Stochastik

Barth, Friedrich

München, [20]03

17. 4. 2. Zusammengesetzte Hypothesen beim einseitigenTest

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

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

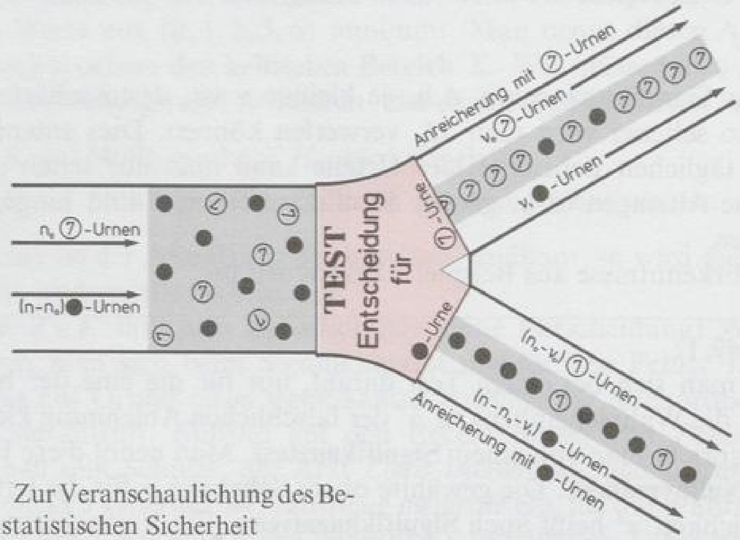

Fig. 350.1 Zur Veranschaulichung des Begriffs der statistischen Sicherheit

(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.«

Lady X. probiert und macht 8mal eine richtige Angabe. Können wir Lady X. die von ihr behauptete geradezu übernatürliche Fähigkeit zugestehen?
Im Gegensatz zu Beispiel 1 aus **17.3.1.** ist das Ergebnis der Stichprobe bereits bekannt. Eine solche Situation ist in der Praxis auch oft anzutreffen. Man könnte nun zwar auch hier vorgehen wie in Beispiel 1, zu einem vorgegebenen Signifikanzniveau α einen kritischen Bereich bestimmen und überprüfen, ob das bekannte Ergebnis des Zufallsexperiments zur Ablehnung der Nullhypothese hinreicht. Statt dessen geht man oft anders vor und bestimmt zu dem eingetretenen Stichprobenergebnis das niedrigste Signifikanzniveau, auf dem man gerade noch die Nullhypothese ablehnen könnte. Wir wollen diese andere Art eines Signifikanztests hier weiter verfolgen. Dazu legen wir uns wieder ein mathematisches Modell für dieses reale Zufallsexperiment zurecht. Das Probieren der Tassen entspricht einer *Bernoulli*-Kette der Länge 10; Treffer beim i-ten Versuch ist das Ereignis »Lady X. beurteilt die i-te Tasse richtig«. Wenn Lady X. sich aufs bloße Raten verlegte, könnte sie genausogut mit einer Laplace-Münze werfen. In diesem Fall hätte also der Parameter der *Bernoulli*-Kette den Wert $\frac{1}{2}$. Besitzt Lady X. hingegen eine Begabung der behaupteten Art, so ist die Wahrscheinlichkeit p für einen Treffer verschieden von $\frac{1}{2}$. $p < \frac{1}{2}$ würde bedeuten, daß Lady X. den Sachverhalt zwar mit gewisser Sicherheit richtig erkennen kann, ihn aber verkehrt benennt. Das hätte sie wohl bei eigenen Versuchen längst selbst bemerkt. Es ist somit sinnvoll, als zulässige Hypothese die Menge $H := \{p \mid \frac{1}{2} \leqq p \leqq 1\}$ zu nehmen. Der Wert p ist also ein Maß für die Begabung von Lady X.; je größer p ist, um so begabter ist sie. Wir wählen als Nullhypothese »Lady X. hat keine Begabung«, kurz »Lady X. rät blind«, also $H_0 := \{\frac{1}{2}\}$, da uns hier ein Fehler 1. Art, nämlich eine unbegabte Dame für begabt zu halten, schlimmer erscheint als ein Fehler 2. Art, nämlich einer begabten Dame die Begabung abzusprechen. Nehmen wir als Testgröße Z die Anzahl der richtig geratenen Tassen, so besagt H_0, Z besitzt die Wahrscheinlichkeitsverteilung $\mathrm{B}(10; \frac{1}{2})$. Die Gegenhypothese lautet »Lady X. ist begabt« also $H_1 := H \backslash H_0$. Sie läßt sich nicht mehr durch endlich viele Parameterwerte beschreiben; alle Zahlen $p \in]\frac{1}{2}; 1]$ sind möglich. Es gibt somit für die Zufallsgröße Z unendlich viele Wahrscheinlichkeitsverteilungen zu dieser Hypothese, nämlich alle $\mathrm{B}(10; p)$ mit $p > \frac{1}{2}$. Da alle p-Werte der Gegenhypothese H_1 auf derselben Seite bezüglich der Nullhypothese »$p = \frac{1}{2}$« liegen, wählt man sinnvollerweise als kritischen Bereich ein Intervall $K := [k; 10]$, so daß das Ereignis »$Z \geqq k$« zur Ablehnung der Nullhypothese führt. Würde man nämlich als kritischen Bereich das Ereignis $K' := [0; k_1] \cup [k_2; 10]$ wählen, so würde man im Falle $Z \in K'$ die Nullhypothese ablehnen, also Lady X. auch dann Begabung bescheinigen, wenn sie nur wenige oder gar keine Tasse richtig benannt hat, was sicherlich nicht erwünscht ist. Da K aus einem einzigen Intervall besteht, handelt es sich also um einen einseitigen Test.
Unser Stichprobenergebnis lautet »$Z = 8$«. Wir müssen somit einen kritischen Bereich wählen, der 8 enthält. Ein möglichst niedriges Signifikanzniveau erzielt man, wenn man den kritischen Bereich möglichst klein wählt. Also entschließen wir uns zu $K := [8; 10]$. Für die Wahrscheinlichkeit α', einen Fehler 1. Art zu begehen, ergibt sich damit
$$\alpha' = P_{H_0}(Z \in K) = P_{0,5}^{10}(Z \geqq 8) = 1 - F_{0,5}^{10}(7) \approx 5{,}5\,\%.$$

Beim üblichen Signifikanzniveau 5% können wir die Nullhypothese »Lady X. rät blind« nicht ablehnen. Ist man jedoch mit einem Signifikanzniveau von 5,5% oder höher zufrieden, so kann man die Nullhypothese »Lady X. rät blind« ablehnen und der Dame Begabung bescheinigen. Die statistische Sicherheit unseres Urteils »Lady X. ist begabt« beträgt dann höchstens 94,5%. Was heißt das? Wenn viele Ladies sich unserer Prüfung unterzögen, attestierten wir ca. 5,5% dieser Damen fälschlicherweise eine gewisse Begabung, weil sie 8 oder mehr Tassen richtig benennen, obwohl sie blind raten.
Was ist aber mit den begabten Damen? Dieser Frage wollen wir im nächsten Abschnitt nachgehen.

17.4.3. Die Operationscharakteristik eines Tests

Beispiel 3: Dem Teetassentest aus Beispiel 2 stellt sich eine Lady, die tatsächlich über eine gewisse Begabung verfügt und mit der Wahrscheinlichkeit $p = 0{,}6$ die Tassen richtig benennt. Mit welcher Wahrscheinlichkeit wird man ihre Begabung verkennen, wenn wir wie in Beispiel 2 als kritischen Bereich die Menge $K = [8; 10]$ nehmen?
Die Wahrscheinlichkeit β', einen solchen Fehler 2. Art zu begehen, ergibt sich zu

$$\beta' = P_{0,6}^{10}(Z \in \bar{K}) = P_{0,6}^{10}(Z \leqq 7) = F_{0,6}^{10}(7) \approx 83{,}3\%.$$

Solchen schwach begabten Damen wird mit unserem Test also oft unrecht getan! Wäre die Begabung der Dame größer, z. B. $p = 0{,}9$, so würden wir sie auch besser erkennen; es ergäbe sich nämlich $\beta' = F_{0,9}^{10}(7) \approx 7{,}0\%$. Weil wir aber über die Begabung der Damen, die sich dem Test unterziehen, nichts wissen, müssen wir uns einen Überblick über alle Wahrscheinlichkeiten für einen Fehler 2. Art verschaffen. Da diese Wahrscheinlichkeiten offensichtlich von p abhängen, betrachten wir die Funktion

$$\beta': p \mapsto P_p^{10}(Z \in \bar{K}), \ D_{\beta'} =]\tfrac{1}{2}; 1].$$

Mit Hilfe einer Wertetabelle können wir den Graphen dieser Funktion zeichnen (Tabelle 353.1 und Figur 353.1).

Man erkennt, daß die Wahrscheinlichkeit β' für einen Fehler 2. Art um so größer wird, je weniger sich die Begabung vom blinden Raten ($p = \frac{1}{2}$) unterscheidet. Da die Definitionsmenge $D_{\beta'}$ links offen ist, gibt es keine größte Irrtumswahrscheinlichkeit 2. Art. Als Ersatz dafür nimmt man das Supremum aller Irrtumswahrscheinlichkeiten 2. Art, also den Wert $1 - \alpha'$. Er ist in unserem Fall etwa 94,5%. Man riskiert also, mit einer Wahrscheinlichkeit bis zu 94,5% begabte – wenn auch sehr schwach begabte – Damen zu Unrecht für unbegabt zu halten. Wir können trotzdem zufrieden sein: Der unangenehme Fall, daß eine Dame nur flunkert und wir ihr dennoch hohe Sensibilität bescheinigen, tritt nur mit 5,5% Wahrscheinlichkeit ein. Daß wir andererseits u. U. einer wirklich begabten Dame ein Unrecht antun, nehmen wir in Kauf in der Gewißheit, daß sich das Genie so oder so eines Tages durchsetzen wird.