Stochastik

Barth, Friedrich

München, [20]03

17. 4. 1. Zusammengesetzte Hypothesen beim zweiseitigen Test

urn:nbn:de:hbz:466:1-83580

17.4.1. Zusammengesetzte Hypothesen beim zweiseitigen Test

Der einfachste Fall eines Signifikanztests besteht zunächst einmal darin, daß die Nullhypothese, über die entschieden werden soll, einfach ist, wogegen als Gegenhypothese mehrere, meist sogar unendlich viele Hypothesen in Frage kommen. Eine solche aus mehreren einfachen Hypothesen bestehende Hypothese heißt **zusammengesetzt.**

Beispiel 1: Zweiseitiger Test einer einfachen Nullhypothese über eine unbekannte Wahrscheinlichkeit. Eine Urne enthalte 10 Kugeln, darunter womöglich auch rote. Theodor behauptet, die Urne enthalte genau 7 rote Kugeln. Diese Behauptung ist also die einfache Nullhypothese. Die Gegenhypothese besteht aus 10 möglichen einfachen Hypothesen; es können nämlich weniger oder mehr als 7 rote Kugeln in der Urne sein. Bezeichnet man den Anteil der roten Kugeln in der Urne mit p, dann kann man diese beiden Hypothesen folgendermaßen kurz charakterisieren:
Nullhypothese $H_0 : p = \frac{7}{10}$
Zusammengesetzte Gegenhypothese H_1: $p \in \{0, \frac{1}{10}, \frac{2}{10}, \ldots, \frac{6}{10}, \frac{8}{10}, \frac{9}{10}, 1\}$
Noch kürzer lassen sich die beiden Hypothesen abstrakt als Mengen von Parametern darstellen; in unserem Fall

$$H_0 = \{\tfrac{7}{10}\} \qquad \text{und} \qquad H_1 = \{0, \tfrac{1}{10}, \tfrac{2}{10}, \ldots, \tfrac{6}{10}, \tfrac{8}{10}, \tfrac{9}{10}, 1\}.$$

Die Menge $H := H_0 \cup H_1$ ist die Menge aller zulässigen Parameter; sie heißt **zulässige Hypothese.**
Zur Durchführung des Tests ziehen wir eine Stichprobe von 6 Kugeln, der Einfachheit halber mit Zurücklegen. Testgröße Z ist die Anzahl der roten Kugeln in der Stichprobe, für die 11 Wahrscheinlichkeitsverteilungen $B(6;p)$ möglich sind. Damit läßt sich die zulässige Hypothese H auch als Menge aller Binomialverteilungen $B(6;p)$ mit $p \in \{0, \frac{1}{10}, \ldots, \frac{9}{10}, 1\}$ schreiben. Da $\mathscr{E}Z = 4{,}2$ ist, falls H_0 vorliegt, halten wir die Ergebnisse »4 rote« bzw. »5 rote Kugeln« in der Stichprobe für verträglich mit H_0. Größere Abweichungen vom Erwartungswert $\mathscr{E}Z$ bezeichnet man als **signifikante Abweichungen***. Wir halten sie normalerweise nicht mehr für verträglich mit H_0. Da die Gegenhypothese sowohl kleinere als auch größere p-Werte als $\frac{7}{10}$ enthält, wird man als Annahmebereich für H_1 zwei getrennt liegende Intervalle wählen. Tests mit solchen Annahmebereichen heißen **zweiseitig.** In unserem Beispiel liegt somit folgende Entscheidungsregel nahe:

$$\delta: \begin{cases} Z \in \{0, 1, 2, 3\} \cup \{6\} & \Rightarrow \text{Entscheidung für } H_1 \\ Z \in \{4, 5\} & \Rightarrow \text{Entscheidung für } H_0 \end{cases}$$

Wie beim Alternativtest haben wir auch hier 2 Möglichkeiten, Fehlentscheidungen zu treffen.

Fehler 1. Art: Die Nullhypothese H_0 trifft tatsächlich zu, aber $Z \in \{0, 1, 2, 3, 6\}$, d.h., es hat sich trotzdem eine signifikante Abweichung ergeben. Man würde

* significare (lat.) = anzeigen, verkünden.

sich also fälschlicherweise für H_1 entscheiden. Die Wahrscheinlichkeit für einen derartigen Fehler 1. Art ergibt sich zu

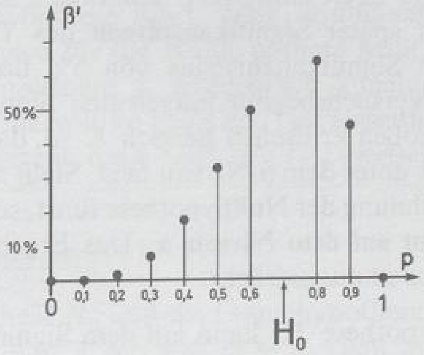
$$\begin{aligned}\alpha' = P_{0,7}^6(\{0, 1, 2, 3, 6\}) &= F_{0,7}^6(3) + \mathrm{B}(6; \tfrac{7}{10}; 6) = \\ &= 0{,}25569 + 0{,}11765 = \\ &= 0{,}37334 \approx 37{,}3\%.\end{aligned}$$

Fehler 2. Art: Eine der 10 einfachen Hypothesen aus der zusammengesetzten Gegenhypothese H_1 trifft tatsächlich zu, aber $Z \in \{4; 5\}$. Man müßte sich für H_0 entscheiden. Und wie groß ist der Fehler, den man dann begeht? Das ist gar nicht so leicht zu beantworten! Denn die Wahrscheinlichkeit für einen Fehler 2. Art hängt nun davon ab, welche der einfachen Hypothesen, die die zusammengesetzte Hypothese H_1 bilden, tatsächlich vorliegt. Diese möglichen Fehlerwahrscheinlichkeiten β' hängen also von p ab:

$$\beta'(p) = P_p^6(\{4; 5\}) = F_p^6(5) - F_p^6(3).$$

Eine leichte Rechnung liefert Tabelle 347.1, deren graphischer Ausdruck Figur 347.1 ist.

p	$\beta'(p)$
0	0
0,1	0,00127
0,2	0,01690
0,3	0,06974
0,4	0,17510
0,5	0,32813
0,6	0,49766
0,8	0,63898
0,9	0,45271
1	0



Tab. 347.1 und Fig. 347.1 Abhängigkeit der Wahrscheinlichkeit für einen Fehler 2. Art von der tatsächlich vorliegenden einfachen Gegenhypothese zur Nullhypothese »$p = 0{,}7$«

Weil man mit dem Schlimmsten rechnen muß, interessiert man sich für den Maximalwert der Wahrscheinlichkeit für einen Fehler 2. Art. In unserem Fall ist dies

$$\beta'(\tfrac{8}{10}) = 0{,}63898 \approx 63{,}9\%.$$

Dieser Wert ist so groß, daß man sich trotz der oben aufgestellten Entscheidungsregel guten Gewissens nicht für H_0 entscheiden kann. Dieses schlechte Gewissen bringt der Statistiker dadurch zum Ausdruck, daß er in diesem Fall sagt: »Man kann die Nullhypothese H_0 nicht ablehnen (nicht verwerfen).« *Ronald Aylmer Fisher* (1890–1962) schreibt dazu 1935 in *The Design of Experiments:*

»[...] it should be noted that the null hypothesis is never proved or established, but is possibly disproved in the course of experimentation. Every experiment may be said to exist only in order to give the facts a chance of disproving the null hypothesis.«

Die Entscheidung eines Signifikanztests besteht also nicht in der Entscheidung für H_0 oder für H_1, sondern nur in der Ablehnung der Nullhypothese H_0. Eine solche Entscheidung fällt man genau dann, wenn die Testgröße Z einen der signifikanten Werte aus $\{0, 1, 2, 3, 6\}$ annimmt. Man nennt diesen Annahmebereich der Gegenhypothese den **kritischen Bereich** K. Wir müssen also die oben aufgestellte Entscheidungsregel revidieren! Bei einem Signifikanztest lautet sie

$$\delta\colon \begin{cases} Z \in K \Rightarrow \text{Nullhypothese } H_0 \text{ wird abgelehnt.} \\ Z \in \bar{K} \Rightarrow \text{Nullhypothese } H_0 \text{ kann nicht abgelehnt werden.} \end{cases}$$

In Worten: Ist der Ausfall der Stichprobe signifikant, so wird die Nullhypothese abgelehnt, andernfalls beibehalten.
Im Falle $Z \in \bar{K}$ fällt also eigentlich gar keine Entscheidung! Weil dem so ist, interessiert man sich beim Signifikanztest nur für den Fehler 1. Art, die Nullhypothese auf Grund eines signifikanten Ausfalls der Stichprobe zu verwerfen, obwohl sie zutrifft. Fußend auf den Erkenntnissen von *Poisson* (1781–1840) führte 1840 sein Schüler, der Arzt *Louis-Dominique-Jules Gavarret**, in seinem Werk *Principes généraux de statistique médicale* ein, für die Wahrscheinlichkeit α' dieses Fehlers 1. Art eine obere Schranke α festzulegen. Diese obere Schranke α nannte man später **Signifikanzniveau** des Tests. Die heute besonders häufig verwendeten Signifikanzniveaus von 5% und 1% führte *R. A. Fisher* ein. Zu einem vor Versuchsbeginn festgelegten Signifikanzniveau α wählt man einen möglichst großen kritischen Bereich K so, daß die Wahrscheinlichkeit für einen Fehler 1. Art unter dem α-Niveau liegt. Stellt sich dann ein Versuchsergebnis ein, das zur Ablehnung der Nullhypothese führt, so sagt man, dieses Versuchsergebnis sei **signifikant auf dem Niveau α**. Das Ergebnis des Tests wird in diesem Fall üblicherweise so ausgedrückt:

»Die Nullhypothese H_0 kann auf dem Signifikanzniveau α abgelehnt werden.«

Die statistische Sicherheit des Urteils hat dann mindestens den Wert $1 - \alpha$.

Versuchen wir nun zu $\alpha = 25\%$ einen kritischen Bereich K für Theodors Vermutung $H_0 = \{\frac{7}{10}\}$ bzw. $H_0 =$ »Z ist nach $B(6; \frac{7}{10})$ verteilt« zu konstruieren. Dem Problem angemessen setzt sich der kritische Bereich K aus zwei Intervallen $[0; k_1]$ und $[k_2; 6]$ zusammen. Es gäbe viele Möglichkeiten, die Fehlerwahrscheinlichkeit α' auf die beiden Teilintervalle aufzuteilen. Üblich ist es, k_1 und k_2 so zu bestimmen, daß in jedem Teilbereich die Fehlerwahrscheinlichkeiten höchstens $\frac{1}{2}\alpha$ sind. Das führt zu

$$P_{H_0}(Z \leqq k_1) \leqq 12{,}5\% \qquad \text{und} \qquad P_{H_0}(Z \geqq k_2) \leqq 12{,}5\%.$$
$$\Leftrightarrow F_{0,7}^{6}(k_1) \leqq 12{,}5\% \qquad \text{und} \qquad 1 - F_{0,7}^{6}(k_2 - 1) \leqq 12{,}5\%.$$

Das ergibt mit Hilfe der *Stochastik-Tabellen* die Bedingungen

$$k_1 \leqq 2 \qquad \text{und} \qquad k_2 \geqq 6, \qquad \text{also } K = [0; 2] \cup [6; 6] = \{0, 1, 2, 6\}.$$

* 28. 1. 1809 Astaffort – 31. 8. 1890 Valmont. Vor seinem Medizinstudium Artillerie-Offizier; 1843 wurde er auf den Lehrstuhl für Physique médicale der Medizinischen Fakultät von Paris berufen.

Hätte Theodors Stichprobe beispielsweise 2 rote Kugeln geliefert, so könnte man seine Vermutung H_0, die Urne enthalte 7 rote Kugeln, auf dem 25%-Niveau ablehnen. Die Sicherheit des Urteils »Ablehnung von H_0« beträgt mindestens 75%.

Je niedriger das Signifikanzniveau, d.h., je kleiner α ist, desto schärfer ist der Test, aber desto seltener wird man H_0 verwerfen können. Dies entspricht der Erfahrung des täglichen Lebens: Klare Urteile kann man nur selten abgeben, verschwommene Aussagen (d.h. großes Signifikanzniveau!) sind hingegen sehr leicht zu machen.
Wir fassen die Erkenntnisse aus Beispiel 1 zusammen in

Definition 349.1:
Beschränkt man sich bei einem Test darauf, nur für die eine der beiden Hypothesen die Wahrscheinlichkeit α' der fälschlichen Ablehnung klein zu machen, so spricht man von einem **Signifikanztest**. Man nennt diese Hypothese dann **Nullhypothese**. Die gewählte obere Schranke α für die Irrtumswahrscheinlichkeit α' heißt auch **Signifikanzniveau**. Ein Versuchsergebnis, das zur Ablehnung der Nullhypothese führt, heißt **signifikant auf dem Niveau** $\boldsymbol{\alpha}$. Der Ablehnungsbereich für die Nullhypothese heißt **kritischer Bereich** $\boldsymbol{K}$ des Tests, sein Komplement $\overline{K}$ gelegentlich Annahmebereich. Besteht K aus einem einzigen Intervall, so heißt der Test **einseitig**. Wird K durch $\overline{K}$ in zwei Intervalle aufgeteilt, dann heißt der Test **zweiseitig**.

Wie konstruiert man einen Signifikanztest?

1. Man formuliert eine Nullhypothese H_0 und die Gegenhypothese H_1 bzw. die zulässige Hypothese H. Dabei – so *J. Neyman* 1939 in Genf auf einer vom Völkerbund veranstalteten Tagung –

   »hat sich mehr oder weniger eingebürgert, als Nullhypothese diejenige Hypothese zu wählen, bei der die Fehler 1. Art von größerer Bedeutung sind als die Fehler 2. Art.«

2. Man legt eine Testgröße Z fest.
3. Man legt das Signifikanzniveau α des Tests fest.
4. Man konstruiert einen möglichst großen kritischen Bereich K so, daß $P_{H_0}(Z \in K) \leqq \alpha$.
   Besteht K aus zwei Teilintervallen K_1 und K_2, dann bestimmt man sie so, daß $P(Z \in K_1) \leqq \frac{1}{2}\alpha$ und $P(Z \in K_2) \leqq \frac{1}{2}\alpha$ erfüllt sind.
5. Man entscheidet nach folgender Regel:
   $$\delta\colon \begin{cases} Z \in K \Rightarrow H_0 \text{ wird abgelehnt.} \\ Z \in \overline{K} \Rightarrow H_0 \text{ kann nicht abgelehnt werden.} \end{cases}$$
6. Sicherheit des Urteils:
   $1 - \alpha$ heißt **statistische Sicherheit** des Urteils »Ablehnung von H_0«, weil mindestens mit der Wahrscheinlichkeit $1 - \alpha$ das Vorliegen von H_0 erkannt würde.

Zur Veranschaulichung der statistischen Sicherheit stellen wir uns vor, daß n Urnen zum Testen vorliegen. n_0 dieser Urnen enthalten tatsächlich 7 rote Kugeln.

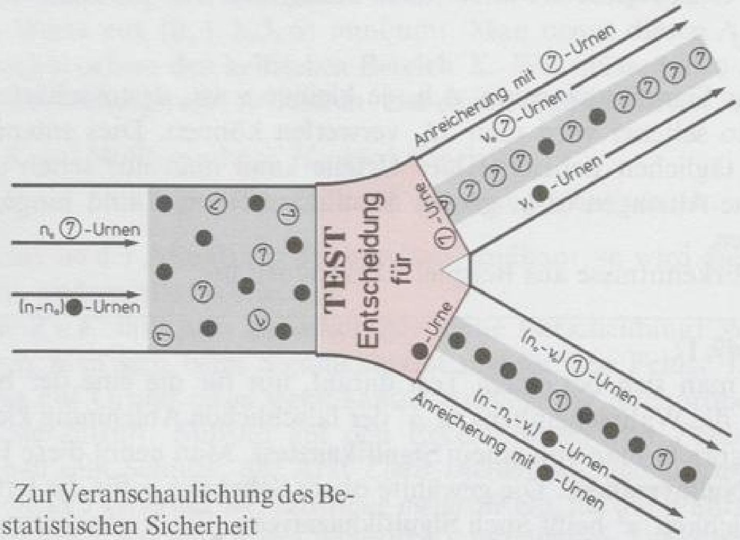


Fig. 350.1 Zur Veranschaulichung des Begriffs der statistischen Sicherheit

(In Figur 350.1 mit ⑦ gekennzeichnet.) Auf Grund der Interpretationsregel für Wahrscheinlichkeiten werden etwa $\alpha' = 37{,}3\%$ dieser Urnen falsch bezeichnet. Der Anteil der falsch bezeichneten Urnen des anderen Typs hängt davon ab, wie viele rote Kugeln die Urne jeweils enthält.

Natürlich ist ein Test kein todsicheres Verfahren zur Trennung der beiden Hypothesen; denn man muß immer Fehlermöglichkeiten in Kauf nehmen. Hören wir dazu *J. Neyman und E.S. Pearson:*

»The tests themselves give no final verdict, but as tools help the worker who is using them to form his final decision; […]. What is of chief importance in order that a sound judgment may be formed is that the method adopted, its scope and its limitations, should be clearly understood«.*

17.4.2. Zusammengesetzte Hypothesen beim einseitigen Test

Beispiel 2: Einseitiger Test einer einfachen Nullhypothese über eine unbekannte Wahrscheinlichkeit. Der Teetassen-Test von *R. A. Fisher***: Lady X. behauptet, sie könne es am Geschmack erkennen, ob der Tee zuerst in der Tasse war und die Milch dazugegeben wurde oder ob man umgekehrt den Tee auf die Milch gegossen habe.
Wir glauben das nicht. Wir setzen, anders als *R. A. Fisher*, Lady X. 10 Tassen Tee mit Milch vor, die in beliebiger – uns bekannter – Weise gefüllt worden sind.

* *On the use and interpretation of certain test criteria for purposes of statistical inference.* Biometrika **20 A** (1928).

** *Sir Ronald Aylmer Fisher* (1890–1962) wählte in *The Design of Experiments* (1935) dieses Beispiel zur Einführung: »A lady declares that by tasting a cup of tea made with milk she can discriminate whether the milk or the tea infusion was first added to the cup. We will consider the problem of designing an experiment by means of which this assertion can be tested.«