UNIVERSITÄTS-
BIBLIOTHEK
PADERBORN

Stochastik

Barth, Friedrich

München, [20]03

17. 4. Signifikanztest

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

Visual Library

Figur 345.1 zeigt in einer vereinfachten Darstellung die Fehlerwahrscheinlichkeiten und die Sicherheiten, je nachdem, welche der beiden Hypothesen vorliegt.

Zum Abschluß geben wir noch einen Überblick über wichtige Aufgabentypen beim Alternativtest. Der Einfachheit halber handle es sich um Hypothesen über den Parameter p einer Binomialverteilung.

Typ 1: Stichprobenlänge n und kritischer Wert k sind gegeben; gesucht sind die Fehlerwahrscheinlichkeiten α' und β'.

Typ 2: Gegeben sind die Stichprobenlänge n und eine obere Schranke α für die Wahrscheinlichkeit α', einen Fehler 1. Art zu begehen. Gesucht ist der sog. beste kritische Wert k, für den α' höchstens α und β' möglichst klein werden.

Typ 3: Gegeben ist je eine obere Schranke α bzw. β für die Fehlerwahrscheinlichkeiten α' bzw. β'. Gesucht ist eine möglichst kleine Stichprobenlänge n und ein dazu passender kritischer Wert k. (Oft wird sich keine eindeutige Lösung ergeben.)

Typ 4: Gegeben sind die Stichprobenlänge n, die jeweiligen Schäden bei den Fehlern 1. bzw. 2. Art und die Wahrscheinlichkeiten für das tatsächliche Vorliegen der beiden Hypothesen. Gesucht ist derjenige kritische Wert k, für den der zu erwartende Schaden minimal wird.

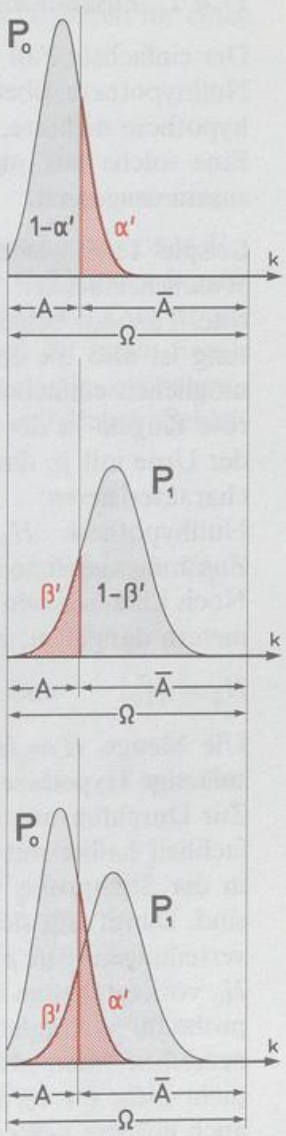


Fig. 345.1 Schematische Skizze für die Wahrscheinlichkeiten der Fehler und der Sicherheiten beim Alternativtest.

17.4. Signifikanztest

Die Situation eines Alternativtests, sich zwischen zwei einfachen Hypothesen entscheiden zu müssen, kommt in der Praxis selten vor, weil die Welt um uns dafür zu kompliziert ist. Sehr viel häufiger stellt sich einem jedoch das folgende **Problem:** Auf Grund irgendwelcher Erfahrungen oder Überlegungen hegt man *eine* Vermutung, die nun durch einen Test, den sog. Signifikanztest, entweder bestätigt oder widerlegt werden soll. Für diese Vermutung prägte *R. A. Fisher* (1890–1962) den Ausdruck **Nullhypothese.** Der Signifikanztest dient – wie sich zeigen wird – dazu, die Frage zu beantworten, ob man mit gutem Grund eine solche Nullhypothese ablehnen kann oder nicht.

UNIVERSITÄTS-
BIBLIOTHEK
PADERBORN

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.

UNIVERSITÄTS-
BIBLIOTHEK
PADERBORN

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

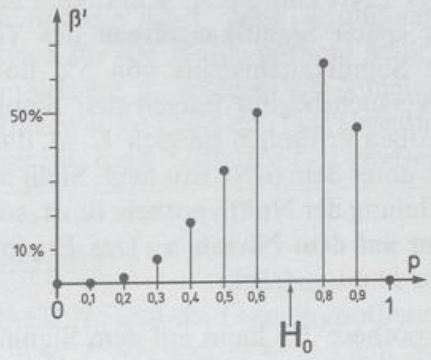
$$\begin{aligned}\alpha' = P^6_{0,7}(\{0,1,2,3,6\}) &= F^6_{0,7}(3) + \mathrm{B}(6;\tfrac{7}{10};6) = \\ &= 0{,}25569 + 0{,}11765 = \\ &= 0{,}37334 \approx 37{,}3\%.\end{aligned}$$

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:

$$\beta'(p) = P^6_p(\{4;5\}) = F^6_p(5) - F^6_p(3).$$

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.«

UNIVERSITÄTS-
BIBLIOTHEK
PADERBORN

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

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

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

$$\begin{aligned} & P_{H_0}(Z \leqq k_1) \leqq 12{,}5\% \quad \text{und} \quad P_{H_0}(Z \geqq k_2) \leqq 12{,}5\%. \\ \Leftrightarrow\ & F^6_{0,7}(k_1) \leqq 12{,}5\% \quad \text{und} \quad 1 - F^6_{0,7}(k_2 - 1) \leqq 12{,}5\%. \end{aligned}$$

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

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

* 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.

UNIVERSITÄTS-
BIBLIOTHEK
PADERBORN

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 α**. Der Ablehnungsbereich für die Nullhypothese heißt **kritischer Bereich 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.

UNIVERSITÄTS-
BIBLIOTHEK
PADERBORN

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

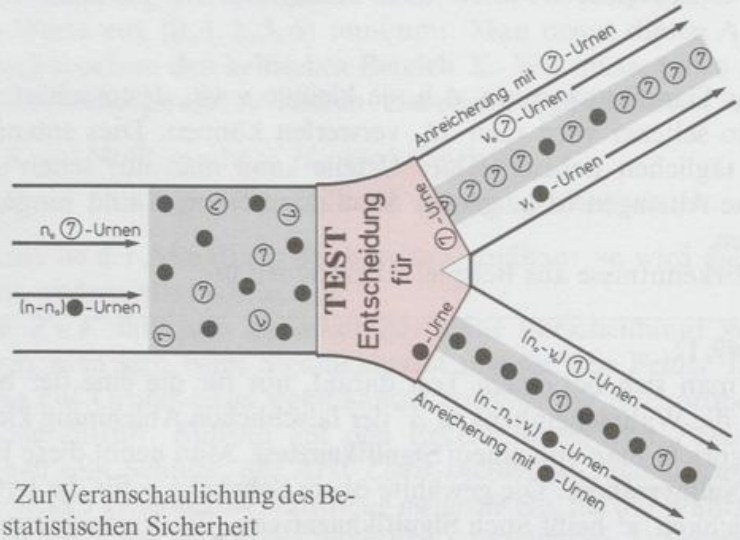


Fig. 350.1 Zur Veranschaulichung des Begriffs der statistischen Sicherheit

(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.«

UNIVERSITÄTS-
BIBLIOTHEK
PADERBORN

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 $B(10; \frac{1}{2})$. 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 $B(10; p)$ 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
$\alpha' = P_{H_0}(Z \in K) = P^{10}_{0,5}(Z \geqq 8) = 1 - F^{10}_{0,5}(7) \approx 5{,}5\%$.

UNIVERSITÄTS-
BIBLIOTHEK
PADERBORN

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

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

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 $\beta' = F^{10}_{0,9}(7) \approx 7{,}0\%$. 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

$$\beta' \colon p \mapsto P^{10}_p(Z \in \bar{K}),\ D_{\beta'} = \left]\tfrac{1}{2}; 1\right].$$

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.

UNIVERSITÄTS-
BIBLIOTHEK
PADERBORN

p	$\beta' = P_p^{10}(Z \leqq 7)$
0,51	0,94
55	90
60	83
65	74
70	62
75	47
80	32
85	18
90	07
95	01
99	0001
1	0

Tab. 353.1 Wahrscheinlichkeit β' für einen Fehler 2. Art beim kritischen Bereich $K = [8; 10]$

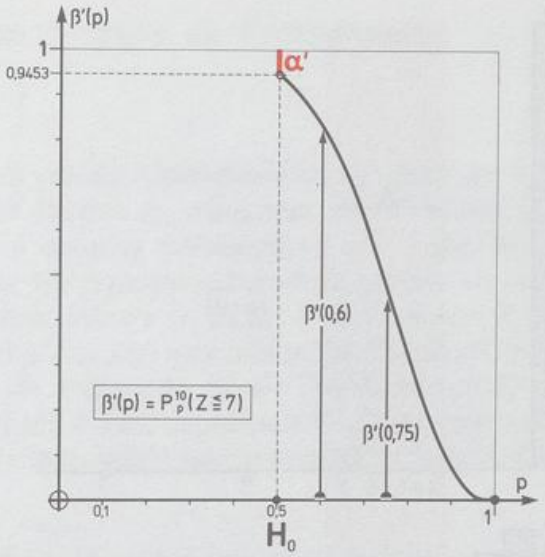


Fig. 353.1 Graph der Funktion $\beta' : p \mapsto P_p^{10}(Z \in \overline{K})$

Setzt sich die Gegenhypothese nur aus endlich vielen einfachen Hypothesen zusammen wie bei Theodors Urne in Beispiel 1 von Seite 346, dann besteht der Graph von β' nur aus diskreten Punkten, so wie ihn Figur 347.1 zeigt. In einem solchen Fall gibt es natürlich eine größte Irrtumswahrscheinlichkeit 2. Art.
Es hat sich in der Statistik eingebürgert, die auf der Gegenhypothese H_1 definierte Funktion $p \mapsto \beta'(p)$ auf die Menge *aller* beim Test betrachteten Hypothesen, d. h. auf die zulässige Hypothese $H := H_0 \cup H_1$ fortzusetzen. Diese Funktion heißt dann Operationscharakteristik des Tests, kurz OC des Tests.

Definition 353.1: Es sei auf dem Ergebnisraum Ω der Testgröße Z eine Menge von Wahrscheinlichkeitsverteilungen als zulässige Hypothese H gegeben. Diese Verteilungen lassen sich durch einen Parameter p kennzeichnen. $A \subset \Omega$ sei ein Ereignis. Dann heißt die Funktion

$$\text{OC}: p \mapsto P_p(A),\ D_{\text{OC}} = H$$

die **Operationscharakteristik des Ereignisses A bezüglich H.** Ihr Graph heißt **OC-Kurve.***

Bemerkung: Der Parameter p muß nicht unbedingt eine Wahrscheinlichkeit sein. So werden z. B. *Poisson*-Verteilungen durch den Parameter »Erwartungswert μ«, Normalverteilungen durch die Parameter μ und σ^2 gekennzeichnet.
Figur 354.1 veranschaulicht am Beispiel des Ereignisses $A := [4; 7]$ und an der Schar $B(16; p)$, $p \in [0; 1]$, als zulässiger Hypothese das Zustandekommen der

* In der Literatur verwendet man vielfach noch die ursprünglich von *Jerzy Neyman* und *E. S. Pearson* zur Kennzeichnung der Güte oder Macht eines Tests eingeführte *power function* = **Gütefunktion g**. Für sie gilt $g(p) := 1 - \text{OC}(p)$.

UNIVERSITÄTS-
BIBLIOTHEK
PADERBORN

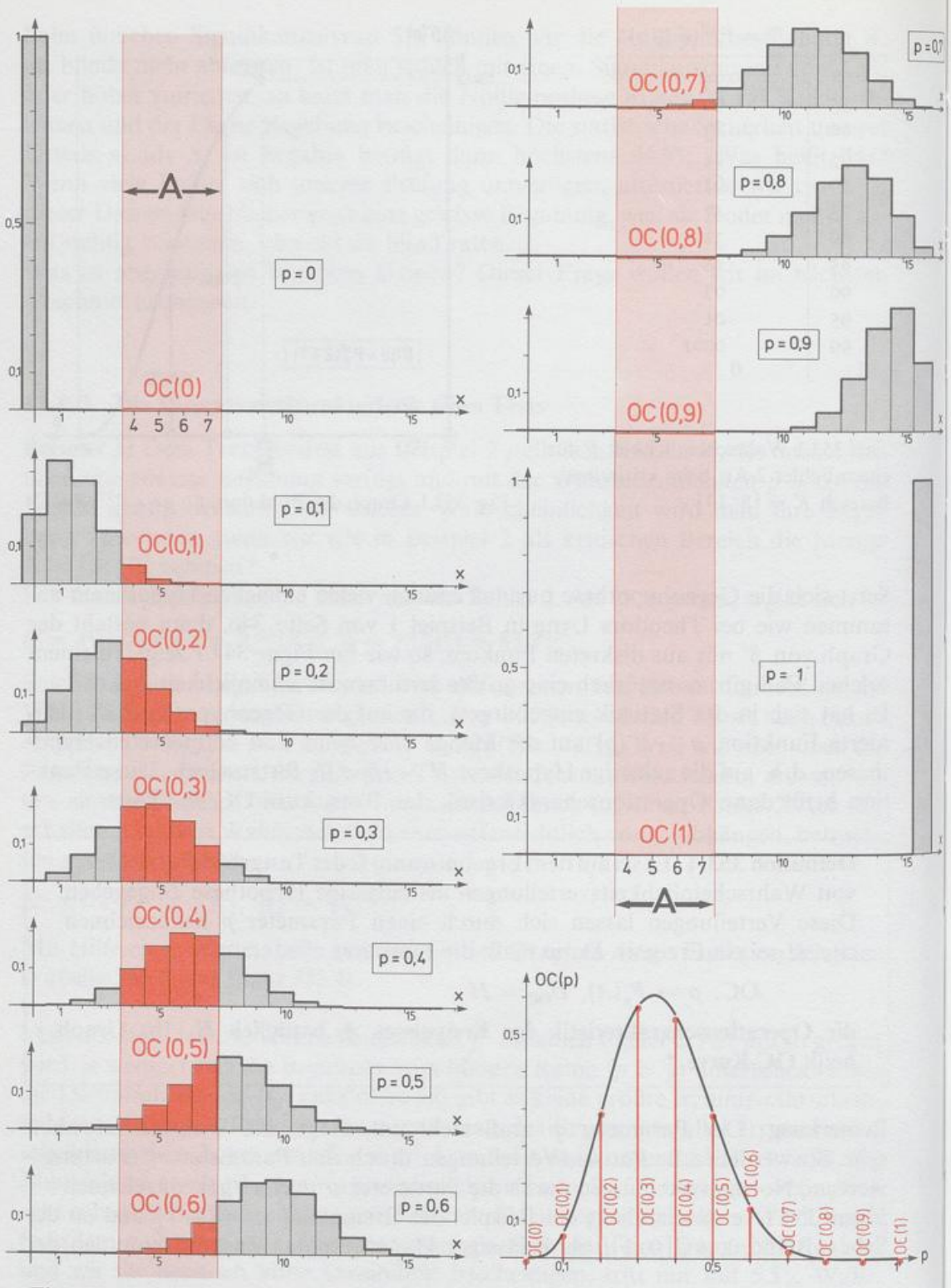


Fig. 354.1 Veranschaulichung der Entstehung der OC des Ereignisses $A := [4; 7]$ bezüglich der zulässigen Hypothese $H := \{B(16; p) | p \in [0; 1]\}$. Bedeutet f_p die Dichtefunktion der Binomialverteilung $B(16; p)$, so läßt sich die Operationscharakteristik mittels eines Integrals schreiben, nämlich OC: $p \mapsto \int_{3,5}^{7,5} f_p(t)\,dt$.

UNIVERSITÄTS-
BIBLIOTHEK
PADERBORN

Operationscharakteristik: Zu jedem p gehört als Funktionswert

$$\mathrm{OC}(p) = P_p^{16}(Z \in A) = \sum_{i=4}^{7} \mathrm{B}(16; p; i).$$

Bei einem Signifikanztest spricht man von der Operationscharakteristik der Entscheidungsregel δ mit dem kritischen Bereich K, wenn man $A = \bar{K}$ wählt. Ihre Funktionswerte $\mathrm{OC}(p) = P_p(\bar{K})$ sind dann in Abhängigkeit von p die Wahrscheinlichkeiten, mit denen man die Nullhypothese beibehält, gleich, ob diese Entscheidung die richtige ist oder nicht. Für $p \in H_1$ ist der Funktionswert $P_p(\bar{K})$ jeweils die Irrtumswahrscheinlichkeit 2. Art, daß man nämlich die Nullhypothese nicht ablehnt, obwohl sie nicht zutrifft. Für $p \in H_0$ ist der Funktionswert $P_p(\bar{K})$ jeweils gleich der Sicherheit $1 - \alpha'(p)$, mit der die zutreffende Nullhypothese nicht abgelehnt wird. Dabei ist $\alpha'(p)$ die zu $p \in H_0$ gehörende Irrtumswahrscheinlichkeit 1. Art.

Übrigens kann auch die Nullhypothese H_0 selbst zusammengesetzt sein. Nehmen wir etwa im Teetassentest von Beispiel 2 (**17.4.2.**) als zulässige Hypothese $H := [0; 1]$ und als Nullhypothese $H_0 := [0; \frac{1}{2}]$, dann ergäbe sich als Operationscharakteristik des Ereignisses »$Z \leqq 7$« die Funktion OC: $p \mapsto F_p^{10}(7)$, $D_{\mathrm{OC}} = [0; 1]$, deren Graph Figur 355.1 wiedergibt. Nun gibt es auch unendlich viele Irrtumswahrscheinlichkeiten 1. Art. Zur Charakterisierung des Tests genügt es offenbar, die größte dieser Wahrscheinlichkeiten anzugeben.
Je nach Lage des kritischen Bereichs K haben die Graphen der Operationscharakteristik, kurz OC-Kurven genannt, eine typische Gestalt. Nehmen wir als zulässige Hypothese die Menge aller Binomialverteilungen $\mathrm{B}(n; p)$ mit $p \in [0; 1]$, so gibt es 4 besonders wichtige Typen. Der Nachweis der aufgeführten Eigenschaften wird Aufgabe 372/**48** vorbehalten.

1) $K := [0; k] \Rightarrow$ OC: $p \mapsto 1 - F_p^n(k)$
Ist K linksbündig, so ist die OC-Kurve echt monoton steigend.

2) $K := [k; n] \Rightarrow$ OC: $p \mapsto F_p^n(k-1)$
Ist K rechtsbündig, so ist die OC-Kurve echt monoton fallend.

3) $K := [0; k_1] \cup [k_2; n] \Rightarrow$
OC: $p \mapsto F_p^n(k_2 - 1) - F_p^n(k_1)$
Ist K getrennt, so hat die OC-Kurve einen inneren Hochpunkt.

4) $K := [k_1; k_2] \Rightarrow$
OC: $p \mapsto F_p^n(k_1 - 1) + 1 - F_p^n(k_2)$
Ist K ein inneres Intervall, so hat die OC-Kurve einen inneren Tiefpunkt.

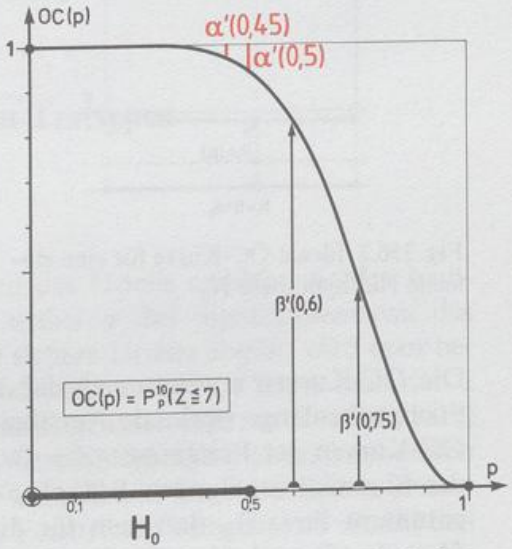


Fig. 355.1 Operationscharakteristik des Ereignisses »$Z \leq 7$« bezüglich $H = [0; 1]$. Vgl. Fig. 353.1

UNIVERSITÄTS-
BIBLIOTHEK
PADERBORN

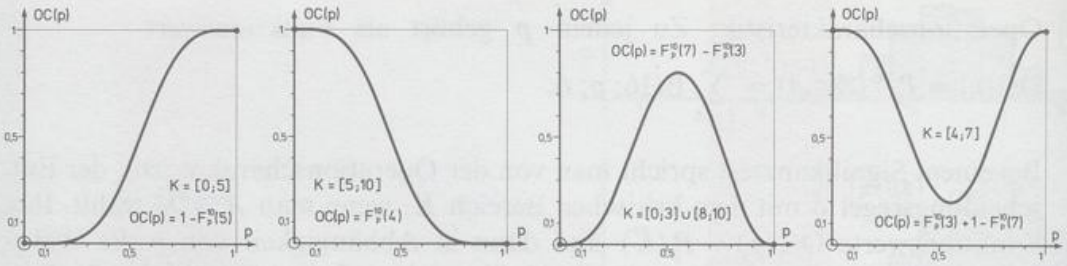


Fig. 356.1 Die 4 wichtigen Typen von OC-Kurven bezüglich $H = \{B(n;p) | p \in [0;1]\}$, veranschaulicht mittels Binomialverteilungen $B(10;p)$

Wie man sich leicht überlegt, sind diese 4 Operationscharakteristiken Polynome n-ten Grades in p. Figur 356.1 veranschaulicht sie für $n = 10$.

Die OC-Kurve gibt uns einen Hinweis auf die Güte des Tests. Je steiler sie nämlich in ihren Flanken ist, desto schneller werden die Irrtumswahrscheinlichkeiten 2. Art klein. Im Idealfall wären für jedes $p \in H_0$ die Irrtumswahrscheinlichkeit $\alpha'(p) = 0$ und für jedes $p \in H_1$ die Irrtumswahrscheinlichkeit $\beta'(p) = 0$. Dann würde man nur richtige Urteile abgeben! Die zugehörige OC-Kurve hätte über H_0 konstant den Wert 1 und über H_1 konstant den Wert 0. Figur 356.2 zeigt die ideale OC-Kurve für eine einfache Nullhypothese, Figur 356.3 für eine zusammengesetzte Nullhypothese.

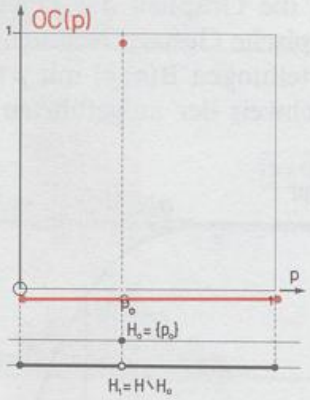


Fig. 356.2 Ideale OC-Kurve für eine einfache Nullhypothese H_0

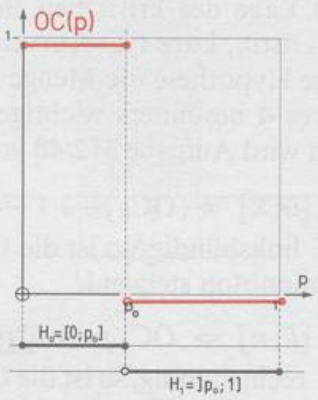


Fig. 356.3 Ideale OC-Kurve für eine zusammengesetzte Nullhypothese H_0

Die OC-Kurven erweisen sich daher als praktisches Hilfsmittel, bei gegebener Stichprobenlänge optimale Annahmebereiche zu finden. Figur 357.1 zeigt die OC-Kurven der Ereignisse »$Z = 0$«, »$Z \leqq 1$«, …, »$Z \leqq 5$« bezüglich der Schar der Binomialverteilungen $B(5;p)$, $p \in [0;1]$, als zulässiger Hypothese H. Man entnimmt ihr z. B., daß man für die Entscheidung zwischen den Hypothesen $H_0 = \{0{,}15\}$ und $H_1 = \{0{,}4\}$ am besten das Ereignis »$Z \leqq 2$« heranzieht, wenn die Wahrscheinlichkeit für einen Fehler 1. Art unter 5% liegen soll. Ohne diese Bedingung würde man sich für »$Z \leqq 1$« entscheiden, weil dann $\alpha' + \beta'$ minimal

UNIVERSITÄTS-
BIBLIOTHEK
PADERBORN

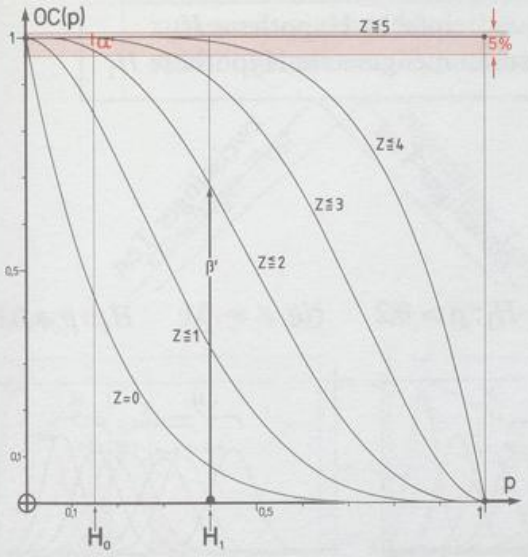


Fig. 357.1 Alternativtest für $H_0 = \{0{,}15\}$, $H_1 = \{0{,}4\}$ und $A = [0; k]$ mit $k \in \{0, 1, 2, 3, 4, 5\}$. Auswahl des optimalen Tests für die Schranke $\alpha = 5\%$

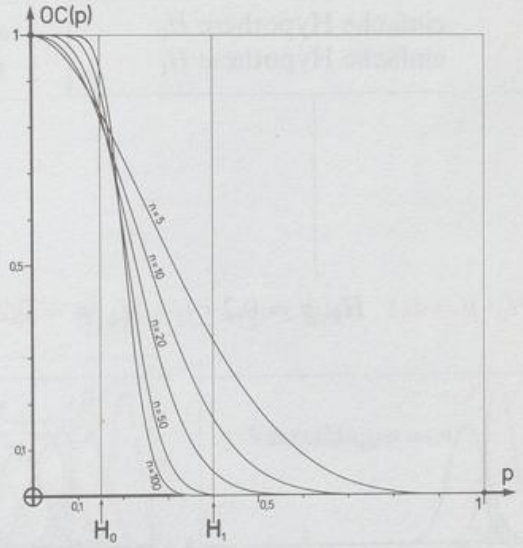


Fig. 357.2 Illustration des Einflusses der Stichprobenlänge n auf die Trennschärfe $H_0 = \{0{,}15\}$; $H_1 = \{0{,}4\}$; $A = [0; 0{,}2n]$; $n \in \{5; 10; 20; 50; 100\}$.

wird. Ein Ereignis ist desto besser für eine Entscheidungsregel geeignet, je stärker die OC-Kurve von dem einen der beiden in Frage kommenden p-Werte bis zum anderen abfällt. Andererseits läßt sich der Einfluß der Stichprobenlänge n auf die **Trennschärfe** des Tests an Hand der zugehörigen OC-Kurven beobachten (Figur 357.2). Wie erwartet fallen die OC-Kurven für größere n steiler von 1 auf 0 ab und trennen daher die Hypothesen besser. Für $n \to \infty$ hätte man einen idealen Test mit senkrecht abfallender OC-Kurve. Die Trennung ist perfekt, die Fehler haben die Wahrscheinlichkeit 0.

17.5. Überblick über die behandelten Testtypen

Siehe Seite 358 f.

17.6. Verfälschte Tests

Bei einem Signifikanztest hat die Sicherheit des Urteils »Ablehnung der Nullhypothese« mindestens den Wert $1 - \alpha$, wobei α das Signifikanzniveau des Tests ist. Da man natürlich gern möglichst sichere Urteile abgibt, wird man bestrebt sein, das Signifikanzniveau α möglichst klein zu halten. Wählt man nun α und damit auch den kritischen Bereich K sehr klein, dann muß man leider in Kauf nehmen, daß nur noch in seltenen Fällen die Nullhypothese abgelehnt werden kann; d. h., der Test wird sehr häufig kein brauchbares Ergebnis liefern. Dieser Sachverhalt könnte einen Tester nun in die Versuchung bringen, erst einmal den Ausfall der Stichprobe abzuwarten und dann den kritischen Bereich K möglichst eng um das Stichprobenergebnis herumzulegen und damit das Signifikanzniveau recht klein zu machen. Der Versuchsausgang erschiene dann in einem besonders

UNIVERSITÄTS-
BIBLIOTHEK
PADERBORN