. häufiger sterben als Frauen, müsste die göttliche Vorsehung dafür sorgen, dass mehr Männer als Frauen geboren werden. Man kann diese Hypothese gegen die H_0 stellen: „Es werden zufällig mal mehr, mal weniger Männer als Frauen geboren“.

Die Geburtsstatistiken über viele Jahrzehnte zeigten tatsächlich, dass Jahr für Jahr mehr Männer als Frauen geboren werden. Im Lichte der H_0 war das Ergebnis außerordentlich unwahrscheinlich. Das beweist: Es gibt (bei einer verschwindend kleinen Irrtumswahrscheinlichkeit) eine göttliche Vorsehung.

Es handelt sich hierbei um den ersten bekannten statistischen Signifikanztest.

Literatur

Arbuthnot, J. (1710). An argument for divine providence taken from the constant regularity in the births of both sexes. Philosophical Transactions of the Royal Society, 27, 186-190.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2. Aufl.). Hillsdale: Erlbaum.

Cohen, J. (1990). Things I have learned (so far). American Psychologist, 45, 1304-1312.

Fisher, R. A. (1973b). Statistical Methods and Scientific Inference (3. Auflage). New York/London: Hafner/Collier McMillan (hier zitiert nach Fisher, R. A.: Statistical Methods, Experimental Design, and Scientific Inference, Oxford: Oxford University Press, 1990).

Gigerenzer, G., Swijtink, Z., Porter, T., Daston, L., Beatty, J., & Krüger, L. (1989). The empire of chance: How probability changed science and everyday life. Cambridge, UK: Cambridge University Press.

Wilkinson, L. (1999). Statistical methods in psychology journals: Guidelines and explanations. American Psychologist, 54(8), 594-604.