Heinz Holling
Günther Gediga

Statistik – Testverfahren

psychlehrbuchplus
hogrefe.de

Bachelorstudium
Psychologie

hogrefe

Statistik – Testverfahren

Bachelorstudium Psychologie

Statistik – Testverfahren

Prof. Dr. Heinz Holling, PD Dr. Günther Gediga

Herausgeber der Reihe:

Prof. Dr. Eva Bamberg, Prof. Dr. Hans-Werner Bierhoff,
Prof. Dr. Alexander Grob, Prof. Dr. Franz Petermann

Heinz Holling
Günther Gediga

Statistik – Testverfahren

hogrefe

Prof. Dr. Heinz Holling, geb. 1950. 1969–1976 Studium der Mathematik, Psychologie und Soziologie in Würzburg und Berlin. 1974–1987 Wissenschaftlicher Mitarbeiter an der FU Berlin und der Universität Osnabrück. Promotion 1980 (Dr. phil.) und 1987 (Dr. rer. nat.). 1987 Habilitation. 1987–1993 Vertretungsprofessor an den Universitäten Oldenburg, Münster und Mannheim. Seit 1993 Professor für Statistik und Quantitative Methoden am Institut für Psychologie Münster.

PD Dr. Günther Gediga, geb. 1953. 1971–1979 Studium der Informatik und Mathematik in Dortmund und Osnabrück. 1986 Promotion in Psychologie. 1994 Habilitation. 1979–2000 wissenschaftlicher Mitarbeiter im Fachgebiet Methodenlehre an der Universität Osnabrück. Seit 2001 außerplanmäßiger Professor am Department of Computer Science an der Brock University in St. Catherines, Kanada und freiberufliche Tätigkeit als Statistical Consultant. Seit 2008 akademischer Rat an der Universität Münster im Institut für Statistik und Methoden.

Informationen und Zusatzmaterialien zu diesem Buch finden Sie unter
www.hogrefe.de/buecher/lehrbuecher/psychlehrbuchplus

Hogrefe Verlag GmbH & Co. KG
Merkelstraße 3
37085 Göttingen
Tel.: +49 551 999 50 0
Fax: +49 551 999 50 111
E-Mail: verlag@hogrefe.de
Internet: www.hogrefe.de

Umschlagabbildung: © Marc Fischer – istockphoto.com
Format: PDF

1. Auflage 2016

(E-Book-ISBN [PDF] 978-3-8409-2302-9)
ISBN 978-3-8017-2302-6
http://doi.org/10.1026/02302-000

Nutzungsbedingungen:

Der Erwerber erhält ein einfaches und nicht übertragbares Nutzungsrecht, das ihn zum privaten Gebrauch des E-Books und all der dazugehörigen Dateien berechtigt.

Der Inhalt dieses E-Books darf von dem Kunden vorbehaltlich abweichender zwingender gesetzlicher Regeln weder inhaltlich noch redaktionell verändert werden. Insbesondere darf er Urheberrechtsvermerke, Markenzeichen, digitale Wasserzeichen und andere Rechtsvorbehalte im abgerufenen Inhalt nicht entfernen.

Der Nutzer ist nicht berechtigt, das E-Book – auch nicht auszugsweise – anderen Personen zugänglich zu machen, insbesondere es weiterzuleiten, zu verleihen oder zu vermieten.

Das entgeltliche oder unentgeltliche Einstellen des E-Books ins Internet oder in andere Netzwerke, der Weiterverkauf und/oder jede Art der Nutzung zu kommerziellen Zwecken sind nicht zulässig.

Das Anfertigen von Vervielfältigungen, das Ausdrucken oder Speichern auf anderen Wiedergabegeräten ist nur für den persönlichen Gebrauch gestattet. Dritten darf dadurch kein Zugang ermöglicht werden.

Die Übernahme des gesamten E-Books in eine eigene Print- und/oder Online-Publikation ist nicht gestattet. Die Inhalte des E-Books dürfen nur zu privaten Zwecken und nur auszugsweise kopiert werden.

Diese Bestimmungen gelten gegebenenfalls auch für zum E-Book gehörende Audiodateien.

Anmerkung:

Sofern der Printausgabe eine CD-ROM beigefügt ist, sind die Materialien/Arbeitsblätter, die sich darauf befinden, bereits Bestandteil dieses E-Books.

Inhaltsverzeichnis

Kapitel 1
Über dieses Buch

Inhaltsübersicht

1.1 Zum Inhalt dieses Buches

Das vorliegende Buch ist der dritte Teil eines insgesamt dreibändigen Lehrbuchs der Statistik. Band 1 behandelt die deskriptive Statistik, Band 2 die Wahrscheinlichkeitstheorie und Parameterschätzung und dieser Band statistische Verfahren zur Hypothesentestung. Die Lektüre dieses Bandes setzt nicht unbedingt die Lektüre der vorhergehenden Bände voraus. Leserinnen und Leser[1] können diesen Band unabhängig von den beiden anderen Bänden rezipieren, wenn sie über grundlegende Kenntnisse der deskriptiven Statistik, Wahrscheinlichkeitstheorie und Parameterschätzung verfügen.

Wie bei den beiden anderen Bänden ist es unser grundlegendes Ziel, die Statistik umfassend und verständlich darzustellen und dabei den Sinn der hier behandelten Hypothesentestung und einzelnen Testverfahren aufzuzeigen. Dazu werden die einzelnen Themen anschaulich und fundiert vermittelt und stets wird der Bezug zu inhaltlichen Problemen hergestellt.

Da Studierende der Psychologie eine wesentliche Zielgruppe dieses Buches bilden, stammen viele Beispiele aus diesem Bereich. Nichtsdestotrotz ist dieser Band ebenso für Studierende anderer Studiengänge aus den Natur-, Sozial- und Erziehungswissenschaften geeignet.

In diesem Band erfolgt zunächst eine ausführliche Darstellung der Hypothesentestung. Sie bildet die Voraussetzung für die dann behandelten statistischen Testverfahren. Neben den wichtigsten Einstichprobentests, Zweistichprobentests und Anpassungstests werden die ein- und zweifaktorielle Varianzanalyse mit sowie ohne Messwiederholung behandelt. Im letzten Kapitel werden Tests für das allgemeine lineare Modell vorgestellt, ebenso wie die Einbettung varianzanalytischer Verfahren in diese statistische Familie. Neben den häufig eingesetzten parametrischen Verfahren werden auch die entsprechenden nichtparametrischen Verfahren behandelt. Dabei wird viel Wert darauf gelegt, auch die den nichtparametrischen Verfahren zugrunde liegende Logik anschaulich zu erläutern. Statistische Analysen beginnen immer vor der Datenerhebung, d. h. sind bereits bei der Planung empirischer Studien von hoher Relevanz. Dementsprechend wird der Power statistischer Verfahren in diesem Band viel Beachtung geschenkt.

[1] Wir werden jedoch im Folgenden immer dann, wenn beide Geschlechter gemeint sind, lediglich die männliche Form benutzen. Damit wird u. E. die Darstellung sehr vereinfacht. Wenn wir explizit die weibliche Form nutzen, sind nur Personen weiblichen Geschlechts gemeint.

Ein Lehrbuch der Statistik ist einfacher zu lesen, wenn es einheitlich und übersichtlich strukturiert ist. Dazu haben wir analog zum ersten und zweiten Band dieser Lehrbuchreihe die folgenden Gestaltungsrichtlinien angewendet.

Um die Darstellung der statistischen Konzepte nicht zu überfrachten, wird auf die Darstellung mathematischer Ableitungen und Beweise, die für das grundlegende Verständnis nicht unbedingt erforderlich sind, verzichtet. Für besonders interessierte Leser werden diese Beweise auf der folgenden Website zur Verfügung gestellt:

http://www.hogrefe.de/buecher/lehrbuecher/psychlehrbuchplus

Im Text wird jeweils auf diese Website hingewiesen, zusätzlich werden die entsprechenden Stellen durch das „Internetsymbol“ am Seitenrand gekennzeichnet.

Um die Inhalte dieses Lehrbuchs übersichtlich darzustellen, verwenden wir im Text verschiedene Arten von Kästen, die Beispiele, Anwendungen von Software oder Zusammenfassungen enthalten. Wichtige Formeln werden ebenfalls durch eine farbige Hinterlegung hervorgehoben, damit sie unmittelbar als solche zu erkennen sind. Bei der Einführung neuer Begriffe werden diese Begriffe im Text kursiv gesetzt und zudem am Seitenrand aufgeführt.

Die Vermittlung von Statistik ist ohne die Verwendung von EDV-Programmen undenkbar und es gibt mittlerweile eine Vielzahl sehr leistungsfähiger Statistikprogrammpakete. Im vorliegenden Lehrbuch nutzen wir die Softwarepakete SPSS und R.

Das Programmpaket IBM SPSS Statistics, kurz SPSS genannt, ist eine seit jeher sehr weit verbreitete Statistiksoftware. In vielen Organisationen, in denen Natur-, Sozial- und Erziehungswissenschaftler beschäftigt sind, wird dieses Programmpaket verwendet. SPSS hat allerdings den Nachteil, dass es kostenpflichtig ist.

Das Open-Source-Programm R ist ebenfalls weit verbreitet, hingegen kostenlos verfügbar. R ist allerdings nicht ganz so benutzerfreundlich wie SPSS. Jedoch bietet auch R eine Menüsteuerung mittels des sogenannten R-Commanders an, der für Anfänger eine einfache Benutzung der wesentlichen Funktionen erlaubt. Ein großer Vorteil von R besteht darin, dass neueste statistische Verfahren schnell zur Verfügung gestellt werden.

Im vorliegenden Buch behandeln wir am Ende der Kapitel jeweils die notwendigen Schritte zur Datenanalyse mittels SPSS und R. Studierende, die diese Programmsysteme nicht kennen, finden auf der oben genannten Website Hinweise auf eine Einführung in R sowie Anleitungen

zur Benutzung von R. Für die Einführung in die Arbeit mit SPSS möchten wir auf das Buch von Leonhart (2010) verweisen.

Auf der Website befinden sich neben den oben bereits angesprochenen Inhalten viele weitere Ressourcen zu diesem Lehrbuch. Dabei handelt es sich um die in diesem Buch verwendeten Datensätze und die entsprechenden Kommandos für die Programme SPSS und R. Des Weiteren enthält die Website ergänzende Inhalte und Literaturhinweise.

1.2 Danksagung

Zur Entstehung dieses Buches haben neben den Autoren zahlreiche weitere Personen beigetragen. Viele wertvolle Kommentare erhielten wir nach dem Studium einer vorläufigen Endversion von Frau Corinna Brauner (B.Sc. Psych), Herrn Jonatan Buhl (B.Sc. Psych), Herrn Paul-Christian Bürkner (M.Sc. Psych, B.Sc. Math), Herrn Dr. Philipp Doebler, Frau Meltem Dogan (B.Sc. Psych), Herrn Dipl.-Psych. Boris Forthmann, Frau Anne Gerwig (B.Sc. Psych), Frau Katharina Hoferichter, Herrn Ruben Kleinkorres (B.Sc. Psych), Frau Kathrin Klute (B.Sc. Psych), Herrn Dr. Jörg-Tobias Kuhn, Frau Julia Raddatz (M.Sc. Psych), Frau Marianna Rusche (B.Sc. Psych) und Frau Jana Scharfen (M.Sc. Psych). Dafür möchten wir ihnen an dieser Stelle unseren großen Dank aussprechen. Ein ganz besonderer Dank gebührt Herrn Dipl.-Math. Carsten Szardenings, Frau Dr. Anna Doebler, Frau Kathrin Gediga (M.Sc. Eval) und Frau Christin Schwenk (M.Sc. Psych). Ihre sorgfältige Überprüfung des gesamten Textes in formaler wie inhaltlicher Hinsicht war sehr hilfreich.

Weiterhin gaben uns Herr Prof. Dr. Hans-Werner Bierhoff und Herr Prof. Dr. Franz Petermann wertvolle Hinweise für die Erstellung dieses Buches. Dafür sei ihnen hier ganz herzlich gedankt. Schließlich gilt unser Dank Frau Dipl.-Psych. Susanne Weidinger, Herrn Dr. Michael Vogtmeier und Herrn Stefan Reins M. A. vom Hogrefe Verlag. Sie haben uns wie gewohnt sehr kompetent betreut.

Kapitel 2
Hypothesentestung

Inhaltsübersicht

2.1 Einleitende Übersicht

Gegenstand dieses letzten Bandes des dreiteiligen Statistiklehrbuches ist die Hypothesentestung, die häufig als wichtigste Aufgabe der Statistik angesehen wird. Im ersten Band (Holling & Gediga, 2011) wurde die

deskriptive Statistik

deskriptive Statistik behandelt. In diesem Teilgebiet der Statistik geht es um die übersichtliche Darstellung der in empirischen Studien erhobenen Daten. Diese Daten werden so gut wie immer an Stichproben erhoben, da die Untersuchung gesamter Populationen zumeist nicht möglich ist oder nicht ökonomisch wäre. Eine weitere wichtige Thematik der deskriptiven Statistik ist die zusammenfassende Beschreibung der Daten mittels Kennwerten, z. B. Mittelwert oder Varianz, die auch Stichprobenstatistiken oder kurz Statistiken genannt werden.

Im zweiten Band (Holling & Gediga, 2013) wurde zunächst die Wahrscheinlichkeitsrechnung behandelt, die die Grundlage für die anschließend dargestellte Punkt- und Intervallschätzung von Parametern bildet. Die Parameterschätzung bildet den ersten von zwei wesentlichen Bereichen der *Inferenzstatistik*, die zur Generalisierung der Ergebnisse aus Stichproben auf die Population dient.

Inferenzstatistik

Parameterschätzung

Bei der *Parameterschätzung* geht es um die Schätzung von Kennwerten der Verteilung von Merkmalen in der Population anhand von Stichprobenstatistiken. So gilt es etwa, in der PISA-Studie für ein bestimmtes Land, z. B. Deutschland, den Erwartungswert der mathematischen Kompetenz zu schätzen. Der Erwartungswert stellt dann den Parameter dar. Eine geeignete Stichprobenstatistik für die Punktschätzung dieses Parameters ist das arithmetische Mittel. Eine solche Punktschätzung ist mit einer gewissen Unsicherheit behaftet. Daher werden sogenannte Konfidenzintervalle berechnet, die mit einer bestimmten, vorher festgelegten Wahrscheinlichkeit einen Parameter enthalten. Punkt- und Intervallschätzungen wurden insbesondere für die folgenden Parameter vorgestellt: Erwartungswert, Varianz bzw. Standardabweichung, Produkt-Moment-Korrelationskoeffizient und Regressionskoeffizienten in linearen Modellen. Diese Parameter sind auch Gegenstand der Hypothesentestung, dem zweiten großen Bereich der Inferenzstatistik, der eng verwandt ist mit der Parameterschätzung.

Hypothesentestung

Bei der *Hypothesentestung* geht es um die Überprüfung von vorher aufgestellten Hypothesen, die sich auf die gesamte Population und nicht auf Stichproben beziehen. So mag eine Hypothese lauten, dass der Erwartungswert der mittleren mathematischen Kompetenz für finnische Schüler höher ist als für die Schüler aller an der PISA-Studie teilnehmenden Staaten. Statistische Testverfahren erlauben es dann, eine Entscheidung über solche Hypothesen zu treffen. Aufgrund einer solchen Ent-

scheidung kann die Gültigkeit einer Hypothese angenommen bzw. nicht angenommen werden. Auch für die Hypothesentestung werden Stichprobenstatistiken verwendet, wie für das obige Beispiel der Mittelwert der mathematischen Kompetenz erhoben an einer Stichprobe finnischer Schüler.

Wir werden in diesem Band die wichtigsten Testverfahren zur Hypothesentestung vorstellen. Dabei geht es zumeist um Hypothesen zu Parametern von Wahrscheinlichkeitsverteilungen. Denn in den meisten Fällen der Hypothesentestung werden bestimmte Wahrscheinlichkeitsverteilungen für die Beschreibung der interessierenden Merkmale vorausgesetzt, z. B. die Normalverteilung für Merkmale, wie Intelligenz, Extraversion etc. Die Hypothesen beziehen sich dann auf bestimmte Parameter dieser Wahrscheinlichkeitsverteilungen, z. B. den Erwartungswert oder die Varianz. Es gilt aber, auch allgemeinere Hypothesen zu überprüfen, insbesondere zur Frage, ob in einer Population eine bestimmte Wahrscheinlichkeitsverteilung für ein interessierendes Merkmal vorliegt. Ein Beispiel für eine solche Hypothese ist: „Die Anzahl von Verkehrsunfällen in Deutschland folgt einer Poisson-Verteilung."

Für eine möglichst einfache Einführung in dieses Gebiet beschränken wir uns zunächst lediglich auf Hypothesen zum Erwartungswert einer normalverteilten Zufallsvariablen mit bekannter Varianz.

2.2 Einführung in die Hypothesentestung

Betrachten wir nun die Grundlagen der Hypothesentestung anknüpfend an die Ausführungen zur Parameterschätzung in Band 2 (Holling & Gediga, 2013). Wir gehen hier zunächst von einem einfachen Fall aus, der Schätzung des Erwartungswertes bei normalverteilten Merkmalen mit bekannter Varianz. Zur Schätzung dieses Parameters ist, wie in Band 2 ausgeführt wurde, das arithmetische Mittel eine optimale Stichprobenstatistik. Während die Stichprobenstatistiken im Kontext der Parameterschätzung *Schätzer* genannt werden, heißen sie im Kontext der Hypothesentestung *Teststatistik* oder auch *Prüfgröße*.

Schätzer
Teststatistik
Prüfgröße

Betrachten wir als Beispiel die Schätzung des Erwartungswertes der Scores eines Tests für die mathematische Kompetenz, wobei alle Schüler einer bestimmten Altersklasse in Deutschland als Population fungieren. Der Test sei für diese Population so normiert, dass diese Scores die Realisierung einer normalverteilten Zufallsvariablen Y darstellen mit $Y \sim N(100, 10)$. Aus der Population all dieser Schüler ziehen wir dann n unabhängige Stichprobenelemente. Die zugehörigen Stichprobenvariablen $Y_1, \ldots, Y_n$ sind wie die Zufallsvariable Y verteilt und zu-

i.i.d.-Annahme

dem unabhängig. Es gilt also die *i.i.d.-Annahme* (i.i.d = independent and identically distributed). Anhand dieser Stichprobenvariablen wird die Stichprobenstatistik bzw. Statistik gebildet. Als Statistik zur Schätzung des Erwartungswertes der normalverteilten Scores zur mathematischen Kompetenz dient das arithmetische Mittel der n Stichprobenelemente $\bar{Y}$, das ebenfalls normalverteilt ist mit der Stichprobenverteilung $N(\mu, \sigma/\sqrt{n})$. $\bar{Y}$ dient als Punktschätzung für μ. Aufgrund der geringeren Standardabweichung von $\bar{Y}$ sind die Mittelwerte, die sich durch wiederholtes Ziehen von Stichproben des Umfangs n ergeben, enger um den zu schätzenden Parameter verteilt als die Werte der Variablen Y und damit zur Schätzung des unbekannten Parameters geeignet.

In Abbildung 2.1 wird die Ziehung von Stichproben aus einer normalverteilten Population veranschaulicht, sowie die Bildung der Stichprobenverteilung für den Mittelwert. Hier werden (vgl. oberer Teil der Abb. 2.1) Stichproben mit einem Umfang von $n = 3$ gezogen. Zieht man nun alle (theoretisch) möglichen Stichproben mit einem Umfang von $n = 3$ und berechnet jeweils den Mittelwert, resultiert die im unteren Teil der Abbildung dargestellte Dichte der normalverteilten Stichprobenstatistik $\bar{Y}$, wobei $\bar{Y} \sim N(\mu, \sigma/\sqrt{3})$. Das symmetrische Intervall um den wahren Wert μ, in dem $(1-\alpha)\cdot 100\,\%$ aller Punktschätzungen liegen, wird auch zentrales $(1-\alpha)$-Schwankungsintervall für Mittelwerte genannt.

Stichprobenstatistiken bilden auch den Ausgangspunkt für die Berechnung von Konfidenzintervallen. Bei Konfidenzintervallen für den Erwartungswert normalverteilter Zufallsvariablen mit bekannter Varianz wird wiederum der Schätzer $\bar{Y}$ verwendet (vgl. Abb. 2.2). Bei zweiseitigen Konfidenzintervallen, die wir hier lediglich betrachten, verteilen sich die Punktschätzungen symmetrisch um den wahren Wert μ. Ausgehend von $\bar{Y}$ bildet man nun ein Konfidenzintervall, indem man symmetrisch um $\bar{Y}$ ein Intervall legt. Die Länge dieses Intervalls entspricht der Länge des entsprechenden Schwankungsintervalls. Damit sind die untere und die obere Grenze des Konfidenzintervalls so festgelegt, dass es in $(1-\alpha)\cdot 100\,\%$ aller möglichen Stichprobenziehungen den wahren Parameter μ überdeckt.

Als Konfidenzintervall resultiert im vorliegenden Fall:

$$\left[\overline{Y} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{Y} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right],$$

wobei $z_{1-\alpha/2}$ das entsprechende Quantil der Standardnormalverteilung darstellt (vgl. Anhang zur Berechnung der Quantile).

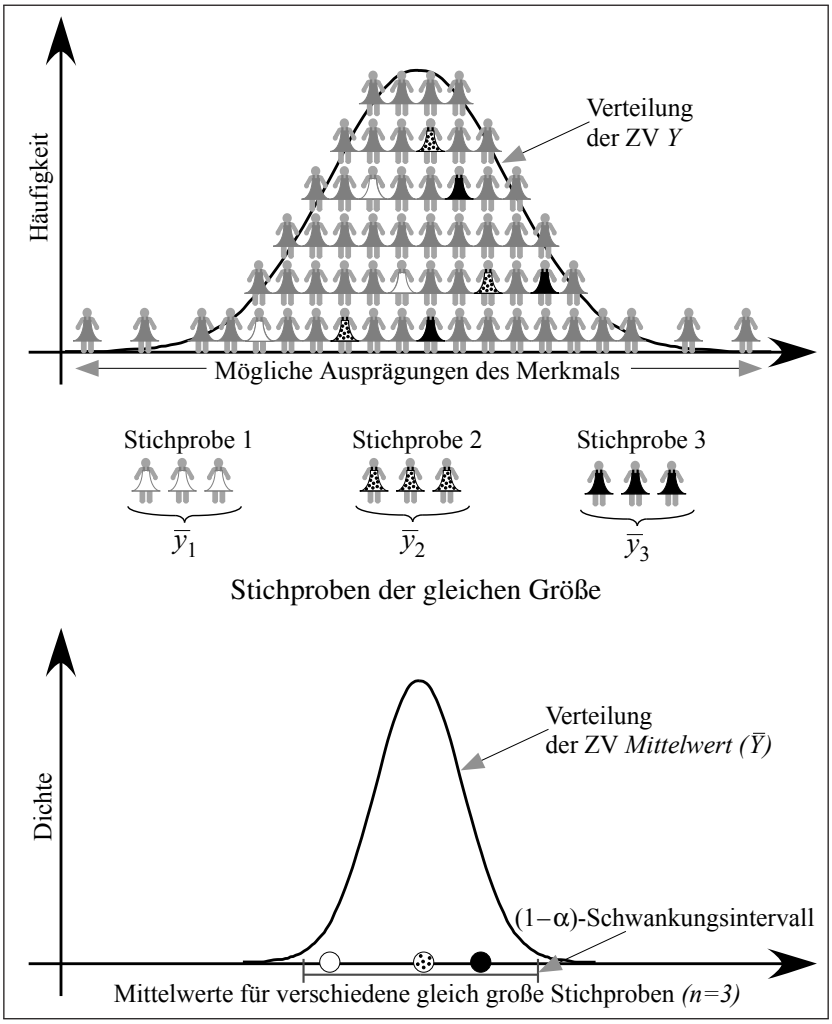

Abbildung 2.1: Veranschaulichung der Stichprobenverteilung des Mittelwertes. Die verschiedenen Stichproben der Größe $n = 3$ sind farbig markiert (ZV=Zufallsvariable).

Während es sich bei Schätzern um *eine* Zufallsvariable handelt, werden zweiseitige Konfidenzintervalle aus *zwei* Zufallsvariablen, d. h. Stichprobenstatistiken, gebildet, die die Grenzen des Konfidenzintervalls beschreiben. Damit handelt es sich bei Konfidenzintervallen um zufällige Intervalle, d. h. um Intervalle, deren Grenzen aus Zufallsvariablen bestehen.

Analog zur Unterscheidung von Schätzern als Zufallsvariablen und den Realisierungen der Schätzer, den Schätzungen, müssen wir auch hier zwischen Konfidenzintervallen als zufällige Intervalle und ihren Realisierungen, den sogenannten realisierten Konfidenzintervallen, unter-

scheiden. Bei einem realisierten Konfidenzintervall verwendet man die Realisierungen der Schätzer als Grenzen. Die Grenzen eines realisierten Konfidenzintervalls bestehen also nicht aus Zufallsvariablen, sondern aus Realisierungen von Zufallsvariablen. Die Unterscheidung zwischen Konfidenzintervallen und realisierten Konfidenzintervallen wird oft nicht getroffen, womit sich eine gewisse begriffliche Unschärfe ergibt, die aber im Kontext einer konkreten Anwendung kein großes Problem darstellen dürfte.

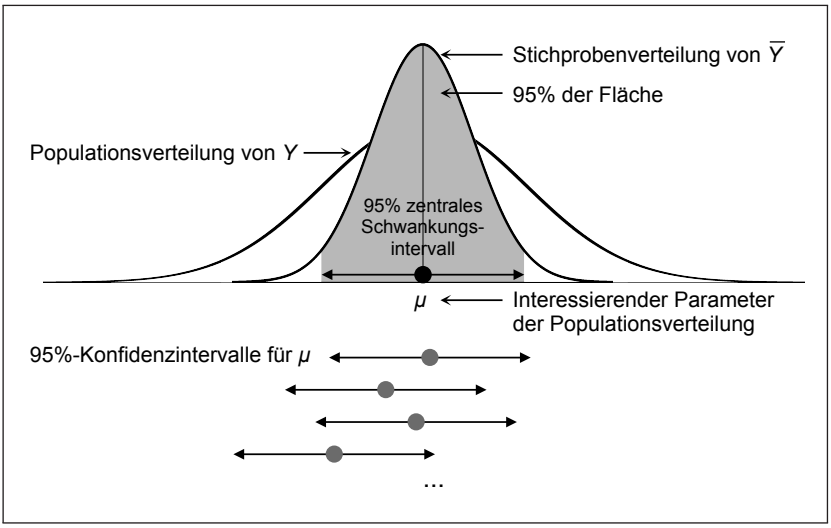

Abbildung 2.2: Veranschaulichung von Konfidenzintervallen für normalverteilte Zufallsvariablen mit μ unbekannt, σ^2 bekannt: $\bar{Y} \pm z_{.975}\frac{\sigma}{\sqrt{n}}$

Abbildung 2.2 zeigt einige realisierte Konfidenzintervalle für den Erwartungswert einer Normalverteilung mit bekannter Varianz σ^2, die auf unterschiedlichen Stichproben mit identischem Stichprobenumfang n beruhen. Dabei beträgt die vorher festgelegte Wahrscheinlichkeit, dass der Parameter von dem Konfidenzintervall überdeckt wird: $1-\alpha = .95$. Wie Abbildung 2.2 zeigt, überdeckt ein $(1-\alpha)$-Konfidenzintervall, das anhand der Daten einer Stichprobe berechnet wurde, den wahren (unbekannten) Wert des Parameters immer dann, wenn die entsprechende Punktschätzung innerhalb des $(1-\alpha)$-Schwankungsintervalls liegt.

Es gibt eine enge Beziehung zwischen der Hypothesentestung und der Intervallschätzung, auf die wir nun eingehen wollen.

Gehen wir von einem (fiktiven) Problem aus, in dem es ebenfalls um den Erwartungswert einer Normalverteilung mit bekannter Varianz geht. Ausgangspunkt sei die von Bildungsforschern geäußerte *inhaltliche Hypothese*, dass die mittlere mathematische Kompetenz der Schüler in

inhaltliche Hypothese

Nordrhein-Westfalen (NRW) nicht mit der mittleren mathematischen Kompetenz aller Schüler in Deutschland übereinstimmt. Es wird dabei nicht gesagt, ob die mittlere Kompetenz der Schüler in NRW nach oben oder unten abweicht.

Untersucht man diese Hypothese in einer empirischen Studie anhand eines normierten Tests, dessen Scores Y für alle deutschen Schüler normalverteilt sind mit $Y \sim N(100, 10)$, so kann die obige Hypothese weiter spezifiziert werden. Es gilt nun, dass der Erwartungswert der mathematischen Kompetenz für die Schüler aus NRW nicht gleich 100 ist. Dabei sei bekannt, dass die Testwerte auch für die Schüler des Landes NRW normalverteilt sind mit einer Standardabweichung von 10. Lediglich der Erwartungswert ist unbekannt und wird entsprechend der Hypothese als ungleich 100 angenommen. Diese nachzuweisende Hypothese wird *Alternativhypothese* genannt und im Allgemeinen mit H_1 bezeichnet.

Alternativhypothese

H_1

Wie kann man nun zu einer Entscheidung über die Hypothese gelangen, dass der Erwartungswert der mathematischen Kompetenz in NRW ungleich 100 ist? Dazu könnte man die mittlere mathematische Kompetenz aus der empirischen Studie heranziehen. Würde z. B. auf der Basis einer Stichprobe mit dem Umfang $n = 100$ ein arithmetisches Mittel von 100.2 oder 99.7 resultieren, wäre der Alternativhypothese wohl nicht zuzustimmen. Anders sähe es aus, wenn ein arithmetisches Mittel von 91 oder 112 auftreten würde. Wann ist also ein bestimmter Abstand von Punkten nach oben oder unten groß genug, um der Hypothese, dass der Erwartungswert in NRW ungleich 100 ist, beizupflichten?

Ausgangspunkt der weiteren Überlegungen seien nun die Ergebnisse einer empirischen Untersuchung an einer repräsentativen Stichprobe von 25 Schülern aus NRW. Der erzielte Mittelwert für die mathematische Kompetenz sei $\bar{y} = 105$. Der Wert von 105 liegt 5 Punkte von dem Erwartungswert 100 entfernt, der für alle deutschen Schüler gilt. Ist nun ein solcher Abstand von 5 Punkten nach unten oder oben groß genug, um davon auszugehen, dass in dem Bundesland NRW ein anderer Erwartungswert gegeben ist? Zur Bewertung der Größe dieses Abstands verwenden wir seine Wahrscheinlichkeitsverteilung.

Unterscheiden sich die Erwartungswerte der mathematischen Kompetenz in NRW und ganz Deutschland nicht, gilt auch für NRW ein Erwartungswert von 100. Diese Annahme, die das Gegenteil der Alternativhypothese darstellt, wird üblicherweise *Nullhypothese* genannt und mit H_0 bezeichnet. Um die Gültigkeit der Alternativhypothese nachzuweisen, gilt es, die Nullhypothese zu widerlegen. Den Wert des Parameters, der unter der Nullhypothese angenommen wird, bezeichnet man mit μ_0 und es gilt hier $\mu_0 = 100$.

Nullhypothese

H_0

Ausgehend von dem Erwartungswert $\mu_0 = 100$ und der bekannten Standardabweichung $\sigma = 10$ bestimmen wir die Stichprobenverteilung des arithmetischen Mittels $\bar{Y}$ für einen Stichprobenumfang von $n = 25$. Für ein normalverteiltes Merkmal mit $Y \sim N(100, 10)$ gilt dann $\bar{Y} \sim N(100, 10/\sqrt{25}) = N(100, 2)$. Aufgrund dieser Stichprobenverteilung ist es recht unwahrscheinlich, einen Mittelwert zu erzielen, der 5 oder mehr Punkte von dem Wert 100 abweicht. Der Abstand von zwei Punkten für $\bar{Y}$ entspricht einer Standardabweichung, und somit entspricht ein Abstand von 5 Punkten 2.5 Standardabweichungen. Dass eine Beobachtung 2.5 oder mehr Standardabweichungen vom Erwartungswert einer normalverteilten Variablen abweicht, entspricht der Wahrscheinlichkeit $P(-2.5 < Z < 2.5) = .012$, wobei Z eine standardnormalverteilte Zufallsvariable darstellt. Somit würde man nur für 1.2% aller möglichen Stichproben von Schülern aus NRW eine Abweichung des Mittelwertes um 5 oder mehr Punkte erhalten, wenn der wahre Erwartungswert der mathematischen Kompetenz $\mu = 100$ wäre. Damit erscheint es plausibler, von einem anderen Erwartungswert auszugehen, der nicht zu einer solch geringen Wahrscheinlichkeit führt, und insofern der Alternativhypothese zuzustimmen, dass μ ungleich 100 ist.

Abbildung 2.3 veranschaulicht die Hypothesentestung für das obige Beispiel. Dargestellt sind im oberen Teil dieser Abbildung die Verteilung der mathematischen Kompetenz Y unter der Annahme $Y \sim N(100, 10)$ und die daraus folgende Stichprobenverteilung des Mittelwertes $\bar{Y}$ für $n = 25$ mit $\bar{Y} \sim N(100, 2)$.

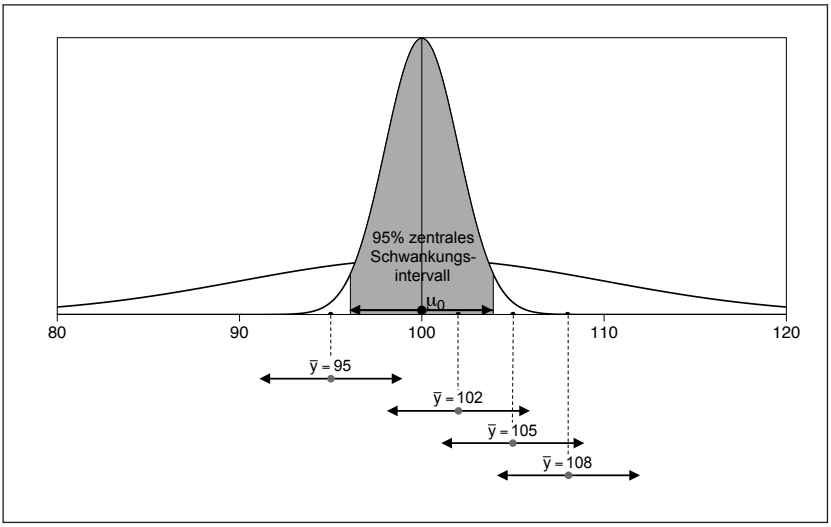

Abbildung 2.3: Konfidenzintervalle und Hypothesentestung: Veranschaulichung

Weiterhin sind im unteren Teil der Abbildung 2.3 einige Schätzungen $\bar{y}$ als Punkte dargestellt. Diese Mittelwerte stammen aus der wahren, aber unbekannten Stichprobenverteilung, die nicht mit der angenommenen Stichprobenverteilung mit $\mu_0 = 100$ übereinstimmen muss. Die wahre, hier nicht dargestellte Stichprobenverteilung besitzt ebenfalls eine Normalverteilung mit $\sigma = 10$ und einem unbekannten Erwartungswert. Ausgehend von der Nullhypothese wird angenommen, dass der wahre Erwartungswert mit dem angenommenen Wert $\mu_0 = 100$ übereinstimmt.

Liegt nun ein Mittelwert $\bar{y}$ zu weit weg von dem angenommenen Erwartungswert $\mu_0 = 100$, ist es wenig plausibel, von einem Erwartungswert gleich 100 auszugehen. So ist – ausgehend von der Nullhypothese $\mu_0 = 100$ – auch ein Mittelwert $\bar{y} = 105$ als sehr extrem zu werten. Wenn ein solcher Mittelwert in einer empirischen Studie erzielt wird, spricht dieses Ergebnis gegen die Hypothese, dass der Erwartungswert $\mu_0 = 100$ für die mathematische Kompetenz der Schüler in NRW zutrifft.

Nun stellt sich die Frage, ab wann allgemein betrachtet ein Mittelwert so extrem ist, dass die Nullhypothese nicht mehr zu halten ist. Dazu werden ein unterer und oberer kritischer Wert, k_u und k_o, festgelegt. Liegt ein Mittelwert jenseits eines kritischen Wertes, d. h. unterhalb von k_u oder oberhalb von k_o, gilt er als extrem. Dann wird die Nullhypothese abgelehnt und die Alternativhypothese angenommen.

Die Menge aller Mittelwerte jenseits der kritischen Werte, die den beiden Intervallen $(-\infty, k_u)$ und (k_o, ∞) in Abbildung 2.3 entspricht, bildet den Ablehnungsbereich für die Nullhypothese, das Intervall $[k_u, k_o]$ entspricht dem Annahmebereich. Fällt der Mittelwert einer bestimmten Stichprobe in den Ablehnungsbereich, wird die Nullhypothese verworfen, ansonsten beibehalten. Für unser Beispiel haben wir für die kritischen Werte $k_u = 96.08$ und $k_o = 103.92$ festgelegt. Diese Werte wurden so gewählt, dass ausgehend von der Nullhypothese $\mu_0 = 100$ die Wahrscheinlichkeit dafür, dass eine Schätzung in den Ablehnungsbereich fällt, 5% beträgt. Der Wert 5%, der definiert, welche Mittelwerte als extrem gelten und somit zur Ablehnung der Nullhypothese führen, heißt Signifikanzniveau und wird mit α bezeichnet. Anstelle von $\alpha = .05$ hätten wir auch einen anderen Wert, z. B. $\alpha = .01$, wählen können. Dann müsste ein Mittelwert noch extremer ausfallen, damit die Alternativhypothese angenommen wird.

Der Annahmebereich für die Nullhypothese korrespondiert mit dem $(1 - \alpha)$-Schwankungsintervall für den wahren Wert des Parameters. Der Mittelpunkt des Schwankungsintervalls ist der wahre Wert des Parameters μ, beim Annahmebereich ist es der unter der Nullhypothese angenom-

mene Wert μ_0. Damit wird die enge Beziehung zwischen den Konfidenzintervallen und der Hypothesentestung offenbar. Im unteren Teil der Abbildung 2.3 sind für die Schätzungen $\bar{y}$ die zweiseitigen 95%-Konfidenzintervalle dargestellt. Aufgrund des Konstruktionsprinzips der Konfidenzintervalle wird der unter der Nullhypothese angenommene Erwartungswert $\mu_0 = 100$ immer dann vom Konfidenzintervall überdeckt, wenn die entsprechende Schätzung im Annahmebereich des Tests liegt. Liegt eine Schätzung im Ablehnungsbereich, überdeckt das Konfidenzintervall nicht den unter der Nullhypothese angenommenen Wert des Parameters. Dann ist es auf der Grundlage der gezogenen Stichprobe nicht plausibel, dass der wahre Wert des Parameters mit dem unter der Nullhypothese angenommenen Wert $\mu_0 = 100$ übereinstimmt.

Gemäß der oben dargestellten Zusammenhänge führen Konfidenzintervalle und Testverfahren in der Regel zu äquivalenten Ergebnissen. In diesen Fällen werden zumeist die Ergebnisse der Testverfahren im Rahmen der Hypothesentestung berichtet.

Das oben eingeführte Konzept der Hypothesentestung, das es nun systematisch zu definieren und zu erweitern gilt, geht auf die testtheoretischen Ansätze von Fisher (1925) sowie Neyman und Pearson (1933) aus den 1920er und 1930er Jahren zurück. Dabei werden Konzepte aus beiden Ansätzen verwendet.

2.3 Statistische Tests

Betrachten wir nun detaillierter die einzelnen Schritte der Hypothesentestung angelehnt an das obige Beispiel. Ausgangspunkt ist eine inhaltliche Hypothese: „Die mathematische Kompetenz der Schüler in NRW unterscheidet sich von der aller Schüler in Deutschland“. Diese sehr allgemeine Hypothese gilt es zunächst zu spezifizieren. Dazu ist u. a. das Konstrukt *mathematische Kompetenz* zu operationalisieren, so dass es gemessen werden kann. Dann sind Annahmen über die Verteilung dieses so spezifizierten und hier mit Y bezeichneten Merkmals in der Population festzulegen, z. B. $Y \sim N(\mu, 10)$.

statistische Hypothese

Nun kann die entsprechende *statistische Hypothese* bestimmt werden. Statistische Hypothesen richten sich auf Wahrscheinlichkeitsverteilungen. Zumeist werden bestimmte (Familien von) Wahrscheinlichkeitsverteilungen vorausgesetzt, im obigen Beispiel waren es Normalverteilungen mit der Standardabweichung $\sigma = 10$. Die statistischen Hypothesen richten sich dann auf die interessierenden Parameter, wie im vorliegenden Beispiel auf den Erwartungswert μ. Bei einem statistischen Test unterscheidet man zwischen der *Nullhypothese* und der *Alternativhypo-*

Alternativhypothese

these, die durch H_0 bzw. H_1 symbolisiert werden. Statt „Nullhypothese“ sagt man auch kurz „Hypothese“ und statt „Alternativhypothese“ kurz „Alternative“. H_1

Null- und Alternativhypothese schließen sich wechselseitig aus, d. h., diese beiden Hypothesen können nicht gleichzeitig gelten. Sie unterteilen die Menge aller möglichen in Frage kommenden Parameter in zwei disjunkte Mengen, die vereint wiederum die gesamte Menge der Parameter ergeben. Dabei kann die gesamte Menge der Parameter lediglich aus zwei Werten, einer Teilmenge der reellen Zahlen oder der gesamten Menge der reellen Zahlen bestehen. Im obigen Beispiel handelte es sich um die Gesamtheit der reellen Zahlen, unter der Nullhypothese wurde ein Wert von 100 und unter der Alternativhypothese ein Wert ungleich 100 angenommen. Bei Testproblemen heißt es, dass man die Nullhypothese gegen die Alternativhypothese testet.

Beinhaltet eine Hypothese H lediglich einen Wert des Parameters μ_0, d. h. $H : \mu = \mu_0$, dann spricht man von einer einfachen Hypothese. Lässt eine Hypothese mehr als einen Wert zu, spricht man von einer zusammengesetzten Hypothese. Damit können die Nullhypothese wie die Alternativhypothese jeweils eine einfache bzw. zusammengesetzte Hypothese darstellen.

Eine weitere Unterscheidung betrifft die Richtung von Hypothesen. *Gerichtete Hypothesen* sind zusammengesetzte Hypothesen, die einen Bereich entweder mit der Größer- oder mit der Kleiner-Relation beschreiben. Beispiele sind die Hypothesen $H : \mu > 100$ oder $H : \mu \leq 100$. *Ungerichtete Hypothesen* sind hingegen zusammengesetzte Hypothesen, die einen Bereich mit der Ungleich-Relation beschreiben. Das wichtigste Beispiel für eine ungerichtete Hypothese ist die Hypothese der Form $H : \mu \neq 100$, die alternativ auch mit $H : \mu < 100$ oder $\mu > 100$ beschrieben werden kann.

gerichtete Hypothesen

ungerichtete Hypothesen

Gegenstand von Studien sind beispielsweise die folgenden Kombinationen von Null- und Alternativhypothesen, auch Testprobleme genannt:

Beispiel: Testprobleme mit Null- und Alternativhypothesen

1) $H_0 : \mu = \mu_0$ gegen $H_1 : \mu \neq \mu_0$
2) $H_0 : \mu \leq \mu_0$ gegen $H_1 : \mu > \mu_0$
3) $H_0 : \mu \geq \mu_0$ gegen $H_1 : \mu < \mu_0$
4) $H_0 : \mu = \mu_0$ gegen $H_1 : \mu > \mu_0$
5) $H_0 : \mu = \mu_0$ gegen $H_1 : \mu < \mu_0$
6) $H_0 : \mu = \mu_0$ gegen $H_1 : \mu = \mu_1$

Das erste Testproblem ist ein zweiseitiges Testproblem, die Testprobleme 2 bis 5 sind gerichtete (einseitige) Testprobleme. Beim letzten Testproblem werden zwei einfache Hypothesen gegeneinander getestet. Fast immer geht es in empirischen Untersuchungen um eines der ersten drei der oben angeführten Testprobleme, da die möglichen Ausprägungen des Parameters in der Regel nicht auf eine Teilmenge eingeschränkt werden können, wie es bei den Hypothesen 4 bis 6 nötig ist. Anzumerken ist weiterhin, dass die Testprobleme 2 und 4 sowie die Testprobleme 3 und 5 jeweils hinsichtlich der Testdurchführung äquivalent sind (s. u.).

Bei der Spezifikation eines Testproblems stellt sich die Frage, ob die inhaltliche Hypothese als die Null- oder die Alternativhypothese umzusetzen ist. In der Regel entspricht die inhaltliche Hypothese der Alternativhypothese. Die Nullhypothese spezifiziert dann das logische „Gegenteil" der inhaltlichen Hypothese bzw. Alternativhypothese. Die Bezeichnung Nullhypothese geht zurück auf den englischen Begriff to nullify (= widerlegen). Die Nullhypothese ist zu widerlegen bzw. zu verwerfen, damit die Alternativhypothese angenommen werden kann. Der Grund für diese Vorgehensweise besteht darin, dass eine irrtümliche Ablehnung der Nullhypothese in der Regel größere negative Konsequenzen mit sich bringt als die irrtümliche Beibehaltung der Nullhypothese. So soll z. B. eine neue Lehrmethode nicht ungerechtfertigterweise eine bewährte Methode ersetzen.

Zusammenfassend können wir ein statistisches Testproblem folgendermaßen definieren:

Definition: Statistisches Testproblem
Ein statistisches Testproblem besteht aus einer Nullhypothese H_0 und einer Alternativhypothese H_1. Diese beiden Hypothesen richten sich auf Wahrscheinlichkeitsverteilungen oder Parameter von Wahrscheinlichkeitsverteilungen des interessierenden Merkmals in der Grundgesamtheit. H_0 und H_1 schließen sich wechselseitig aus. In der Regel werden die folgenden Testprobleme untersucht, wobei θ für einen beliebigen Parameter, z. B. μ oder σ, steht:

$$H_0 : \theta = \theta_0 \text{ gegen } H_1 : \theta \neq \theta_0$$
$$H_0 : \theta \leq \theta_0 \text{ gegen } H_1 : \theta > \theta_0$$
$$H_0 : \theta \geq \theta_0 \text{ gegen } H_1 : \theta < \theta_0$$

Das erste Testproblem ist ein ungerichtetes Testproblem, die beiden letzten Testprobleme sind gerichtet.

Ist das statistische Testproblem definiert, gilt es, eine Stichprobenstatistik bzw. Teststatistik zu wählen, anhand derer eine Entscheidung ge-

troffen werden kann, welche Hypothese anzunehmen ist. Dazu ist eine Regel zu entwickeln, für welche Werte der Teststatistik H_0 bzw. H_1 akzeptiert werden soll. Jedoch sind die Teststatistik und die Regel so zu wählen, dass eine „vernünftige“ Entscheidung getroffen wird. Entsprechende Kriterien für die Bestimmung der Teststatistik und der Entscheidungsregel werden in den folgenden Kapiteln diskutiert.

Zusammenfassend wird ein statistischer Test wie folgt definiert.

Definition: Statistischer Test
Ein statistischer Test besteht aus einer Regel, die es anhand einer Teststatistik erlaubt, eine Entscheidung über ein Testproblem zu treffen. Es gilt also, die Entscheidung zu treffen, ob H_0 beizubehalten ist oder ob H_0 abzulehnen ist und damit zugleich H_1 angenommen wird.

2.4 Signifikanztests

In diesem Kapitel wird das Konzept des Signifikanztests vorgestellt, das auf Fisher (1925) zurückgeht. Ein *Signifikanztest* ist nach Fisher ein spezieller Test, der die Kontrolle einer fälschlichen Ablehnung der H_0 gestattet. Wie oben bereits ausgeführt, wird die inhaltliche Hypothese als Alternativhypothese spezifiziert. Die Alternativhypothese wird erst dann angenommen, wenn das logische Gegenteil, die Nullhypothese, abgelehnt wird. Damit werden hier die Null- und Alternativhypothese asymmetrisch behandelt.

Signifikanztest

Um den Fehler einer fälschlichen Ablehnung der Nullhypothese zu kontrollieren, wird vor der Testung eine Wahrscheinlichkeit für diesen Fehler festgelegt. Die Wahrscheinlichkeit, die Nullhypothese abzulehnen, wenn sie in Wirklichkeit wahr ist, $p(H_0 \text{ ablehnen} | H_0 \text{ wahr})$, wird wie bereits oben ausgeführt *Signifikanzniveau (significance level)* genannt. Das Signifikanzniveau wird mit α bezeichnet und kann (im Prinzip) frei gewählt werden.

Signifikanzniveau α

Wird die Nullhypothese bei Verwendung des Signifikanzniveaus α abgelehnt, so sagt man, dass der Test (zum Niveau α) signifikant ist. Zumeist wird $\alpha = .05$ oder $\alpha = .01$ gewählt.

Im Beispiel in Kapitel 2.2 wurden ausgehend von einem Signifikanzniveau $\alpha = .05$ der Ablehnungsbereich sowie der Annahmebereich für die Nullhypothese festgelegt. Dabei ist die Wahrscheinlichkeit, dass die Teststatistik einen Wert in dem Ablehnungsbereich annimmt, unter Gültigkeit der Nullhypothese kleiner als $\alpha = .05$ und die entsprechende Wahrscheinlichkeit für den Annahmebereich beträgt mindestens $1 - \alpha = .95$. Die Entscheidungsregel wurde wie folgt definiert:

Lehne H_0 ab bzw. nimm H_1 an, wenn der Wert der Teststatistik in den Ablehnungsbereich fällt, ansonsten behalte H_0 bei.

Die Anwendung dieser Entscheidungsregel führt zur Einhaltung der geforderten Bedingung[1]:

$$P(\text{H}_0 \text{ ablehnen} \mid \text{H}_0 \text{ wahr}) < \alpha$$

Ablehnungsbereich

Die Gesamtheit aller möglichen Werte der Teststatistik, die zu einer Ablehnung der Nullhypothese führen, heißt der *Ablehnungsbereich* des Tests. Entsprechend nennt man die Menge aller möglichen Werte der Teststatistik, bei denen die Nullhypothese beibehalten wird, den *Annahmebereich* des Tests.

Annahmebereich

In dem Beispiel aus Kapitel 2.2 hatten wir die Teststatistik $\bar{Y} - \mu_0$ verwendet, wobei diese Teststatistik normalverteilt ist mit der Standardabweichung $\sigma/\sqrt{n}$. Es ist einfacher, anstelle dieser Teststatistik die standardisierte Form dieser Teststatistik zu verwenden, die einer Standardnormalverteilung folgt:

$$Z = \frac{\bar{Y} - \mu_0}{\sigma/\sqrt{n}}$$

Abbildung 2.4 illustriert diese Situation für die standardnormalverteilte Teststatistik Z und $\alpha = .05$.

kritische Werte

Die beiden Werte, die die Teststatistik unterschreiten bzw. übertreffen muss, die also den Ablehnungs- vom Annahmebereich abgrenzen, werden *kritische Werte* genannt. Die kritischen Werte berechnen sich aus der Verteilung der Teststatistik und dem Signifikanzniveau. Im Beispiel ergeben sich die kritischen Werte aus dem $\alpha/2$-Quantil und dem $(1 - \alpha/2)$-Quantil der Stichprobenverteilung der Teststatistik Z und betragen -1.96 bzw. 1.96. Bei einem zweiseitigen Testproblem liegen damit links von einem kritischen Wert und rechts von einem kritischen Wert jeweils $\alpha/2 \cdot 100\,\%$ der Fläche unter der Verteilung der Teststatistik. Die kritischen Werte gehören nicht zum Ablehnungsbereich, sondern bilden die Grenze des Annahmebereichs.

[1] In vielen Statistikbüchern wird für die Definition des Signifikanztests die Bedingung $P(\text{H}_0 \text{ ablehnen} \mid \text{H}_0 \text{ wahr}) \leq \alpha$ genutzt. Gemäß dieser Bedingung gehören die Randpunkte zum Ablehnungsbereich. Damit ist diese Definition nicht konsistent mit der Definition eines Signifikanztests anhand von Konfidenzintervallen, wenn die Konfidenzintervalle wie üblich als geschlossene Intervalle definiert werden. Für die praktische Anwendung von Tests ist es jedoch unerheblich, ob man die Bedingung $P(\text{H}_0 \text{ ablehnen} \mid \text{H}_0 \text{ wahr}) \leq \alpha$ oder die Bedingung $P(\text{H}_0 \text{ ablehnen} \mid \text{H}_0 \text{ wahr}) < \alpha$ wählt.