LEHRBUCH

Henning Kreis
Raimund Wildner
Alfred Kuß

Marktforschung

Datenerhebung und Datenanalyse

8. Auflage

MOREMEDIA

Springer Gabler

Marktforschung

SN Flashcards Microlearning

Schnelles und effizientes Lernen mit digitalen Karteikarten – für Arbeit oder Studium!

Diese Möglichkeiten bieten Ihnen die SN Flashcards:

- Jederzeit und überall auf Ihrem Smartphone, Tablet oder Computer **lernen**
- Den Inhalt des Buches lernen und Ihr Wissen **testen**
- Sich durch verschiedene, mit multimedialen Komponenten angereicherte Fragetypen **motivieren lassen** und zwischen drei Lernalgorithmen (Langzeitgedächtnis-, Kurzzeitgedächtnis- oder Prüfungs-Modus) **wählen**
- Ihre eigenen Fragen-Sets **erstellen**, um Ihre Lernerfahrung zu **personalisieren**

So greifen Sie auf Ihre SN Flashcards zu:

1. Gehen Sie auf die **1. Seite des 1. Kapitels** dieses Buches und folgen Sie den Anweisungen in der Box, um sich für einen SN Flashcards-Account anzumelden und auf die Flashcards-Inhalte für dieses Buch zuzugreifen.
2. Laden Sie die SN Flashcards Mobile App aus dem Apple App Store oder Google Play Store herunter, öffnen Sie die App und folgen Sie den Anweisungen in der App.
3. Wählen Sie in der mobilen App oder der Web-App die Lernkarten für dieses Buch aus und beginnen Sie zu lernen!

Sollten Sie Schwierigkeiten haben, auf die SN Flashcards zuzugreifen, schreiben Sie bitte eine E-Mail an **customerservice@springernature.com** und geben Sie in der Betreffzeile **SN Flashcards** und den Buchtitel an.

Henning Kreis · Raimund Wildner · Alfred Kuß

Marktforschung

Datenerhebung und Datenanalyse

8., überarbeitete Auflage

Springer Gabler

Henning Kreis
Fachbereich Wirtschaftswissenschaften
HWR Berlin School of Economics and Law
Berlin, Deutschland

Alfred Kuß
Marketing-Department, Freie Universität Berlin
Berlin, Deutschland

Raimund Wildner
Vorstand MIR e. V. (vormals GfK Verein)
und Honorarprofessor an der Universität
Erlangen-Nürnberg
Fürth, Deutschland

ISBN 978-3-658-44455-6 ISBN 978-3-658-44456-3 (eBook)
https://doi.org/10.1007/978-3-658-44456-3

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über https://portal.dnb.de abrufbar.

Planung/Lektorat: Barbara Roscher
Springer Gabler ist ein Imprint der eingetragenen Gesellschaft Springer Fachmedien Wiesbaden GmbH und ist ein Teil von Springer Nature.
Die Anschrift der Gesellschaft ist: Abraham-Lincoln-Str. 46, 65189 Wiesbaden, Germany

Vorwort zur 8. Auflage

Mit dem vorliegenden Lehrbuch soll Studierenden und interessierten Praktikern eine Einführung und ein Überblick zum großen und für das Marketing bedeutsamen Gebiet der Marktforschung gegeben werden. Der Schwerpunkt des Buchs liegt bei den zentralen und allgemein gültigen Prinzipien und Methoden der Marktforschung, nicht bei technischen Details. Das Buch ist hinsichtlich seines Inhalts und Umfangs so konzipiert, dass es eine einführende Lehrveranstaltung zur Marktforschung begleiten kann. Daneben kann es Marketing-Praktikern dazu dienen, Maßstäbe für die Leistungsfähigkeit und Aussagekraft von Marktforschungsuntersuchungen kennen zu lernen. Für die selbstständige Durchführung von Untersuchungen bedarf es aber in der Regel methodischer Detailkenntnisse und Erfahrungen, die deutlich über den Rahmen dieses Buches hinausgehen.

Das sehr umfangreiche Gebiet der Datenanalyse mit statistischen Methoden, wozu ja reichhaltige Spezial-Literatur existiert, wird im vorliegenden Lehrbuch nur in seinen Grundzügen dargestellt. Hier werden die Grundideen und Anwendungsmöglichkeiten von statistischen Methoden, die in der Marktforschung gängig sind, in möglichst leicht verständlicher Weise skizziert. Dabei werden nur begrenzte Vorkenntnisse der Statistik vorausgesetzt. Dieses ist auch dadurch begründet, dass zwar in vielen (insbesondere wirtschaftswissenschaftlichen) Studiengängen Grundkenntnisse der Statistik vermittelt werden, dass aber die dort erworbenen Kenntnisse oftmals nur begrenzt nachhaltig sind.

Die drei Autoren haben mit der achten Auflage die kooperative und harmonische Zusammenarbeit an diesem Buch fortgesetzt. Dabei liegt die Federführung jetzt bei Henning Kreis. Im Vergleich zur siebten Auflage ist das gesamte Buch überarbeitet und aktualisiert worden. An einigen Stellen sind Ergänzungen hinzugekommen, z. B. im Hinblick auf die Analyse qualitativer Daten (thematic analysis) und relevante Entwicklungen im Bereich künstlicher Intelligenz. An anderen Stellen gab es auch Straffungen. In den laufenden Text eingefügt und von diesem etwas abgesetzt finden sich zahlreiche Beispiele und weiterführende Hinweise, die die Lektüre leichter, abwechslungsreicher und ergiebiger machen sollen. Außerdem wurde häufig auf spezielle Literatur verwiesen, um eine vertiefende Beschäftigung mit den Methoden der Marktforschung zu erleichtern.

Im Buch wird immer wieder zwischen weiblichen und männlichen Formen von Begriffen (z. B. „die Managerin“ oder „der Kunde“) gewechselt. Das soll andeuten, dass die im Buch enthaltenen Aussagen natürlich geschlechtsneutral sind.

Die Autoren haben versucht, grundlegende Aspekte der Marktforschung in gut verständlicher Weise auf knappem Raum darzustellen. Das ist nicht immer leicht und manche Kompromisse sind unvermeidlich. Vor diesem Hintergrund bleiben die Autoren für kritische Anmerkungen und Hinweise zur Weiterentwicklung des Lehrbuchs dankbar.

Birgit Borstelmann und Barbara Roscher vom Verlag Springer Gabler haben auch die Vorbereitung der 8. Aufl. wieder kompetent und konstruktiv begleitet. Diese Zusammenarbeit war stets sehr angenehm. Dafür herzlichen Dank. Für verbliebene Fehler und Unzulänglichkeiten tragen natürlich die Autoren die Verantwortung.

Berlin/Fürth
Januar 2024

Henning Kreis
Raimund Wildner
Alfred Kuß

Inhaltsverzeichnis

Einführung

1

Zusammenfassung

In diesem Kapitel werden zunächst die wesentlichen Merkmale der Marktforschung kurz umrissen. Es folgt ein Überblick über typische Anwendungen der Marktforschung bei Marketing-Problemen.

Mit der kostenlosen Flashcard-App „SN Flashcards“ können Sie Ihr Wissen anhand von Fragen überprüfen und Themen vertiefen. Für die Nutzung folgen Sie bitte den folgenden Anweisungen:

1. Gehen Sie auf https://flashcards.springernature.com/login
2. Erstellen Sie ein Benutzerkonto, indem Sie Ihre Mailadresse angeben und ein Passwort vergeben.
3. Verwenden Sie den folgenden Link, um Zugang zu Ihrem SN Flashcards Set zu erhalten: https://sn.pub/7sEmtw

Sollte der Link fehlen oder nicht funktionieren, senden Sie uns bitte eine E-Mail mit dem Betreff „SN Flashcards“ und dem Buchtitel an customerservice@springernature.com.

H. Kreis et al., *Marktforschung,* https://doi.org/10.1007/978-3-658-44456-3_1

1.1 Kennzeichnung der Marktforschung

Die Marktforschung gehört zu den am längsten etablierten Teilgebieten der Marketingwissenschaft. Sie ist untrennbar mit dem Marketing verbunden, weil die Ausrichtung von Angeboten der verschiedenen Unternehmen auf Kundenwünsche ebenso wie die Beeinflussung der Kunden durch die Anbieterunternehmen natürlich angemessene Informationen über Kunden und Märkte voraussetzt.

Während ein Handwerksmeister oder ein Einzelhändler vor Ort oftmals noch unmittelbar von seinen Kunden erfährt, was diese wünschen, und direkt beobachten kann, wie diese auf seine Angebote reagieren, bestehen bei vielen größeren Unternehmen, insbesondere wenn sie ihre Produkte auf internationalen Märkten absetzen, kaum noch direkte Kontakte zu Endkunden, die eine Informationsgewinnung ermöglichen. Weil in diesen Fällen typischerweise unterschiedliche Absatzmittler (z. B. Groß- und Einzelhandel, Importeure, Exporteure) den Absatzweg vom Hersteller zum Endkunden bestimmen, sind die Beziehungen zu den Kunden für die Herstellerunternehmen weitgehend anonym und ein leistungsfähiges System zur Sammlung und Aufbereitung von Marktinformationen – eben die Marktforschung – wird notwendig. Meist geht eine kontinuierliche Marktbeobachtung (z. B. der Entwicklung von Marktgröße und Marktanteilen) mit eher anlassbezogenen speziellen Untersuchungen (z. B. Produkttests, Segmentierungsanalysen) einher. Dafür steht ein recht umfassendes methodisches Instrumentarium zur Verfügung, dessen wichtigste Bestandteile in ihren Grundzügen im vorliegenden Lehrbuch dargestellt werden sollen.

Nun ist es kaum möglich, ein umfassendes und komplexes Fachgebiet wie die Marktforschung durch eine kurze Definition vollständig zu kennzeichnen. Manche in der Literatur zu findenden **Definitionen der Marktforschung** knüpfen direkt an pragmatische Überlegungen an und konzentrieren sich auf den Aspekt der Bereitstellung von Informationen für das Marketing. Allerdings bieten solche Definitionen wenig Trennschärfe im Hinblick auf Informationen, die zwar für Marketing-Entscheidungen höchst relevant sind, die aber üblicherweise kaum der Marktforschung zugerechnet werden (z. B. Auswertungen von Patent-Anmeldungen, Analysen von Konkurrenzprodukten). Die beiden folgenden Definitionen mögen einer Charakterisierung der Marktforschung dienen. Die Darstellung von Methoden und Anwendungen der Marktforschung in den folgenden Kapiteln wird aber diese Kennzeichnung sicher noch erweitern und abrunden.

▶ **Definition** Christian Homburg (2020, S. 270) fasst das Wesen der Marktforschung sehr prägnant zusammen:

> „Unter Marktforschung verstehen wir die systematische Sammlung, Aufbereitung, Analyse und Interpretation von Daten über Märkte (Kunden und Wettbewerber) zum Zweck der Fundierung von Marketingentscheidungen."

Eine etwas umfassendere Darstellung enthält die international breit akzeptierte **Definition der American Marketing Association** (AMA) von 2017 (www.ama.org), bei der allerdings der größere Gehalt mit größerem Umfang „erkauft“ werden muss:

> „Marktforschung ist die Funktion, die den Konsumenten, Kunden und die Öffentlichkeit durch Informationen mit dem Anbieter verbindet – Informationen, die benutzt werden
> zur Identifizierung von Marketing-Chancen und -Problemen,
> zur Entwicklung, Modifizierung und Überprüfung von Marketing-Maßnahmen,
> zur Überprüfung des Marketing-Erfolges und
> zur Verbesserung des Verständnisses des Marketing-Prozesses.
> Die Marktforschung bestimmt die zur Untersuchung dieser Gesichtspunkte notwendigen Informationen, entwickelt die Methoden zur Sammlung der Informationen, plant die Datenerhebung und führt diese durch, analysiert die Ergebnisse und präsentiert diese und die Schlussfolgerungen daraus.“

Die Definition der AMA enthält drei deutlich abgegrenzte Teile. Am Anfang steht eine kurze Kennzeichnung der Marktforschung. Es folgt eine Übersicht über vier wesentliche Anwendungsbereiche der Marktforschung im Marketing. Der Begriff „Marktforschung“ wird sowohl auf eine bestimmte Funktion als auch auf Tätigkeiten bezogen. Eine ähnliche – etwas differenziertere – Darstellung des Forschungsprozesses wird Gegenstand des Abschn. 2.2.1 des vorliegenden Buches sein.

1.2 Anwendungen der Marktforschung

In der im vorigen Abschnitt dargestellten Kennzeichnung der Marktforschung durch die AMA sind schon Kernaufgaben genannt worden. Der letzte der dort aufgeführten Aspekte („Verbesserung des Verständnisses des Marketing-Prozesses“) ist eher auf Grundlagenforschung ausgerichtet. Bei den anderen drei Aufgabenbereichen ist aber der Bezug zum Marketing deutlich erkennbar und kann mithilfe der Abb. 1.1 leicht veranschaulicht werden.

Man erkennt in der Abb. 1.1 auf der linken Seite die drei Aufgabenbereiche der Marktforschung aus der AMA-Definition. Auf der rechten Seite sind Aufgabenbereiche des Marketings aufgeführt, die entsprechende Informationen aus der Marktforschung verwenden bzw. deren Ergebnisse durch die Marktforschung gemessen werden. Das Zusammenwirken von Marktforschung und Marketing-Management ist in Abb. 1.1 durch ein einfaches Beispiel eines Markenartikel-Herstellers im Lebensmittelsektor illustriert.

In den im vorigen Abschnitt angesprochenen Definitionen der Marktforschung ist schon angeklungen, dass bei der Marktforschung die **Verbesserung von Marketing-Entscheidungen** durch Reduktion der Unsicherheit über Reaktionen von Kunden und Wettbewerbern, Marktentwicklungen etc. im Mittelpunkt steht. Daneben gibt es in der Praxis eine ganze Reihe anderer Gründe, Marktuntersuchungen durchzuführen, die zumindest vereinzelt eine Rolle spielen können:

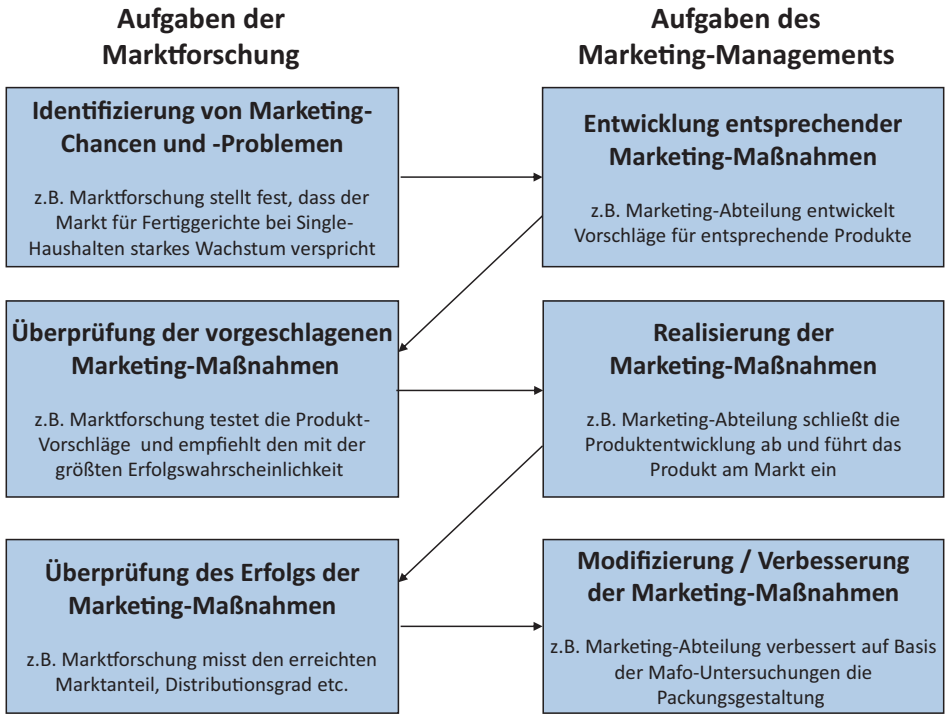


Abb. 1.1 Zusammenwirken von Marktforschung und Marketing und Marketing-Management. (Nach Assael, 1993, S. 219)

- **Unterstützung der eigenen Meinung:** (Ausgewählte) Untersuchungsergebnisse werden verwendet, um im eigenen Unternehmen eine bestimmte Entscheidung (z. B. über ein neues Produkt oder eine Werbekampagne) zu begründen und durchzusetzen.
- **Absicherung der Verantwortlichen:** Eine (Fehl-)Entscheidung lässt sich nachträglich leichter rechtfertigen, wenn man belegen kann, dass die Alternativen sorgfältig untersucht worden sind und die Ergebnisse die getroffene Entscheidung begründet haben.
- **Hilfsmittel bei rechtlichen Auseinandersetzungen:** Zum Beispiel lassen sich Fragen der Verwechslung von Marken, Irreführung durch Werbung etc. am ehesten durch entsprechende Messungen (meist Umfragen) bei den Konsumenten klären (Jacoby, 2013).
- **Argumente für Public Relations und Werbung:** Z. B. „Deutschlands meistgekaufter Energy-Drink", „Europas beliebtester Kleinwagen".

In einer weiteren Übersicht zur Anwendung der Marktforschung sollen im Folgenden einige Nutzer von Marktforschungsdaten kurz gekennzeichnet werden (die prozentualen Anteile am Umsatz der großen deutschen Marktforschungsinstitute im Jahr 2021 sind jeweils eingetragen; Quelle: www.adm.ev.de):

- **Hersteller von Konsumgütern** (einschl. Automobilherst.) 24 %: Hier sind große Märkte mit einer Vielzahl von Konsumenten, zu denen keine direkten Kontakte bestehen, typisch. Laufende Änderungen der Marktverhältnisse durch Konkurrenzaktivitäten, Geschmacksveränderungen etc. machen kontinuierliche Messungen („Tracking“) erforderlich.
- **Investitionsgüter/Energiewirtschaft** 2 %**:** Wegen verbreiteter direkter Marktkontakte (z. B. durch persönlichen Verkauf) spielt hier die Marktforschung eine relativ geringe Rolle.
- **Dienstleistungsunternehmen** (einschl. Banken, Versicherungen, Transport, Verkehr, Tourismus) 10 %: Bei Dienstleistungen sind direkte Kundenkontakte typisch und somit Informationen von den Kunden erhältlich. Im Dienstleistungsbereich spielen regelmäßige Messungen der Kundenzufriedenheit eine besonders große Rolle.
- **Einzelhandel** 9 %**:** Der Einzelhandel kann sich zum großen Teil auf seine selbst erhobenen Daten (nicht zuletzt Scanner-Daten) über Abverkäufe, Reaktionen auf Sonderangebote etc. stützen und hat naturgemäß zahlreiche direkte Kundenkontakte
- **Medien** (Verlage, Fernsehsender etc.) 16 %: Medienunternehmen nutzen Marktforschung – wie Anbieter anderer Produkte auch – zur Bestimmung von Zielgruppen, zur Messung von Verbreitungsgraden und Einschaltquoten, bei der Entwicklung neuer Zeitschriften, Sendeformate usw. Daneben wird eine intensive Medienforschung (Anzahl und Zusammensetzung von Lesern und Zuschauern) betrieben, die nicht zuletzt dazu dient, Werbekunden Daten für ihre Entscheidungen zu liefern.
- **Öffentliche Auftraggeber** (Ministerien, Kommunen etc.) 3 %: Hier wird Marktforschung eher selten betrieben, u. a. aber im Zusammenhang mit Stadtmarketing und Tourismuswerbung.

Letztlich sei in diesem Abschnitt noch kurz die Frage angesprochen, wer Marktforschung in der Regel durchführt. Eine zentrale Rolle spielen hier sogenannte **Allround-Marktforschungsinstitute**. Das sind (mittlere bis größere) Unternehmen, die entsprechende Untersuchungen mit unterschiedlichen Methoden (Umfragen, Produkttests etc.) zu einem weiten Spektrum von Themen (z. B. Image-Untersuchungen, Messungen von Kundenzufriedenheit oder Bekanntheitsgrad, Produkttests) anbieten und durchführen. In der folgenden Übersicht sind einige prominente Allround-Institute im deutschsprachigen Raum mit Angaben zum Gründungsjahr und Standort in Deutschland genannt (Stand: März 2023). Die aufgeführten Internet-Adressen ermöglichen den Zugang zu genaueren Informationen, oftmals auch über berufliche Möglichkeiten bei diesen Instituten.

Institutsname	Standort	Internetadresse
GfK-Gruppe	Nürnberg	www.gfk.com
Ipsos	Hamburg	www.ipsos.de
Kantar	München	www.kantardeutschland.de
AC Nielsen GmbH	Frankfurt/a. M.	www.nielseniq.com

Im Herbst 2023 wurde die GfK-Gruppe von NielsenIQ freundlich übernommen, wodurch die fusionierte Firma sicher an Stelle 1 der Liste kommt. Aufgrund von Auflagen der Kartellbehörden wurde das Haushaltspanel der GfK im Januar 2024 an die Firma You Gov cerkauft. Wie sich die obige Liste dadurch ändert bleibt abzuwarten. Daneben sind am „Marktforschungs-Markt" zahlreiche (meist kleinere) **Spezial-Institute** tätig, die sich auf ein besonderes Anwendungsgebiet (z. B. Werbeforschung, Automobil-Marktforschung) oder eine besondere Methodik (z. B. Online-Befragungen, qualitative Marktforschung) spezialisiert haben. Als Beispiel für solche Spezial-Institute sei hier „Eye-Square" (www.eye-square.com) genannt – ein Institut, das sich auf Blickregistrierung („eye tracking") und andere biometrische sowie indirekte Methoden (z. B. Reaktionszeitmessung) spezialisiert hat.

Einen Überblick über die Marktforschungsbranche in Deutschland und auch über entsprechende berufliche Möglichkeiten erhält man u. a. mithilfe der Internet-Angebote des Arbeitskreises Deutscher Markt- und Sozialforschungsinstitute (www.adm-ev.de) sowie des Berufsverbandes Deutscher Markt- und Sozialforscher (bvm.org).

Die sogenannten „**betrieblichen Marktforscher"** sind ein wichtiges Bindeglied zwischen Marktforschung (im engeren Sinne) und der Marketing-Praxis. Sie sind – wie die Bezeichnung schon andeutet – bei Unternehmen tätig, die Untersuchungen meist bei Instituten in Auftrag geben. Die betrieblichen Marktforscher sind dort für die Zusammenarbeit mit Marktforschungsinstituten zuständig, sie bereiten Ergebnisse auf, beraten intern in Marktforschungsfragen und führen kleine Untersuchungen sowie Sekundärforschung (Abschn. 2.4.2.1) selbst durch.

Hintergrundinformationen

Die folgenden Internet-Adressen bieten nützliche Informationen zur Marktforschungsbranche:
www.adm-ev.de (Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute)
www.bvm.org (Berufsverband Deutscher Markt- und Sozialforscher)
www.esomar.org (European Society for Opinion and Marketing Research)
www.vmoe.at (Verband der Marktforscher Österreichs)
www.swiss-insights.ch (Swiss Data Insights Association)
www.dgof.de (Deutsche Gesellschaft für Online-Forschung e. V.)
www.asi-ev.org (Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e. V.)
www.rat-marktforschung.de (Rat der Deutschen Markt- und Sozialforschung e. V.)
www.planung-analyse.de (Marktforschungszeitschrift „planung&analyse")

Literatur

Assael, H. (1993). *Marketing – Principles and Strategy* (2. Aufl.). Dryden Press.
Homburg, C. (2020). *Marketingmanagement* (7. Aufl.). SpringerGabler.
Jacoby, J. (2013). *Trademark surveys – Designing, implementing, and evaluating surveys* (Bd. 1). American Bar Association.

Grundlagen

2

Zusammenfassung

Im 2. Kapitel werden die verschiedenen Teil-Aufgaben bei Untersuchungen der Marktforschung überblicksartig dargestellt. Dazu werden einerseits der typische Ablauf einer Marktforschungsstudie in der Praxis und andererseits ein Modell der zentralen Aspekte theorieorientierter Marketingforschung erläutert. Es folgt ein erster Blick auf die zentralen Gütekriterien von Untersuchungen (Reliabilität, Validität und Generalisierbarkeit), die in den folgenden Kapiteln eine wesentliche Rolle spielen. Am Ende des Kapitels werden die wichtigsten Untersuchungsarten der Marktforschung kurz vorgestellt.

2.1 Überblick

Im vorliegenden Kapitel sollen – dem Titel „Grundlagen" entsprechend – zentrale Aufgaben, methodische Grundfragen und Vorgehensweisen der Marktforschung sowie Anforderungen an Marktforschungsuntersuchungen gekennzeichnet, erläutert und diskutiert werden. Dazu wird von zwei verschiedenartigen Darstellungen des Forschungsablaufs der Marktforschung, in die die relevanten Konzepte in ihrem Zusammenhang eingeordnet sind, ausgegangen. Die eine dieser Darstellungen (siehe Abschn. 2.2.1) ist auf einen typischen **Untersuchungsablauf in der Marktforschungs*praxis*** ausgerichtet. Hier lässt sich bei den verschiedenen Schritten von der Definition des Untersuchungsproblems bis zum Bericht jeweils aufzeigen, welche Entscheidungen zu treffen und welche Probleme zu lösen sind.

Das in Abschn. 2.2.2 dargestellte „Grundmodell der empirischen Marketingforschung" ist weniger auf praktische Fragestellungen fokussiert, sondern auf **die grundlegenden gedanklichen Schritte,** die mit Untersuchungen, die wissenschaftlichen

H. Kreis et al., *Marktforschung,* https://doi.org/10.1007/978-3-658-44456-3_2

Zielen (vorrangig Entwicklung und Test von Theorien) entsprechen sollen, verbunden sind. Insofern ist hier der Abstraktionsgrad etwas höher als im Abschn. 2.2.1.

Aus beiden Darstellungen lassen sich zentrale Anforderungen bezüglich der **Qualität von Marktforschungsuntersuchungen** ableiten. Dies sind die Gültigkeit (Validität), die Verlässlichkeit (Reliabilität) und die Generalisierbarkeit einer Untersuchung. Diese in Abschn. 2.3 vorgestellten Kriterien sind die wichtigsten Maßstäbe für die Überlegungen und Entscheidungen bei der Ausgestaltung der verschiedenen Teile einer Untersuchung (z. B. Stichprobenziehung, Frageformulierung, statistische Tests).

Gewissermaßen als Überleitung zu den stärker den Einzelheiten von Methoden gewidmeten folgenden Kapiteln findet sich in Abschn. 2.4.2.2 ein Überblick über verschiedene Grundtypen von Untersuchungsdesigns. In einem **Untersuchungsdesign** werden die Festlegungen hinsichtlich der Art der Datenerhebung (z. B. Befragung oder Beobachtung), der Untersuchungsgegenstände, der Erhebungseinheiten (z. B. Stichprobe von Personen oder Haushalten einer bestimmten Region) und der durchzuführenden Analysen (z. B. Signifikanztests, Schätzung von Anteilswerten in der Grundgesamtheit) zusammengefasst. Einzelheiten der verschiedenen Untersuchungsdesigns und insbesondere die dabei jeweils relevanten methodischen Aspekte werden im Anschluss (ab Kap. 3) dargestellt.

2.2 Zwei Sichtweisen des Forschungsprozesses – Theorie und Praxis

2.2.1 Untersuchungsablauf in der Marktforschungspraxis

Im vorliegenden Abschnitt soll der typische Ablauf einer Marktforschungsuntersuchung relativ grob skizziert werden, um auf diese Weise einen ersten Überblick über Techniken und methodische Probleme der Marktforschung zu geben. In den folgenden Kapiteln werden dann die einzelnen dabei angesprochenen Aspekte genauer erörtert.

Abb. 2.1 zeigt ein Schema mit typischen Phasen einer Marktforschungsuntersuchung. Natürlich ist der dort wiedergegebene Ablauf gegenüber der Praxis der Marktforschung vereinfacht und verallgemeinert. In der Realität findet man sicher häufig Studien, bei denen einzelne der hier angegebenen Schritte ausgelassen oder andere hinzugefügt werden. Außerdem treten normalerweise vielfältige Rückkoppelungen im Forschungsprozess auf. Beispielsweise kann man sich leicht vorstellen, dass man bei der Planung der Datenanalyse feststellt, dass im vorher liegenden Schritt der Entwicklung von Messinstrumenten Daten entstehen, die im Hinblick auf das Messniveau (siehe Abschn. 7.2) nicht den Anforderungen eines vorgesehenen Analyseverfahrens entsprechen. In einem solchen Fall müsste also die Entwicklung der Messinstrumente erneut aufgenommen werden.

Durch das in Abb. 2.1 dargestellte Phasenschema soll auch angedeutet werden, dass die einzelnen Schritte im Untersuchungsablauf stark voneinander abhängen. Schwächen

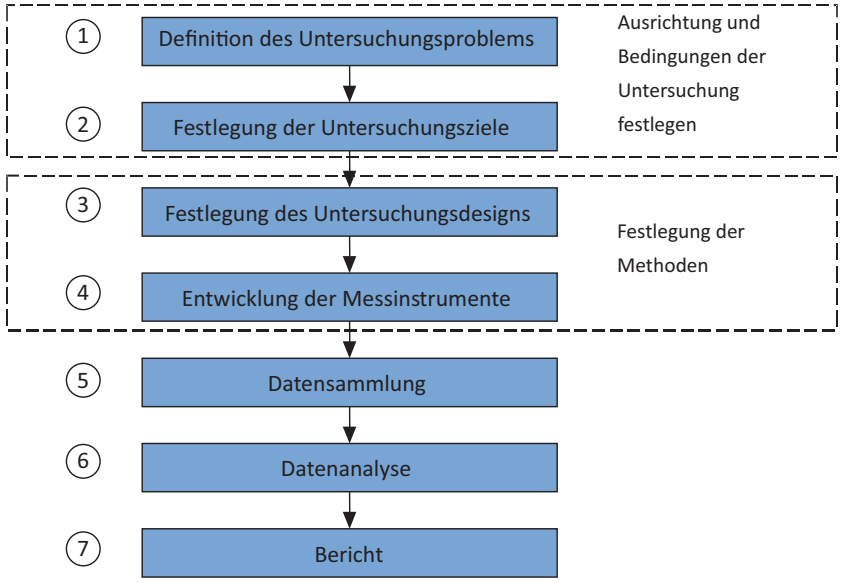


Abb. 2.1 Typische Phasen einer Marktforschungsuntersuchung

und Fehler in frühen Phasen können nicht durch besondere Sorgfalt und großen Aufwand in späteren Phasen ausgeglichen werden. Wenn bei der Datenerhebung gravierende methodische Mängel aufgetreten sind (z. B. durch unzureichende Stichprobenausschöpfung oder durch systematisch verzerrende Messmethoden), so kann man diese eben nicht durch anspruchsvolle Verfahren der Datenanalyse kompensieren (allenfalls verschleiern). In diesem Sinne ist das Ergebnis einer Marktforschungsuntersuchung nur so stark wie das schwächste Glied in der Kette der Untersuchungsschritte.

Bei der ersten Phase der Untersuchung, der **Definition des Untersuchungsproblems** (1), werden gewissermaßen die Weichen für deren Erfolg gestellt. Eine unpräzise Beschreibung des Untersuchungsgegenstandes kann eben dazu führen, dass man – möglicherweise mit großem Aufwand – am relevanten Problem „vorbeiforscht“. Ausschlaggebend ist die Kommunikation zwischen dem Marketing-Management (als Auftraggeber einer Untersuchung) und den Marktforschern, die eine Untersuchung konzipieren und durchführen. Das Management muss die Möglichkeiten und Grenzen der einschlägigen Forschungsmethoden kennen; das Marktforschungsinstitut bzw. die betriebliche Marktforschungsabteilung muss das anstehende Entscheidungsproblem des Managements und den damit verbundenen Informationsbedarf kennen.

Damit ist schon angedeutet, dass das Entscheidungsproblem keineswegs mit dem Untersuchungsproblem identisch sein muss. Häufig wird eben nur ein Teilaspekt des Entscheidungsproblems mit den Methoden der Marktforschung untersucht werden können. In der Regel ist das Untersuchungsproblem auch konkreter und präziser formuliert als das Entscheidungsproblem, das den Ausgangspunkt dafür bildete. In Abb. 2.2

Entscheidungsproblem	Untersuchungsproblem
Entwicklung der Packung für ein neues Produkt	Überprüfung der Wirkung alternativer Packungsentwürfe
Geografische Aufteilung des Werbebudgets	Bestimmung des gegenwärtigen Umfangs des Marktdurchdringung in den entsprechenden Gebieten
Einführung eines neuen Produkts	Entwicklung eines Testmarkts, mit dessen Hilfe die voraussichtliche Akzeptanz des Produkts ermittelt werden kann

Abb. 2.2 Beispiele für Entscheidungs- und Untersuchungsprobleme in Marketing und Marktforschung. (Nach Churchill & Iacobucci, 2005, S. 51)

findet sich eine Gegenüberstellung von Beispielen für Entscheidungs- und Untersuchungsprobleme, die diesen Aspekt illustrieren.

Trotz der angesprochenen Bedeutung einer angemessenen Definition des Untersuchungsproblems kann dieser Teil des Marktforschungsprozesses hier nicht vertiefend behandelt werden, weil die jeweilige Vorgehensweise so stark situationsspezifisch ist, dass man kaum generelle Aussagen dazu machen kann. Die Kommunikation zwischen Marketing-Management und Marktforschung und einschlägige Erfahrungen der Beteiligten haben dabei zentrale Bedeutung.

Beispiel

Ein inzwischen schon „klassisches" Beispiel für grob irreführende Marktforschung durch eine falsche Problemdefinition bietet die – seinerzeit fürchterlich gescheiterte – Veränderung des Geschmacks von Coca-Cola in den USA Mitte der 1980er Jahre. Angeregt durch das bessere Abschneiden von Pepsi-Cola in sog. „Blindtests" (vergleichende Geschmackstests mit verdeckten Markennamen), veränderte man die Zusammensetzung von Coca-Cola und gab dieser einen deutlich süßeren Geschmack. Entsprechende Blindtests ergaben dann deutlich bessere Werte für die „neue Coca-Cola". Die Markteinführung des veränderten Produkts geriet bekanntlich zum Desaster: Konsumenten protestierten, der Marktanteil von Coca-Cola sackte deutlich ab. Was war geschehen? Man hatte vonseiten der Marktforschung ausschließlich den Geschmack getestet und völlig ignoriert, dass in den USA lange verfestigte Gewohn-

heiten und auch emotionale Bindungen an den wohlvertrauten Coke-Geschmack bei vielen Menschen bestehen. Das für die Marktforschung zu definierende Problem hätte nicht darauf beschränkt werden dürfen, die Reaktionen auf eine Geschmacksänderung zu messen, sondern hätte breiter definiert werden müssen im Hinblick auf die Akzeptanz der Veränderung eines Produkts, mit dem man schon lange vertraut ist. (Quelle: Burns & Bush, 2006, S. 86 f.) ◀

Mit der **Festlegung der Untersuchungsziele** (2) wird die Aufgabenstellung für eine Untersuchung, die mit der Problemdefinition bereits umrissen wurde, konkretisiert und präzisiert. Im Rahmen einer allgemeinen Problemdefinition, die z. B. darin bestehen könnte, dass für ein Produkt die Qualitätseinschätzung im Vergleich zu entsprechenden Produkten von Wettbewerbern ermittelt werden soll, könnte man ein Untersuchungsziel so formulieren, dass die für die in Frage kommende Zielgruppe wichtigsten Produkteigenschaften (→ Entscheidungskriterien beim Kauf) und die Einschätzung der auf dem Markt angebotenen Produkte hinsichtlich dieser Eigenschaften ermittelt werden sollen. Aus der Festlegung der Untersuchungsziele ergibt sich die Art der jeweils vorgesehenen Untersuchung. Dabei sind folgende typischen Ziele zu unterscheiden:

- **Explorative Untersuchungen**
  Dabei geht es um das Ziel, Ursachen für Probleme oder Zusammenhänge zwischen Variablen zu *entdecken*. Derartige Untersuchungen stehen oft am Anfang eines Projekts und dienen dann der Vorbereitung weiterer Untersuchungen.
- **Deskriptive Untersuchungen**
  Hier steht das Ziel der Kennzeichnung bzw. *Beschreibung* einer interessierenden Grundgesamtheit (z. B. Personen, Haushalte) hinsichtlich für das Untersuchungsproblem relevanter Merkmale (z. B. Markenpräferenzen, Verbrauchshäufigkeit) im Vordergrund.
- **Kausal-Untersuchungen**
  Kausal-Untersuchungen haben das Ziel festzustellen, welches die *Ursachen* (Gründe) für beobachtete Phänomene sind. Die Überprüfung und gegebenenfalls Bestätigung von Ursache-Wirkungs-Beziehungen stellt allerdings besonders hohe Anforderungen an die Untersuchungsanlage (siehe Abschn. 6.1).

Auf die Festlegung von Untersuchungszielen und die Arten von Untersuchungen wird im Abschn. 2.4 noch ausführlicher eingegangen.

Die **Festlegung des Untersuchungsdesigns** (3) ist eine komplexe Aufgabe, bei der die grundlegenden Entscheidungen über die anzuwendenden Methoden getroffen werden. Dabei ist zunächst zu entscheiden, ob das Untersuchungsziel durch angemessene Aufbereitung und Analyse vorhandener (früher oder von anderen Institutionen gesammelter) Daten erreicht werden kann **(Sekundärforschung)** oder ob dazu eine neue, gezielte Datenerhebung notwendig ist **(Primärforschung)**. Einzelheiten dazu finden sich im Abschn. 2.4.2.1. Die in der Marktforschung gängigen Untersuchungsdesigns (bei

Primärforschung) können in bestimmte Grundtypen eingeteilt werden, die wiederum bestimmten Forschungszielen (s. o.) entsprechen. Es lassen sich die folgenden vier Grundtypen identifizieren:

- **Qualitative Untersuchungen**
  Hier geht es nicht darum, quantifizierende und repräsentative Aussagen zu machen, sondern eher darum, Arten, Zusammenhänge und Wirkungen problemrelevanter Variablen kennen zu lernen (bzw. zu entdecken → *explorativ*). Durch entsprechende (qualitative) Untersuchungsmethoden (siehe Kap. 3) versucht man, dieser Aufgabenstellung gerecht zu werden.
- **Querschnitts-Untersuchungen**
  Dabei ist an Studien zu denken, bei denen auf einen Zeitpunkt bezogene quantifizierende Aussagen über eine bestimmte Grundgesamtheit (z. B. Einkommensverteilung in einer bestimmten Bevölkerungsgruppe) gemacht werden sollen. Es werden also Merkmale dieser Grundgesamtheit gewissermaßen beschrieben (→ *deskriptiv*).
- **Längsschnitt-Untersuchungen**
  Damit kann man dynamische Phänomene (z. B. Markenwechsel von Konsumenten oder Veränderungen von Marktanteilen) im Zeitablauf durch die Erhebung entsprechender (gleichartiger) Daten an mehreren Zeitpunkten messen. Hier wird die Entwicklung und Veränderung von Merkmalen im Zeitablauf beschrieben (→ *deskriptiv*).
- **Experimente**
  Experimente (siehe Kap. 6) sind dadurch gekennzeichnet, dass eine oder mehrere (so genannte *unabhängige*) Variablen so manipuliert wird/werden, dass die Wirkungen dieser Manipulationen auf eine oder mehrere andere (so genannte *abhängige*) Variable gemessen werden können. Es geht also in den meisten Fällen um Ursache-Wirkungs-Beziehungen (→ *kausal*), manchmal auch um die Ermittlung der Art des Zusammenhangs zwischen Variablen (z. B. bei der Bestimmung einer Preis-Absatz-Funktion).

Detailliertere Ausführungen zu den verschiedenen Untersuchungsdesigns finden sich im Abschn. 2.4.2.2.

Nachdem die Konzeption einer Untersuchung (Untersuchungsdesign) festliegt, müssen die **Messinstrumente** entwickelt (4) werden, mit deren Hilfe die im jeweiligen Zusammenhang interessierenden Merkmalsausprägungen von Untersuchungsobjekten (z. B. Ausgaben eines Haushalts für Urlaubsreisen) ermittelt – also gemessen – werden sollen. Typische Beispiele für Messinstrumente sind einzelne Fragen oder so genannte Multi-Item-Skalen in Fragebögen. Die Entwicklung von exakten und zuverlässigen Messtechniken für Zwecke der Marktforschung ist oftmals mit besonderen Schwierigkeiten verbunden, auf die im Abschn. 4.3 ausführlich eingegangen wird.

Die **Datensammlung** (5) ist bei vielen Studien der Marktforschung die Untersuchungsphase, in der die meisten Ressourcen (zeitlich, personell, finanziell) in Anspruch genommen werden. Dieser Bereich ist oft weniger durch theoretische Fundierung als durch Erfahrung und „handwerkliche" Sorgfalt der Verantwortlichen geprägt. Die durch geringfügig erscheinende technische Probleme und menschliche Schwächen bei der Datensammlung entstehenden Fehlermöglichkeiten dürfen aber nicht unterschätzt werden (siehe Abschn. 4.5).

Die **Datenanalyse** (6) ist bestimmt durch den Einsatz statistischer Methoden für die Verdichtung der typischerweise großen Menge gesammelter Daten und für Schlüsse von Ergebnissen in einer Stichprobe auf die in der Regel eher interessierenden Verhältnisse in der entsprechenden Grundgesamtheit. Bei der Erörterung der Datenanalyse wird unterschieden in:

- einfache deskriptive Verfahren (statistische Maßzahlen, Häufigkeitstabellen, graphische Darstellungen etc., → Kap. 7),
- Schätzungen und statistische Tests (Schlüsse von Stichproben auf Grundgesamtheiten, → Kap. 8) sowie
- multivariate Verfahren (gleichzeitige und zusammenhängende Analyse einer Vielzahl von Variablen, → Kap. 9).

Am Ende einer Untersuchung steht die Erstellung des **Berichts** (7), in dem die wesentlichen Ergebnisse, Schlussfolgerungen und Handlungsempfehlungen enthalten sind. Dadurch sollen die durch die Problemdefinition und die Festlegung der Untersuchungsziele gestellten Fragen beantwortet werden. Üblicherweise enthält ein Untersuchungsbericht mindestens vier Teile:

1. Kurze Zusammenfassung von Problemdefinition und Untersuchungszielen
2. Erläuterung der Untersuchungsmethode
3. Darstellung der Untersuchungsergebnisse
4. Schlussfolgerungen und Empfehlungen

Hinsichtlich der Form des Berichts soll versucht werden, einen Kompromiss zwischen **Genauigkeit** der Darstellung von Methoden und Ergebnissen, die häufig eine spezifische Fachsprache (z. B. „Cronbach's α" oder „erklärte Varianz") erfordert, und **Verständlichkeit** zu erzielen.

Inhalt und Gestaltung eines Untersuchungsberichts hängen natürlich in starkem Maße von der jeweiligen Untersuchung und den dabei angewandten Methoden ab, sodass darüber wenig generelle Aussagen gemacht werden können. Auf den Untersuchungsbericht wird deswegen im vorliegenden einführenden Lehrbuch nicht weiter eingegangen.

2.2.2 Ein Grundmodell der empirischen Marketingforschung

Im vorliegenden Abschnitt sollen für die wissenschaftliche (also auf Theorien ausgerichtete) Marketingforschung zentrale Begriffe (wie „Theorien", „Hypothesen", „Operationalisierung") erläutert und deren Zusammenhang erklärt werden. Damit sollen einerseits die (teilweise etwas abstrakten, aber keineswegs unverständlichen) allgemein gültigen gedanklichen Grundlagen empirischer Marketingforschung umrissen und andererseits die Basis für später folgende methodische Überlegungen geschaffen werden.

In einer sehr allgemeinen Weise kann man die empirische Forschung als eine von mehreren Möglichkeiten des Menschen zur Betrachtung und zum Verständnis der Marketing-Realität ansehen. Für die wissenschaftliche Betrachtungsweise von **Realität** ist es typisch, dass versucht wird, in sich widerspruchsfreie Systeme von Aussagen, die man unter bestimmten Voraussetzungen (s. u.) als **Theorie** bezeichnet, aufzustellen, deren Entsprechung zur Realität systematisch überprüft wird bzw. werden kann bzw. werden sollte. Da diese Aussagensysteme normalerweise einen Komplexitäts- und/oder Abstraktionsgrad aufweisen, der eine unmittelbare Prüfung nicht zulässt, bedient man sich dazu in der Regel geeigneter **Methoden.** Beispielsweise bedarf es für die Untersuchung des Zusammenhangs zwischen Einstellungen zu einem Produkt und Markentreue in der Regel eines recht aufwendigen Designs. Durch bloßen Augenschein kann man diese Überprüfung nicht vornehmen.

Die drei Grundelemente empirischer Forschung (Realität, Theorie, Methoden) seien zunächst kurz vorgestellt, bevor die Beziehungen dieser Elemente untereinander erläutert werden.

Realität Unabhängig vom jeweiligen Forschungsinteresse ist immer nur die Betrachtung von entsprechenden Ausschnitten der Realität möglich. Ihre vollständige Beschreibung oder gar Erklärung ist wegen einiger genereller Eigenschaften von Realität ausgeschlossen. Sie ist nach Jaccard und Jacoby (2020, S. 10 f.)

- komplex,
- dynamisch,
- (teilweise) verdeckt und
- einzigartig.

Beispiel

Diese Gesichtspunkte seien anhand eines Beispiels illustriert. Man stelle sich dazu einen Supermarkt am Nachmittag eines Werktags vor.

Komplexität: Der Versuch einer vollständigen Beschreibung dieses Supermarkts muss schnell scheitern. Eine Erfassung aller Details der hier angebotenen Produkte und der Ladenausstattung (Form der Regale, Farben, Beleuchtung etc.) überfordert jeden auch extrem geduldigen Forscher.

Dynamik: Selbst wenn es gelänge, die Einrichtung und die Produkte des Supermarkts weitgehend zu beschreiben, wäre damit wenig gewonnen, denn währenddessen verändert sich die Realität: Neue Kunden treten ein, Regale werden nachgefüllt, es wird dunkler etc.

Verdecktheit: Zahlreiche (auch wesentliche) Einzelheiten sind nicht direkt beobachtbar. Beispielsweise ist es für die Situation in dem Supermarkt wichtig, welche Bedürfnisse oder Wünsche bei den Kunden vorhanden sind, obwohl diese selbst mit anspruchsvollen Messmethoden nicht immer eindeutig feststellbar sind.

Einzigartigkeit: Da eine bestimmte Situation in dem Supermarkt mit gleichem Regalbestand, gleichen Kunden mit gleichen Wünschen etc. sich so nie wiederholt, wäre eine vollständige Beschreibung oder Erklärung auch nutzlos, weil eben keine Situation genauso wieder auftritt, in der man so detailliertes Wissen gebrauchen könnte. ◄

Theorie Wegen der skizzierten Aussichts- und Sinnlosigkeit des Versuchs, Realität vollständig zu erfassen, ist die Zielrichtung empirischer Forschung in der Wissenschaft eine ganz andere. Man bedient sich dabei bestimmter Abstraktionen einzelner Erscheinungen, die für die jeweilige Betrachtungsweise zweckmäßig sind. Diese nennt man **Konzepte** (vgl. Jaccard & Jacoby, 2020, S. 11 ff.). Ähnlich wie man sich in der Regel nicht mit der ungeheuren Vielfalt von Gegenständen mit vier Rädern und einem Motor in allen Einzelheiten befasst, sondern das Konzept „Auto“ verwendet, kann man sich in dem Supermarkt-Beispiel auch auf im jeweiligen Untersuchungszusammenhang wichtige Konzepte wie z. B. Sortimentstiefe, Verkaufsfläche oder Umsatz konzentrieren.

Hier ein Beispiel zur Verwendung von Konzepten im Alltag

Wenn eine Person frei von Beschwerden ist und keine Anzeichen für Fehlfunktionen von Organen, zu hohen Blutdruck, Herzrhythmusstörungen etc. erkennbar sind, dann spricht man davon, dass diese Person *gesund* ist. Man verwendet also das Konzept *„Gesundheit“*, um auszudrücken, dass bei dieser Person alle wesentlichen (Frage an die Leserin oder den Leser: Was ist hier wesentlich?) Funktionen des Körpers „in Ordnung“ sind und abstrahiert von allerlei Details verschiedener Personen mit den unterschiedlichsten Körperfunktionen und Messwerten.

Dabei wird schon erkennbar, dass die Kennzeichnung (Definition) beim Beispiel „Gesundheit“ keineswegs trivial und gewissermaßen automatisch einheitlich ist. So kann man unterschiedlicher Auffassung sein, ob Gesundheit sich nur auf physische Merkmale bezieht oder ob auch Aspekte des psychischen Wohlbefindens einbezogen werden müssen. Was ist damit gemeint, dass bei gesunden Menschen alle Körperfunktionen „in Ordnung“ sein sollen? Sollen die entsprechenden Messwerte den Idealwerten entsprechen oder nur in einem Bereich tolerierbarer Abweichungen von diesen Idealwerten liegen? ◄

Konzepte dienen dazu, eine Vielzahl von Objekten, Ereignissen, Ideen etc. im Hinblick auf einzelne oder mehrere gemeinsame Charakteristika und unter Zurückstellung sonstiger Unterschiede zusammenzufassen. Sie ermöglichen also eine Kategorisierung bzw. Klassifizierung und damit eine Vereinfachung des Bildes von der Realität.

Diese Kategorisierung erlaubt es, unabhängig von der Einzelsituation zu generalisieren, Beziehungen zwischen Objekten, Ereignissen etc. zu erkennen und (bei Übereinstimmung über den Gebrauch von Konzepten) Gedanken zwischen Menschen auszutauschen. Das gleiche Objekt kann – je nach Sichtweise – sehr verschiedenen Konzepten zugeordnet werden. Beispielsweise kann ein und derselbe Mensch als Stammkunde, als Ehemann, als Schlosser usw. betrachtet werden. Gelegentlich (aber nicht im vorliegenden Buch) wird zwischen Konzepten und **Konstrukten** unterschieden, wobei unter letzteren abstraktere Konzepte verstanden werden (Hildebrandt, 2008, S. 86 f.). Weil gleichen (gedanklichen) Konzepten (z. B. in verschiedenen Sprachen) verschiedene *Begriffe bzw. Worte* (z. B. „Werbung“ bzw. „Advertising“) gegenüberstehen können, wird auch hier oftmals eine gedankliche Trennung vorgenommen. Im Zusammenhang dieses Buches sind aber derartige Unterscheidungen nicht so wichtig; es wird *synonym* von Konzepten, Konstrukten und Begriffen gesprochen.

Wenn man durch Konzepte gewissermaßen „die Umwelt gedanklich vereinfacht und geordnet“ hat, kann man bestimmte Regelmäßigkeiten und Zusammenhänge entdecken. Diese können sehr konkrete (z. B. „Platzierung eines Produkts in Augenhöhe führt zu höheren Verkaufszahlen“), aber auch abstraktere Phänomene (z. B. „Kundenzufriedenheit ist eine Voraussetzung für Kundenbindung“) betreffen.

Besonders leistungsfähig sind natürlich Systeme von Aussagen, die eine größere Zahl von Konzepten und Beziehungen zwischen diesen umfassen. Ein solches Aussagesystem besteht aus drei Arten von Komponenten (Eisend & Kuß, 2023, S. 35)

- *Konzepte* mit den zugehörigen Definitionen,
- Aussagen über die *Beziehungen* zwischen diesen Konzepten und
- *Argumente* zur Begründung dieser Aussagen

und wird als **Theorie** bezeichnet.

Hintergrundinformation

Frederick Suppe (1977, S. 223) fasst die Funktion von Theorien knapp und prägnant zusammen:

> „Wissenschaftliche Theorien haben eine Menge von Phänomenen zum Gegenstand, die der *angestrebte Aussagebereich* einer Theorie sind. Die Aufgabe einer Theorie besteht darin, eine verallgemeinerte Beschreibung der Phänomene aus diesem Aussagebereich zu geben, die es ermöglicht, eine Vielzahl von Fragen zu diesen Phänomenen und den ihnen zugrunde liegenden Mechanismen zu beantworten. Diese Fragen betreffen typischerweise Prognosen, Erklärungen und Darstellungen dieser Phänomene.“
>
> Günther Schanz (1988, S. VII) zur Relevanz von Theorien:
>
> „Theorien sind die Hauptinformationsträger der wissenschaftlichen Erkenntnis (….).“

Jede Einzelaussage einer Theorie verwendet mehrere Konzepte (beispielsweise „Kundenzufriedenheit“ und „Kundenbindung“). Insofern bilden Konzepte die *Bausteine von Theorien.* Theorien sind wichtige Hilfsmittel zum Verständnis von Realität. Im Einzelnen dienen sie dazu,

- wichtige von unwichtigen Konzepten bei der Betrachtung von Ausschnitten der Realität zu trennen,
- Ausschnitte der Realität zu erklären und reale Entwicklungen zu prognostizieren,
- Kommunikation zwischen Fachleuten zu erleichtern und
- Erkenntnisfortschritt durch weitere empirische Forschung zu stimulieren.

Hier einige Beispiele zur Funktion von Theorien

Wichtige und unwichtige Konzepte trennen: Bei der Betrachtung von Kundenbindungen gilt das Konzept der Kundenzufriedenheit als *besonders wichtig.*

Ausschnitte der Realität erklären/prognostizieren: Zwischen Kontakthäufigkeit bei einer Werbung und deren Wirkung besteht offenbar ein Zusammenhang, der auch dazu dient, die Werbewirkung bei einer bestimmten Werbeplanung (Budget, Anzahl von Anzeigen, Spots, etc.) zu prognostizieren.

Kommunikation erleichtern und Erkenntnisfortschritt stimulieren: Beispielsweise gilt die Einstellungstheorie vielen Forschern als Basis für ihre Untersuchungen und ist vielfach untersucht, überprüft und diskutiert worden (siehe z. B. Kroeber-Riel & Gröppel-Klein, 2019, S. 198 ff.). ◀

Die Prozesse der Theoriebildung und -prüfung (siehe z. B. Eisend & Kuß, 2023) lassen sich anhand einer Darstellung von de Vaus (2001, S. 6) illustrieren (Abb. 2.3). Dabei wird dem Prozess der **Theoriebildung** durch Induktion der Vorgang der **Theoriebildung** durch den Test von (deduktiv) aus der Theorie abgeleiteten Hypothesen gegenübergestellt. Unter **Induktion** versteht man die Generalisierung von in der Realität beobachteten Regelmäßigkeiten. Wenn man beispielsweise bei einer Vielzahl von Werbekampagnen beobachtet, dass Bilder stärkere emotionale Wirkungen hervorrufen als Texte, dann wird man vielleicht vermuten, dass *generell* ein Zusammenhang zwischen Bildanteilen in der Werbung und emotionaler Werbewirkung besteht und entsprechende theoretische Vorstellungen entwickeln. Wenn eine Theorie vorliegt, dann besteht ein üblicher Weg zu deren Überprüfung darin, daraus Aussagen (Hypothesen) abzuleiten (→ **Deduktion**), deren Zutreffen man dadurch überprüft, dass man die auf dieser Basis *erwarteten* Ergebnisse und die *tatsächlichen* Beobachtungen gegenüberstellt. Bei weitgehender Übereinstimmung spricht man von einer Bestätigung der Theorie, anderenfalls kommt man zur Ablehnung (Falsifikation) bzw. zur Modifikation der Theorie.

In diesem Sinne sind Hypothesen also aus theoretischen Aussagen abgeleitet und sind gleichzeitig konkreter als diese. Sie ermöglichen damit den empirischen Test einer Aussage (Jaccard & Jacoby, 2020, S. 96 f.).

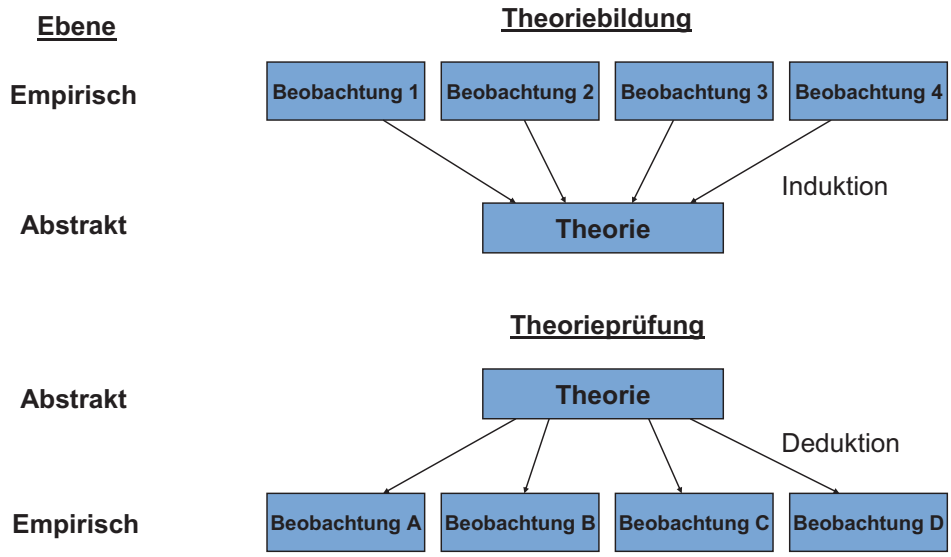


Abb. 2.3 Theoriebildung und Theorieprüfung. (Nach: de Vaus, 2001, S. 6)

Im Zusammenhang mit der Überprüfung von Theorien (oder Teilen von Theorien), aber auch bei praktischen Fragestellungen, spielen also **Hypothesen** eine bedeutsame Rolle. Man versteht darunter (zu überprüfende) Vermutungen über:

- *Ausprägungen* von Variablen (z. B. „Mindestens 10 % der Konsumenten werden das neue Produkt X probieren". „Höchstens 20 % aller Werbebotschaften werden länger als 2 Tage erinnert.")
- *Zusammenhänge* von Variablen (z. B. „Junge Konsumenten sind aufgeschlossener für aktuelle Mode". „Je positiver die Einstellung zu einem Produkt ist, desto größer ist die Kaufneigung").

Wie kommen nun derartige Hypothesen zu Stande? Ganz direkt ist die Beziehung von Hypothesen zu Theorien, wie vorstehend skizziert. Daneben können einschlägige Erfahrungen des Managements und bisherige Untersuchungen (einschließlich speziell für diesen Zweck durchgeführter *explorativer* Untersuchungen, siehe Abschn. 2.4.1) als Ausgangspunkt für Hypothesen genannt werden. Im Bereich der (eher auf Theorien ausgerichteten) Grundlagenforschung ist die Hypothesenbildung ein wesentlicher Schritt im Forschungsprozess. Bei anwendungsorientierten Untersuchungen für die Marketing-Praxis ergeben sich dagegen Untersuchungshypothesen meist mehr oder weniger direkt aus den Wünschen und Vorgaben der Auftraggeber einer Untersuchung.

Die Entwicklung von Hypothesen ist wichtig im Hinblick auf die für die Entwicklung bzw. Auswahl von Methoden erforderliche Konkretisierung zu untersuchender Fragestellungen. Wenn man beispielsweise an das oben skizzierte Beispiel einer Hypothese

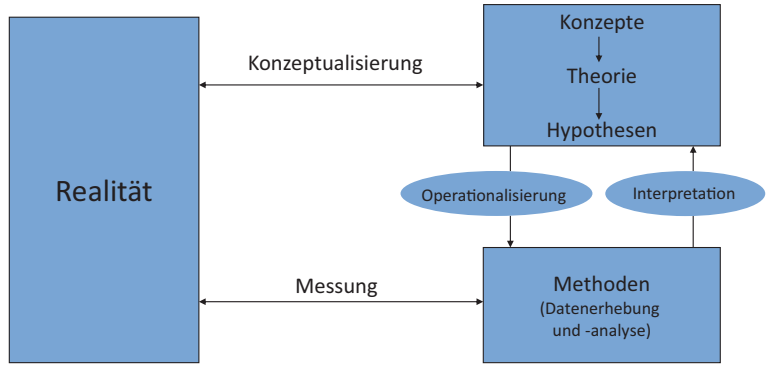


Abb. 2.4 Grundmodell der empirischen Marketingforschung

denkt, so erkennt man, dass sich daraus direkt ableiten lässt, welche Variablen (hier: „Bildanteil in der Werbung“, „Emotionale Werbewirkung“) gemessen werden müssen. Dazu benötigt man geeignete Methoden, deren Festlegung Gegenstand des nächsten gedanklichen Schrittes im Forschungsprozess ist.

Methoden Wenn Theorien oder Teile davon im Hinblick auf ihre Übereinstimmung mit der Realität getestet werden sollen, bedarf es dazu also in der Regel spezieller Methoden. Gerade bei Theorien, die Konzepte hohen Abstraktionsgrades betreffen, ist mit besonders schwierigen methodischen Problemen zu rechnen.

Es geht also darum, eine *Verbindung* zwischen den abstrakteren Elementen von **Theorien** und der **Realität** herzustellen. Man kann auch die Methoden der empirischen Marketingforschung als Hilfsmittel betrachten, um trotz aller Komplexität (s. o.) die jeweils interessierenden Aspekte der Realität beobachten zu können. Beispielsweise geben die Verfahren der Stichprobenziehung an, welche (zahlenmäßig beschränkte) Teilmenge von Untersuchungsobjekten betrachtet wird. Viele Befragungsverfahren dienen dazu, sehr unterschiedliche Personen, Meinungen, Verhaltensweisen zu Kategorien (z. B. Personen mit hohem Bildungsgrad, negativer Haltung zur Fernsehwerbung, Markentreue) zusammenzufassen oder auf Skalen einzuordnen. Die Verfahren der Datenanalyse haben u. a. den Zweck, eine große Menge von Einzeldaten zu verdichten (z. B. zu Maßzahlen oder graphischen Darstellungen).

Die drei Elemente der empirischen Marketingforschung sind in Abb. 2.4 dargestellt. Die verbindenden Pfeile kennzeichnen grundlegende Teilaufgaben im Prozess der empirischen Forschung, auf die anschließend eingegangen wird.

Als **Konzeptualisierung** bezeichnet man den Vorgang, interessierende Teile der Realität abstrahierend zu kennzeichnen und Vermutungen über die Beziehungen dieser Elemente von Theorien aufzustellen. Dabei kann es sich um einen kreativen Prozess der Theoriebildung oder um die Anwendung existierender Theorien auf ein gerade interessierendes Problem handeln. Bei der Konzeptualisierung wird gewissermaßen in zwei Richtungen vorgegangen: Einerseits findet eine Abstrahierung von der Realität statt,

andererseits bestimmt diese Abstrahierung auch die Betrachtungsweise der Realität (siehe dazu auch Hildebrandt, 2008).

Hintergrundinformation
Deborah MacInnis (2011) hat Wesen und Prozess der Konzeptualisierung umfassend diskutiert und kommt zu folgender Kennzeichnung (S. 140):

> „Konzeptualisierung ist der Prozess des abstrakten Verständnisses einer Situation oder eines Problems durch die Identifizierung von Zusammenhängen und Regelmäßigkeiten sowie jeweils typischer Eigenschaften."

Ergebnisse des Konzeptualisierungsvorgangs sind also wohldefinierte Konzepte und darauf aufbauende theoretische Vorstellungen (z. B. „Kontakt zur Werbung führt zu positiveren Einstellungen und dann zu entsprechenden Kaufabsichten."). Die Formulierung entsprechender Hypothesen (s. o.) erlaubt dann die empirische Überprüfung einer Theorie in den folgenden Schritten des Forschungsprozesses.

Hier ein Beispiel zur Konzeptualisierung

Kundenzufriedenheit wird meist durch die Übereinstimmung von Erwartungen des Kunden gegenüber einem Produkt mit den Erfahrungen des Kunden nach dem Kauf gekennzeichnet (siehe z. B. Homburg, 2008, S. 20 ff.).

Man stelle sich für einige Kunden deren Erwartungen und Erfahrungen vor. Es folgen einige Beispiele dafür:

Erwartung	Erfahrung
„Auto X ist zuverlässig."	„Der Wagen muss oft in die Werkstatt."
„Auto Y ist komfortabel."	„Das Auto ist meist zu laut."
•	•
•	•
•	•
„Auto Z ist sehr langlebig."	„An den Türen sind schon Rostflecken."

In allen genannten Fällen stimmen Erwartungen und Erfahrungen nicht überein. Es ist also offenbar Unzufriedenheit (und nicht Zufriedenheit) der Kunden entstanden. Welcher gedankliche Schritt ist erfolgt? Man hat von den Einzelheiten (Zuverlässigkeit, Komfort, Langlebigkeit etc.) *abstrahiert* und nur noch das Maß der Übereinstimmung von Erwartungen (welcher auch immer) und Erfahrungen betrachtet. Andererseits bestimmt das Interesse am (eher abstrakten) Konzept „Kundenzufriedenheit" wegen seiner Bedeutung für den dauerhaften Markterfolg den hier interessierenden Ausschnitt der Realität, der hinsichtlich des Verhaltens von Kunden nach dem Kauf betrachtet wird. ◄

Zur Überprüfung von Theorien auf ihre Übereinstimmung mit der Realität sind geeignete Methoden auszuwählen. Beispielsweise muss man entscheiden, mit welcher Skala man Einstellungen misst, die man vielleicht als Ursache für Kaufverhalten (wie zu messen?) ansieht. Ein statistisches Verfahren muss gewählt werden, mit dem man die vermutete Beziehung zwischen Einstellungen und Kaufverhalten überprüfen kann. Diesen ganzen Vorgang nennt man **Operationalisierung**, weil man Vorgehensweisen („Operationen") festlegt, mit denen man den interessierenden Sachverhalt erfassen kann. Hier werden also abstrakten Konzepten konkrete Messverfahren, statistische Verfahren etc. zugeordnet. Damit verbunden ist in der Regel auch eine Einengung recht allgemeiner Konzepte auf konkrete Untersuchungsgegenstände. So kann man wohl kaum ganz allgemein den Zusammenhang zwischen Einstellungen und Verhalten empirisch untersuchen, sondern muss sich auf deutlich konkretere – und damit weniger allgemeine, aber eher beobachtbare – entsprechende Zusammenhänge konzentrieren (z. B. den Zusammenhang „Einstellung zu einer bestimmten Marke" → „Kaufverhalten hinsichtlich dieser Marke").

Ein Beispiel zur Operationalisierung

Der Zusammenhang zwischen Kundenzufriedenheit und Wiederkauf-Wahrscheinlichkeit soll untersucht werden. Die Operationalisierung besteht z. B. darin festzulegen, dass dieser Zusammenhang durch eine Befragung bei 1000 repräsentativ ausgewählten KonsumentInnen ermittelt werden soll, dass die Kundenzufriedenheit durch eine Rating-Skala und die Wiederkauf-Wahrscheinlichkeit durch eine verbale Skala mit den Werten „sicher nicht", „extrem unwahrscheinlich", …„sehr wahrscheinlich", „sicher" gemessen werden sollen und für die Bestimmung des Zusammenhanges ein Korrelationsmaß für ordinalskalierte (siehe Abschn. 7.2) Daten verwendet werden soll. ◀

Die Anwendung der ausgewählten Verfahren in der Realität bezeichnet man als **Messung.** Auch dieser Vorgang ist ein zweiseitiger: Versuchspersonen, Objekte etc. werden mit Messinstrumenten konfrontiert; Messwerte (Daten) fließen zurück. Je nach Untersuchungsgegenstand und Messverfahren werden die Untersuchungsobjekte (z. B. KonsumentInnen) bestimmten Gruppen/Kategorien (z. B. weiblich/männlich; StudentIn, Lehrling etc.) zugeordnet oder ein Messwert (z. B. monatl. Einkommen; Präferenzstärke hinsichtlich verschiedener Urlaubsziele) wird für jedes Untersuchungsobjekt festgestellt (siehe dazu Abschn. 7.2 „Messniveau von Daten").

▶ **Definition** Nunnally und Bernstein (1994, S. 3) definieren: „Messungen bestehen aus Regeln für die Zuordnung von Symbolen zu Objekten dergestalt, dass (1) quantifizierbare Eigenschaften numerisch repräsentiert werden (Skalierung) oder (2) definiert wird, ob Objekte in gleiche oder verschiedene Kategorien im Hinblick auf eine bestimmte Eigenschaft gehören (Klassifikation)."

Durch die Operationalisierung wird also eine Verbindung zwischen theoretischen Konzepten und beobachtbaren Sachverhalten hergestellt. Das führt zu **Variablen,** die ein spezifisches Merkmal oder Verhalten (z. B. Monatseinkommen, Kaufhäufigkeit) von Objekten (Personen, Organisationen etc.) sind, das unterschiedliche Ausprägungen haben kann, die wiederum durch Messungen festgestellt werden. Bei der Operationalisierung werden also gleichzeitig die in der jeweiligen Untersuchung verwendeten Variablen festgelegt (Eisend & Kuß, 2023, S. 194). Durch die Datenerhebung entsteht dann eine sogenannte *Datenmatrix,* deren Zeilen üblicherweise für die Untersuchungsobjekte stehen und deren Spalten jeweils einer Variablen entsprechen (siehe dazu Abschn. 4.5.3). Bereits hier sei hervorgehoben, dass die Frage der Entsprechung von theoretischen Konzepten und gemessenen Variablen einen der zentralen Problembereiche empirischer Sozial- und Marketingforschung darstellt. Wenn man z. B. das Konzept „Bildung" messen will und dafür eine Variable „Schulbildung" („Abitur", „Mittl. Reife" etc.) heranzieht, dann kann man sich schon ernsthaft fragen, ob diese Messung adäquat ist. Das Konzept „Bildung" hat – je nach Sichtweise – vielleicht mehr und andere Facetten als nur die Art des Schulabschlusses. Damit ist hier erstmals das Problem der *Validität* angesprochen, das – beginnend im folgenden Abschnitt – im Rahmen dieses Buches noch ausführlicher beleuchtet wird.

Messungen in der Marktforschung sind mit einigen spezifischen Problemen verbunden, die in anderen Bereichen der Anwendung wissenschaftlicher Messmethoden (z. B. Physik, Astronomie, Chemie) so nicht auftreten (Chang & Cartwright, 2008):

- Ein sehr großer Anteil der Messungen in der Marktforschung erfolgt gewissermaßen *„indirekt"*. Damit ist gemeint, dass oftmals verbale Angaben (in Fragebögen) von ManagerInnen, KonsumentInnen etc. zu den eigentlich interessierenden Untersuchungsgegenständen (z. B. Produktqualität, Mediennutzung, Verkaufszahlen) erhoben werden. Nun können solche Angaben (z. B. durch mangelnde Erinnerung oder persönliche Interessen) deutlich verzerrt sein (siehe dazu Kap. 4).
- Sehr viele Messungen in der Marktforschung sind „aufdringlich", d. h. dass die jeweilige Versuchs- oder Auskunftsperson – typischerweise bei Befragungen – die Datenerhebung bemerkt, was wiederum zu einer systematischen Veränderung des Verhaltens bei der Beantwortung von Fragen führen kann. Bei einem solchen Vorgang spricht man von „Reaktivität**",** weil der Messvorgang den Untersuchungsgegenstand beeinflusst. Typische Beispiele sind Fragen, bei denen gesellschaftliche Normen (z. B. Umweltbewusstsein) oder individuelle Werte (z. B. persönlicher Erfolg) angesprochen werden.
- Gerade bei repräsentativen Untersuchungen mit großen Stichproben gibt es bestimmte Einschränkungen bei den Messmethoden der Marktforschung. Zunächst müssen die Messungen breit und unkompliziert anwendbar sein, weil sie bei unterschiedlichsten Zielpersonen in unterschiedlichsten Situationen „funktionieren" sollen. Die Anforderungen an diese Personen müssen sich in Grenzen halten, weil die Untersuchungsteilnahme in aller Regel freiwillig ist. Weiterhin sind ethische Maßstäbe zu

respektieren, die es verbieten, Versuchs- oder Auskunftspersonen starkem Stress auszusetzen oder in deren Intimsphäre einzudringen. Zum Vergleich: Ein Chemiker, der bestimmte Kunststoffe untersucht, kennt solche Begrenzungen kaum.

Die beim Messvorgang entstehenden Messwerte für die untersuchten Merkmale bei allen Untersuchungsobjekten (z. B. Personen) bilden den *Datensatz* einer Untersuchung. Erhobene Daten werden mit statistischen Methoden verdichtet, dargestellt und weitergehend analysiert. Den Vergleich von Ergebnissen der Datenanalyse mit den Aussagen der Theorie und die damit verbundenen Überlegungen nennt man **Interpretation**. Dabei stellt man fest, ob die Theorie bzw. Teile damit bestätigt wurden oder nicht und ob Modifizierungen der Theorie vorgenommen werden müssen. Insbesondere bei praktischen Anwendungen der Marktforschung geht die Interpretation deutlich über diesen strikten Theoriebezug hinaus. Hier sind auch kreative Überlegungen (z. B. im Hinblick auf die Definition von Zielgruppen) auf Basis der erhobenen Daten üblich.

2.3 Anforderungen an Marktforschungsuntersuchungen: Reliabilität, Validität, Generalisierbarkeit

Sowohl für den im Abschn. 2.2.1 dargestellten Untersuchungsablauf der Marktforschungspraxis als auch für das Grundmodell der empirischen Marketingforschung gilt, dass **Untersuchungsergebnisse,** die einem Untersuchungsproblem bzw. einer Hypothese entsprechen sollen, natürlich nur aussagekräftig sein können, wenn die Datenerhebung und Datenanalyse (mit Stichprobenziehung, Messungen, Datenaufbereitung etc.) *tatsächlich den zu untersuchenden Phänomenen gerecht werden.* Das mag trivial klingen. Wer würde schon, wenn er sein Körpergewicht messen will, ein Zentimetermaß verwenden? Letztlich geht es um die Frage, ob Untersuchungsergebnisse der im jeweiligen Fall untersuchten Realität möglichst weitgehend entsprechen, ob die Ergebnisse also in diesem Sinne (annähernd) *wahr* sind. In der einschlägigen Literatur spricht man in diesem Zusammenhang von der Verlässlichkeit (**Reliabilität**) und Gültigkeit **(Validität)** von Untersuchungsergebnissen.

Daneben stellt sich die Frage, inwieweit Ergebnisse, die oft nur auf Angaben weniger hundert Auskunftspersonen in einer oftmals recht künstlichen Untersuchungssituation beruhen, aussagekräftig für das reale Denken, Fühlen, Handeln etc. von Millionen von KonsumentInnen, WählerInnen etc. sind. Es geht also um das Problem der **Generalisierbarkeit** von Untersuchungsaspekten. Darauf wird im letzten Teil dieses Abschnitts etwas näher eingegangen.

Hintergrundinformation

Kerlinger und Lee (2000, S. 474) kennzeichnen die Relevanz des Generalisierbarkeitsproblems:

> „Können wir die Ergebnisse einer Untersuchung im Hinblick auf andere Teilnehmer, andere Gruppen oder andere Bedingungen generalisieren? Vielleicht ist die Frage so besser formuliert: In welchem Maße können wir die Ergebnisse der Untersuchung generalisieren? Dieses ist wahrscheinlich die komplexeste und schwierigste Frage, die bezüglich einer Untersuchung gestellt werden kann, weil sie nicht nur technische Aspekte betrifft (wie Stichprobenziehung oder Untersuchungsdesign), sondern wesentliche Probleme von Grundlagenforschung und angewandter Forschung."

Bei sozialwissenschaftlichen Messungen ist das Problem von Reliabilität und Validität alles andere als trivial. Jeder kennt die laufenden Umfragen zu Parteipräferenzen und Wahlverhalten. Aber kann man tatsächlich von geäußerten Präferenzen auf späteres Verhalten schließen? Was sagt eine heute geäußerte Parteipräferenz über tatsächliches Wahlverhalten einige Wochen oder Monate später aus? Entsprechende Probleme entstehen in der Marktforschung: Wenn ein Konsument äußert, dass er eine Marke „gut findet", kann man dann tatsächlich daraus schließen, dass er sie auch (immer, meist, gelegentlich?) kaufen wird? Kann man von der Angabe von Konsumenten zu der beim letzten Einkauf gekauften Waschmittelmarke auf die tatsächlich gekaufte Marke schließen oder muss man damit rechnen, dass Erinnerungslücken, Anpassungen an die Erwartungen eines Interviewers oder bewusst geäußerte Falschangaben hier zu Messfehlern führen? Es geht also im Wesentlichen darum, ob die Umsetzung einer Problemstellung in ein Untersuchungsdesign (mit Stichprobenziehung, Messmethoden etc.) und dessen Realisierung angemessen, also der Problemstellung entsprechend ist. Dabei geht es im Grunde um diese beiden Aspekte:

- Führt die Untersuchung mit allen ihren methodischen Einzelheiten zu einer *systematischen* Abweichung vom „wahren Wert" des zu untersuchenden Gegenstandes? Beispiel: Führt die Messung des Alkoholkonsums in der Bevölkerung durch eine entsprechende Befragung zu einer systematisch zu niedrigen Einschätzung, weil viele Menschen (wegen der eher geringen sozialen Akzeptanz von Alkoholkonsum) tendenziell zu niedrige Angaben über ihren eigenen Alkoholkonsum machen?
- Wird das Untersuchungsergebnis durch *Zufälligkeiten* (und Nachlässigkeiten) bei der Untersuchungsdurchführung beeinflusst? Beispiel: Kann es sein, dass der Befragungszeitpunkt (morgens oder abends) zu unterschiedlichen Angaben der Auskunftspersonen zu ihren Präferenzen hinsichtlich (alkoholischer) Getränken führt?

Damit kommt man zu den beiden grundlegenden Kriterien für die Qualität und Aussagekraft von Untersuchungen (nicht nur) der Marktforschung: **Validität,** die sich auf (nach Möglichkeit nicht vorhandene oder sehr geringe) *systematische* Abweichungen des Untersuchungsergebnisses von der Realität bezieht, und **Reliabilität,** bei der es um die Unabhängigkeit eines Untersuchungsergebnisses von einem (von verschiedenen *Zufälligkeiten* der jeweiligen Untersuchungssituation beeinflussten) einmaligen Messvorgang geht. Bei hoher Reliabilität, also bei geringen situativen Einflüssen, müssten gleichartige Messungen zu gleichen (zumindest sehr ähnlichen) Ergebnissen führen.

▶ **Definition** Die **Validität** eines Untersuchungsergebnisses lässt sich also folgendermaßen kennzeichnen: Ein Untersuchungsergebnis wird als valide (gültig) angesehen, wenn es den Sachverhalt, der ermittelt werden soll, tatsächlich wiedergibt.

Auch die **Reliabilität** sei charakterisiert: Als Reliabilität bezeichnet man die Unabhängigkeit eines Untersuchungsergebnisses von einem einmaligen Untersuchungsvorgang und den jeweiligen situativen (zufälligen) Einflüssen.

Hintergrundinformation
David de Vaus (2002) charakterisiert die Relevanz von Reliabilität und Validität:

> *Reliabilität:* „Wenn wir uns nicht auf die Antworten zu Fragen aus dem Fragebogen verlassen können, dann ist jede Analyse auf der Grundlage solcher Daten problematisch. Wenn die Ergebnisse, die wir auf Basis einer Stichprobe erhalten, genauso gut anders sein könnten, wenn wir die Befragung erneut durchführen, wie viel Vertrauen sollen wir zu diesen Ergebnissen haben?“ (S. 17)

> *Validität:* „Weil die meisten sozialwissenschaftlichen Untersuchungen relativ konkrete Messungen für abstraktere Konzepte verwenden, stehen wir vor der Frage, ob unsere Messinstrumente tatsächlich das messen, was wir glauben. Dieses ist das Problem der Validität. Wir müssen uns irgendwie darauf verlassen können, dass unsere relativ konkreten Fragen tatsächlich die Konzepte treffen, für die wir uns interessieren.“ (S. 25)

Bedeutung und Zusammenhang von Validität und Reliabilität lassen sich in Anlehnung an Churchill (1979) durch eine einfache Formel illustrieren

$$X_B = X_w + F_s + F_z$$

mit

X_B	=	Gemessener, beobachteter Wert
X_w	=	„wahrer“ (normalerweise nicht bekannter) Wert des zu messenden Konstrukts
F_s	=	Systematischer Fehler bei einer Messung (z. B. durch Frageformulierungen, die eine bestimmte Antworttendenz begünstigen)
F_Z	=	Zufälliger Fehler bei einer Messung (z. B. durch situative, kurzfristig veränderliche Faktoren, wie Interviewereinfluss, Zeitdruck etc., die längerfristig konstante Meinungen, Absichten, Präferenzen etc. überlagern)

Eine Messung wird als *valide* angesehen, wenn keine systematischen und keine zufälligen Fehler vorliegen. Es gilt dann:

$F_s = 0$ und $F_Z = 0$ und deswegen $X_B = X_w$

Aus der Reliabilität einer Messung ($F_Z = 0$) muss also keineswegs folgen, dass die Messung auch valide ist, da ja $F_s \neq 0$ sein kann.

Die grundlegende Bedeutung von Reliabilität und Validität für empirische Untersuchungen dürfte leicht einsehbar sein. Wenn diese Anforderungen nicht erfüllt sind, dann spiegeln die Untersuchungsergebnisse eben nicht die Realität wider und haben deswegen keine Aussagekraft bzw. sind zur Vorbereitung und Unterstützung von Marketing-Entscheidungen unbrauchbar. Die vorstehend umrissene Aussage, dass die *Reliabilität eine notwendige – aber keineswegs hinreichende – Voraussetzung der Validität* ist, lässt sich leicht nachvollziehen, wenn man bedenkt, dass Untersuchungsergebnisse mit geringer Reliabilität bei Wiederholungen starken Schwankungen unterworfen sind, dass es also gewissermaßen einen „Glücksfall" darstellt, unter diesen Umständen den „wahren Wert" hinreichend genau zu treffen.

Beispiel

Wesen und Zusammenhang von Validität und Reliabilität lassen sich durch ein sehr, sehr einfaches (nicht aus der Marktforschung stammendes) Beispiel illustrieren. Man stelle sich vor, dass nicht ein Marktanteil, Bekanntheitsgrad o.ä. gemessen werden soll, sondern ganz einfach das Körpergewicht einer Person. Dazu werden natürlich keine Fragebögen, Stichproben etc. benötigt, sondern eine simple Badezimmer-Waage. Das „wahre" Körpergewicht der Person sei bekannt (80 kg). Die Person tritt sechsmal kurz nacheinander auf die Waage und erhält mehr oder weniger unterschiedliche Messergebnisse. Derartige Ergebnisse sind für unterschiedliche Qualitäten des „Messinstruments Waage" in der Abb. 2.5 eingetragen.

Wie lassen sich die unterschiedlichen Ergebniskonstellationen interpretieren?

Feld links oben: Die Messergebnisse weichen vom wahren Wert 80 kg kaum, jedenfalls nicht systematisch, ab. Geringe Unterschiede von Messung zu Messung sind vielleicht durch eine etwas ausgeleierte oder angerostete Feder in der Waage zu erklären, beeinträchtigen die Aussagekraft des Ergebnisses aber nur wenig. Hier sind also hohe Reliabilität und hohe Validität gegeben.

Feld links unten: Auch hier schwanken die verschiedenen Ergebnisse nur wenig (hohe Reliabilität), aber die Waage ist offenbar systematisch verstellt und zeigt immer etwa 20 kg zu viel an (geringe Validität). Hier zeigt sich, dass Reliabilität nur notwendige, nicht hinreichende Voraussetzung der Validität ist.

Feld rechts unten: Hier geht alles schief. Die Waage zeigt um die 20 kg zu viel an (geringe Validität) und außerdem schwanken die Ergebnisse wegen einer ausgeleierten Feder sehr stark (geringe Reliabilität). Eine Person, die sich mit einer Diät quält und eine solche Waage benutzt, wäre wohl zu bedauern.

Feld rechts oben: Besonders interpretationsbedürftig ist dieser Fall. Die Messwerte schwanken stark (geringe Reliabilität) um den richtigen Wert von 80 kg (scheinbar hohe Validität). Da man aber normalerweise nicht mehrere Messungen durchführt, sondern nur einen Messwert (z. B. 74,2 oder 85,6 kg) verwendet, muss man wegen der geringen Reliabilität damit rechnen, dass dieser deutlich vom wahren Wert abweicht und die Messung deswegen nicht valide ist. Die durch die Tabelle nahe gelegte Aussage einer hohen Validität ist also falsch, was durch die Durchstreichung dieses Tabellenfeldes gekennzeichnet ist. ◀

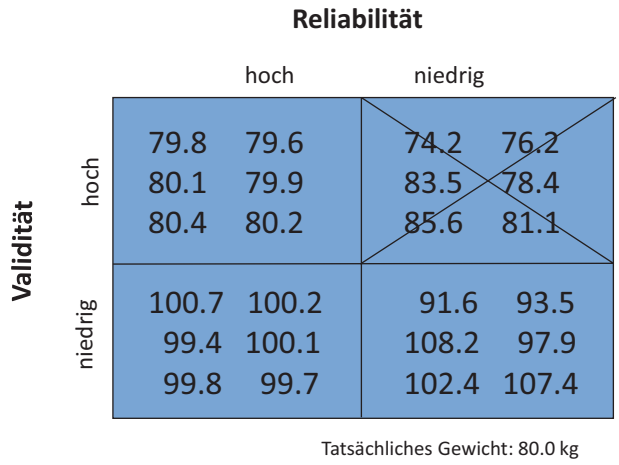


Abb. 2.5 Ein sehr einfaches Beispiel zu Reliabilität und Validität von Messungen

In späteren Teilen dieses Lehrbuchs werden Validität und Reliabilität insbesondere bei der Entwicklung von Messinstrumenten der Marktforschung (Frageformulierung, Skalenentwicklung) eine wesentliche Rolle spielen. Viele Beispiele (siehe Abschn. 4.3) belegen, dass missverständliche, unklare, tendenziöse oder die Auskunftsperson überfordernde Fragetechniken zu (erstaunlich großen) Abweichungen eines Marktforschungsergebnisses vom „wahren Wert" führen können.

Die vorstehend skizzierten Ideen lassen sich zusammenfassen, indem man versucht zu kennzeichnen, was man unter „Validierung" verstehen kann. In Anlehnung an Jacoby (2013, S. 218) kann man unter **Validierung** den *Ausschluss alternativer Erklärungsmöglichkeiten* für ein Untersuchungsergebnis verstehen. Was ist mit dieser zunächst etwas abstrakt wirkenden Kennzeichnung gemeint? Man stelle sich vor, eine Untersuchung habe zu einem bestimmten Ergebnis geführt, beispielsweise zu dem Ergebnis, dass ein hoher Bildanteil in der Werbung dazu führt, dass diese besser erinnert wird. Wenn man diese Untersuchung methodisch gründlich und sorgfältig durchgeführt hat, wenn man also ausschließen kann, dass das Untersuchungsergebnis ein Artefakt von verzerrenden Fragetechniken, nicht repräsentativer Stichprobenauswahl etc. ist, wenn man also alle derartigen alternativen Erklärungsmöglichkeiten ausschließen kann, dann kann das Ergebnis offenkundig nur dadurch zu Stande gekommen sein, dass die Verhältnisse in der Realität *tatsächlich* so sind und sich unverzerrt in dem Ergebnis widerspiegeln. Ein solches Untersuchungsergebnis bezeichnet man also als valide.

Für eine Zusammenfassung der Überlegungen in diesem Abschnitt sei noch einmal auf das in Abb. 2.4 vorgestellte „Grundmodell der empirischen Marketingforschung" zurückgegriffen. Bei der Betrachtung der Beziehungen zwischen dessen Elementen erkennt man in Abb. 2.6 bestimmte Schwerpunkte. Wenn die Methoden möglichst gut den Elementen der Theorie entsprechen, wenn also hier nur geringe systematische Abweichungen und Fehler existieren, dann spricht man von der *Validität* einer Untersuchung. Die Anwendung von (validen) Methoden auf Phänomene aus der Realität kann

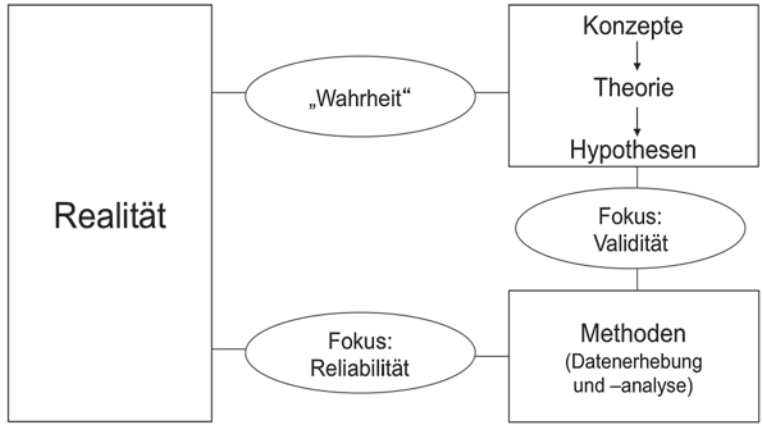


Abb. 2.6 Validität, Reliabilität und Wahrheit im Grundmodell empirischer Marketingforschung

aber durch vielfältige Störungen und zufällige Einflüsse beeinträchtigt werden. Wenn eine Methode relativ robust gegenüber solchen Faktoren ist, wenn das Ergebnis einer Messung also relativ unabhängig vom einzelnen Messvorgang und damit wiederholbar ist, liegt *Reliabilität* vor. Neu hinzugekommen ist hier der in der Wissenschaftstheorie höchst umstrittene Begriff „Wahrheit“. In der Sichtweise von Shelby Hunt (2010, S. 287) wird **Wahrheit** folgendermaßen gekennzeichnet: „Wenn man mit irgendeiner Theorie konfrontiert wird, dann stelle man die grundlegende Frage: *Ist die Theorie wahr?* Weniger knapp gesagt: In welchem Maße ist die Theorie übereinstimmend mit der Realität? Ist die reale Welt tatsächlich so aufgebaut, wie es die Theorie unterstellt oder nicht?“. In dieser Perspektive geht es also um die Übereinstimmung von Theorie und Realität, wie in Abb. 2.6 dargestellt. In der Wissenschaftstheorie spricht man bei diesem Ansatz von der „Korrespondenztheorie“ der Wahrheit (Eisend & Kuß, 2021, S. 49).

Nun zum Aspekt der **Generalisierbarkeit** von Untersuchungsergebnissen. Allgemein strebt man natürlich eine weitreichende Generalisierbarkeit an. Bei anwendungsorientierten Untersuchungen für die Marketing-Praxis und bei theorieorientierten Studien der Grundlagenforschung gibt es allerdings etwas unterschiedliche Schwerpunkte. In der Grundlagenforschung ist man typischerweise an Aussagen interessiert, die (möglichst weitgehend) generalisiert werden können, die also deutlich über den Einzelfall hinaus Bedeutung haben, weil eben die entsprechenden Theorien einen eher allgemeinen Charakter haben sollen. Bei Untersuchungen für praktische Zwecke (z. B. Test einer Verpackung) gilt dieses Ziel nur begrenzt. Man ist dort in der Regel eher am jeweiligen Spezialproblem (z. B. Auswahl der geeignetsten Packung für ein bestimmtes Produkt) und weniger an generellen Aussagen (z. B. „Werden Kaufentscheidungen durch Verpackung stärker beeinflusst als durch Verkaufsförderung?“) interessiert. Gleichwohl ist zumindest im Hinblick auf zwei Aspekte auch für die Praxis die Möglichkeit zur Generalisierung eines entsprechenden Untersuchungsergebnisses ganz wesentlich:

- Das Ergebnis muss von der (relativ kleinen) Zahl von Untersuchungsteilnehmern (Auskunfts- oder Versuchspersonen) auf die interessierende *Grundgesamtheit* (z. B. Millionen von Kunden in einem Marktsegment) generalisiert werden können. Maßgeblich dafür ist die Stichprobenziehung (siehe Abschn. 4.2).
- Weiterhin muss es möglich sein, von Auskünften und Verhaltensweisen der untersuchten Personen in der (oft etwas künstlichen) Untersuchungssituation (z. B. bei standardisierten Befragungen oder bei Laborexperimenten, siehe Abschn. 6.2) auf *reales Verhalten* beim Einkauf, bei der Mediennutzung etc. schließen zu können. Dieses Problem wird in der Literatur oft unter dem Stichwort „externe Validität" erörtert (siehe Abschn. 6.2).

Im Zusammenhang mit eher theorieorientierten Untersuchungen ist im Abschn. 2.2.2 beim Stichwort „Operationalisierung" erläutert worden, dass es für die Überprüfung von Theorien in der Realität notwendig ist, den (abstrakten) theoretischen Konzepten durch den Einsatz entsprechender Methoden konkrete Messungen zuzuordnen und die Ergebnisse dieser Messungen im Hinblick auf die verwendeten Hypothesen zu analysieren. Der Prozess der Operationalisierung ist also gleichzeitig ein Prozess der Konkretisierung und damit der Einengung der Untersuchung. Beispielsweise wird auf diesem Weg aus einer allgemeinen Frage nach dem Zusammenhang von Involvement und Informationsnachfrage vor dem Kauf eine konkrete Untersuchungsfrage zum Zusammenhang zwischen Involvement *gegenüber bestimmten Produkten* und der Informationsnachfrage *beim Kauf dieser Produkte.* Darüber hinaus wird die betreffende Untersuchung zu einem bestimmten Zeitpunkt, in einem bestimmten kulturellen Umfeld, mit bestimmten Methoden etc. durchgeführt. Es stellt sich die Frage, welche Aussagekraft eine solche konkrete Untersuchung für die allgemeinere Fragestellung hat, die am Anfang stand. Aus entsprechenden Überlegungen ergibt sich die folgende **Definition der Generalisierbarkeit:**

▶ **Definition** Die Generalisierbarkeit von Untersuchungsergebnissen bezieht sich auf die Frage, inwieweit von einem bestimmten Ergebnis auf andere *Objekte* (z. B. Stichprobe → Grundgesamtheit), *Untersuchungsgegenstände* (z. B. Einstellung zu einem Produkt → Einstellungen generell), *Kontexte* (z. B. EU-Länder → Zentral-Afrika) einschließlich anderer Zeitpunkte mit ihren jeweiligen Kontexten und mögliche Ergebnisse bei Anwendung anderer *Methoden* (z. B. Labor- → Feldexperiment) geschlossen werden kann.

Wenn man von einem einzelnen Untersuchungsergebnis zu (ganz) generellen Aussagen kommen will, dann muss man die folgenden fünf Fragen positiv beantworten können:

- Lassen sich die Ergebnisse von der relativ geringen Zahl untersuchter Objekte (z. B. Personen, Verkaufsstellen) auf die entsprechende Grundgesamtheit übertragen? Im Hinblick auf diese Fragestellung nutzt man in der Forschungspraxis standardmäßig das Instrumentarium der Stichprobenziehung (siehe Abschn. 4.2) und der Inferenzstatistik (siehe Kap. 8).

- Lassen sich die Ergebnisse im Hinblick auf andere (entsprechende) Untersuchungsgegenstände verallgemeinern?
- Lassen sich die Ergebnisse auf andere Kontexte (z. B. anderes kulturelles oder soziales Umfeld; andere Märkte) übertragen? Wenn man nach der Übertragbarkeit von Ergebnissen auf andere Zeitpunkte fragt, dann meint man damit, dass sich im Zeitablauf der jeweilige Kontext geändert haben könnte.
- Erhält man bei der Anwendung anderer Methoden entsprechende Ergebnisse oder sind die Ergebnisse von der in der jeweiligen Studie angewandten Methode beeinflusst?

In der wissenschaftlichen Marketingforschung werden vor allem die folgenden Ansätze verfolgt, um die Generalisierbarkeit von Aussagen zu prüfen:

- Durchführung von **Replikationsstudien**
  Unter Replikationsstudien versteht man Wiederholungen von Untersuchungen, die sich nicht im Untersuchungsgegenstand, meist aber hinsichtlich einiger Aspekte der Vorgehensweise von der Originalstudie unterscheiden. Dadurch erreicht man eine größere Unabhängigkeit der Ergebnisse von den Stichprobenfehlern, den Spezifika der Untersuchungsmethoden, den Einflüssen einzelner Personen und zumindest vom Untersuchungszeitpunkt. Für eine etwas detailliertere Darstellung sei auf Eisend/Kuß (20) und Nosek et al. (2022) verwiesen.
- **Meta-Analysen**
  Als Meta-Analysen bezeichnet man „quantitative Methoden der Zusammenfassung und Integration einer Vielzahl empirischer Befunde zu einem bestimmten Problem oder Phänomen“ (Franke, 2002, S. 233). Informativ ist auch die Kurz-Bezeichnung „Analyse von Analysen“. Man geht dabei so vor, dass man möglichst viele (im Idealfall alle) einschlägigen empirischen Ergebnisse zusammenfasst und unter Berücksichtigung der unterschiedlichen Stichprobengrößen und Effektstärken (z. B. Größe der Korrelationskoeffizienten) gewissermaßen ein „gemeinsames“ Ergebnis berechnet. Die Unterschiedlichkeit der verwendeten Studien gilt dabei auch als Vorteil, weil auf diese Weise das Gesamtergebnis unabhängiger von den Spezifika einzelner Studien wird bzw. der Einfluss der Unterschiede der Studien geprüft werden kann. Zu Einzelheiten der Methode existiert inzwischen umfangreiche Literatur; knappe Einführungen bieten Eisend (2009) und Eisend und Kuß (2023); eine umfassende Darstellung von Grundlagen und Methoden der Meta-Analyse findet man z. B. bei Borenstein et al. (2009) und Eisend (2014).

Fazit

Bei *Reliabilität* und *Validität* von Untersuchungen steht die *„Wahrheit“* der Untersuchungsergebnisse im Mittelpunkt, also der Aspekt, dass die Ergebnisse die Realität im Wesentlichen korrekt widerspiegeln.

Bei der *Generalisierbarkeit* geht es um die *Übertragbarkeit* von Untersuchungsergebnissen auf andere Personen(-gruppen), Kontexte, Zeitpunkte etc.

2.4 Untersuchungsziele und -designs

2.4.1 Untersuchungsziele

An die im Abschn. 2.2.1 schon erläuterte Problemdefinition, mit der der Zweck einer Untersuchung bereits grob (und treffend!) umrissen werden soll, schließt sich also die Festlegung von Untersuchungszielen an, die so genau sein sollen, dass davon ausgehend ein Untersuchungsdesign und später Messinstrumente entwickelt werden können. Die Art des Untersuchungsproblems und das Ausmaß vorhandenen problembezogenen Vorwissens bestimmen im Wesentlichen den Typ (explorativ, deskriptiv, kausal) der zu planenden Untersuchung.

Explorative Untersuchungen
Wenn über das interessierende Problem vor Beginn der Untersuchung wenige Informationen vorliegen, z. B. weil das Management nicht auf einschlägige Erfahrungen zurückgreifen kann oder weil dazu noch keine Ergebnisse früherer Studien vorliegen, so wird eine *explorative* Untersuchung angemessen sein. Explorative Untersuchungen dienen vor allem dazu,

- die für ein Problem überhaupt relevanten Einflussfaktoren zunächst zu identifizieren,
- Zusammenhänge zwischen Variablen zu entdecken,
- das Untersuchungsproblem zu präzisieren,
- die Vertrautheit des Forschers mit dem Untersuchungsgegenstand wachsen zu lassen,
- eine komplexere Fragestellung in übersichtliche (und der Forschung besser zugängliche) Einzelfragen aufzubrechen,
- anschließende (deskriptive oder kausale) Untersuchungen (z. B. durch die Generierung von Hypothesen) vorzubereiten und entsprechende Prioritäten zu setzen.

Der Begriff „explorativ“ (= erkundend, entdeckend) drückt die zentrale Idee aus; der Schwerpunkt liegt „bei der Gewinnung von Ideen und Einsichten“. (Iacobucci & Churchill, 2010, S. 58) In diesem Sinne kann man explorative Untersuchungen auch als Hilfsmittel im Prozess der Konzeptualisierung (siehe Abschn. 2.2.2) ansehen.

Hintergrundinformation
Die beiden Genforscher Patrick Brown und David Botstein (1999, S. 33) geben eine kurze und anschauliche Beschreibung von explorativer (entdeckender) Forschung:

> „Entdeckung bedeutet, sich umsehen, beobachten, beschreiben und unbekanntes Gebiet vermessen, nicht Test von Theorien oder Modellen. Das Ziel ist es, Dinge herauszufinden, die wir weder kennen noch erwarten, und Beziehungen und Verbindungen zwischen den Elementen festzustellen, egal ob vorher erwartet oder nicht. Daraus folgt, dass dieser Prozess nicht durch Hypothesen geleitet wird und dass er so unabhängig von existierenden Vermutungen sein sollte wie möglich.“

Als wesentliche Charakteristika explorativer Untersuchungen sind folgende Gesichtspunkte zu nennen:

- Es geht weniger um quantifizierende Angaben als darum, möglichst *vielfältige und tiefgehende Einsichten* in den Untersuchungsgegenstand zu gewinnen. Dem entsprechend werden typischerweise die Untersuchungseinheiten (z. B. Auskunftspersonen, Beobachtungsobjekte) *gezielt* (und nicht repräsentativ) ausgewählt (Yin, 2011, S. 88 f.). Es geht bei dieser Auswahl in erster Linie darum, relativ wenige besonders „interessante" Untersuchungseinheiten zu finden, die typische oder auch extreme Einsichten erlauben. Die Ergebnisse explorativer Studien haben deswegen meist eher impressionistischen als definitiven Charakter. Daraus wird auch die enge Bindung zwischen explorativen Untersuchungs*zielen* und qualitativen Untersuchungs*designs* (siehe Abschn. 2.4.2.2) erkennbar.
- Damit der Zweck explorativer Forschung erreicht werden kann, ist eine enge *Einbindung des Forschers in den Prozess der Informationssammlung* hilfreich. Wenn diese Tätigkeit vom/von der ForscherIn delegiert wird, verliert er/sie die Möglichkeit, seine/ihre wachsende Vertrautheit mit dem Untersuchungsgegenstand durch spezifische Fragestellungen in den Forschungsprozess einzubringen, und muss damit rechnen, dass Einzelheiten und Nuancen der von anderen Personen gesammelten Informationen nicht zu ihm/ihr gelangen. Die Forscherin bzw. der Forscher soll in der Lage sein, eine Vielzahl erfasster Einzelheiten zu einem Gesamtbild zusammenzufügen.
- Zentrale Bedeutung für den Erfolg explorativer (!) Forschung hat die Vorgehensweise des Forschers bzw. der Forscherin: Er/sie muss in erster Linie *Ideen und Einsichten suchen*, nicht (bereits existierende) Vermutungen bestätigen wollen. Die *Aufgeschlossenheit* für neue Informationen und für laufende Veränderungen eines sich entwickelnden Bildes ist die Voraussetzung für eine sinnvolle Anwendung qualitativer Methoden mit explorativen Zielen.

Wie oben schon erwähnt sind qualitative Methoden bei explorativen Untersuchungen dominierend. Auf diese Art von Methoden wird im 3. Kapitel ausführlicher eingegangen. Hier mag eine kurze Kennzeichnung gängiger Methoden genügen:

- **Gruppendiskussion** (Focus Group Interview): Gleichzeitige Befragung (meist 6–10) Auskunftspersonen, die auch untereinander kommunizieren (→ Diskussion) zu einem interessierenden Thema (z. B. Wünsche und Entscheidungskriterien bei Urlaubsreisen).
- **Qualitatives Interview**: Ausführliches Gespräch zwischen einem besonders geschulten Interviewer und einer Auskunftsperson über ein interessierendes Thema (z. B. Motive beim Kauf von Bekleidung), das weitgehend frei geführt wird und nur einem groben Leitfaden folgt.

- **Fallstudie**: Bei Fallstudien werden Abläufe/Ereignisse (z. B. Entscheidungsprozesse) oder Verhaltensweisen von Personen und Organisationen mithilfe verschiedener Informationsquellen (z. B. Interviews, Protokolle) beschrieben und analysiert.
- **Ethnografische Studie**: Direkte Beobachtung von Verhaltensweisen und Gebräuchen in einem natürlichen Umfeld (z. B. Produktnutzung, Handhabung von Verpackungen).
- **Social Media-Analysen** Analyse von Verbraucheräußerungen zu Marken oder bestimmten Verwendungszusammenhängen in sozialen Medien, z. B. zur Früherkennung von Qualitätsproblemen oder neuen Verwendungszusammenhängen.

Seit einiger Zeit ist **Data Mining** als Ansatzpunkt für explorative Untersuchungen hinzugekommen. Dabei geht es darum, Verfahren einzusetzen, mit deren Hilfe große Datenbestände (z. B. Kundendatenbanken) im Hinblick auf bestimmte Merkmalszusammenhänge („Muster") automatisch analysiert werden. Mithilfe (multivariater) Verfahren (siehe Kap. 9) werden Zusammenhänge zwischen Kundenmerkmalen und Verhaltensweisen *entdeckt.* Beispielsweise kann man aus Kundendaten die Merkmale identifizieren, die bei Intensivverwendern bestimmter Produkte besonders stark ausgeprägt sind. Damit hat man Ansatzpunkte für eine genauere gedankliche Durchdringung und die gezielte Untersuchung von Faktoren, die zur Intensivverwendung führen. Neu ist dabei nicht die Anwendung entsprechender Analysemethoden, sondern deren Anwendung auf sehr, sehr große Datensätze, beispielsweise Scanner-Daten über viele Millionen Einkaufsvorgänge. Weil hier eben sehr große Datenmengen die Basis der Analysen bilden, wird in der Praxis in diesem Zusammenhang von „Big Data" gesprochen.

Beispiel

Iacobucci und Churchill (2010, S. 26) berichten über das Beispiel einer US-Versicherungsgesellschaft, die nach der Anwendung von Data Mining ihre Tarife für die Versicherung von Sportwagen marktgerechter gestalten konnte. Meist denkt man bei Sportwagenfahrern zunächst an jüngere Männer mit einer gewissen Risikofreude. Dem entsprechend waren Versicherungen für Sportwagen relativ teuer. Durch Anwendung von Data Mining auf die Daten von 10 Mio. (!) Kunden hatte man dann festgestellt, dass es eine zweite große Gruppe von Sportwagenbesitzern mit deutlich anderen Merkmalen gab: Familienväter, die einen Sportwagen als Zweit- oder Dritt-Auto besaßen und wesentlich weniger Schadensfälle hatten. Dieser Kundengruppe konnte man auf Basis der Analyse günstigere Tarife anbieten und erreichte eine höhere Kundenbindung. ◀

Deskriptive Untersuchungen

Deskriptive Untersuchungen dürften in der Praxis der kommerziellen Marktforschung dominierend sein. Sie betreffen Fragestellungen, die das „tägliche Brot" des Marktforschers ausmachen. Typisch sind folgende Arten von Problemen**:**

- **Charakterisierung von Märkten und Marktsegmenten**
  (Wie groß ist das Marktpotenzial für Streaming-Dienste? Welches sind die typischen Merkmale der Käufer von Elektro-Autos? etc.)
- **Analyse von Zusammenhängen zwischen Variablen**
  (Nutzen freizeitaktive Konsumenten andere Medien als andere Gruppen der Bevölkerung? Wie verändert eine bestimmte Werbekampagne die Einstellung zu einem Produkt? etc.)
- **Prognosen**
  (Wie groß wird ein bestimmter Markt in 2 Jahren sein? Wie werden sich wachsende Einkommen auf die Nachfrage nach Kreuzfahrten auswirken? etc.)

Dabei deutet sich schon an, dass es bei deskriptiven Untersuchungen in der Regel darauf ankommt, möglichst genaue Aussagen zu machen, sei es über die Größe von Märkten, über die Merkmale von Kundengruppen oder über das Wachstum von Märkten. Daraus ergeben sich zwei charakteristische methodische Anforderungen:

- Deskriptive Untersuchungen sind typischerweise *repräsentativ* angelegt, da man ja möglichst präzise Angaben über eine Grundgesamtheit gewinnen will. Die Fehler beim Schluss von einer Stichprobe auf die jeweilige Grundgesamtheit sollen also möglichst gering sein.
- An Anlage und Durchführung von deskriptiven Untersuchungen werden insofern hohe Anforderungen gestellt, als *systematische* (durch Mängel der Untersuchung begründete) *Fehler* im Interesse exakter Ergebnisse *möglichst* geringgehalten werden sollen. Deswegen findet man hier typischerweise (im Kontrast zur explorativen Forschung) genaue Festlegungen des Vorgehens und sorgfältige Kontrollen des Untersuchungsablaufs.

Auf einige verbreitete Formen von deskriptiven Untersuchungen – insbesondere auf repräsentative Befragungen – wird noch genauer eingegangen. Dann wird auch hinsichtlich zeitpunkt- und zeitraumbezogener Aussagen in **Querschnitts**- und **Längsschnitt-Untersuchungen** unterschieden.

Kausal-Untersuchungen

Kausal-Untersuchungen stellen besonders hohe Anforderungen an die methodische Vorgehensweise. Sie führen aber in Wissenschaft und Praxis zu besonders gehaltvollen Aussagen. Wenn ein Wissenschaftler z. B. festgestellt hat, dass eine bestimmte Merkmalskombination die Ursache für ein bestimmtes Konsumentenverhalten ist, dann ist er eben seinem Ziel, Realität zu verstehen und erklären zu können, ein gutes Stück nähergekommen. Wenn der Praktiker feststellt, dass bestimmte qualitative Mängel die Ursache für sinkende Marktanteile eines Produkts sind, dann hat er eben einen entscheidenden Ansatzpunkt gefunden, um das Problem der sinkenden Marktanteile zu lösen.

Hintergrundinformation
Lehmann et al. (1998, S. 143) zur praktischen Relevanz von Kausal-Untersuchungen:

> „Das Konzept der Kausalität impliziert, dass bei der Veränderung einer bestimmten Variablen (z. B. Werbung) sich eine andere Variable (z. B. Absatzmenge) als Ergebnis der getroffenen Maßnahme verändert. Deswegen werden bei fast jeder Marketing-Entscheidung (…) implizit deren Konsequenzen bedacht. Wenn Manager Verständnis für die kausalen Beziehungen in ihrem Markt entwickeln, dann können sie „optimale" Entscheidungen treffen. Deswegen sind Kausal-Untersuchungen wesentlich für wirkungsvolle Entscheidungen."

Was sind nun die Merkmale einer **Kausalbeziehung?** Wann spricht man davon, dass ein Merkmal die *Ursache* für das Auftreten eines anderen ist? Zunächst muss natürlich eine theoretisch überzeugende, zumindest aber plausible *Begründung* für einen Kausalzusammenhang gegeben sein. Dem entsprechend nennt Shelby Hunt (2002, S. 127 f.) als Kriterium für die Feststellung einer Kausalbeziehung das *Vorliegen einer entsprechenden theoretischen Begründung.* Das bezieht sich darauf, dass es sich ja bei einer Kausalbeziehung um einen *systematischen* Zusammenhang zwischen Variablen handeln soll, also um einen begründeten und nachvollziehbaren Zusammenhang, der nicht nur ein (möglicherweise zufällig) empirisch beobachtbares Phänomen darstellt

Hintergrundinformation
David de Vaus (2001, S. 36) zur Bedingung des Vorliegens einer theoretischen Begründung für eine Kausalbeziehung:

> „Die Behauptung von Kausalität muss Sinn haben. Wir sollten in der Lage sein, zu erläutern wie X Einfluss auf Y ausübt, wenn wir auf eine Kausalbeziehung zwischen X und Y schließen wollen. Selbst wenn wir empirisch nicht zeigen können, wie X Einfluss auf Y hat, müssen wir eine plausible Erläuterung für den Zusammenhang geben können (plausibel im Sinne von anderer Forschung, aktuellen Theorien etc.)."

Im Hinblick auf eine empirische Prüfung ist zu fordern, dass die beiden Merkmalsausprägungen, zwischen denen man einen kausalen Zusammenhang vermutet, auch *gemeinsam auftreten.* Wenn man beispielsweise annimmt, dass hohe Werbebudgets die Ursache hoher Marktanteile sind, dann müsste man, wenn man verschiedene Produkte oder Märkte betrachtet, einen großen Anteil von Fällen finden, bei denen die Merkmalskombination „hoher Marktanteil und hoher Werbeetat" auftreten. Damit wäre aber noch nicht bestätigt, dass der Werbeetat tatsächlich der *Grund* für die Entwicklung des Marktanteils ist. Der Zusammenhang zwischen den beiden Variablen könnte ja auch auf andere Weise erklärt werden, z. B: „Bei hohen Marktanteilen sind die Umsätze entsprechend hoch und es ist genügend Geld für Werbung vorhanden". Demnach wäre also der Marktanteil die Ursache für entsprechende Werbebudgets. Man stellt deshalb neben der Voraussetzung der gemeinsamen Variation von „Grund" und „Effekt" die Forderung auf, dass die *Variation des Grundes der (entsprechenden) Variation des Effekts vorauszugehen* hat. Wenn man also in dem genannten Beispiel feststellt, dass in der Regel erst der Werbeetat und dann der Marktanteil gestiegen ist, dann kann man die oben genannte zweite mögliche Kausalbeziehung (Marktanteil als Ursache für Werbeetat) ausschließen.

Auch jetzt kann man noch nicht davon sprechen, dass ein Kausalzusammenhang bestätigt wurde. Es wäre ja denkbar, dass mit der Erhöhung des Werbeetats üblicherweise eine Verstärkung der Außendienstanstrengungen, der Aktivitäten dem Handel gegenüber etc. einhergeht. Dann könnte es sein, dass nicht der erhöhte Werbeetat, sondern der stärkere Außendiensteinsatz die Ursache für steigenden Marktanteil ist. Man muss also – das ist die vierte Anforderung bei der Überprüfung von Kausalbeziehungen – *alternative Erklärungsmöglichkeiten* für die gemeinsame Variation von Grund und Effekt in der vorgegebenen zeitlichen Abfolge *ausschließen* können. In diesem Zusammenhang sei an die Kennzeichnung der *Validierung* als „Ausschluss alternativer Erklärungsmöglichkeiten" im Abschn. 2.3 erinnert.

▶ **Wichtig** Die **Merkmale einer Kausalbeziehung** lassen sich also in vier Gesichtspunkten zusammenfassen:

1. Theoretische Begründung des Zusammenhanges
2. Gemeinsame Variation von „Grund" und „Effekt"
3. Veränderung des „Grundes" geht der Veränderung des „Effekts" voraus
4. Ausschluss alternativer Erklärungsmöglichkeiten für den beobachteten Zusammenhang

Wegen dieser recht strengen Anforderungen an die Feststellung von Kausalzusammenhängen ist dafür ein bestimmtes *Untersuchungsdesign* typisch, das **Experiment** (siehe dazu Kap. 6).

2.4.2 Festlegung des Untersuchungsdesigns

2.4.2.1 Primärforschung und Sekundärforschung

Mit der Festlegung des Untersuchungsdesigns werden die wesentlichen Entscheidungen über die anzuwendenden Forschungsmethoden getroffen, indem die Art der Datenerhebung (z. B. Befragung oder Beobachtung), das Vorgehen bei der Stichprobenziehung etc. bestimmt werden. Zuvor steht aber häufig die Frage an, ob es überhaupt notwendig ist, für einen bestimmten Untersuchungszweck Daten neu zu erheben und auszuwerten. In manchen Fällen könnte es ausreichen, vorhandene Daten im Hinblick auf das aktuelle Problem neu zu analysieren. Es geht also um die Entscheidung zwischen Primär- und Sekundärforschung.

Als **Primärforschung** bezeichnet man die *Neu*-Erhebung von Daten für ein anstehendes Untersuchungsproblem. Im Jargon der Sozialwissenschaften spricht man dabei auch von Feldforschung („Field Research"). Dagegen ist die **Sekundärforschung** dadurch gekennzeichnet, dass bereits erhobene und gespeicherte Daten für einen gegebenen Untersuchungszweck neu aufbereitet und analysiert werden. Da auf diese Weise Untersuchungen am Schreibtisch des Forschers bzw. der Forscherin durchgeführt wer-

den, findet man in diesem Zusammenhang gelegentlich auch den Begriff „Desk Research“.

Normalerweise ist Sekundärforschung deutlich weniger aufwendig als Primärforschung (siehe unten). Deswegen wird in der Regel vor der Entscheidung über die Durchführung einer Primäruntersuchung die Frage gestellt, ob die Auswertung vorhandener Daten für die Bearbeitung des anstehenden Problems ausreichend sein könnte. Wenn man diese Frage bejaht, dann wäre in dem Fall die Sekundärforschung ein *Ersatz für Primärforschung.*

Gelegentlich basiert Primärforschung auf Daten, die durch Sekundärforschung gewonnen wurden. Man denke hier z. B. an die bei Primäruntersuchungen erforderlichen Stichprobenziehungen, wozu oft Sekundärdaten (z. B. von statistischen Ämtern) herangezogen werden. In diesen Fällen dient Sekundärforschung also der *Vorbereitung von Primärforschung.*

Manchmal werden Ergebnisse der Primärforschung auch zu Daten der Sekundärforschung in Beziehung gesetzt. Beispielsweise kann man sich vorstellen, dass man in einer Primäruntersuchung die soziodemografischen Daten einer Personengruppe feststellt, die für den Kauf eines neuen Produkts am ehesten infrage kommt, und dann mithilfe von Sekundärdaten abschätzt, wie groß der Anteil dieser Gruppe an der Gesamtbevölkerung ist, um Anhaltspunkte für die Größe des entsprechenden Marktes zu bekommen. Hier dient die Sekundärforschung also der *Ergänzung der Primärforschung.*

Beispiel

Hier einige Beispiele für Internetadressen, unter denen man Zugang zu einer großen Menge unterschiedlicher Sekundärdaten für die Marktforschung bekommt:

- www.kaggle.com
  Kaggle ist ein Online-Portal, welches laut eigenen Angaben die größte Data-Science-Community weltweit beheimatet. Regelmäßig werden hier von den Nutzern große Datenmengen, z. T. mit Fragestellungen und Analysevorschlägen hochgeladen. Das Portal verzeichnet nahezu 150.000 Einreichungen pro Monat und stellt Ressourcen (Daten) und Tools (Software-Codes) bereit, um Fortschritte in der Datenwissenschaft umfassend zu unterstützen.
- www.destatis.de
  Statistisches Bundesamt in Wiesbaden. Enthält Daten zu einer Vielzahl von Statistiken, u. a. das komplette Statistische Jahrbuch.
- www.google.de/trends
  Google Trends liefert, wie häufig wann und wo bestimmte Begriffe gesucht werden. Dient z. B. zur zeitlichen und räumlichen Planung von Werbekampagnen.
- http://epp.eurostat.ec.europa.eu
  Homepage von eurostat, des Statistischen Amtes der Europäischen Union mit einer Vielzahl von amtlichen Daten auf europäischer Ebene.
- www.kba.de

Kraftfahrtbundesamt mit Bestands- und Zulassungsdaten zu Kraftfahrzeugen.
- https://gik.media/best-4-planning
  Best for Planning Studie: Umfangreicher Datenbestand zu einer Vielzahl von Warengruppen, basierend auf einer Markt-Media-Studie verschiedener Zeitschriften-Verlage mit einer Stichprobe von über 30.000 Personen. ◀

Die Sekundärforschung hat im Vergleich zur Primärforschung typischerweise einige **Vorteile,** wobei man heute davon ausgehen kann, dass Sekundärforschung – zumindest mit externen Daten – weitgehend am PC bzw. über das Internet erfolgt:

- Der Rückgriff auf Daten der amtlichen Statistik, aus firmeninternen Quellen, aus kommerziellen Datenbanken etc. ist im Vergleich zur Primärforschung normalerweise mit erheblichen *Kostenvorteilen* verbunden.
- Viele Primäruntersuchungen dauern mit der Problemdefinition über die Methodenentwicklung, Datensammlung und -analyse und Vorlage eines Berichts bis zu mehreren Monaten. Wenn Zugang zu vorhandenen vergleichbaren Daten besteht, deren Aufbereitung und Auswertung vielleicht einige Tage dauert, ergibt sich eine deutliche *Zeitersparnis.*
- Ein erheblicher Teil der erhältlichen Sekundärdaten beruht auf *Totalerhebungen,* ist also in seiner Aussagekraft nicht durch Stichprobenfehler eingeschränkt (wohl aber durch andere Probleme, siehe unten).
- Sekundärdaten sind oftmals auch für die *Vergangenheit* verfügbar, wodurch man die Möglichkeit erhält, Veränderungen im Zeitablauf zu beobachten. Wie sollte man sonst die Entwicklung von Einkommensverteilungen, Marktanteilen, Marktvolumina etc. untersuchen können? Vergangenheitsbezogene Primärforschung ist dagegen ein seltener Ausnahmefall.

Mit der Verwendung von Sekundärdaten sind andererseits oftmals auch bestimmte **Probleme** verbunden, die im Folgenden kurz charakterisiert werden sollen:

- **Erhältlichkeit**
  Für viele Marketing-Probleme, insbesondere wenn es um Informationen geht, die direkt auf das Marketing-Mix für ein Produkt bezogen sind (z. B. Daten zur Erinnerungswirkung einer bestimmten Werbemaßnahme), sind in allgemein zugänglichen Quellen keine Daten vorhanden.
- **Maßeinheiten**
  Gelegentlich sind die in bestimmten Statistiken verwendeten Maßeinheiten für die Vorbereitung von Marketing-Entscheidungen wenig geeignet. Beispielsweise ist die Angabe der Anzahl verkaufter PKWs ohne Informationen über deren Preise nur begrenzt aussagekräftig.
- **Klassengrößen**

Statistische Daten werden meist in bestimmten Größenklassen (z. B. Jahreseinkommen bis € 20.000,–, über € 20.000,– bis € 30.000,– usw.) ausgewiesen, deren Angemessenheit für den jeweiligen Untersuchungszweck fraglich sein kann.

- **Aktualität**
  Daten – nicht zuletzt aus der amtlichen Statistik – werden teilweise mit erheblicher Verzögerung gegenüber dem Erhebungszeitpunkt publiziert.
- **Genauigkeit**
  Hinsichtlich der Genauigkeit von Sekundärdaten ist Vorsicht geboten, da man oft nicht nachvollziehen kann, wie die Daten erhoben worden sind, was ja erhebliche Auswirkungen haben kann. Bei Publikationen, die bestimmten Interessen der herausgebenden Organisationen dienen sollen, muss man besonders mit systematisch verzerrten Angaben rechnen.
- **Repräsentanz**
  Manche Sekundärdaten sind für die interessierende Grundgesamtheit nicht repräsentativ. Beispielsweise kommt es vor, dass von Verbänden publizierte Daten nur hinsichtlich der Mitglieder dieser Verbände Aussagekraft haben.
- **Aggregation**
  In publizierten Statistiken sind Daten z. B. in regionaler oder branchenmäßiger Hinsicht teilweise so stark aggregiert (Beispiel: „Umsatz der deutschen Werkzeugmaschinenindustrie"), dass sie für detailliertere Marketing-Fragestellungen nicht mehr aussagekräftig sind.

Sekundärdaten können aus sehr unterschiedlichen **Quellen** stammen; je nach Problem und Branche wird man versuchen müssen, angemessene Informationen zu finden. Der Versuch, auch nur annähernd vollständige Angaben über Sekundärquellen zu machen, wäre aussichtslos (und für die LeserInnen auch ermüdend und wenig ergiebig). Deswegen reicht es hier aus, bestimmte Arten von Sekundärquellen zu charakterisieren und mit Beispielen zu illustrieren:

- **Unternehmensinterne Quellen** (z. B. Umsatz- und Auftragsstatistik, Außendienstberichte, Reklamationsstatistik)
- **Amtliche und halbamtliche nationale Quellen** (z. B. statistische Bundes- und Landesämter, Kraftfahrzeugbundesamt, s. o.)
- **Amtliche und halbamtliche internationale Quellen** (z. B. Weltbank, Eurostat, International Labour Organization)
- **Nichtstaatliche Quellen** (z. B. Verbände, Industrie- und Handelskammern)

Die Möglichkeiten des Zugriffs zu Datenbanken (z. B. Literatur- oder Unternehmensdatenbanken) sind allgemein bekannt und bedürfen deswegen keiner besonderen Erläuterung, zumal entsprechende Angaben rasch veralten. Entsprechendes gilt für die Informationssammlung mit Suchmaschinen (z. B. Google).

2.4.2.2 Typen von Untersuchungsdesigns

Nach dem vorstehenden kurzen Ausblick auf die Sekundärforschung geht es jetzt um die Untersuchungsdesigns, die der Neu-Erhebung von Daten – also der Primärforschung – dienen. „Die Funktion von Untersuchungsdesigns ist es sicherzustellen, dass die gesammelten Daten uns in die Lage versetzen, dem Untersuchungsziel möglichst eindeutig zu entsprechen." (de Vaus, 2001, S. 9). Hier werden die zentralen Ideen der im Abschn. 2.2.1 schon kurz angesprochenen vier Typen von Untersuchungsdesigns

- Qualitative Untersuchungen,
- Querschnitts-Untersuchungen,
- Längsschnitt-Untersuchungen und
- Experimente

eingehender dargestellt. Die konkreten Methoden und Techniken, die oft nicht eindeutig bestimmten Designs zugeordnet werden können, werden dann in den folgenden Kapiteln vorgestellt.

Hintergrundinformation

David de Vaus (2001, S. 9) illustriert die zentrale Aufgabe des Untersuchungsdesigns:

> „Beim Untersuchungsdesign geht es um ein logisches Problem, nicht um ein logistisches Problem. Bevor ein Bauherr oder Architekt einen Arbeitsplan erstellen oder Material bestellen kann, müssen sie zunächst die Art des Bauwerks festlegen, seine Nutzung und die Bedürfnisse der Mieter. Der Arbeitsplan ergibt sich daraus. In ähnlicher Weise sind in der Sozialforschung die Probleme der Stichprobenziehung, der Datenerhebung (z. B. Fragebogen, Beobachtung, Dokumentenanalyse) und der Frageformulierung dem Problem untergeordnet, welche 'Beweise' benötigt werden"

Qualitative Untersuchungen

Die Kennzeichnung qualitativer Marktforschung umfasst sowohl deren typische Methodik als auch – in Verbindung damit – deren typische Ausrichtung auf bestimmte Erkenntnisziele. Charakteristisch sind in methodischer Hinsicht kleine Fallzahlen ohne repräsentative Auswahl, nicht standardisierte Datenerhebung und eher interpretierende als statistische Analyse. Qualitative Untersuchungen in der Grundlagenforschung sind typischerweise auf Theoriebildung ausgerichtet; bei der Anwendung auf Praxis-Probleme stehen exploratorische und diagnostische Zwecke im Vordergrund. Der Stellenwert **explorativer Forschung** ist im Abschn. 2.4.1 schon erläutert worden. Die Diagnose-Funktion hat ihre Bedeutung insbesondere im Zusammenhang mit der Entwicklung, Erprobung und Verbesserung von Marketing-Maßnahmen. Man denke nur an die Entwicklung eines neuen Produkts, einer neuen Packung oder eines neuen Werbemittels. Dabei kommt der Marktforschung nicht nur die Aufgabe zu, die Akzeptanz, Wirkung usw. des Produkts, der Packung oder des Werbemittels insgesamt zu testen, sondern auch zu ermitteln, welche einzelnen Elemente Schwächen haben und weiterentwickelt wer-

Tab. 2.1 Hier einige Beispiele für Fragestellungen, bei denen eher eine qualitative bzw. eine quantitativ-deskriptive Untersuchung infrage kommt

	Eher qualitative Untersuchung	Eher quantitativ-deskriptive Untersuchung
Produktpolitik	Warum und in welchen Situationen trinken KonsumentInnen „Red Bull“?	Wie groß ist der Anteil der zufriedenen KonsumentInnen bei „Red Bull“ in der Gruppe der 16–25 jährigen?
Kommunikationspolitik	Zu welchen Assoziationen führt die ‚lila Kuh‘ in einer Milka-Anzeige?	Um wie viel Prozent ist der Bekanntheitsgrad von Milka nach einer Werbekampagne gestiegen?
Distributionspolitik	Welche Gründe gibt es für die Händler, dass sie mein Produkt nicht in Griffhöhe platzieren?	Wie hoch ist der Umsatzzuwachs bei einer Marke nach der Ausweitung des Distributionsgrades um 10 %?

den sollen. Letztlich können Ergebnisse qualitativer Untersuchungen die Grundlage für kreative Entwicklungen neuer Produkte, Kommunikationsmittel etc. bilden (Kent, 2007, S. 89 f.).

Beispielsweise kann man vor allem durch qualitative Methoden feststellen, welche Probleme die Handhabung eines Produkts bereitet oder welche Teile einer Anzeige kaum wahrgenommen werden. Dieses ließe sich durch quantitative Untersuchungen (repräsentative Querschnitts-Untersuchungen) nicht mit vergleichbarer Differenziertheit erreichen. Wenn man versucht, anknüpfend an die obige Kennzeichnung wesentliche Merkmale qualitativer Marktforschung zu nennen, so stehen dabei folgende Aspekte im Vordergrund (siehe auch Tab. 2.1):

- **Geringe Festlegung des Forschungsprozesses**
  Phasen der Datenerhebung und -analyse wechseln sich ab und beeinflussen sich gegenseitig, insofern als Ergebnisse von Zwischen-Analysen Einfluss auf weitere Schritte der Datenerhebung haben.
- **Kleine, nicht repräsentative Stichproben**
  Die Zahl der Probanden bei qualitativen Untersuchungen liegt meist deutlich unter 100. Wegen der eine hohe Bereitschaft zur Mitwirkung erfordernden Untersuchungsmethoden ist eine repräsentative Auswahl kaum möglich und in der Regel auch nicht erforderlich. Vor diesem Hintergrund ist die Generalisierbarkeit der Ergebnisse (siehe Abschn. 2.3) in der Regel äußerst begrenzt.
- **Eher verbal beschreibende und interpretierende Analyse**
  Wegen der Eigenheiten der Auswahl von Auskunfts- oder Versuchspersonen und des Vorgehens bei der Datenerhebung werden statistische Analysen nur in Ausnahmefällen und mit großer Vorsicht gemacht (z. B. „fast alle Auskunftspersonen waren der

Meinung …“ oder „die Mehrheit der Untersuchungsteilnehmer…“). Im Vordergrund steht die verbal beschreibende und interpretierende Darstellung im Untersuchungsprozess gewonnener Einsichten.

- **Anwendung freier (nicht standardisierter) Formen von Befragungen und Beobachtungen**
  Während bei gut vorbereiteten (→ Untersuchungsdesign, Entwicklung der Messinstrumente) quantitativen Untersuchungen der Prozess der Datensammlung (z. B. Durchführung von Interviews) von geschulten Laien nach exakt vorgegebenen Regeln durchgeführt werden kann, sind bei qualitativen Studien typischerweise themenspezifisch und methodisch kompetente InterviewerInnen für die Datensammlung nötig, weil sie diese weitgehend selbstständig gestalten müssen.

In einem durch wissenschaftliche Methodik geprägten Arbeitsgebiet wie der Marktforschung gibt es hinsichtlich der qualitativen Forschung immer noch einen gewissen „Rechtfertigungsbedarf“. Welchen Stellenwert sollen z. B. Gruppendiskussionen haben, die zu impressionistischen Ergebnissen führen (s. o.), im Vergleich zu Umfrageverfahren, bei denen Ergebnisse (z. B. Bekanntheitsgrade von Produkten) auf Kommastellen genau ausgewiesen werden? Deswegen sollen hier drei Gesichtspunkte genannt werden, die bestimmte *Vorteile* der qualitativen im Vergleich zur quantitativen Forschung aufzeigen:

- Der Komplexität menschlichen Verhaltens, das ja in vielen Marktforschungsprojekten untersucht werden soll, wird man oftmals durch die mit quantitativer Forschung verbundenen standardisierten Messtechniken und stark aggregierenden Techniken der statistischen Datenanalyse nicht voll gerecht.
- Mit den Hilfsmitteln der quantitativen Forschung erreicht man vielfach nur eine Schein-Genauigkeit. So ist es beispielsweise üblich, mit Hilfe der induktiven Statistik sehr präzise erscheinende Angaben über Vertrauensbereiche und Sicherheitswahrscheinlichkeiten von Ergebnissen zu machen. Die Erfahrung und entsprechende Studien (vgl. z. B. Assael & Keon, 1982) lehren aber, dass die dabei überhaupt nicht berücksichtigten systematischen Fehler (z. B. durch verzerrende Fragestellungen, Interviewereinfluss) weitaus größeres Gewicht haben können.
- Für diagnostische Zwecke ist qualitative Forschung oftmals besser geeignet als quantitative. Wenn man z. B. an Werbe-Pretests denkt, dann kann man mit quantitativen Untersuchungsmethoden sicher die Wirksamkeit eines Werbemittels (im Vergleich zu alternativen Entwürfen) feststellen. Wenn man aber aus unbefriedigenden Resultaten eines solchen Tests Konsequenzen hinsichtlich der Gestaltung des Werbemittels ziehen will, dann muss man häufig auf qualitative Untersuchungsergebnisse zurückgreifen. So können beispielsweise qualitative Interviews (siehe Kap. 3) Hinweise auf die Anmutungsqualität von einzelnen Elementen (z. B. Teile von Abbildungen, Überschriften) des Werbemittels geben.

Querschnitts-Untersuchungen

Im Abschn. 2.2.1 ist bereits darauf hingewiesen worden, dass Querschnitts-Untersuchungen am ehesten dem Bereich der **deskriptiven Forschung** zuzurechnen sind. Damit verbunden ist typischerweise eine repräsentative Untersuchungsanlage, die die angestrebten präzisen Angaben über eine Grundgesamtheit ermöglicht, und das schwerpunktmäßige Bemühen, systematische Fehler gering zu halten.

Kennzeichnendes Merkmal einer Querschnitts-Untersuchung ist ihre *Zeitpunktbezogenheit*. Die Datenerhebung findet also an einem Zeitpunkt (in der Praxis allerdings in einem Zeitraum, der einige Wochen umfassen kann) statt, was aber nicht ausschließt, dass ein Teil der Aussagemöglichkeiten darüber hinausreicht. Beispielsweise kann man im Rahmen einer Konsumentenbefragung natürlich nach früherem Kaufverhalten oder (zukunftsbezogen) nach Kaufabsichten fragen. Allerdings stellt sich dann oft die Frage, inwieweit die Präzision von vergangenheitsbezogenen Angaben durch Erinnerungsmängel und die Aussagekraft von zukunftsbezogenen Angaben durch Änderungen im Zeitablauf (z. B. Veränderungen von Bedürfnissen, Präferenzen) beeinträchtigt werden.

Über den Untersuchungszeitpunkt hinaus reichen die Interpretationsmöglichkeiten von Querschnitts-Untersuchungen aber auch, wenn deren Ergebnisse im Vergleich zu früheren oder für später geplanten (möglichst genau) entsprechenden Messungen analysiert werden (z. B. Vergleich gemessener Bekanntheitsgrade im Abstand mehrerer Jahre).

Hintergrundinformation

David de Vaus (2001, S. 170) kennzeichnet das Wesen von Querschnitts-Untersuchungen („Cross-Sectional Designs“):

> „In den üblichen Querschnitts-Untersuchungen werden die Daten zu einem Zeitpunkt erhoben. In dieser Hinsicht unterscheidet sich dieses Design von normalen Panel-Designs und von experimentellen Designs mit Vor- und Nachmessung, bei denen Daten zu verschiedenen Zeitpunkten erhoben werden. Deswegen kann man mit Querschnitts-Untersuchungen nur Unterschiede zwischen Gruppen feststellen und nicht Veränderungen im Zeitablauf.“

Typische Aussagemöglichkeiten von Querschnitts-Untersuchungen beziehen sich auf:

- Schätzung von Anteilswerten und anderen statistischen Maßzahlen (z. B. Mittelwerte, Mediane) in Grundgesamtheiten (Beispiele: Bekanntheitsgrad von Marken, Durchschnittseinkommen der Bevölkerung, Anteil von Rauchern an der männlichen Bevölkerung)
- Vergleich unterschiedlicher Gruppen im Hinblick auf interessierende Merkmale (Beispiele: Einstellungen zu einer Marke bei Verwendern und Nicht-Verwendern dieser Marke, Ausgaben für Bekleidung bei Männern und Frauen)

- Zusammenhänge zwischen Variablen (Beispiele: Mit zunehmendem Einkommen steigt in der Regel der Anteil gesparten Einkommens; je zufriedener jemand mit einem Produkt ist, desto stärker wird seine Bindung an dieses Produkt)

Die beiden zuletzt genannten Arten von Aussagen werden in der Praxis oftmals wie Kausal-Aussagen interpretiert. So findet man häufig Schlussweisen, bei denen z. B. aus einem deutlichen („signifikanten“) Unterschied der Einstellungen zu einer Marke bei Männern und Frauen gefolgert wird, dass das Geschlecht die Ursache für diesen Unterschied sei. Dieser Schluss entspricht natürlich nicht den deutlich strengeren Anforderungen an Aussagen über Kausal-Beziehungen (siehe Abschn. 2.4.1). Gleichwohl kann man in solchen Fällen nicht ignorieren, dass ein – wie auch immer gearteter – Zusammenhang zwischen den Merkmalen vorliegt. Eine detaillierte Diskussion von Kausal-Aussagen auf der Basis von Querschnitts-Untersuchungen findet sich bei de Vaus (2001, S. 170 ff.).

Gegenüber den nachstehend erörterten Längsschnitt-Untersuchungen und Experimenten haben Querschnitts-Untersuchungen naturgemäß den Vorteil einer kürzeren Untersuchungsdauer und in Verbindung damit oft auch Kostenvorteile.

Querschnitts-Untersuchungen sind der Untersuchungstyp, der in der kommerziellen Marketing-Forschung am meisten eingesetzt wird. Ihre gängigste Form sind repräsentative Umfragen, auf die im Kap. 4 des vorliegenden Buches wegen ihrer großen Bedeutung besonders ausführlich eingegangen wird.

Längsschnitt-Untersuchungen

Bei Längsschnitt-Untersuchungen geht es um Aussagen, die auf **Zeiträume** oder zumindest auf **verschiedene Zeitpunkte** bezogen sind. Damit entspricht man einerseits einem der wichtigsten Informationsbedürfnisse der Praxis. Häufig geht es dort nicht so sehr darum festzustellen, welche Werte bestimmte relevante Messgrößen (z. B. Bekanntheitsgrad, Marktanteil) haben, sondern eher um deren Entwicklung im Zeitablauf. So ist eben ein sinkender Marktanteil typischerweise ein Anlass, Gegenmaßnahmen ins Auge zu fassen, und ein wachsender Marktanteil möglicherweise ein Indikator dafür, dass bestimmte Marketing-Maßnahmen Wirkung zeigen. Andererseits vermindert sich das Problem der Interpretation prinzipiell fehlerbehafteter Daten dadurch, dass gleichartige Messungen wiederholt durchgeführt und im Vergleich interpretiert werden und damit die irreführende Wirkung systematisch verzerrter Einzel-Messungen relativiert wird.

Das weitaus bedeutsamste Instrument der Marktforschung für Längsschnitt-Studien sind **Panel-Untersuchungen** (siehe Kap. 5). Als **Panel** bezeichnet man eine festgelegte, gleich- bleibende Menge von Erhebungseinheiten, bei denen über einen längeren Zeitraum wiederholt oder kontinuierlich die gleichen Merkmale erhoben werden. Eine solche Untersuchungsanlage erlaubt es nicht nur, die Veränderungen aggregierter Größen (z. B. Marktanteile) im Zeitablauf zu analysieren, sondern auch, Veränderungen auf der Ebene der einzelnen Erhebungseinheiten (z. B. Änderungen des Markenwahlverhaltens von Haushalten) zu beobachten.

Hintergrundinformation
Günther et al., (2006, S. 3) zur Bedeutung von Panels für die Marketing-Praxis:

> „Neben der aktuellen Beschreibung des Marktes sind es für das Marketingmanagement in aller Regel die Veränderungen, die Maßnahmen auslösen oder Beurteilungskriterien für in der Vergangenheit durchgeführte Maßnahmen bieten. Den Veränderungen im Marktgeschehen gilt daher das besondere Interesse des Marketingmanagements. Von daher ist es verständlich, dass jedes Panel als Stichprobenuntersuchung charakterisiert werden kann, die gleich in mehrfacher Hinsicht auf die möglichst genaue Messung von Marktveränderungen hin optimiert ist."

Daneben gibt es die Möglichkeit, mithilfe einer *zeitlichen Abfolge von (gleichartigen!) Querschnitts-Untersuchungen* („Wellenbefragungen", siehe Abschn. 5.6) Veränderungen im Zeitablauf zu messen. So lässt sich die Entwicklung des Bekanntheitsgrades einer Marke z. B. durch entsprechende repräsentative Befragungen im Halbjahres-Abstand ermitteln. Hervorzuheben sind bei dieser Art von Untersuchungsdesigns zwei Aspekte:

- Es werden (im Unterschied zum Panel) immer wieder neue Stichproben gezogen, die jeweils für die (gleichbleibende) Grundgesamtheit repräsentativ sein sollen. Dadurch lassen sich Veränderungen der *aggregierten* Werte von Bekanntheitsgrad, Marktanteil etc. feststellen, *aber nicht* Veränderungen bei einzelnen Personen oder Haushalten, da diese ja nur zu einem Zeitpunkt befragt oder beobachtet werden.
- Die *Erhebungstechniken* (z. B. die Fragetechniken) müssen bei allen Einzel-Untersuchungen *identisch* sein. Da schon kleine Veränderungen bei der Datenerhebung die Ergebnisse wesentlich beeinflussen können, wäre ansonsten nicht feststellbar, ob ein Ergebnisunterschied bei zwei Zeitpunkten auf eine Veränderung im Zeitablauf oder auf eine Änderung der Messmethode zurückzuführen ist.

Experimentelle Untersuchungen

▶ **Definition** In der Marktforschung versteht man unter einem Experiment eine Methode, bei der eine oder mehrere *unabhängige Variable* dergestalt *manipuliert* werden, dass die Auswirkungen dieser Manipulation auf eine oder mehrere abhängige Variablen gemessen werden können. Je nach Problemstellung können die entsprechenden Messungen durch Befragungs- oder Beobachtungsverfahren vorgenommen werden. Die Anwendung experimenteller Designs ist also nicht an eines dieser Erhebungsverfahren gebunden.

Kennzeichnend für eine experimentelle Vorgehensweise sind also Maßnahmen *(Manipulationen)*, die vorgenommen werden, um *unterschiedliche Werte* (Ausprägungen) *unabhängiger Variablen* bei den Versuchspersonen zu *schaffen.* Die Auswirkung dieser Manipulation (z. B. Kontakt zu einer Werbebotschaft) auf die abhängige Variable (z. B. Einstellung zu der beworbenen Marke) wird dann überprüft bzw. gemessen (Jaccard & Becker, 2002, S. 241). Jacoby (2013, S. 206) fasst die zentrale Idee in einem Satz

zusammen: „Das charakteristische Merkmal aller Arten von Experimenten (…) besteht darin, dass die Teilnehmer einem Ereignis oder Stimulus (unabhängige Variable bzw. vermuteter Grund) ausgesetzt werden und danach deren Reaktionen auf dieses Ereignis oder diesen Stimulus (abhängige Variable …) festgestellt werden." Bei „einfachen" Befragungen/Beobachtungen („nicht-experimentelle" Vorgehensweise) werden dagegen Methoden angewandt, die es lediglich erlauben, *gegebene* Ausprägungen von Variablen zu messen.

Beispiel

Der wesentliche Unterschied zwischen beiden Vorgehensweisen (experimentell und nicht-experimentell) sei durch ein Beispiel illustriert. Man stelle sich vor, der Zusammenhang zwischen Kundenzufriedenheit und Markentreue soll untersucht werden. Bei einer *nicht-experimentellen* Vorgehensweise könnte man eine Gruppe von Kunden nach ihrer Zufriedenheit und ihrer Markentreue befragen. Wenn sich dabei ein positiver Zusammenhang (→ Korrelation) dergestalt zeigt, dass Personen mit hoher Zufriedenheit meist auch markentreu sind (und umgekehrt), dann wäre die entsprechende Hypothese bestätigt. Allerdings wären die (strengen) Anforderungen an die Feststellungen von Kausal-Zusammenhängen (siehe Abschn. 2.4.1) noch nicht erfüllt. Man hätte ja nur bestätigt, dass tatsächlich ein Zusammenhang existiert, aber noch keine Kausalbeziehung (Ursache → Wirkung).

Dazu bedürfte es einer *experimentellen* Untersuchung. Diese könnte beispielsweise so aussehen, dass man bei einer Gruppe von Kunden (z. B. Kunden einer Region) versucht, die Kundenzufriedenheit (als mögliche Ursache) durch besonders hohe Qualität der Leistungen, exzellenten Nachkauf-Service etc. positiv zu beeinflussen (also zu „manipulieren") und dann zu beobachten, ob sich die Markentreue (Wirkung) hier deutlich („signifikant") positiver entwickelt als bei ansonsten gleichartigen anderen Kunden. Wenn es keine systematischen Unterschiede zwischen diesen („Experiment-„ und „Kontroll-„) Gruppen gibt (z. B. wegen zufälliger Zuordnung zu diesen Gruppen), dann kann die unterschiedliche Markentreue nur durch die Veränderung der Kundenzufriedenheit verursacht worden sein. ◀

Experimente werden also typischerweise für Kausal-Untersuchungen eingesetzt. Es geht dabei darum zu überprüfen, ob eine bestimmte Ausprägung einer Variablen tatsächlich die *Ursache* für eine gewisse Ausprägung einer anderen Variablen *(Wirkung)* ist. Dafür müssen besondere „experimentelle Designs" entwickelt werden, durch die andere mögliche Erklärungen für ein Ergebnis ausgeschlossen (→ Validierung) werden können (siehe Kap. 6).

Hier ein Beispiel für ein einfaches Experiment in der Marktforschung

Man stelle sich vor, ein Unternehmen will die Wirkung (gemessen durch die Zahl von Reaktionen per Mouseclick) unterschiedlich gestalteter Online-Werbung messen und damit feststellen, ob z. B. eine aggressive Farbgestaltung die *Ursache*

(→ Kausalbeziehung) für eine höhere Zahl von Clicks (siehe Abschn. 4.4.3) ist. Dazu könnte man insgesamt 20.000 Zielpersonen auswählen und diese nach dem *Zufallsprinzip* auf zwei Gruppen à 10.000 aufteilen.

Durch die zufällige Aufteilung wäre weitgehend sichergestellt, dass Unterschiede bei der Zahl der Clicks durch die Verschiedenheit der Online-Werbung, und nicht durch systematische Unterschiede (z. B. unterschiedliches Alter) der beiden Gruppen erklärt werden können. Wenn nun die eine Gruppe eine zurückhaltend farbige Botschaft und die andere Gruppe eine aggressiv farbige Botschaft bekommt und sich bei letzterer eine deutlich („signifikant") höhere Zahl von Clicks ergibt, dann kann das wohl nur an der unterschiedlichen Gestaltung der Botschaften liegen, da ja alle anderen Einflussfaktoren (Gruppenzusammensetzung, Untersuchungszeitpunkt etc.) konstant gehalten worden sind. Für diese Schlussweise ist also die *zufällige Zuordnung* der Versuchspersonen zu den beiden Gruppen entscheidend, *nicht die zufällige Auswahl* der Personen. ◀

2.4.3 Zusammenfassung

Auf die grundlegende Bedeutung einer angemessenen Problemdefinition sowie entsprechender Untersuchungsziele und -designs für die Aussagekraft von Marktforschungsuntersuchungen ist schon eingegangen worden. Dabei stellt sich die besondere Schwierigkeit, dass es dafür – im Unterschied zu anderen methodischen Problemen (z. B. Stichprobenziehung, statistische Analyse) – wenig erprobte und direkt anwendbare „Rezepte" gibt. Besonders augenfällig ist das bei der Problemdefinition, für die es naturgemäß keine standardisierte Vorgehensweise geben kann. Aber auch bei den folgenden Schritten gibt es kein eindeutiges und generalisierbares Vorgehen. Die folgende Darstellung soll lediglich den Ablauf verdeutlichen und Leitlinien für gängige Untersuchungsziele und -designs aufzeigen, wobei natürlich im Einzelfall Abweichungen davon sinnvoll sein können.

Im Abschn. 2.4.1 sind drei grundlegend verschiedene Arten von Untersuchungs*zielen* herausgearbeitet worden:

- „Entdeckung" von Marketing-Chancen und –Problemen und deren Einflussfaktoren, von Zusammenhängen zwischen Variablen und von Grundlagen für weitere (genauere) Untersuchungen (explorative Untersuchungen)
- „Beschreibung" von Märkten, von Zusammenhängen zwischen Variablen und von Trends (deskriptive Untersuchungen)
- „Begründung" und Bestätigung von Ursache-Wirkungs-Beziehungen (Kausal-Untersuchungen)

Im Abschn. 2.4.2 folgte die Kennzeichnung von vier Arten von Untersuchungs*designs*: qualitative Untersuchungen, Querschnitts-Untersuchungen, Längsschnitt-Untersuchungen

Tab. 2.2 Untersuchungsziele und Untersuchungsdesigns

Untersuchungsziel	Typische Art des Untersuchungsdesigns
Explorativ („entdecken“)	Qualitative Untersuchung
Deskriptiv („beschreiben“)	Querschnitts- oder Längsschnitt-Untersuchung
Kausal („begründen“)	Experiment

und Experimente. Durch die Festlegung auf eine dieser Arten sind allerdings zahlreiche weitere methodische Schritte noch nicht bestimmt. So ist mit der Entscheidung für eine Querschnitts-Untersuchung noch nicht geklärt, ob die Daten durch Befragung oder Beobachtung gewonnen werden sollen und welche Art der Stichprobenziehung verwendet werden soll. Typischerweise sind Methoden der Datenerhebung (Befragung, Beobachtung, Ziehung von Zufallsstichproben etc.) und Datenanalyse (Signifikanztests, Varianzanalyse etc.) *nicht eindeutig* bestimmten Typen von Untersuchungsdesigns zuzuordnen. Diese Beziehung ist vielmehr in beiden Richtungen mehrdeutig. Beispielsweise kann bei Experimenten die Datenerhebung durch Befragung, Beobachtung oder eine Kombination von beiden stattfinden; andererseits kann die Befragung mit ihren unterschiedlichen Spielarten in allen vier Typen von Untersuchungsdesigns zur Anwendung kommen.

Immerhin sind in Abschn. 2.4.1 Zusammenhänge zwischen Untersuchungszielen und dafür typischen Untersuchungsdesigns schon angesprochen worden. Diese lassen sich in Tab. 2.2 zusammenfassend darstellen.

Die Entwicklung von Untersuchungsdesigns und -methodik aus der Problemdefinition lässt sich – wie gesagt – kaum eindeutig und generell bestimmen. Deswegen soll hier die Vorgehensweise wenigstens für ein Beispiel illustriert werden. Dazu wird das in Abb. 2.7 dargestellte Schema verwendet. Es zeigt die verschiedenen Schritte einschließlich der Festlegung von Art des Untersuchungsdesigns und der Bestimmung der anzuwendenden Methoden.

Beim Beispiel in Abb. 2.8 steht am Beginn die Feststellung, dass nach Vorinformationen (z. B. Außendienstberichte, Informationen vom Handel) das eigene Produkt in der Wahrnehmung der Kunden Qualitätsprobleme hat. Es soll festgestellt werden, welcher Art diese Probleme und wie gravierend diese sind (→ deskriptive Untersuchung). Zum aktuellen Zeitpunkt (→ Querschnitts-Untersuchung) soll also eine entsprechende Befragung repräsentativ ausgewählter Kunden durchgeführt werden.

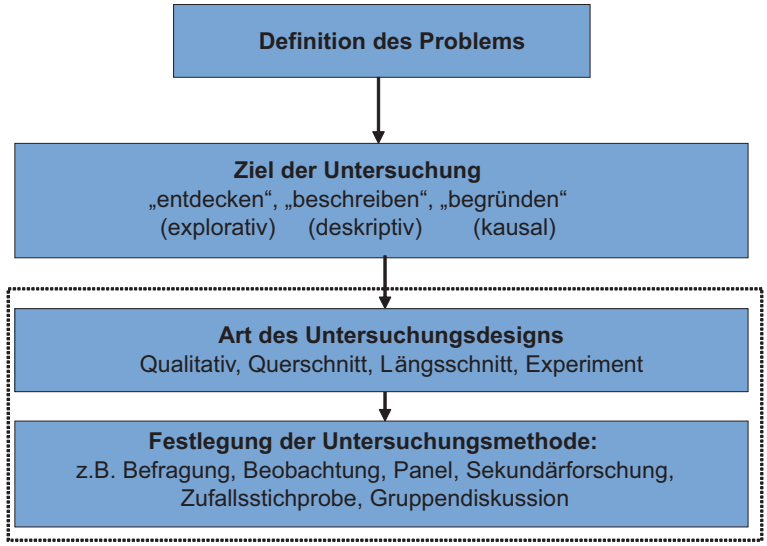


Abb. 2.7 Von der Problemdefinition zum Untersuchungsdesign

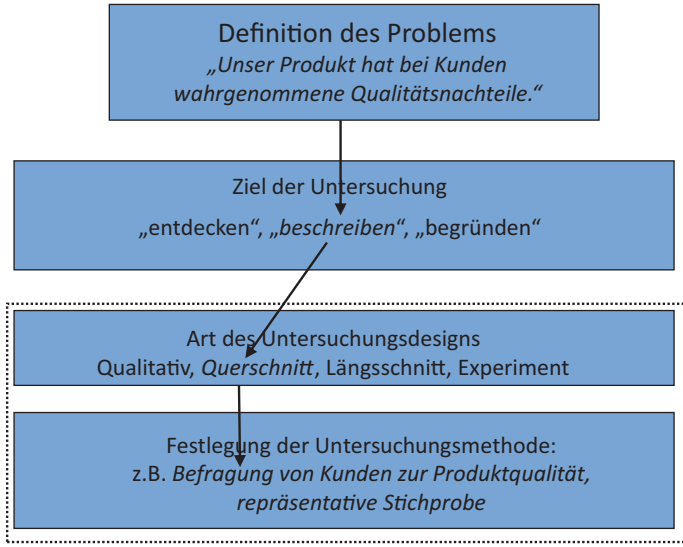


Abb. 2.8 Beispiel zum Untersuchungsdesign

Literatur

Assael, H., & Keon, J. (1982). Nonsampling vs. sampling errors in survey research. *Journal of Marketing, 46*, 114–123.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis*. Wiley.

Brown, P., & Botstein, D. (1999). Exploring the new world of the genome with DNA microarrays. *Nature Genetics, 21*, 33–37.

Burns, A., & Bush, R. (2006). *Marketing research* (5. Aufl.). Prentice Hall.

Chang, H., & Cartwright, N. (2008). Measurement. In S. Psillos & M. Curd (Hrsg.), *The Routledge companion to philosophy of science* (S. 367–375). Routledge.

Churchill, G., & Iacobucci, D. (2005). *Marketing research – methodological foundations* (9. Aufl.). Thomson South West.

Churchill, G. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research, 16*, 64–73.

Eisend, M. (2009). Metaanalyse. In C. Baumgarth, M. Eisend, & H. Evanschitzky (Hrsg.), *Empirische Mastertechniken* (S. 433–456). Gabler.

Eisend, M. (2014). *Metaanalyse*. Hampp.

Eisend, M., & Kuß, A. (2023). *Grundlagen empirischer Forschung – Zur Methodologie in der Betriebswirtschaftslehre* (3. Aufl.). SpringerGabler.

Franke, N. (2002). *Realtheorie des Marketing*. Mohr Siebeck.

Günther, M., Vossebein, U., & Wildner, R. (2006). *Marktforschung mit Panels* (2. Aufl.). Gabler.

Hildebrandt, L. (2008). Hypothesenbildung und empirische Überprüfung. In A. Herrmann, C. Homburg, & M. Klarmann (Hrsg.), *Handbuch Marktforschung* (3. Aufl., S. 81–105). Gabler.

Homburg, C. (Hrsg.). (2008). *Kundenzufriedenheit* (7. Aufl.). Gabler.

Hunt, S. (2002). *Foundations of marketing theory*. M. E. Sharpe.

Hunt, S. (2010). *Marketing theory – Foundations, controversy, strategy, resource-advantage theory*. M. E. Sharpe.

Iacobucci, D., & Churchill, G. (2010). *Marketing research – Methodological foundations* (10. Aufl.). Cengage.

Jaccard, J., & Becker, M. (2002). *Statistics for the behavioral sciences* (4. Aufl.). Wadsworth.

Jaccard, J., & Jacoby, J. (2020). *Theory construction and model-building skills – A practical guide for social scientists* (2. Aufl.). Guilford.

Jacoby, J. (2013). *Trademark surveys Vol. 1 – Designing, implementing, and evaluating surveys*. American Bar Association.

Kent, R. (2007). *Marketing research – Approaches, methods and applications in Europe*. Thomson Learning.

Kerlinger, F., & Lee, H. (2000). *Foundations of behavioral research* (4. Aufl.). Wadsworth.

Kroeber-Riel, W., & Gröppel-Klein, A. (2019). *Konsumentenverhalten* (11. Aufl.). Vahlen.

Lehmann, D., Gupta, S., & Steckel, J. (1998). *Marketing research*. Addison Wesley.

MacInnis, D. (2011). A framework for conceptual contributions in marketing. *Journal of Marketing, 75*(4), 136–154.

Nosek, B., Hardwicke, T., & Vazire, S. (2022). Replicability, robustness, and reproducibility in psychological science. *Annual Review of Psychology, 73*, 719–748.

Nunnally, J., & Bernstein, I. (1994). *Psychometric theory* (3. Aufl.). McGraw-Hill.

Schanz, G. (1988). *Methodologie für Betriebswirte* (2. Aufl.). Poeschel.
Suppe, F. (1977). *The structure of scientific theories* (2. Aufl.). University of Illinois Press.
de Vaus, D. (2001). *Research design in social research*. Sage.
de Vaus, D. (2002). *Analyzing social science data*. Sage.
Yin, R. (2011). *Qualitative Research from Start to Finish*. Guilford.

Explorative Untersuchungen mit qualitativen Methoden

3

Zusammenfassung

In diesem Kapitel geht es um Untersuchungen, die in der Marktforschung oft als Vorstudien für andere Untersuchungen verwendet werden, die sogenannten „explorativen Untersuchungen“. Hier geht es beispielsweise darum, vor einer groß angelegten Untersuchung zu Kriterien bei einer Kaufentscheidung oder zur Kundenzufriedenheit zu „entdecken“, welche Aspekte für die Kunden dabei überhaupt relevant sind. Dazu werden typischerweise so genannte „qualitative“ Methoden angewandt, die solche Einsichten ermöglichen, aber keine exakten Messungen und Schlussweisen erlauben.

3.1 Kennzeichnung

Die im Abschn. 2.4.1 erläuterten explorativen Untersuchungen sind die Domäne des Einsatzes qualitativer Untersuchungsmethoden (siehe auch Abschn. 2.4.2.2) in der Marktforschung. Das gilt sowohl für den Bereich anwendungsorientierter Untersuchungen in der Praxis als auch für die Grundlagenforschung. Wie die Bezeichnung „explorativ“ schon erkennen lässt, geht es hier also um *Entdeckung* von Zusammenhängen, Verhaltensmustern etc. Dem entsprechend liegt der Fokus qualitativer Methoden im Hinblick auf Grundlagenforschung bei der Theoriebildung und Entwicklung von Hypothesen (siehe dazu Abschn. 2.2.2); bei angewandter Forschung steht die Entwicklung von Problemverständnis sowie die Vorbereitung weiterer Untersuchungen im Vordergrund. In der kommerziellen Marktforschungspraxis sind aber auch explorative Studien auf Fragestellungen und Ziele ausgerichtet, die vom Auftraggeber vorgegeben werden und sind in diesem Sinne nicht so frei wie vorstehend skizziert.

Die Diskussion über Vor- und Nachteile und sogar die Existenzberechtigung sogenannter „quantitativer“ (z. B. repräsentative Befragungen, Experimente) und quali-

H. Kreis et al., *Marktforschung,* https://doi.org/10.1007/978-3-658-44456-3_3

tativer Methoden wird seit Jahren in den Sozialwissenschaften intensiv – teilweise erbittert – geführt. Dabei geht es nicht zuletzt um grundlegend verschiedene wissenschaftstheoretische Standpunkte und Wissenschaftsauffassungen (siehe dazu z. B. Hunt, 2010; Eisend & Kuß, 2023). Man findet in der Literatur hauptsächlich drei Argumentationsstränge zur Begründung der Anwendung qualitativer Methoden in der Markt- und Sozialforschung:

Ein **„pragmatischer Ansatz“**, der auch in diesem Buch vertreten wird, geht davon aus, dass in frühen Phasen einer anwendungsorientierten Untersuchung ein Problemverständnis entwickelt werden muss (s. o.), wozu eben qualitative Methoden besser als quantitative geeignet sind. Das gilt analog auch für die auf *Entwicklung* von Theorien ausgerichtete wissenschaftliche Forschung. Bei diesem Ansatz folgen im Forschungsprozess oftmals weitere („quantitative“) Untersuchungen. Diese „pragmatische“ Sichtweise ist in der internationalen Marketingforschung seit langer Zeit dominierend.

Hintergrundinformation

Der international renommierte Wissenschaftsphilosoph Gerhard Schurz (2014, S. 37) formuliert eine entsprechende Position:

> „Die ideologische Polarisierung bezüglich quantitativer und qualitativer Methoden, die von wenigen qualitativen Forschern (...) vertreten wird, erscheint als überflüssig und übertrieben. Es ist eher so, dass sich qualitative und quantitative Methoden ergänzen. Die Stärke qualitativer Methoden (z. B. Fallstudien, narrative Interviews) liegt bei einer Vorstufe quantitativer Methoden – bei der Entdeckung relevanter Einflussfaktoren und bei der Entwicklung vielversprechender Hypothesen. Aber einer qualitativen Erkundung muss eine quantitativ-statistische Analyse folgen, die der einzige Weg ist, um die generelle Aussagekraft einer Hypothese zu testen, insbesondere in einer Situation, in der man noch nicht über bewährtes Hintergrundwissen verfügt. Dass sich qualitative und quantitative Methoden ergänzen, ist eine verbreitete Auffassung unter empirischen Forschern in den Sozialwissenschaften, obwohl diese Ansicht nicht unumstritten ist.“

Der **„relativistische Ansatz“** hat in der Marketingwissenschaft einige Zeit eine gewisse Rolle gespielt. Er knüpft an die seinerzeit stark beachtete und diskutierte Analyse wissenschaftlicher Entwicklungen von Thomas Kuhn (1970) vor dem Hintergrund sogenannter „Paradigmen“ an. Dabei geht es wesentlich darum, dass solche Paradigmen („Weltanschauungen“, Hunt, 2003, S. 115) die herrschenden Theorien und die typische Forschungsmethodik einer wissenschaftlichen Disziplin in einer Epoche prägen. Andere Paradigmen führen danach zu ganz anderen Sichtweisen. Insofern gibt es in dieser Sichtweise nicht eine bestimmte Forschungsmethode, die zu „objektiven“, der Wahrheit entsprechenden Erkenntnissen führt. Wissenschaft ist in dieser Perspektive „relativ“, also vom jeweils herrschenden Paradigma abhängig (siehe dazu z. B. Hunt, 2010; Eisend & Kuß, 2023). Deswegen kommt es hier auch nicht auf Methoden an, die – wie das für quantitative Methoden typisch ist – mit dem Bestreben verbunden sind, zu objektiven Aussagen über die Annahme oder Ablehnung von Hypothesen zu führen. Qualitative Methoden sind in *diesem Kontext* für die Forschung ausreichend und bieten auch den Vorteil größerer Ergiebigkeit bei der Theorie*bildung*.

Der **„interpretative Ansatz“** spielt in Teilen der Soziologie und der Konsumentenforschung eine Rolle (siehe z. B. Neuman, 2011; Hirschman, 1989). Dem liegt eine Auffassung zugrunde, dass man soziale Phänomene und menschliches Verhalten in ihrem Kontext analysieren und verstehen müsse. In dieser Perspektive haben qualitative Methoden – z. B. Fallstudien – deutliche Vorteile gegenüber „quantitativen“ Methoden, bei denen man in der Regel nur wenige ausgewählte Variable unter Ausklammerung des jeweiligen Kontexts analysiert.

Die **qualitative Methodik** ist also auf die *Gewinnung* von Ideen und Hypothesen ausgerichtet, nicht auf exakte Messungen und Analysen. Bei aller Vielfalt qualitativer Untersuchungen (siehe die folgenden Abschnitte) kann man einige typische Gemeinsamkeiten identifizieren, die nachstehend in Anlehnung an Yin (2011, S. 7 ff.) und Kuß (2010) überblicksartig gekennzeichnet werden:

- Geringe Festlegung und Strukturierung des Forschungsprozesses in dem Sinne, dass hier die Untersuchung nicht in vorher genau festgelegten Schritten folgt
- Enge Verflechtung von Datenerhebung, Analyse und Hypothesengenerierung im Forschungsprozess
- Kleine (meist nicht repräsentative) Stichproben von Auskunfts- oder Versuchspersonen (in der Regel: $n \ll 100$; oft: $n < 10$)
- Vorwiegend beschreibende und interpretierende Analyse, kaum quantifizierende Aussagen (wie z. B. Prozent-Angaben, statistische Maßzahlen)
- Anwendung freier (nicht standardisierter) Formen von Befragungen und Beobachtungen unter Einbeziehung der ForscherInnen in die Datenerhebung, die Vertrautheit der ForscherInnen mit dem Untersuchungsgegenstand erfordert
- Untersuchung menschlichen Verhaltens unter möglichst natürlichen Bedingungen (z. B. eher Beobachtungen in realer Situation oder Interviews in natürlicher Gesprächssituation als Laborexperimente oder standardisierte Interviews)
- Einbeziehung der Rahmenbedingungen, unter denen ein Verhalten stattfindet (und nicht isolierte Betrachtung nur weniger ausgewählter Variabler)
- Ausrichtung auf umfassendes Verstehen menschlichen Verhaltens und nicht nur Prüfung isolierter Hypothesen
- Verwendung mehrerer Informationsquellen (z. B. Dokumente, Interviews, Beobachtungen) in einer Untersuchung und nicht nur Anwendung einer einzigen Untersuchungsmethode (z. B. ein Experiment)

Trotz der eher explorativen Ausrichtung qualitativer Forschung wird auch von Vertretern dieses Ansatzes zunehmend auf Reliabilität und Validität solcher Untersuchungen geachtet, um den Eindruck der Beliebigkeit von Ergebnissen zu vermeiden (siehe z. B. Silverman, 2005; Yin, 2011). Eine zentrale Rolle spielt hier das Stichwort **„Triangulation“**. Die entsprechende Grundidee bezieht sich darauf, einen Gegenstand aus verschiedenen Blickwinkeln zu betrachten. Das sind im Zusammenhang mit Forschungsmethoden insbesondere verschiedene Beobachter und verschiedene Informationsquellen (Neuman,

2011, S. 164). Hier sei bereits auf eine Analogie zu dem im Kap. 4 unter dem Stichwort „Konvergenzvalidität“ diskutierten Ansatz verwiesen. Dabei geht es um übereinstimmende Ergebnisse von Messungen eines Konstrukts mit möglichst unterschiedlichen Methoden; hier – bei qualitativer Forschung – geht es darum, dass übereinstimmende Informationen von verschiedenen Personen oder aus verschiedenen Quellen als Bestätigung für die Korrektheit dieser Informationen angesehen werden.

3.2 Gruppendiskussionen

Die heute wohl gängigste Form qualitativer Marktforschung ist die **Gruppendiskussion (Focus Group Interview**). Darunter versteht man die gleichzeitige Befragung von mehreren (meist 6–10) Auskunftspersonen, denen Interaktionen untereinander zumindest gestattet sind (vgl. z. B. Iacobucci & Churchill, 2010, S. 63; Yin, 2011, S. 141). Die Gruppendiskussion steht in der Regel unter der Leitung eines psychologisch geschulten Diskussionsleiters und konzentriert sich auf ein vom Auftraggeber der Untersuchung vorgegebenes Thema. Als spezifische **Vorteile** von Gruppendiskussionen insbesondere im Vergleich zu Einzelinterviews (siehe Abschn. 3.3) werden genannt:

- Stimulierung der Teilnehmer im Hinblick auf Reflexion und Meinungsäußerungen durch die – möglichst rege – Diskussion innerhalb der Gruppe.
- Annäherung an natürliche Gesprächssituation in der Gruppe, bei der Hemmungen der Teilnehmer abgebaut werden.
- *Relativ* geringe Kosten pro teilnehmender Person.
- Gute Möglichkeiten, Ablauf und Inhalt der Diskussion an das Management zu übermitteln, weil typischerweise (Audio- oder Video-) Aufzeichnungen gemacht werden und/oder Vertreter des Managements hinter einem Einwegspiegel die Diskussion live mitverfolgen können.

Beim Vergleich zum qualitativen Interview (siehe Abschn. 3.3) ergeben sich aber auch drei **Nachteile:**

- Weniger Information pro teilnehmender Person.
- Geringere Vollständigkeit der Angaben der Einzel-Personen, weil sich nicht jeder Teilnehmer zu jedem Aspekt äußert.
- Vor allem bei kleinen, viel beschäftigten Zielgruppen, die räumlich weit verstreut sind (z. B. Einkäufer in einer bestimmten Branche) ist es schwierig, diese an einem Ort zu einer Zeit zusammenzubringen. Dieser Nachteil verliert jedoch durch die Möglichkeit, Gruppendiskussionen online durchzuführen, an Bedeutung.

Hintergrundinformation
Einen Vor- und einen Nachteil von Gruppendiskussionen kennzeichnet Yin (2011, S. 142):

> „Ein klarer ‚trade-off' im Vergleich zu individuellen Interviews ist der Zugewinn an Effizienz (Gespräche mit mehreren Leuten zur gleichen Zeit) auf der einen Seite und die Einbuße beim Tiefgang (weniger Informationen von jeder einzelnen Auskunftsperson) auf der anderen Seite. Allerdings hat der wesentliche Grund für die Durchführung von Gruppendiskussionen damit nichts zu tun. Vielmehr sind Gruppendiskussionen nützlich, wenn man vermutet, dass bestimmte Leute (z. B. Jugendliche oder Kinder) sich eher artikulieren, wenn sie in einer Gruppe sind, als wenn sie mit einem Einzelinterview konfrontiert werden."

Während einer Gruppendiskussion achtet der Leiter/die Leiterin vor allem auf die Einhaltung und möglichst vollständige „Abarbeitung" des vorher festgelegten Themas und der damit verbundenen Einzel-Aspekte und darauf, dass der Gesprächsfluss in Gang gehalten wird. Oftmals ist es auch notwendig, einen Ausgleich zwischen „starken" und „schwachen" Gruppenmitgliedern herbeizuführen, um eine Dominanz einzelner Personen zu verhindern und die Artikulationen zurückhaltender Untersuchungsteilnehmer zu ermöglichen.

Beispiel

Hier ein Beispiel eines Gesprächsleitfadens für eine Gruppendiskussion über Urlaubsreisen:

1. Begrüßung und Einführung
2. Frage nach dem letzten und nächsten Urlaubsziel
3. Wann und wo werden Reisen gebucht?
4. Auswahl von Reiseangeboten (Art der Unterbringung, Anreise, Voll- oder Halbpension, Unterhaltungsangebot, Preis)
5. Welche Rolle spielen Last-Minute-Angebote?
6. Wie wichtig sind Vertrautheit des Reiseziels und Vertrauen zum Veranstalter? ◀

Der entscheidende Gesichtspunkt hinsichtlich der Zusammensetzung der Gruppe ist der, dass alle Teilnehmer eine **Beziehung zum vorgegebenen Untersuchungsthema** haben müssen, weil anderenfalls keine oder keine hinreichend substanziellen Äußerungen zu erwarten sind. Wenn man sich beispielsweise vorstellt, dass Gruppendiskussionen in frühen Phasen der Produktentwicklung des iPad oder von Kosmetikartikeln zur Entwicklung oder Überprüfung entsprechender Konzepte eingesetzt werden, so dürfte es einleuchten, dass die Einbeziehung von technisch desinteressierten Rentnern (beim iPad) bzw. von im Hinblick auf Ästhetik gleichgültigen Männern (bei Kosmetika) in solche Gruppendiskussionen wohl wenig ergiebig wäre. Bei der Zusammenstellung von Gruppen stellt sich noch die Frage, ob man deren **Homogenität** oder **Heterogenität** anstreben soll. In hinsichtlich sozialer und psychischer Merkmale relativ homogenen Gruppen (z. B. berufstätige Ehefrauen im Alter von 20–30 Jahren mit gehobener Bildung und modernem Lebensstil aus städtischen Wohngebieten) findet man häufig ähnliche oder

schnell konvergierende Meinungen. Dagegen werden in heterogenen Gruppen die Teilnehmer durch sehr unterschiedliche Meinungen und Erfahrungen stärker gefordert und zur intensiven Auseinandersetzung mit dem Untersuchungsthema gereizt. Allerdings sollte man auch darauf achten, dass die Teilnehmer nicht zu heterogen sind, weil dann die Diskussion in unergiebige Bahnen geraten kann. So ist zu fürchten, dass bei einer Diskussion über Wäschepflege, an der Frauen und Männer teilnehmen, die Diskussion sich mehr um eine faire Aufteilung der Hausarbeit dreht als um das eigentliche Thema.

Es erfolgt in der Regel eine Audio- oder Video-Aufzeichnung des Gesprächsverlaufs, wobei letztere Art natürlich die umfassendere **Auswertung** auch im Hinblick auf nicht-verbale Reaktionen (z. B. Mimik) erlaubt. Teilweise werden Gruppendiskussionen auch in speziellen Teststudios durchgeführt, in denen der Auftraggeber durch einen Einwegspiegel den Gesprächsverlauf beobachtet und über eine Mikrofon-Ohrhörer-Verbindung mit dem Diskussionsleiter auch beeinflussen kann. Bei der Auswertung der Gespräche in Form von Protokollen und schriftlichen Zusammenfassungen steht normalerweise nicht die Verdichtung von Einzelaussagen zu einem nur scheinbar eindeutigen Ergebnis, sondern die Wiedergabe der auftretenden Vielfalt von Gesichtspunkten und Argumenten im Vordergrund. Inzwischen haben **Online-Gruppendiskussionen,** bei denen die Kommunikation über das Internet erfolgt, wegen ihrer praktischen Vorteile (kein physisches Zusammentreffen nötig, automatische Aufzeichnung des „Gesprächs"-Verlaufs, regionale Streuung der Teilnehmer möglich) deutlich wachsende Bedeutung erlangt (Burns & Bush, 2006, S. 212).

Beispiel

Sudman und Blair (1998, S. 190) und Burns und Bush (2006, S. 214 ff.) stellen einige Beispiele für sinnvolle Anwendungen von Gruppendiskussionen in der Marktforschung zusammen:

- Generierung und erster Test von Ansätzen für neue Produkte
- Untersuchung von Arten und Zwecken der Produktnutzung
- Untersuchung von Problemen bei der Nutzung von Produkten
- Entdeckung von Bedürfnissen, Motiven, Wahrnehmungen etc. der Konsumenten
- Ergebnisse „quantitativer" Untersuchungen illustrieren und besser verstehen
- Sprache der Konsumenten verstehen
- Analyse von Einstellungen und ihrer Gründe
- Identifizierung von Gesichtspunkten, die später in eine standardisierte Untersuchung einbezogen werden sollen ◄

3.3 Qualitative Interviews

Bei qualitativen Untersuchungen spielen (neben Gruppendiskussionen) auch Einzel-Interviews eine Rolle. Diese sind typischerweise nicht – jedenfalls nicht im Kern – standardisiert. Allenfalls einige Passagen (z. B. Ermittlung soziodemographischer Merk-

male) können standardisiert sein. Auch hier werden Ablauf und Inhalt des Gesprächs nur in einem Leitfaden relativ grob festgelegt. Die in der Sozial- und Marktforschung prominenteste Form eines solchen Interviews stellt das qualitative Interview dar.

▶ **Definition** Nicola Döring und Jürgen Bortz (2016, S. 365) kennzeichnen *qualitative Interviews* folgendermaßen:

> „Qualitative Interviews (….) arbeiten mit offenen Fragen, so dass sich die Befragten mündlich in eigenen Worten äußern können. Zudem wird der Gesprächsverlauf weniger von den Interviewenden und ihren Fragen vorstrukturiert, sondern stärker von den Befragten mitgestaltet. Auf diese Weise sollen die individuellen Sichtweisen der Befragten nicht nur oberflächlich, sondern detailliert und vertieft erschlossen werden."

Qualitative Interviews unterscheiden sich grundlegend von den allseits bekannten standardisierten Interviews, die im 4. Kapitel dieses Buches ausführlich behandelt werden. Robert Yin (2011, S. 134 f.) hebt dabei drei Gesichtspunkte hervor:

a) Die Interaktion zwischen InterviewerIn und Auskunftsperson ist nicht wie bei standardisierten Befragungen in (möglichst) allen wesentlichen Einzelheiten (z. B. Fragereihenfolge und -formulierung) festgelegt.
b) Das Interview wird – auf der Basis eines thematischen Leitfadens – an den Charakter eines normalen zwischenmenschlichen Gesprächs angepasst.
c) Bei den wichtigen Themen werden keine festgelegten Antwortkategorien vorgegeben, vielmehr können die Auskunftspersonen frei mit ihren eigenen Worten formulieren.

Vor dem Hintergrund einer solchen Interview-Situation kann man sich gut vorstellen, dass bei bestimmten Fragen und Nachfragen den Auskunftspersonen Motive, Verhaltensweisen etc. bewusst werden, die sie selbst bisher kaum wahrgenommen haben bzw. an die sie sich ansonsten kaum erinnern würden. Zu den Einzelheiten der Durchführung qualitativer Interviews sei hier auf Döring und Bortz (2016, S. 365 ff.) verwiesen, die ausführlich deren typischen Ablauf skizzieren.

Hier einige spezifische Vorteile und Probleme von qualitativen Interviews:

Vorteile

- Man erhält in den entsprechenden Gesprächsprotokollen vollständige Gedanken- und Argumentationsketten, die sehr viele Einzel-Aspekte enthalten. Damit können komplexe psychische Zusammenhänge – z. B. bei der Entwicklung von Markenpräferenzen – relativ gut abgebildet werden.
- Die verschiedenen Aussagen sind einzelnen Personen klar zuzuordnen, was bei Gruppendiskussionen gelegentlich Schwierigkeiten bereitet.
- Man kann von Auskunftspersonen Informationen erhalten, die diesen ohne das Interview nicht bewusst geworden wären (siehe oben).

Probleme

- Die Anforderungen an die Interviewer sind hoch. Sie müssen im Hinblick auf die Interviewtechnik speziell geschult sein und benötigen ein tiefgehendes Verständnis des jeweiligen Untersuchungsgegenstandes.
- Die Auskunftspersonen müssen gewissen intellektuellen Mindestanforderungen genügen, insbesondere hinsichtlich ihrer Verbalisierungsfähigkeit.
- Relativ hohe Kosten pro teilnehmender Person (im Vergleich zur Gruppendiskussion).
- Die Ergebnisse sind recht unübersichtlich und untereinander nicht leicht vergleichbar.

Im Zusammenhang mit qualitativen Interviews werden gelegentlich auch so genannte **projektive Techniken** angewandt (siehe z. B. Gröppel-Klein & Königstorfer, 2009). Diese Techniken der Marktforschung basieren auf der Neigung vieler Menschen, eigene unangenehme Gefühle, Meinungen, Verhaltensweisen etc. auf andere Leute zu übertragen („zu projizieren"). Dem entsprechend werden bei unangenehmen Fragen „Projektionshilfen" angeboten. Mit diesem Ausweg kann die Auskunftsperson über Unangenehmes Auskünfte geben, ohne eine direkte Beziehung zu sich selbst herstellen zu müssen. Beispielsweise gibt es sicher mehr Menschen, die darüber berichten, dass in ihrer Nachbarschaft, Kollegenschaft etc. viel Fast Food verzehrt wird, als Menschen, die das für sich selbst zugeben. Dieses sozial unerwünschte Verhalten wird also auf andere Leute *projiziert.*

3.4 Fallstudien

Bei Problemen der Markt- und Managementforschung – vor allem bei explorativen Untersuchungen – werden gelegentlich umfassende Analysen von Einzelfällen verwendet. Beispielsweise haben für das Verständnis der besonders komplexen organisationalen Beschaffungsprozesse derartige **Fallstudien** eine bedeutsame Rolle gespielt. Fallstudien können sich auf Abläufe/Ereignisse (z. B. Innovationsprozesse), Personen (z. B. Entstehung einer Markenbindung), Organisationen (z. B. Struktur und Strategie) oder soziale Einheiten (z. B. Gruppen, Gemeinden) beziehen. Typisch für eine solche Fallstudie ist die Anwendung unterschiedlicher Datenquellen und Erhebungsmethoden zur umfassenden Beschreibung des jeweiligen Falles. Als Beispiele seien hier Auswertungen von Aufzeichnungen und Dokumenten (z. B. Schriftverkehr, Protokolle), Beobachtungen und Experten-Interviews genannt.

▶ **Definition** Robert Yin (2009, S. 18) definiert Fallstudien in folgender Weise:

> „Eine Fallstudie ist eine empirische Untersuchung, bei der ein aktuell stattfindendes Phänomen in die Tiefe gehend und in seinem Kontext analysiert wird, insbesondere wenn die Grenzen zwischen dem Phänomen und seinem Kontext nicht deutlich sind."

Als weitere Charakteristika hebt Yin (2009) hervor, dass in der Regel eine Vielzahl von Merkmalen bei einer relativ geringen Zahl von Fällen erhoben wird und dass sich die Datenerhebung auf *reale* Abläufe bezieht, nicht auf mehr oder weniger künstliche Untersuchungssituationen. Allerdings gibt es offenbar bei publizierten Fallstudien nicht selten Probleme bei der Sicherung der Validität der Studien (Gibbert & Ruigrok, 2010).

Hintergrundinformation
Iacobucci und Churchill (2010, S. 62) kennzeichnen zentrale Anforderungen bei Fallstudien:

> „Bei der Analyse eines Falles zeichnet der Forscher alle relevanten Daten auf, nicht nur die, die seine anfänglichen Hypothesen stützen. Bei explorativer Forschung besteht das Ziel darin, Einsichten zu gewinnen, nicht Erklärungsmöglichkeiten zu testen. Durch seine Neutralität ist es für den Forscher leichter, flexibel im Hinblick auf neu auftauchende Informationen zu bleiben. Der Forscher muss auch in der Lage sein, bei der Beschäftigung mit vielen Einzelheiten das „große Bild" zu erkennen, Einsichten, die für mehrere Fälle relevant sind, und nicht nur bei einem bestimmten Fall zutreffen."

Besondere Aussagekraft haben Fallstudien in der Marktforschung u. a. in den folgenden Arten von Fragestellungen:

- Darstellung und Verständnis komplexer Prozesse. Oben ist dazu schon das Beispiel organisationaler Beschaffungsprozesse genannt worden. Dabei handelt es sich oftmals um Prozesse, die längere Zeit erfordern und die mit vielfältigen Interaktionen zwischen verschiedenen Personen aus mindestens zwei Organisationen verbunden sind. Für derartige Analysen sind weder die sonst üblichen Datenerhebungsmethoden noch die gängigen statistischen Methoden ausreichend.
- Analyse und Vergleich extremer Fälle. Bei einigen praktischen und wissenschaftlichen Fragestellungen kann der Vergleich von Extremfällen anregend und informativ sein, beispielsweise der Vergleich von Merkmalen und Vorgehensweisen von besonders erfolgreichen und weniger erfolgreichen Verkäufern oder der Vergleich erfolgreicher und nicht erfolgreicher Produktinnovationen. Daraus lassen sich Merkmale identifizieren, die wesentliche Einflussfaktoren sein können.

3.5 Ethnografische Marktforschung

Die zunehmende Entfernung vieler Manager von der Situation ihrer Verbraucher hat in den letzten Jahren das Bedürfnis wachsen lassen, den Verbraucher in seiner Umgebung beim Gebrauch der Produkte zu beobachten. Ziel ist dabei, durch die direkte Beobachtung auch solche Bedürfnisse, Hindernisse und Anwendungen zu ermitteln, die in Interviews nicht direkt angesprochen werden, sei es, weil niemand auf die Idee kommt, danach zu fragen, sei es, weil die Verbraucher sich scheuen, dies anzusprechen. Zu diesem Zweck nahm die Marktforschung Anleihen bei der Völkerkunde. So wie Ethno-

grafen die Gebräuche von bisher unbekannten Völkern studieren, indem sie eine Zeit lang bei diesen leben, so gehen auch der Marktforscher und der Marketing-Manager in die Häuser und Wohnungen ihrer Verbraucher und versuchen so, den Gebrauch ihrer Produkte im Alltagskontext besser zu verstehen.

Dabei soll die Beobachtung der **Nutzung** und des Nutzungszusammenhangs bisher nicht bekannte Nutzungsarten, Verwendungsbarrieren und nicht befriedigte Bedürfnisse offenlegen. Meist geschieht dies, um Ideen für Produktverbesserungen bzw. für Neuprodukte zu generieren oder auch die Kommunikation der Bedürfnisse zu verbessern. So zitiert Mariampolski (2007) den früheren CEO von Procter & Gamble, A.G. Lafley, der aufgrund von Interviews stets der Meinung war, dass es den Verbrauchern leichtfällt, die Waschmittelpackung zu öffnen, der dann aber sehen musste, wie Hausfrauen die Packung mit Schraubenziehern traktierten, um sie aufzubekommen. Dies war für ihn Anlass, einen einfach zu handhabenden Öffner für Waschmittelpackungen zu entwickeln.

Beispiel

Ein Unilever-Manager berichtete in einem Diskussionsbeitrag auf einer MSI Konferenz (MSI Marketing Science Institute) am 11.4.2008 in Boston, dass bei einer ethnografischen Forschung festgestellt wurde, dass Frauen in Bolivien Seife zerhackten und mit Wasser zu einer Waschpaste verarbeiteten, weil nur so ihre Männer nach der Arbeit im Bergwerk sich wieder sauber waschen konnten. Dies lieferte die Idee für ein einfach herzustellendes und erfolgreiches Neuprodukt: Eine in kleinen Plastikeimern verpackte Waschpaste. ◀

Eine besondere Form der ethnografischen Forschung sind **Shopper-Studien,** bei denen Verbraucher beim Einkaufen im Supermarkt beobachtet werden, oft unterstützt durch Videoaufzeichnung und/oder Blickregistrierung. Hintergrund ist die Tatsache, dass ein Produkt oft nur 1 bis 2 s Zeit hat, in den Einkaufswagen zu gelangen. Es ist daher sehr wichtig, zu verstehen, was in dieser kurzen Zeit geschieht. Welches andere Produkt wird noch angeschaut? Wie wird im Regal gesucht? Was wird auf der Packung gelesen? Zusätzliche Erkenntnisse lassen sich oft noch gewinnen, wenn nach dem Einkauf den Verbrauchern das Video vorgeführt wird und sie nach den Gründen für das Verhalten gefragt werden.

Ethnografische Forschung kann nicht nur zum Verständnis der Endverbraucher beitragen, sondern kann auch für die Marktforschung bei Unternehmen hilfreich sein. So kann es für Unternehmen wie SAP oder Microsoft sinnvoll sein, Angestellte bei Kundenunternehmen zu beobachten, wie ihre Software genutzt wird, welche Einsatzgebiete besonders häufig sind und an welchen Punkten es Probleme bei der Anwendung gibt. Hier geht die Ethnografische Forschung in die **„User Experience Forschung“** oder kurz **„UX-Forschung“** über.

Die Rekrutierung von Privatpersonen geschieht meist durch die Zahlung eines Geldbetrags, bei Unternehmen eher durch die Aussicht, dass die Produkte besser auf ihre Bedürfnisse hin weiterentwickelt werden. Die Datenerhebung bei der ethnografischen

Forschung erfordert viel Fingerspitzengefühl. Dabei sollte ein positives und wertschätzendes Klima zum „Gastgeber" geschaffen werden. Die Beobachter sollen freundliches und bestätigendes Interesse deutlich machen, jedoch nicht positiv oder negativ werten, um das Verhalten möglichst wenig zu beeinflussen. Das Verhalten der Probanden wird möglichst detailliert protokolliert, oft unterstützt durch Fotos, Ton- oder Videoaufzeichnungen. Anschließend können diese Aufzeichnungen den Probanden auch vorgeführt werden, um sie z. B. nach den Gründen für das jeweilige Verhalten zu fragen. Ethnografische Studien liefern meist sehr viel Material. Die Sichtung und zielgerichtete Auswertung, oft in einem Workshop gemeinsam mit dem Kunden, stellt einen nicht unerheblichen Aufwand dar. Wie bei jeder qualitativen Forschung kann auch ethnografische Forschung Ideen liefern, die dann in *quantitativen* Studien genauer zu untersuchen und in ihrer Bedeutung zu bewerten sind.

Nicht unproblematisch ist in Deutschland die Einhaltung der Standesregeln (siehe dazu Kap. 10). Will das beauftragende Unternehmen an der Untersuchung selbst teilnehmen, was sehr häufig der Fall ist, so darf die Untersuchung in der Wohnung bzw. im Haus der Probanden nicht durch ein Marktforschungsinstitut organisiert werden, da dann die Anonymität der Befragten nicht gewahrt werden kann. Die Untersuchung wird in diesem Fall entweder vom Unternehmen selbst oder von einer Agentur organisiert, die kein Marktforschungsunternehmen ist.

3.6 Neue, auf dem Internet basierende Formen

Die Entwicklung des Internets ermöglicht neue Formen der internetbasiert qualitativen Forschung mit weiter gehenden Erkenntnismöglichkeiten.

So ist es möglich, dass die Teilnehmer über eine spezielle Internetplattform miteinander und mit der Studienleitung kommunizieren. Ihnen können über mehrere Tage hinweg verschiedene Aufgaben gestellt werden, mit denen das Thema von möglichst umfassend Seiten beleuchtet werden kann. Parallel dazu werden sie dazu angehalten, in einem Forum über das Thema frei zu diskutieren oder auch Anregungen und Ideen zum Thema in einem Inspirationsalbum der Gruppe zur Verfügung zu stellen.

Beispiel

Wildner (2017) schildert ein solches Projekt, mit dem ermittelt werden sollte, was junge Menschen, die vor der Gründung eines eigenen Haushalts stehen, über die künftige Haushaltsführung denken und wie sie diese gestalten wollen.

Dazu wurden 19 junge Männer und 24 junge Frauen im Alter von 19 bis 24 Jahren persönlich in zwei Städten angeworben und bekamen Zugang zu einer speziellen Internetplattform, die nur für diese Studie eingerichtet wurde. Die Datenerhebung über insgesamt fünf Tage begann mit einer kurzen Vorstellung und der Schilderung der aktuellen Haushaltsführung bei den Eltern. Anschließend wurden sie nach Vorbildern und Anregungen für die eigene künftig selbständige Haushaltsführung gefragt.

Ein Ergebnis war, dass die Eltern hier vielfach Vorbild sind. Von einem Generationenkonflikt war nichts zu merken.

In einem zweiten Schritt wurden sie zu ihren Planungen und Vorstellungen für die Zeit nach ihrem Auszug befragt. Was wollen sie gegenüber dem Zustand zu Hause ändern, was beibehalten? Ein Ergebnis: Auch wenn viel vom Vorbild der Eltern beibehalten wird, so soll doch mehr Technik eingesetzt werden, um vor allem lästige Tätigkeiten zu verkürzen.

Die dritte Aufgabe war die Erstellung einer Collage, mit der sie ihren Vorstellungen zu ihrer Haushaltsführung der Zukunft bildlich Ausdruck verliehen. Auch hier drückte sich der Wunsch nach Vereinfachung aus, um Zeit für die interessanten Aspekte des Lebens zu gewinnen. Auf einer Collage waren Putzroboter, ein Lieferdienst und eine Haushaltshilfe im Einsatz. Dadurch sollte gemeinsame Zeit für die Familie am „sozialen Lagerfeuer", also gemeinsames Kochen und Essen, gewonnen werden.

Im nächsten Schritt wurde direkt zu bestimmten Zukunftskonzepten und Geräten gefragt. Klares Ergebnis war, dass solche Geräte gewünscht sind, aber nur dann, wenn sie nicht die eigene Entscheidungsfreiheit beeinträchtigen. Ein Knopf an der Waschmaschine, der immer ein ganz bestimmtes vom Hersteller voreingestelltes Waschmittel bestellt, wird abgelehnt.

Parallel dazu wurden Themen wie die faire Verteilung der Hausarbeit auf die Partner oder auch die Vor- und Nachteile der Verwendung von Bio- bzw. veganen Lebensmitteln diskutiert.

Insgesamt konnte über die fünf Tage ein breites Spektrum an Themen in ausreichender Tiefe behandelt werden. ◀

Ein weiteres wichtiges Instrument sind **Social-Media-Analysen**. Dabei werden Äußerungen von Verbrauchern in sozialen Medien zu einem Thema gesucht und systematisch ausgewertet. Solche Analysen sind eher untypische qualitative Studien, da sie häufig mit großen Fallzahlen arbeiten. Dennoch sind sie der qualitativen Forschung zuzuordnen, da ihr Ziel die Exploration von Ansichten ist und da sie keine quantifizierbaren Ergebnisse über eine wohl definierte Grundgesamtheit zulassen.

Social-Media-Analysen können einmal als Frühwarnsystem dienen, da sie auf Probleme mit einem Produkt oder einer Dienstleistung aufmerksam machen können. Weiter machen sie die schon immer vorhandene, früher aber nur schwer und unvollkommen messbare Word-of-Mouth-Kommunikation der Verbraucher sichtbar. Schließlich sind sie Teil der medialen Äußerungen zu einer Marke. Es gibt Beispiele für markenrelevante Social-Media-Inhalte mit zweistelligen Millionenzahlen von Aufrufen, wie z. B. der auf Youtube hochgeladene Film „United breaks guitars", der ein Serviceproblem bei der Fluggesellschaft United Airlines thematisiert. In diesen Fällen werden Social-Media-Äußerungen zu Massenmedien. Ein Beispiel für eine Social-Media-Analyse findet sich bei Schubert und Unterreitmeier (2017), die Äußerungen zu dem elektrisch betriebenen Fahrzeug BMW i3 untersuchen.

3.7 Zur Analyse Qualitativer Daten

Die Analyse qualitativer Daten bezieht sich auf den Prozess der Untersuchung von nichtnumerischen Informationen, um Muster, Themen und Einsichten zu identifizieren und ggf. auf dieser Grundlage Praxis-Probleme besser zu durchdringen oder evtl. Theorien zu entwickeln. Im Gegensatz zur quantitativen Forschung, die sich auf entsprechende Daten konzentriert, beinhaltet die qualitative Analyse das Verständnis von Kontext und Bedeutungen, abgeleitet, in der Regel aus Textdaten.

Unabhängig davon, auf welche Art qualitative Daten erhoben werden, ergibt sich in der Regel eine komplexe und reichhaltige Datenlage, die verschiedene Herausforderungen mit sich bringt, die anhand der folgenden Übersicht zu den wichtigsten Schritten der Analyse quantitativer Daten deutlich werden. Die Ausführungen lehnen sich an, aber verkürzen die Darstellung von Miles et al. (2018).

1. **Datenerhebung:** Sammeln von Daten durch verschiedene Methoden wie Interviews, Fokusgruppen, Beobachtungen oder Textanalyse (siehe hierzu auch die vorangegangenen Abschnitte).
2. **Transkription:** Bei verbalen Daten wie Interviews müssen diese (i. d. R. Audioaufnahmen) oft transkribiert werden, um sie in schriftlicher Form weiter bearbeiten zu können.
3. **Kodierung:** Identifizieren und Markieren von bestimmten Abschnitten der Daten (Text oder andere Formen), die als bedeutungsvolle Einheiten betrachtet werden.
4. **Kategorisierung und Thematisierung:** Gruppieren von Kodes in breitere Kategorien und Identifizierung von übergeordneten Themen oder Mustern. Dies hilft, Schlüsselthemen oder Trends in den Daten zu erkennen.
5. **Interpretation:** Interpretation der Bedeutung der identifizierten Themen im Kontext der Forschungsfrage oder des Ziels. Dies erfordert oft eine tiefere Reflexion und Verbindung mit der Literatur.
6. **Theorie-Bildung:** Aufbau von theoretischen Konzepten oder Modellen, basierend auf den interpretierten Daten. Dies ermöglicht es, die Ergebnisse in einen größeren theoretischen Rahmen zu integrieren.
7. **Berichterstattung:** Präsentation der Ergebnisse in Form von Berichten, Artikeln oder Präsentationen. Dies beinhaltet oft die Verwendung von Zitaten oder Beispielen aus den Daten, um die Schlussfolgerungen zu stützen.
8. **Überprüfung** der Gültigkeit und Zuverlässigkeit der Analyse durch Peer-Review, Reflexivität und den Einsatz von verschiedenen Analysetechniken.

Die oben genannten Schritte bieten eine allgemeine Struktur, wobei die spezifischen Methoden je nach Forschungsdesign und Fragestellung variieren können (für eine entsprechende Übersicht siehe z. B. Flick, 2013). Es ist wichtig zu betonen, dass qualitative Forschung oft einen explorativen Charakter hat und sich auf die Entdeckung von Mustern und tieferes Verständnis von Phänomenen konzentriert.

Die **thematische Analyse (Thematic Analysis)** hat sich in den letzten Jahren zu einem häufig verwendeten Ansatz zur Analyse qualitativer Daten auch in den Wirtschaftswissenschaften entwickelt. Eine schnelle Suche bei Google Scholar mit den Schlüsselbegriffen „Thematic Analysis“ und „Economics“ (in Anführungszeichen) ergibt mehr als 120.000 Ergebnisse. Braun und Clarke (2019) weisen darauf hin, dass die thematische Analyse eine gemeinsame Geschichte mit der qualitativen Inhaltsanalyse (z. B. Mayring, 2015) hat und seit den 1980er Jahren als qualitativer Datenanalyseansatz in sozialwissenschaftlichen Studien verwendet wird. Allein der Schlüsselartikel „Using Thematic Analysis in Psychology“ von Braun und Clarke (2006) wurde bis Januar 2024 mehr als 185.000-mal zitiert.

Die Thematische Analyse konzentriert sich darauf, Themen und Muster in den gesammelten Daten zu identifizieren, um so ein tiefergehendes Verständnis für die Bedeutung von Texten zu gewinnen. In der Folge soll die Vorgehensweise dieser Methodik anhand eines fiktiven Beispiels kurz genauer vorgestellt werden. Die Ausführungen orientieren sich an dem von Braun und Clarke (2021) vorgeschlagenen Vorgehen:

Beispiel

Im Rahmen einer Forschungsarbeit soll die Verbraucherwahrnehmung von nachhaltigen Produkten im Rahmen einer thematischen Analyse untersucht werden.

1. **Datenerhebung:**
   Online frei verfügbare Kundenrezensionen von nachhaltigen Produkten können als Rohdaten dienen.
2. **Bekanntmachung mit dem Datenmaterial:**
   Die gesammelten Daten sollten mehrmals durchgelesen werden, um ein Verständnis für den Gesamtkontext zu entwickeln. Durch das Lesen der Rezensionen macht sich der Forscher/die Forscherin mit dem Datenmaterial vertraut, um ein erstes Verständnis für die verschiedenen Aspekte der Verbraucherwahrnehmung zu entwickeln. Bereits zu diesem Zeitpunkt werden relevante Textstellen, die auf Kundenmeinungen, Erwartungen oder Bedenken hindeuten, markiert.
3. **Kodierung:**
   Im Anschluss wird eine sogenannte offene Kodierung durchgeführt, bei der spezifische Textstellen identifiziert werden, die auf Schlüsselkonzepte im Zusammenhang mit der Verbraucherwahrnehmung von nachhaltigen Produkten hinweisen. Es werden Kodes erstellt, die verschiedene relevanten Aspekte der Kundenbewertungen repräsentieren. Diese Kodes sollen sich auf konkrete Aussagen, Ideen oder Konzepte beziehen. Im konkreten Fall könnte man sich vorstellen, dass in den Rezensionen unter anderem folgende Aspekte diskutiert werden:
   - inwieweit macht die Nutzung nachhaltiger Produkte überhaupt einen Unterschied?
   - der hohe Preis von nachhaltigen Produkten im Vergleich zu konventionellen Varianten

- der Preis als Qualitätsindikator
- Gefühl etwas Gutes zu tun, wenn man für Nachhaltigkeit einen höheren Preis bezahlt
- Vertrauen in Nachhaltigkeitssiegel und Produktstandards

4. **Kategorisierung:**
   Ähnliche Kodes werden zu breiteren Kategorien gruppiert, die die verschiedenen Dimensionen der Verbraucherwahrnehmung abdecken. Die Kategorisierung kann allein auf Grundlage der erhobenen Daten beruhen oder aber auch einen Bezug zu in der Literatur bereits existierenden Kategorien oder Konzepten haben. Beispielhaft könnten folgende Kategorien genannt werden:
   - Umweltfreundlichkeit: Bewertungen bezüglich der ökologischen Auswirkungen des Produkts
   - Preis als Barriere: Aussagen über einen zu hohen Preis
   - Preis als Anreiz: Einschätzungen des Preis-Leistungs-Verhältnisses von nachhaltigen Produkten
   - Produktqualität: Meinungen zur Qualität und Leistung des nachhaltigen Produkts
5. **Überprüfung der Kategorien:**
   Anschließend werden die Kategorien dahingehend überprüft, ob sie die Vielfalt der in den Kundenrezensionen geäußerten Meinungen angemessen repräsentieren.
6. **Benennung der Themen:**
   Den Kategorien werden dann aussagekräftige Namen gegeben, die die wesentlichen Aspekte der Verbraucherwahrnehmung von nachhaltigen Produkten reflektieren. Hierdurch werden die Kategorien zu den thematischen Elementen der Analyse. In diesem Falle könnten folgende Beispielthemen infrage kommen:
   - Umweltbewusstsein als Treiber: Kundenprioritäten im Zusammenhang mit ökologischen Überlegungen
   - Preis als Barriere und Anreiz: Wie der Preis die Wahrnehmung nachhaltiger Produkte beeinflusst
   - Qualitätsansprüche und Nachhaltigkeit: Die Verbindung von Produktqualität und Nachhaltigkeit
7. **Interpretation:**
   Im Rahmen der Interpretation wird dann untersucht, wie die einzelnen Themen miteinander in Beziehung stehen und welche Erkenntnisse sie über die Faktoren bieten, die die Verbraucherwahrnehmung von nachhaltigen Produkten beeinflussen. Ein Bericht, der die wichtigsten Erkenntnisse, Schlussfolgerungen und Handlungsempfehlungen enthält, wird erstellt. Vielfach wird dabei auch eine Visualisierung der Ergebnisse als sogenannte Thematic Map vorgenommen. Die Abb. 3.1 stellt eine hypothetische Thematic Map für das gewählte Beispiel dar. Die freien Elemente verdeutlichen dabei die Möglichkeit weiterer Kodes, Kategorien und Themen als die hier beispielhaft genannten. ◀

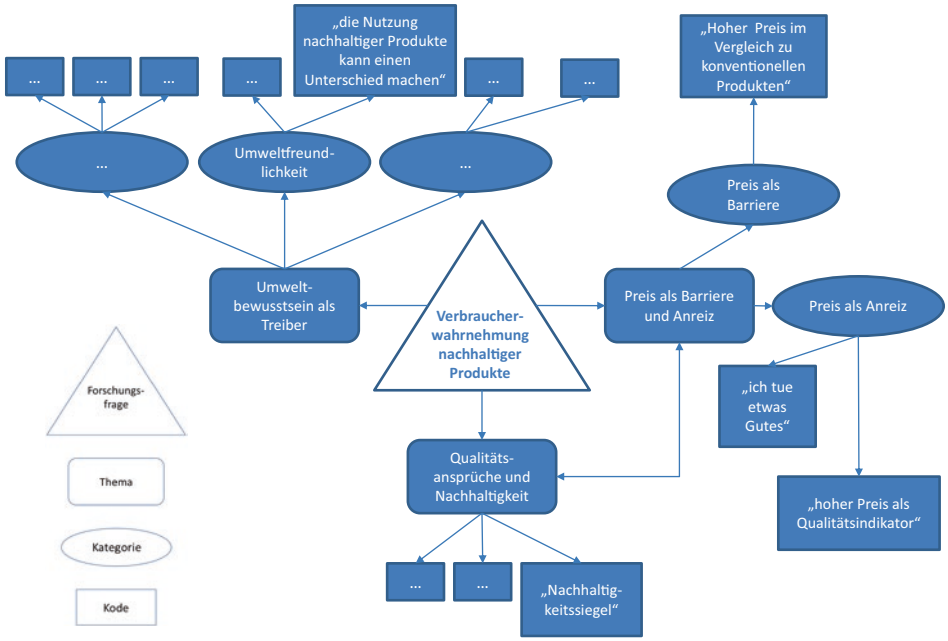


Abb. 3.1 Thematic Map für das Beispiel „Verbraucherwahrnehmung von nachhaltigen Produkten“

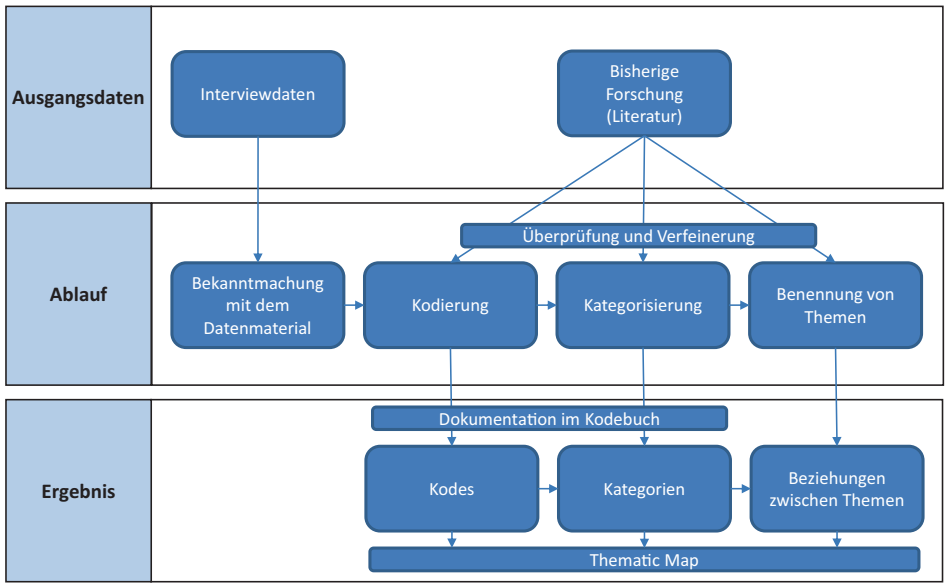


Abb. 3.2 Ein möglicher Ablauf einer Thematic Analysis

Die entwickelten Kodes, Kategorien und Themen werden in einem Kodebuch dokumentiert und klar charakterisiert; es dient als Referenz zur Sicherstellung einer konsistenten Analyse. Der oben dargestellte Ablauf ist in aller Regel nicht linear, sondern eher rekursiver Natur und lässt sich in Anlehnung an Becker et al. (2017) wie in Abb. 3.2 dargestellt verdeutlichen.

Literatur

Becker, M., Kolbeck, A., Matt, C., & Hess, T. (2017). Understanding the continuous use of fitness trackers: A thematic analysis. In *PACIS 2017 Proceedings 40*.

Braun, V. and V. Clarke. 2006. Using Thematic Analysis in Psychology. *Qualitative Research in Psychology* 3(2), 77–101.

Braun, V., & Clarke, V. (2019). Reflecting on reflexive thematic analysis. *Qualitative Research in Sport, Exercise and Health*, 11(4), 589–597.

Braun, V., & Clarke, V. (2021). *Thematic analysis – A practical guide*.

Burns, A., & Bush, R. (2006). *Marketing research* (5. Aufl.). Pearson.

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Aufl.). Springer.

Eisend, M., & Kuß, A. (2023). *Grundlagen empirischer Forschung* (3. Aufl.). SpringerGabler.

Flick, U. (2013). *The SAGE Handbook of qualitative data analysis*. SAGE Publications Ltd.

Gibbert, M., & Ruigrok, W. (2010). The „What" and „How" of case study rigor: Three strategies based on published work. *Organizational Research Methods, 13*, 710–737.

Gröppel-Klein, & Königstorfer, J. (2009). Projektive Verfahren in der Marktforschung. In R. Buber & H. Holzmüller (Hrsg.), *Qualitative Marktforschung – Konzepte, Methoden, Analysen* (2. Aufl., S. 537–554). Gabler.

Hirschman, E. (Hrsg.). (1989). *Interpretive consumer research*. Association for Consumer Research.

Hunt, S. (2003). *Controversy in marketing theory*. Sharpe.

Hunt, S. (2010). *Marketing theory – Foundations, controversy, strategy, resource-advantage theory*. Sharpe.

Iacobucci, D., & Churchill, G. (2010). *Marketing research – Methodological foundations* (10. Aufl.). Cengage.

Kuhn, T. (1970). *The structure of scientific revolutions* (2. Aufl.). University of Chicago Press.

Kuß, A. (2010). Mixed method-designs – Alter Wein in neuen Schläuchen? *Zeitschrift für Betriebswirtschaft, 2010, 2. Aufl.* (Special Issue 5), 115–125.

Mariampolsky, H. (2007). Ethnography and observational research. In M. van Hamersveld & C. de Bont (Hrsg.), *Market research handbook* (S. 435–445). Chichester.

Mayring, P. (2015). *Qualitative Inhaltsanalyse. Grundlagen und Techniken* (12. Aufl.). Beltz.

Miles, M. B., Hubermann, A. M., & Saldana, J. (2018): *Qualitative data analysis – A methods sourcebook* (4. Aufl.). Sage.

Neuman, W. (2011). *Social research methods – Qualitative and qualitative approaches* (7. Aufl.). Pearson.

Schubert, R., & Unterreitmeier, A. (2017). Das Auto elektrisiert das soziale Netz. *planung & analyse* 3/2017, 36–38

Schurz, G. (2014). *Philosophy of Science – A Unified Approach*. Routledge.

Silverman, D. (2005). *Doing qualitative research* (2. Aufl.). SAGE.

Sudman, S., & Blair, E. (1998). *Marketing research – A problem solving approach*. McGraw-Hill.
Wildner, R. (2017). Trends, die auch in Zukunft tragen. In GfK SE und GfK Verein (Hrsg.), *Trends, die tragen – Treiber einer erfolgreichen Markenführung, Nürnberg* (S. 76–93).
Yin, R. (2009). *Case study research – Design and methods* (4. Aufl.). SAGE.
Yin, R. (2011). *Qualitative research from start to finish*. Guilford.

Querschnittsuntersuchungen

4

Zusammenfassung

Das 4. Kapitel ist dem wohl gängigsten Untersuchungsdesign, den Querschnitts-Untersuchungen, gewidmet. Das sind Studien, bei denen auf den jeweiligen Zeitpunkt bezogene quantifizierende Aussagen über bestimmte Grundgesamtheiten (z. B. die Gesamtbevölkerung oder die Kunden eines Unternehmens) gemacht werden sollen. Zunächst geht es um die Stichprobenziehung, also um die Methoden, die dazu dienen, die befragten oder beobachteten Personen so auszuwählen, dass man mit hinreichender Sicherheit von deren Merkmalen auf die interessierende Grundgesamtheit schließen kann. Die am meisten angewandte Form der Datenerhebung ist die Befragung, die zwar sehr vielfältige Anwendungsmöglichkeiten bietet, die aber bei laienhaftem oder nachlässigem Vorgehen zu erheblichen Fehlern und Verzerrungen der Ergebnisse führen kann. Deswegen werden die Methoden zur Entwicklung von Fragebögen einschließlich der entsprechenden Gütekriterien (Reliabilität und Validität) hier ausführlich und mit zahlreichen Beispielen behandelt. Daneben wird auch – deutlich kürzer – die zweite Form der Datenerhebung, die Beobachtung, vorgestellt und durch Beispiele illustriert. Hier sei ausdrücklich darauf hingewiesen, dass die Datenerhebung durch Befragung oder Beobachtung natürlich nicht nur bei Querschnitts-Untersuchungen, sondern auch bei anderen Untersuchungsdesigns (z. B. bei Panels oder bei Experimenten) zum Einsatz kommt.

4.1 Einführung und Überblick

Im vorliegenden Kapitel geht es um eine Art von Untersuchungen, die durch die Zuordnung zu den Querschnitts-Untersuchungen (→ Untersuchungsdesign) und durch die Datenerhebung mittels Befragung oder Beobachtung gekennzeichnet ist.

H. Kreis et al., *Marktforschung,* https://doi.org/10.1007/978-3-658-44456-3_4

▶ **Definition** Unter einer **Befragung** wird „die zielgerichtete, systematische und regelgeleitete Generierung und Erfassung von verbalen Äußerungen einer Befragungsperson (….) zu ausgewählten Aspekten ihres Wissens, Erlebens und Verhaltens (….)“ (Döring & Bortz, 2016, S. 256) verstanden. Nicola Döring und Jürgen Bortz (2016, S. 324) haben auch wissenschaftliche **Beobachtungen** durch eine Definition charakterisiert: „Unter einer wissenschaftlichen Beobachtung (‚scientific observation‘) versteht man die zielgerichtete, systematische und regelgeleitete Erfassung, Dokumentation und Interpretation von Merkmalen, Ereignissen oder Verhaltensweisen mithilfe menschlicher Sinnesorgane und/oder technischer Sensoren zum Zeitpunkt ihres Auftretens.“

In der Markt- und Sozialforschung spielen Befragungen eine dominierende Rolle und werden deshalb hier recht ausführlich behandelt. Auf Spezifika von Beobachtungen wird erst im Abschn. 4.4 eingegangen. Einige der anderen im 4. Kapitel diskutierten Aspekte (z. B. Stichprobenziehung, Validität und Reliabilität von Messungen, Datensammlung und -aufbereitung) können aber analog auf Beobachtungen übertragen werden. Auch die in später folgenden Kapiteln behandelten Prinzipien der Datenanalyse gelten natürlich für Befragungsdaten ebenso wie für Beobachtungsdaten.

Die **Grundidee einer Befragung,** die keineswegs so trivial ist wie sie vielleicht klingt, besteht also darin, dass die gesuchten Informationen von der Auskunftsperson als Reaktion auf entsprechende Fragen mündlich (oder schriftlich oder durch Computer-Eingabe) gegeben werden. Voraussetzungen dafür sind einerseits die *Fähigkeit* und andererseits der *Wille* der Auskunftsperson, die gewünschten Angaben zu machen. So ist beispielsweise nicht jeder *fähig,* sich an Einzelheiten einer früher getroffenen Kaufentscheidung zu erinnern und Angaben über in Betracht gezogene Marken oder akzeptable Preise zu machen. Bei Fragestellungen, die die eigene finanzielle Situation oder Aspekte der Intimsphäre berühren, *wollen* viele Menschen die entsprechenden Informationen nicht geben.

Beispiel

Hier einige Beispiele für Fragen, die viele Konsumenten nicht korrekt beantworten *können* oder *wollen:*

„Wie viele Tassen Kaffee haben Sie in der letzten Woche getrunken?“

„Wohin werden Sie im übernächsten Jahr zum Sommerurlaub fahren?“

„Wie viele Stunden pro Tag sehen Sie durchschnittlich fern?“

„Lesen Sie regelmäßig die BILD-Zeitung?“

„Wie viele Schlaftabletten haben Sie in den letzten drei Monaten eingenommen?“

„Sind Sie in der letzten Woche mindestens einmal nach starkem Alkoholkonsum Auto gefahren?“ ◀

Mit der Fähigkeit und Willigkeit von Auskunftspersonen, die bei einer Befragung zu erhebenden Informationen hinreichend präzise und unverzerrt zu äußern, ist wieder die bereits im Abschn. 2.3 behandelte zentrale Frage der Validität (Gültigkeit) von Untersuchungsergebnissen angesprochen. Inwieweit kann man von den verbalen Angaben einer Person auf ihre tatsächlichen Einstellungen, Absichten, Verhaltensweisen etc. schließen? Wie müssen Frageformulierungen, Fragebögen, Interviewtechniken etc. gestaltet sein, damit ein solcher Schluss begründet ist und nur zu vertretbaren Fehlern führt? Derartige Probleme werden im vorliegenden vierten Kapitel eine zentrale Rolle spielen.

Hintergrundinformation

Eine umfassende und informative Kennzeichnung von repräsentativen Befragungen („Surveys") stammt von Jack Jacoby (2013, S. 202):

> „Im Wesentlichen sind wissenschaftliche Umfragen eine sorgfältig geplante Suche nach Informationen, die durch eine festgelegte Forschungsfrage bestimmt und intersubjektiver Überprüfung und Bestätigung unterworfen ist. Die grundlegenden Elemente der Umfrageforschung umfassen die Bestimmung und Definition der Grundgesamtheit (Wer oder was soll untersucht werden?), die Festlegung, wie die zu befragenden Auskunftspersonen ausgewählt werden, wo und zu welchen Erfahrungsbereichen diese befragt werden, welche Fragen den Auskunftspersonen hinsichtlich dieser Erfahrungsbereiche gestellt werden, wie die Umfrage durchgeführt wird und wie die Daten nach der Erhebung analysiert und interpretiert werden und wie über die Ergebnisse berichtet wird."

Trotz mancher Schwierigkeiten haben Befragungen wesentliche Vorzüge, die wohl ausschlaggebend dafür sind, dass sie über lange Zeit und bis heute das weitaus am stärksten genutzte Erhebungsinstrument der Marktforschung waren und sind. Diese Vorzüge lassen sich vor allem durch die Stichworte **„Flexibilität/Breite des Anwendungsbereichs"** und **„begrenzter Aufwand"** kennzeichnen.

Zunächst zu dem erstgenannten Aspekt. Die große Spannweite der **Anwendungen von Befragungen** lässt sich anhand von *zwei Dimensionen* umreißen. Einerseits geht es um die *Art der zu untersuchenden Aspekte,* also in der Marktforschung in erster Linie um **Verhaltensweisen** (z. B. Markenwahl, Mediennutzung, Gebrauch von Produkten), um **gedankliche Phänomene** (z. B. Einstellungen, Absichten, Wissen, Motive, Präferenzen) und um Angaben zur **Person** und zum **sozialen Umfeld** (z. B. Alter, Einkommen, Familiengröße).

Die zweite Dimension bezieht sich auf die *Zeit*, hier also auf die Frage nach Verhaltensweisen, Gedanken und persönlichen Merkmalen in **Vergangenheit, Gegenwart und Zukunft.** Besonders gängig ist die Erhebung der gegenwärtigen Merkmale. In der Marktforschung kommt häufig die Erhebung von Verhaltensweisen in der Vergangenheit (z. B. „In welchem Jahr haben Sie Ihr Auto gekauft?") und Zukunft (z. B. „Werden Sie bei Ihrem nächsten Autokauf bei derselben Marke bleiben?") hinzu. Dagegen spielen Aussagen zu persönlichen Merkmalen und Gedanken in Vergangenheit und Zukunft eine geringere Rolle bzw. dürften nur begrenzt valide zu ermitteln sein. In beiden Di-

mensionen bieten Befragungsverfahren ein breites (aber nicht unbegrenztes Spektrum) von Möglichkeiten, wobei zentrale dabei auftretende Probleme nicht unbeachtet bleiben dürfen:

- Wie stark und detailliert ist die Erinnerung an früheres Verhalten und Denken (z. B. Einstellungen vor mehreren Jahren)? Welche Validität haben vor diesem Hintergrund verbale Angaben?
- Welche Aussagekraft haben verbale Angaben zu künftigem Verhalten (z. B. zu umweltorientiertem Kaufverhalten) im Hinblick auf (viel) später folgendes tatsächliches Verhalten?
- In welchem Maße sind Auskunftspersonen in der Lage, ihre eigenen Gedanken, Emotionen, Motive etc. korrekt wahrzunehmen (siehe z. B. Nisbett & Wilson, 1977) und zu artikulieren?
- In welchem Maße sind Auskunftspersonen willens, „wahre" Angaben zu ihren Gedanken und Verhaltensweisen zu machen, auch wenn diese für sie selbst unangenehm oder peinlich sind?

Bei Befragungen entsteht im Vergleich zu anderen Methoden typischerweise ein begrenzter (aber nicht unbedingt geringer) Kosten- und Zeitaufwand. Die bei einer Befragung entstehenden Kosten sind natürlich nicht zuletzt vom jeweiligen Qualitätsanspruch abhängig. So führen umfangreiche Vorstudien oder Vergrößerungen der Stichprobe, die die Aussagekraft einer Untersuchung steigern, direkt zu einer entsprechenden Kostensteigerung. Die verständige Anwendung angemessener Methoden und das zur Verfügung stehende Budget bestimmen gemeinsam die Qualität der Untersuchungsergebnisse.

Hintergrundinformation

Groves et al. (2009, S. 30) kennzeichnen das Bemühen der Methodenforschung um Qualitätsverbesserungen bei Umfragen unter Kostenrestriktionen:

> „Die Methoden-Forschung zu Umfragen bemüht sich um die Entwicklung von Vorgehensweisen für die Anlage von Umfragen, Datensammlung bei Umfragen, Durchführung und Analyse von Umfragen im Hinblick auf die Kosten und die Qualität von Umfrageergebnissen. D. h., dass dieses Forschungsgebiet ausgerichtet ist auf die Verbesserung der Qualität unter Kostenrestriktionen bzw. die Kostenreduktion bei einem vorgegebenen Qualitätsanspruch."

Im vorliegenden Kapitel soll es um die Art von Befragungen gehen, die deren gängigste Form ist, nämlich um **repräsentative Befragungen** (als Querschnitts-Untersuchungen). Dazu werden hier vor allem die Auswahl von Auskunftspersonen bzw. die Stichprobenziehung (Abschn. 4.2) die Entwicklung von Fragebögen (Abschn. 4.3.1 bis 4.3.3) sowie die unterschiedlichen Kommunikationsformen (mündlich, telefonisch, schriftlich, elektronisch; Abschn. 4.3.4) bei Befragungen erörtert. Viele der dabei behandelten Gesichtspunkte und Methoden spielen auch bei anderen Untersuchungsdesigns eine Rolle.

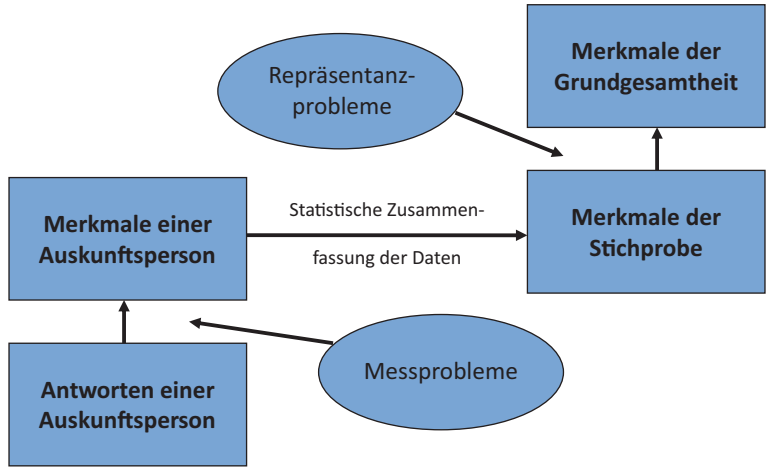


Abb. 4.1 Zwei Schlussweisen bei Umfragen. (Nach Groves et al., 2009, S. 39)

So greift man bei vielen experimentellen Untersuchungen auf Fragebögen als Messinstrument zurück und wendet die Prinzipien der Stichprobenziehung auch bei Beobachtungen an.

Mit den hier behandelten Problembereichen sind zentrale Aspekte der Schlussfolgerungen aus Umfrageergebnissen angesprochen:

- Lassen die Angaben einer Auskunftsperson den Schluss auf deren *tatsächliche Merkmale* (z. B. Einstellungen, Kaufabsichten) zu oder können Messprobleme, verursacht durch Frageformulierungen oder die Kommunikation bei der Befragung, die Aussagekraft der Angaben für das jeweilige Individuum beeinträchtigen?
- Inwieweit sind die befragten Personen *repräsentativ* für die Grundgesamtheit (z. B. deutsche Autofahrer oder Zeitungsleser in Berlin), über die auf Basis der Stichprobe Aussagen gemacht werden sollen? Repräsentativität einer Umfrage gilt als zentrale Voraussetzung für die Generalisierbarkeit (siehe Abschn. 2.3) ihrer Ergebnisse. Die im folgenden Abschnitt behandelte Stichprobenziehung ist wiederum die Grundlage für die Sicherstellung von Repräsentativität.

Abb. 4.1 illustriert diese beiden Aspekte bei Schlussfolgerungen.

4.2 Stichprobenziehung bei repräsentativen Befragungen

4.2.1 Grundlagen

Marktforschungsstudien sollen in der Regel Informationen über eine bestimmte Gesamtheit von Merkmalsträgern, die sogenannte **„Grundgesamtheit“** ermitteln, also eine ent-

sprechende *Generalisierbarkeit* (siehe Abschn. 2.3) der Ergebnisse ermöglichen. Es ist wichtig, diese Grundgesamtheit gleich zu Beginn einer Studie angemessen zu definieren. Die Angemessenheit der Definition ergibt sich zunächst aus dem Ziel der Marktforschungsuntersuchung. Soll ein neues Fahrzeug der Golfklasse auf seine Verbraucherakzeptanz hin überprüft werden, so sind die Neuwagenkäufer von Fahrzeugen dieser Klasse gefragt. Geht es dagegen um die Bekanntheit eines Bekleidungshauses für Damen und Herren, so sind alle Käufer von Bekleidung, also praktisch alle Erwachsenen, die im Einzugsbereich dieses Bekleidungshauses wohnen, relevant.

Wichtig ist auch, dass die *Definition der Grundgesamtheit* vollständig ist. Das bedeutet, dass die Grundgesamtheit sachlich, räumlich und zeitlich abgegrenzt sein muss. Dabei zielt die sachliche Abgrenzung auf das Vorliegen bestimmter für die Untersuchung wichtiger Eigenschaften ab. Die räumliche Abgrenzung zielt auf örtliche Eigenschaften der Merkmalsträger ab und die zeitliche Abgrenzung legt fest, zu welchem Zeitpunkt die sachliche und örtliche Abgrenzung gegeben sein müssen. Dies kann ein Stichtag oder aber der Zeitpunkt des Interviews sein und wird dann nicht explizit vorgenommen. Dabei bestimmt sich die Definition der Grundgesamtheit außer nach dem Untersuchungsziel auch nach erhebungspraktischen Erwägungen.

Beispiel

Beispielsweise ist die Grundgesamtheit der Media-Analyse Radio, deren Gegenstand die Nutzung der Radiosender ist und die per Telefon durchgeführt wird, definiert als die „deutschsprachige Bevölkerung in Privathaushalten am Ort der Hauptwohnung in der Bundesrepublik Deutschland im Alter von 10 und mehr Jahren“ (www.agma-mc.de/media-analyse/ma-radio/datenerhebung, aufgerufen am 8.8.2017 um 14:40 Uhr).

Gemäß der obigen Definition werden nur Personen in Privathaushalten erfasst. Personen, die in Anstaltshausalten wie z. B. Altersheimen oder Justizvollzugsanstalten leben, sind damit von der Untersuchung ausgeschlossen. Dies geschieht einmal aus erhebungspraktischen Gründen, da diese Personen teilweise nur schwer oder überhaupt nicht per Telefon erreichbar sind. Da die Media-Analyse Radio vor allem aber auch genutzt wird, um die Zahl und die soziodemografische Beschreibung der durch die Werbung in den einzelnen Radiosendern erreichten Zielgruppe zu ermitteln, geschieht diese Abgrenzung aber auch, weil diese Personenkreise für die Werbung nicht oder wenig relevant sind. Weiter muss sich die Hauptwohnung in der Bundesrepublik Deutschland befinden. Durch diese räumliche Abgrenzung sind Touristen aus anderen Ländern oder nicht anerkannte Asylbewerber ausgeschlossen. Schließlich wird gefordert, dass die Person deutschsprachig zu sein hat. Dies ist aus zwei Gründen sinnvoll. Einmal ist davon auszugehen, dass Personen, welche die deutsche Sprache nicht beherrschen, nicht zur Zielgruppe deutschsprachiger Radiosender gehören. Weiter ist dadurch sichergestellt, dass das Interview in deutscher Sprache durchgeführt werden kann, was die Organisation und Durchführung der Erhebung sehr vereinfacht. Das bedeutet aber auch, dass dauerhaft in Deutschland lebende Ausländer, welche der deutschen Sprache mächtig sind, zur Grundgesamtheit gehören. Schließlich erfolgt eine

Begrenzung auf Personen, die 10 Jahre oder älter sind, wodurch Interviews mit kleinen Kindern ausgeschlossen werden, was u. a. Verständnisprobleme zur Folge hätte. ◀

Weiter ist wichtig, dass die Definition der Grundgesamtheit anhand von Eigenschaften erfolgt, die schnell und einfach durch den Interviewer noch vor dem eigentlichen Interview festgestellt werden können. Deshalb ist beispielsweise eine Abgrenzung der Grundgesamtheit nach dem Einkommen nicht sinnvoll, da ein erheblicher Teil der Bevölkerung nicht bereit ist, dazu Auskunft zu geben, schon gar nicht als Eingangsfrage. Sollen nur bestimmte Einkommensgruppen befragt werden, dann bietet sich stattdessen eine Abgrenzung nach dem Beruf des Hauptverdieners bzw. der Hauptverdienerin des Haushalts an. In der Praxis erfolgt die Abgrenzung bei Konsumentenstichproben meist durch soziodemografische Merkmale und ggf. noch durch die Nutzung bestimmter Güter oder Dienstleistungen, bei Unternehmensstichproben durch Branche und Beschäftigtenzahl.

Schließlich ist es wichtig, dass eine aktuelle Beschreibung der Grundgesamtheit nach Größe und Struktur vorliegt, weil sonst nicht von der Stichprobe auf die Grundgesamtheit geschlossen werden kann (s. u. zur Hochrechnung). Eine wichtige Quelle dafür sind sekundärstatistische Daten (siehe Abschn. 2.4.2.1). So werden für die Beschreibung der Bevölkerung in Deutschland nach Alter, Geschlecht und Staatsangehörigkeit oft Daten des Statistischen Bundesamtes in Wiesbaden herangezogen (www.destatis.de). Eine weitere wichtige Quelle ist die Markt-Mediastudie „Best for Planning“ (www.b4p.de), die vor allem dann gefragt ist, wenn die Zielgruppe anhand des Verbrauchs oder Besitzes von Gütern definiert ist, wie z. B. Personen, die mindestens einmal pro Monat Tiefkühlpizzas kaufen als Grundgesamtheit für den Test einer neuen Tiefkühlpizza.

Hintergrundinformation

Unter den „Best Practices“ der American Association for Public Opinion Research (www.aapor.org) findet sich eine Kennzeichnung der Relevanz der Definition und Abdeckung von Grundgesamtheiten:

> „Zentrale Elemente einer vorbildlichen Umfrage sind: a) sicherzustellen, dass (zur Bearbeitung der interessierenden Fragestellung) in der Tat die richtige Population für die Stichprobenziehung ausgewählt wird, und b) alle Elemente dieser Population zu lokalisieren, damit sie eine Chance haben, in die Stichprobe aufgenommen zu werden. Die Qualität der Auflistung der Elemente … das heißt die Aktualität und Vollständigkeit der Liste, ist wahrscheinlich die wichtigste Voraussetzung, um eine angemessene Abbildung der zu untersuchenden Population zu erreichen.“

In der Regel wird die Erhebung für Marktforschungsuntersuchungen nur an einem Teil der Elemente der Grundgesamtheit vorgenommen. Dieser Teil heißt **Stichprobe.** Dabei ist die Stichprobe im Vergleich zur Grundgesamtheit oft sehr klein. Stichproben von 1000 oder 2000 Personen zur Abbildung der deutschsprachigen Bevölkerung in Deutschland sind durchaus üblich. Die oben erwähnte Media-Analyse Radio ist mit einer Stichprobe von 70.000 Personen eine der größten repräsentativen Untersuchungen in

Deutschland. Auch die bereits erwähnte „Best for Planning"-Studie hat mit über 30.000 Befragten eine der größten Stichproben der Marktforschung in Deutschland.

Die Erhebung nur in einer Stichprobe anstatt in der Grundgesamtheit hat deutliche *Vorteile.* Sie ist vor allem wesentlich preiswerter und geht wesentlich schneller als eine Vollerhebung der Grundgesamtheit.

Nachteilig ist jedoch zunächst, dass mit der Befragung nur einer Stichprobe eine zusätzliche Unsicherheit einhergeht, die sich im **Stichprobenfehler** ausdrückt (siehe auch Kap. 8). Der Stichprobenfehler entsteht dadurch, dass es nicht ausgeschlossen werden kann, dass zufällig eine „schiefe" Stichprobe gezogen wird. Dies kann man sich sehr einfach anhand der Wahlforschung verdeutlichen. So ist es bei einer 1000er-Stichprobe grundsätzlich möglich, dass nur Wähler der SPD befragt werden. Dies ist zwar extrem unwahrscheinlich, kann aber aufgrund der Tatsache, dass es eine zweistellige Millionenzahl von Wählern dieser Partei gibt, nicht grundsätzlich ausgeschlossen werden.

Ein zweiter Nachteil von Stichproben gegenüber Totalerhebungen besteht in den *begrenzten Möglichkeiten der Aufgliederung von Ergebnissen* in Bezug auf spezielle Teilgruppen. So wird man beispielsweise in einer 1000er Stichprobe aus der Gesamtbevölkerung Deutschlands nur sehr wenige (oder keine) Angehörige der Teilgruppen „katholische Landfrauen im Saarland" oder „Leser der Zeitschrift ‚Opernwelt' in Schleswig-Holstein" finden, so dass man über diese für ein bestimmtes Untersuchungsthema vielleicht interessierenden Gruppen kaum noch Aussagen machen kann. In der Tat ist die Stichprobe der Media-Analyse Radio (s. o.) deswegen so groß, weil auch über kleine Verbreitungsgebiete von Sendern (wie z. B. Radio Bremen) belastbare Aussagen getroffen werden sollen.

Die Ergebnisse der Stichprobe müssen dann auf die Grundgesamtheit übertragen werden. Eine solche **Hochrechnung** ist also der Schluss von der Stichprobe auf die Grundgesamtheit (siehe Kap. 8). Ob dieser Schluss so möglich ist, dass die daraus resultierenden Schätzwerte für die Grundgesamtheit unverzerrt sind, hängt außer von der Erhebung vor allem davon ab, auf welche Art und Weise die Stichprobe gezogen wird. Dabei ist ein Schätzwert dann *unverzerrt*, wenn der Erwartungswert der Stichprobe gleich dem unbekannten Wert der Grundgesamtheit ist.

Bei der Hochrechnung werden die Ergebnisse der Befragung in der Stichprobe jeweils mit einem *Hochrechnungsfaktor* multipliziert und fließen so in die Berichterstattung ein. Dieser Hochrechnungsfaktor für das j-te Stichprobenelement ist 1/p(j), wobei p(j) die Wahrscheinlichkeit ist, mit der dieses Element in die Stichprobe kommt. Im einfachsten Fall ist die Stichprobe („einfache Zufallsstichprobe") proportional, d. h. jedes Element der Grundgesamtheit hat die gleiche Wahrscheinlichkeit in die Stichprobe zu gelangen. Dann ist $p(j) = n/N$ für alle j, wobei n den Umfang der Stichprobe und N den Umfang der Grundgesamtheit bedeutet. Dadurch wird aber auch deutlich, dass eine sachgerechte Hochrechnung voraussetzt, dass die Größe der Grundgesamtheit bekannt ist.

▶ **Definition** Eine Stichprobe heißt dann **repräsentativ**, wenn sich aus der Stichprobe unverzerrte Schätzwerte für die Grundgesamtheit bestimmen lassen. Dies ist dann der Fall, wenn die Hochrechnungsfaktoren bestimmt werden können, wenn also vor der Stichprobenziehung für jedes Element der Grundgesamtheit bekannt ist, mit welcher Wahrscheinlichkeit es in die Stichprobe kommt.

Ein einfaches Beispiel soll das verdeutlichen

Ein Tierpark möchte eine Untersuchung zur Bekanntheit und Attraktivität des Parks vornehmen. Dazu unterscheidet er einen engeren und einen weiteren Einzugsbereich. Im engeren Einzugsbereich wohnt eine Million Menschen der definierten Zielgruppe, im weiteren Einzugsbereich wohnen 4 Mio. Aus jedem der beiden Einzugsbereiche wird eine Stichprobe von 500 Menschen gezogen. Die Stichprobe wurde bewusst disproportional gewählt, weil so genauere Aussagen über den engeren als über den weniger interessanten weiteren Einzugsbereich möglich sind.

Dann ergeben sich folgende Hochrechnungsfaktoren:

- Für den engeren Einzugsbereich $1/(500/1.000.000) = 1.000.000/500 = 2000$.
- Für den weiteren Einzugsbereich $1/(500/4.000.000) = 4.000.000/500 = 8000$.

Nehmen wir weiter an, dass 300 der 500 Befragten des engeren und 100 der 500 Befragten des weiteren Einzugsbereichs sagen, dass sie schon einmal den Tierpark besucht haben. Dann ergibt sich eine Schätzung für beide Einzugsbereiche zusammen wie folgt:

$300 \times 2000 + 100 \times 8000 = 1.400.000$.

Im Ergebnis wird also geschätzt, dass 1,4 Mio. den Tierpark schon besucht haben. ◀

Das Beispiel zeigt, dass es nicht notwendig ist, dass die Stichprobe proportional ist, damit sie repräsentativ ist. Vielmehr können auch disproportionale Stichproben repräsentativ sein, wenn das Ausmaß der Disproportionalität bekannt ist. Dagegen wird eine Stichprobe, die dadurch entsteht, dass auf einer Website zur Teilnahme an einer Umfrage aufgerufen wird, nicht repräsentativ für die Gesamtbevölkerung sein können, schon weil nur Besucher dieser Website erreicht werden können.

Zusammenfassend lassen sich damit die Schritte bei der Durchführung einer Marktforschungsstudie wie folgt beschreiben:

- Es wird eine Grundgesamtheit definiert.
- Aus der Grundgesamtheit wird eine Stichprobe gezogen.
- Die erforderlichen Daten werden in der Stichprobe erhoben.
- Die erhobenen Daten werden auf die Grundgesamtheit hochgerechnet.

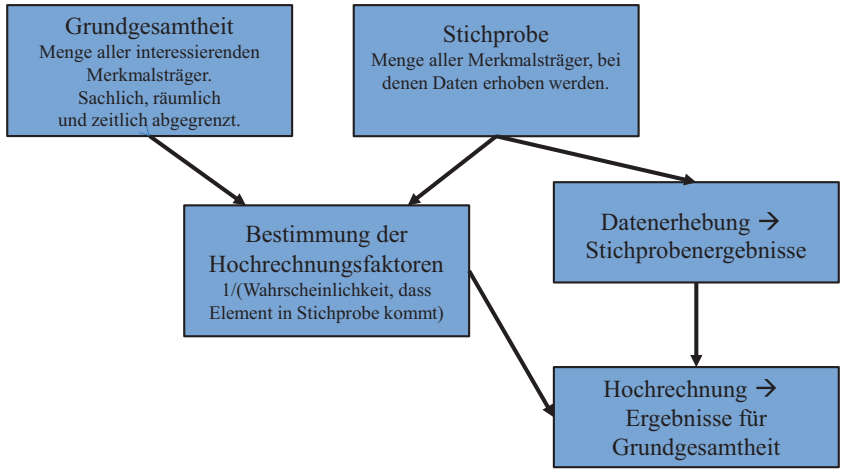


Abb. 4.2 Grundgesamtheit, Stichprobe, Erhebung und Hochrechnung

Die Abb. 4.2 verdeutlicht diese Zusammenhänge.

4.2.2 Arten von Stichproben

Ist die Grundgesamtheit definiert, so muss festgelegt werden, wie die Einheiten bestimmt werden, bei denen die Erhebung durchgeführt werden soll. Hierfür stehen mehrere grundsätzliche Vorgehensweisen zur Verfügung, welche die Abb. 4.3 zeigt und die im Folgenden erläutert werden.

Eine grundlegende Unterscheidung von Stichproben ist die in Zufallsstichproben und andere, nicht zufällige Stichproben. In der Marktforschungspraxis sind dies vor allem Quotenstichproben.

Die entscheidende Anforderung an **Zufallsstichproben** besteht darin, dass jedes Element der Grundgesamtheit eine berechenbare Wahrscheinlichkeit größer Null hat, in die Stichprobe zu kommen. Auf dieser bei der Stichprobenziehung zu realisierenden Annahme beruhen statistische Techniken zur Schätzung von Stichprobenfehlern (siehe hierzu auch Abschn. 8.2). Die Berechenbarkeit der Auswahlchance der Stichprobenelemente wird dadurch gewährleistet, dass die Auswahl zufällig erfolgt und damit den Kalkülen der Wahrscheinlichkeitsrechnung zugänglich ist.

Für eine Zufallsauswahl kann z. B. so verfahren werden, dass jedem Element der Grundgesamtheit eine Zufallszahl zugeordnet wird, d. h. jedem Datensatz in einer entsprechenden Datei wird eine mit unterschiedlichster Software (z. B. in Excel mit der Funktion = zufallszahl()) leicht zu erzeugende Zufallszahl angehängt. Die Datensätze werden dann nach dieser Zufallszahl (auf- oder absteigend) sortiert und die ersten n **(Stichprobengröße)** Elemente gelangen in die Stichprobe. Der Interviewer bzw. die

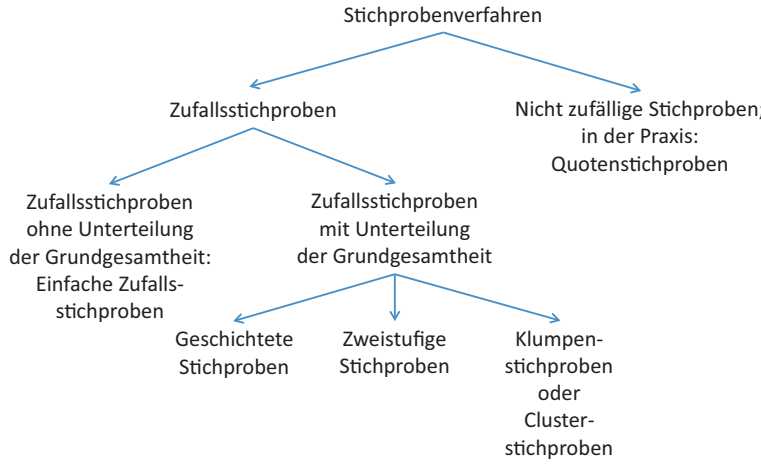


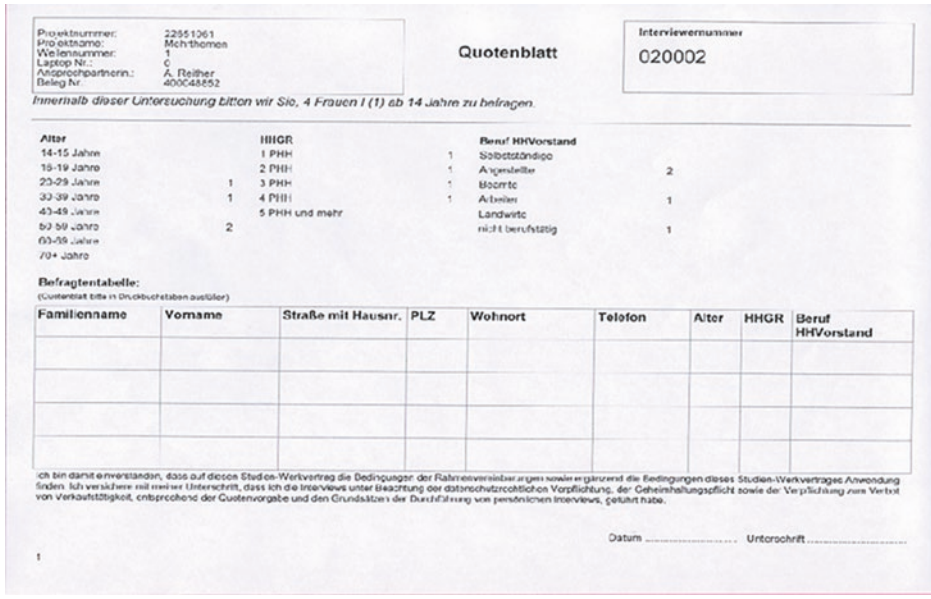
Abb. 4.3 Arten von Stichproben in der Marktforschung

Interviewerin hat also keinerlei Einfluss auf die Auswahl der zu befragenden Person. Nur so kann vor der Auswahl jedem Element der Grundgesamtheit eine feste Wahrscheinlichkeit zugeordnet werden, dass dieses Element in die Stichprobe kommt.

Damit dies funktioniert, müssen mehrere *Voraussetzungen* gegeben sein. Zunächst ist es erforderlich, eine irgendwie geartete Liste der Auswahleinheiten der Grundgesamtheit zu erstellen. Eine solche Liste haben z. B. Banken oder Mobilfunkunternehmen, wenn sie ihre Kunden befragen wollen. In manchen Ländern (z. B. in Schweden) ist auch eine solche Liste der Bevölkerung des Landes erhältlich. In anderen Ländern (z. B. auch in Deutschland) gibt es solche Listen nicht, weshalb mühsam Ausweichverfahren entwickelt werden müssen (vgl. Abschn. 4.2.3).

Eine weitere Voraussetzung ist, dass der Anteil der Antwortverweigerer gering ist. Dies wird deutlich, wenn man sich verdeutlicht, dass eine Antwortverweigerung dazu führt, dass die vorher zugeordnete Auswahlwahrscheinlichkeit einer Zielperson, welche die Antwort verweigert, fehlerhaft ist. Bei einer kleinen Verweigerungsquote kann dieser Fehler hingenommen werden. Allgemein wird eine Verweigerungsquote von bis zu 30 % akzeptiert. Sehr häufig sind die Verweigerungsquoten aber sehr viel größer. So betrug beim gut dokumentierten „Allbus 2014“ (www.gesis.org/allbus/inhalte-suche/methodenberichte) die Ausschöpfungsquote nur 35,5 % und sank bis 2021 weiter auf 29,5 % ab (Baumann et. al. 2024). Mehr als 2 von 3 Befragten waren also entweder nicht erreichbar oder haben die Antwort verweigert. Bei Stichproben, bei denen eine kontinuierliche Mitarbeit gefordert ist, wie z. B. beim Haushaltspanel (siehe Abschn. 5.3), ist die Ausschöpfungsquote mit ca. 5 % noch wesentlich geringer. In der Marktforschung sind daher Zufallsstichproben oft nicht möglich. In diesem Fall können Quotenstichproben gezogen werden.

Bei **Quotenstichproben** werden dem Interviewer Eigenschaften vorgegeben, welche die interviewten Personen erfüllen müssen. Ansonsten ist der Interviewer bzw. die Inter-



Projektnummer: 22551361
Projektname: Mchrthemen
Wellennummer: 1
Laptop Nr.: c
Ansprechpartnerin.: A. Reither
Beleg Nr. 400048852

Quotenblatt

Interviewernummer
020002

Innerhalb dieser Untersuchung bitten wir Sie, 4 Frauen / (1) ab 14 Jahre zu befragen.

Alter		HHGR		Beruf HHVorstand	
14-15 Jahre		1 PHH	1	Selbstständige	
16-19 Jahre		2 PHH	1	Angestellte	2
20-29 Jahre	1	3 PHH	1	Beamte	
30-39 Jahre	1	4 PHH	1	Arbeiter	1
40-49 Jahre		5 PHH und mehr		Landwirte	
50-59 Jahre	2			nicht berufstätig	1
60-69 Jahre					
70+ Jahre					

Befragtentabelle:
(Quotenlauf bitte in Druckbuchstaben ausfüllen)

Familienname	Vorname	Straße mit Hausnr.	PLZ	Wohnort	Telefon	Alter	HHGR	Beruf HHVorstand

Ich bin damit einverstanden, dass auf diesen Studien-Werkvertrag die Bedingungen der Rahmenvereinbarungen sowie ergänzend die Bedingungen dieses Studien-Werkvertrages Anwendung finden. Ich versichere mit meiner Unterschrift, dass ich die Interviews unter Beachtung der datenschutzrechtlichen Vorpflichtung, der Geheimhaltungspflicht sowie der Verpflichtung zum Verbot von Verkaufstätigkeit, entsprechend der Quotenvorgabe und den Grundsätzen der Durchführung von persönlichen Interviews, geführt habe.

Datum ..... Unterschrift .....

1

Abb. 4.4 Beispiel für Quotenplan. (Quelle: GfK)

viewerin in der Auswahl der befragten Person frei. Abb. 4.4 zeigt ein Beispiel für solch einen Quotenplan für einen Interviewer, der auf die Merkmale Geschlecht, Alter, Haushaltsgröße und Berufsgruppe des Haushaltsvorstands abgestellt ist.

Die vorgegebenen Eigenschaften heißen **Quotenmerkmale.** Es ist wichtig, dass die Quotenmerkmale leicht zu erheben sind, weil sie noch vor dem eigentlichen Interview festgestellt werden müssen. Weiter sind aktuelle Daten bezüglich ihrer Verteilung in der Grundgesamtheit erforderlich, damit ihre Verteilung richtig vorgegeben werden kann. Schließlich sollten die Quotenmerkmale mit dem Erhebungsgegenstand korreliert sein.

Beispiel

Für die Auswahl von Haushalten für das Haushaltspanel, in dem die Einkäufe verpackter Verbrauchsgüter erhoben werden, werden z. B. folgende Quotierungsmerkmale herangezogen:

- Haushaltsgröße
- Zahl der Kinder unter 15 Jahren
- Alter der haushaltsführenden Person (das ist die Person, die solche Güter überwiegend einkauft)
- Gemeindegröße
- Bundesland bzw. Teil des Bundeslandes, z. B. Regierungsbezirk ◄

Ein wichtiger *Nachteil* der Quotenstichprobe gegenüber der Zufallsstichprobe besteht darin, dass der Zufallsfehler nicht berechnet werden kann. Experimente – insbesondere des Instituts für Demoskopie in Allensbach (vgl. Noelle-Neumann & Petersen, 2000, S. 263 ff.) – zeigen jedoch, dass eine qualitativ hochwertige Quotenstichprobe mindestens so gut ist, wie eine einfache Zufallsauswahl. Es ist deshalb in der Praxis durchaus üblich, das Vertrauensintervall einer einfachen Zufallsauswahl auch bei der Quotenauswahl als Anhaltspunkt für einen möglichen Zufallsfehler anzugeben, freilich verbunden mit dem Hinweis, dass dies sich nur auf Erfahrung, nicht auf exakte Berechnungen stützt. Dagegen ist es ein eindeutiger *Vorteil* der Zufallsauswahl, dass hier zumindest in der Theorie für ein vorgegebenes Signifikanzniveau (in der Regel 90 oder 95 %) **Vertrauensintervalle** für die errechneten Werte angegeben werden können, also ein Intervall, in dem mit einer Wahrscheinlichkeit von 90 % bzw. 95 % der wahre Wert der Grundgesamtheit liegt.

Die einfachste Form der **Zufallsauswahl** ist (nicht ganz überraschend) die *einfache Zufallsauswahl,* die dadurch gekennzeichnet ist, dass die Grundgesamtheit nicht weiter unterteilt ist und jedes Element der Grundgesamtheit die gleiche Wahrscheinlichkeit hat, in die Stichprobe zu gelangen. Wird – wie oben beschrieben – an jedem der N Elemente der Grundgesamt eine Zufallszahl hinzugefügt, wird anschließend nach diesen Zahlen auf- oder absteigend sortiert und werden die ersten n Einheiten in die Stichprobe genommen, so wird eine einfache Zufallsauswahl realisiert.

Die **Standardabweichung des Mittelwerts einer einfachen Zufallsstichprobe** ergibt sich wie folgt:

$$\sigma(\overline{x}) = \sqrt{\frac{N-n}{N}} \cdot \frac{\sigma(x)}{\sqrt{n}}$$

Dabei ist:

$\sigma(\overline{x})$:	Standardabweichung des Mittelwerts des Merkmals x
$\sigma(x)$:	Standardabweichung des Merkmals x
N:	Umfang der Grundgesamtheit
n:	Umfang der Stichprobe

Das bedeutet:

1. Solange der Umfang der Stichprobe sehr viel kleiner ist als der Umfang der Grundgesamtheit, spielt der Umfang der Grundgesamtheit keine Rolle. Dieser geht ja nur in dem ersten Teil $((N-n)/N)^{0,5}$ ein. Ist z. B. die Grundgesamtheit 20 Mio. und die Stichprobe gleich 2000, so ergibt sich für diesen Teil der Formel der Wert 0,9999 und damit praktisch 1.

2. Wenn n sehr viel kleiner als N, dann nimmt die Standardabweichung des Mittelwerts mit der Wurzel des Stichprobenumfangs ab. Soll die Standardabweichung des Mittelwerts *halbiert* werden, so muss die Stichprobe demnach *vervierfacht* werden.
3. Die Standardabweichung des Mittelwerts verhält sich proportional zur Standardabweichung des Merkmals. Werte von Merkmalen mit geringerer Streuung werden daher genauer geschätzt als Werte mit großer Streuung.

Ist die Standardabweichung berechnet, so ergibt sich für $n > 30$ daraus einfach ein 95 %-Vertrauensintervall für den Mittelwert wie folgt (siehe dazu auch Abschn. 8.1):

$$\bar{x} \pm 1{,}96 \cdot \sigma(\bar{x})$$

Beispiel

Bei einer Befragung von $n = 100$ Leitern eines bestimmten Geschäftstyps (Umfang der Grundgesamtheit $N = 10.000$) zum Abverkauf einer Kamera eines bestimmten Typs im letzten Monat ergab sich ein Mittelwert von 15 mit einer Standardabweichung von 2.

Dann errechnet sich die Standardabweichung des Stichprobenmittelwerts wie folgt:

$$\sigma(\bar{x}) = \sqrt{\frac{10.000 - 100}{10.000}} \cdot \frac{2}{\sqrt{100}} = 0.198$$

Das 95 %-Vertrauensintervall ist wie folgt: $15 \pm 1{,}96$ * 0,198, es geht also von 14,61 bis 15,39.

Etwas vereinfacht lässt es sich wie folgt interpretieren: Die Wahrscheinlichkeit, dass der wahre Mittelwert der Grundgesamtheit zwischen 14,61 und 15,39 liegt, beträgt 95 %. ◀

Für die **komplexeren Methoden der Zufallsauswahl** ist es Voraussetzung, dass die Grundgesamtheit in mehrere Teile unterteilt ist. Diese Teile heißen bei der geschichteten Stichprobe „Schichten", bei der zweistufigen Auswahl „Auswahleinheiten 1. Stufe" (wobei die einzelnen Elemente die Auswahleinheiten 2. Stufe sind) und bei den Klumpen- bzw. Clusterstichproben eben Klumpen bzw. Cluster. Wichtig ist in jedem Fall, dass diese Teile vollständig und nicht überlappend sind, dass also jedes Element der Grundgesamtheit genau einem dieser Teile angehört.

So kann für eine Befragung der deutschen Wohnbevölkerung die Aufteilung dieser Grundgesamtheit in Bundesländer oder Kreise bzw. kreisfreie Städte relevant sein. Für eine Befragung von Bahnreisenden kann dagegen die Aufteilung der Bahnreisenden auf Züge zum Tragen kommen.

Abb. 4.5 zeigt links eine Grundgesamtheit, deren Elemente durch kleine Kreise gekennzeichnet sind. Bei einer einfachen Zufallsauswahl erfolgt keine Unterteilung, alle Elemente sind nur im großen Kasten. Bei den komplexeren Verfahren nutzt man die Tatsache, dass es kleinere Teilgesamtheiten (z. B. Bundesländer) gibt und jeder Punkt zu

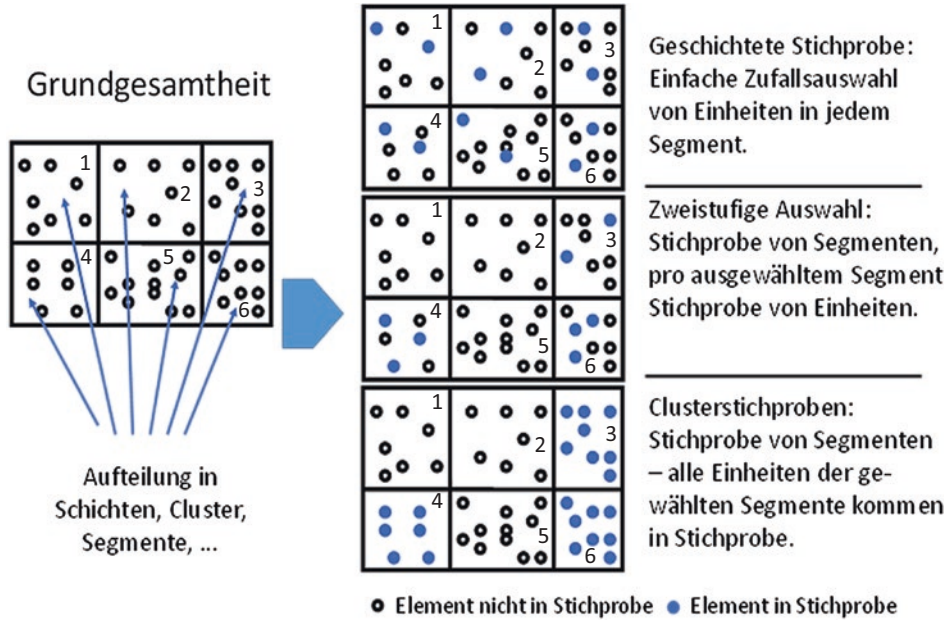


Abb. 4.5 Verschiedene Verfahren komplexer Zufallsstichproben

genau zu einer dieser Teilgesamtheiten gehört. Im Beispiel der Abbildung sind die 47 Einheiten der Grundgesamtheit in 6 Teilgesamtheiten (kleine Felder) aufgeteilt, die mit 1 bis 6 durchnummeriert sind. Dabei müssen die Teilgesamtheiten nicht gleich groß sein (die Bundesländer sind es ja auch nicht). So hat die Teilgesamtheit 4 (links unten) 6 Einheiten, die Teilgesamtheit 5 dagegen 11 Einheiten.

Auf der rechten Seite der Abb. 4.5 ist dargestellt, wie diese Teilgesamtheiten bei den komplexen Auswahlverfahren genutzt werden können. Dabei sind die zufällig ausgewählten Einheiten durch blaue ausgefüllte Kreise dargestellt, nicht ausgewählte Einheiten durch kleine schwarze Kreise.

Bei der **geschichteten Zufallsauswahl** werden alle Teilgesamtheiten (die hier Schichten genannt werden) ausgewählt und aus jeder der Schichten wird eine einfache Zufallsstichprobe mit einem pro Schicht definierten Umfang gezogen. Im Beispiel der Befragung von Bahnreisenden würde das bedeuten, dass aus jedem Zug eine Zufallsauswahl von Bahnreisenden befragt wird. In der Abb. 4.5 ist dies rechts oben dadurch gekennzeichnet, dass aus jeder der sechs Teilgesamtheiten (Schichten) 2 Elemente zufällig gezogen wurden.

Die Aufteilung der Stichprobe auf die Schichten kann proportional (die Anteile der Schichten in der Grundgesamtheit und in der Stichprobe sind gleich) oder disproportional (die Anteile sind verschieden) erfolgen. Disproportionale Aufteilungen sind insbesondere dann sinnvoll, wenn eine kleine Schicht besondere Bedeutung hat und separat ausgewiesen werden soll oder wenn die Streuungen in den Schichten sehr

unterschiedlich sind. Aus den Schichten mit großer Streuung sind dann mehr Elemente auszuwählen als aus den Schichten mit geringer Streuung.

Die geschichtete Zufallsauswahl liefert gegenüber der einfachen Zufallsauswahl insbesondere dann genauere Schätzwerte, wenn die Streuung in den Schichten kleiner ist als die Streuung in der Grundgesamtheit. Ein weiterer wichtiger Vorteil der geschichteten Zufallsstichprobe ist, dass sie sicherstellt, dass auch Werte pro Schicht ausgewiesen werden können, weil pro Schicht ein bestimmter Stichprobenumfang garantiert ist. Bei der einfachen Zufallsauswahl gibt es dagegen eine solche Garantie nicht.

Bei der **Klumpenauswahl** bzw. **Clusterauswahl** wird nur ein Teil der Teilgesamtheiten (- hier als Klumpen oder Cluster bezeichnet -) zufällig ausgewählt, diese werden aber vollständig erhoben. Im Beispiel der Befragung der Bahnreisenden würden zufällig Züge ausgewählt und bei jedem ausgewählten Zug würden alle Reisenden befragt. In der Abb. 4.5 rechts unten wurden die Teilgesamtheiten oder Klumpen 3, 4 und 6 zufällig ausgewählt. Aus den ausgewählten Klumpen werden jeweils *alle* Einheiten erhoben (alle Einheiten der ausgewählten Klumpen sind durch gefüllte blaue Kreise gekennzeichnet).

Klumpenauswahl wird vor allem aus erhebungstechnischen Gründen gewählt. So wird in der Fernsehforschung das Fernsehverhalten von allen Mitgliedern eines Haushalts ab einem bestimmten Alter erhoben. Der Grund ist darin zu suchen, dass die Installation der Messtechnik sehr teuer ist und die Erhebung weiterer Haushaltsmitglieder kaum zusätzliche Kosten verursacht.

Schließlich steht die **zweistufige Zufallsauswahl** in gewisser Weise zwischen der geschichteten Zufallsauswahl und der Klumpenauswahl, weil sie die beiden Methoden kombiniert. Zunächst wird eine Zufallsstichprobe der Teilgesamtheiten (die hier „Einheiten erster Stufe“ heißen) gezogen. Aus jeder der gezogenen Einheiten erster Stufe wird dann jeweils eine einfache Zufallsstichprobe von Einheiten (die hier dann „Einheiten zweiter Stufe“ heißen) erhoben. Bei dem Beispiel der Befragung der Bahnreisenden würden zunächst zufällig Züge ausgewählt und dann in jedem Zug wieder eine Zufallsauswahl von Bahnreisenden ausgewählt, die dann befragt würden. In Abb. 4.5 rechts in der Mitte wurden erst die Teilgesamtheiten oder Einheiten erster Stufe 3, 4 und 6 zufällig gezogen. Aus diesen wurde jeweils eine Stichprobe von 3 Einheiten ebenfalls zufällig gezogen, die durch die blauen gefüllten Kreise gekennzeichnet sind.

Auch diese Methode wird vor allem aus erhebungstechnischen Gründen verwendet, z. B. bei dem im nächsten Abschnitt beschriebenen Random-Route-Verfahren.

Auch für die komplexeren Auswahlverfahren geschichtete Stichprobe, zweistufige Stichprobe und Clusterstichprobe existieren Formeln für den Zufallsfahler, also die Standardabweichung des Mittelwerts. Diesbezüglich muss hier jedoch auf die entsprechende Spezialliteratur verwiesen werden (z. B. Cochran, 1977; Chaudhuri & Stenger, 2005; Thompson, 2012).

4.2.3 Vorgehensweise bei der Stichprobenziehung

Maßgeblich für die Festlegung der **Stichprobengröße** ist die gewünschte **Genauigkeit** bzw. die gewünschte **Sicherheit** der Ergebnisse. Dabei gilt, dass bei einer einmal festgelegten Stichprobenmethode und bei vorgegebener Sicherheit der Stichprobenfehler sich nur reduzieren lässt, indem der Stichprobenumfang erhöht wird.

Das Problem ist, dass oft erst im Nachhinein festgestellt werden kann, wie groß der Stichprobenfehler ist, weil zu seiner Bestimmung die Varianz des analysierten Merkmals benötigt wird und diese Information erst dann geschätzt werden kann, wenn die Erhebung erfolgt ist. Bei einem wichtigen Spezialfall, nämlich bei der Schätzung von Anteilen, lässt sich jedoch bereits vor der Erhebung der **notwendige Stichprobenumfang** n bei einfacher Zufallsauswahl bestimmen, damit der Fehler („Error") maximal einem vorgegebenen Wert E ist. Dazu muss bei einer Sicherheitswahrscheinlichkeit von 95 % und für den Fall, dass die Grundgesamtheit wesentlich größer als die Stichprobe ist, für den Stichprobenumfang n näherungsweise folgende Ungleichung gelten: $n >= 1/E^2$.

Beispiel

Bei einem Produkttest soll der Anteil der Befragten mit Kaufabsicht auf mindestens ±4 % genau geschätzt werden → $E = 0{,}04$ (E muss als Dezimalbruch ausgedrückt werden).

Dann gilt näherungsweise:

$$n \geq \frac{1}{0{,}04^2} = 625$$

Der Stichprobenumfang sollte also mindestens 625 Interviews umfassen. ◀

Bei der Bestimmung des Stichprobenumfangs spielen in der Praxis neben der benötigten Genauigkeit des Totalwerts auch folgende Aspekte eine Rolle:

- **Finanzielle Restriktionen**
  Wegen der mit der Datenerhebung verbundenen Kosten muss oft eine Einschränkung der (eigentlich erwünschten) Stichprobengröße in Kauf genommen werden.
- **Gewünschte Aufschlüsselung der Ergebnisse**
  Sollen bei der Datenanalyse Aussagen über sehr spezielle Teilgruppen gemacht werden, so muss die Stichprobe natürlich groß genug sein, damit dafür noch eine hinreichend breite Basis vorhanden ist. Soll ein kleines Segment extra gezeigt werden, so hilft man sich manchmal mit einem „Boost". Das bedeutet, dass in diesem Segment die Stichprobe größer ist, als es ihrem Anteil in der Grundgesamtheit entspricht. So werden z. B. bei der bereits erwähnten Media-Analyse Radio die Stichproben in den kleinen Sendegebieten erhöht.

Beispiel

Vor dem Hintergrund komplexerer Verfahren der Stichprobenziehung formulieren Günther et al. (2006, S. 20) eine entsprechend angepasste Kennzeichnung der Repräsentativität:

„Eine Stichprobe wird … dann als repräsentativ bezeichnet, wenn sie den Schluss auf die Grundgesamtheit zulässt. Dies ist der Fall, wenn eine Rechenvorschrift existiert, so dass die Mittelwerte der errechneten Werte aller möglichen Stichproben gleich den entsprechenden Mittelwerten der Grundgesamtheit sind (Erwartungstreue oder auch Validität der Schätzung)."

In der Praxis haben sich einige Standardverfahren der Stichprobenziehung in Abhängigkeit von der Auswahlmethode und vom Untersuchungsziel herausgebildet. Eine erste, wichtige Frage ist, ob Quotenauswahl oder Zufallsauswahl angewendet werden soll. Zufallsauswahl wird immer dann angestrebt, wenn die Ergebnisse allgemein anerkannt werden müssen, wie dies z. B. bei der Werbeträgerforschung der Fall ist. Allerdings wird in der Regel Zufallsauswahl mit einer längeren Feldzeit und höheren Kosten erkauft, weil ein aktuell nicht erreichbarer Interviewpartner nicht durch einen anderen ersetzt werden darf, sondern es muss immer wieder versucht werden, diesen zu erreichen.

Ist die Entscheidung für die **Quotenauswahl** gefallen, so sind einige Regeln einzuhalten, um die Qualität der Stichprobe zu sichern:

- **Quoten sind nicht zu leicht und nicht zu schwer zu erfüllen:** Sind sie zu leicht zu erfüllen, so werden vor allem sehr leicht erreichbare Personen angesprochen und es ergeben sich u. U. erhebliche Schiefen. Wird z. B. nur vorgegeben, eine bestimmte Anzahl von Männern und Frauen zu interviewen, so hat der Interviewer die Möglichkeit, seine Befragten am Bahnhof zu rekrutieren und wird daher vor allem Nutzer Öffentlicher Verkehrsmittel und mehr Berufstätige erreichen. Sind die Quoten jedoch zu schwer zu erreichen, so ergeben sich eine zu lange Feldzeit und zu hohe Kosten. Außerdem steigt die Gefahr, dass Interviewer falsche Quoten angeben.
- **Kleine Zahl von Interviews pro Interviewer:** Der Grund für diese Anforderung ist darin zu suchen, dass die Interviewer ihre Interviewpartner in ihrem erweiterten sozialen Netz suchen. Die Interviews eines Interviewers sind also einem sozialen Netz entnommen. Damit sich keine Klumpeneffekte ergeben, ist es erforderlich, die Zahl der Interviews pro Interviewer auf fünf bis acht zu begrenzen. Das bedeutet auch, dass für jede Studie eine große Zahl von Interviewern einzusetzen ist.
- **Heterogener Bestand an Interviewern bezüglich Alter, sozialer Schicht und regionaler Verteilung:** Nur so kann gewährleistet werden, dass verschiedene soziale Umfelder auch tatsächlich erreicht werden.
- **Sicherstellen, dass keine Person zu oft interviewt wird:** Wenn die Interviewer ihre Interviewpartner in ihrem erweiterten sozialen Netz suchen, dann besteht die Gefahr, dass die gleichen Personen immer wieder befragt werden, was auf Dauer die

erhobenen Daten beeinflussen kann. So ist zu erwarten, dass nach einem Interview zu Werbung für Kraftfahrzeuge eine erhöhte Aufmerksamkeit für solche Werbung besteht. In der Praxis wird davon ausgegangen, dass ein weiteres Interview nach frühestens drei Monaten unschädlich ist. Dies kann z. B. durch wechselnde Quoten erreicht werden. So soll ein Interviewer zu einem Projekt junge Männer, zum nächsten Projekt ältere Männer, dann junge Frauen und schließlich ältere Frauen befragen.

Diese Regeln sind aus der Praxis entwickelt. Insgesamt lässt sich feststellen, dass die Quotenstichprobe dagegen in der Wissenschaft nur sehr wenig behandelt wird. So ist von den 399 Seiten des Stichprobenbuches von Cochran (1977) gerade eine halbe Seite den Quotenstichproben gewidmet.

Entscheidet man sich für eine **Zufallsstichprobe**, so ist zunächst zu untersuchen, ob es eine **Liste der Auswahleinheiten der Grundgesamtheit** gibt. Dies ist z. B. in vielen Branchen (u. a. Banken, Telekommunikationsdienstleistungen oder Energie) der Fall, wenn Kunden z. B. zu Ihrer Zufriedenheit befragt werden sollen. Dann lässt sich einfach mit einem Zufallszahlengenerator eine einfache oder auch eine geschichtete Stichprobe zu ziehen (Abschn. 4.2.2). Bei **allgemeinen Bevölkerungsumfragen** in Deutschland ist eine solche Liste in der Regel nicht vorhanden. Hier muss nach der Erhebungsmethode unterscheiden werden:

Bei **Telefonstichproben für Bevölkerungsumfragen** sind die Telefonbücher nicht als Auswahlgrundlage geeignet, da sich viele Menschen nicht mehr eintragen lassen (in Großstädten z. T. über 40 %) und so nicht erreichbar wären. Deshalb wird ein Verfahren (sogen. „Gabler-Häder-Verfahren“) angewendet, mit dem auch nicht eingetragene Telefonnummern erreicht werden können. Dazu werden von der Regulierungsbehörde für die Telekommunikation sämtliche an die Telekommunikationsdienstleister frei gegebenen Nummernkreise zur Verfügung gestellt. Aus diesen werden die z. B. aufgrund der Gelben Seiten bekannten Firmennummern herausgestrichen. Bei den verbliebenen Nummern werden die letzten beiden Ziffern gestrichen. Von diesen „Stämme“ genannten Restnummern werden alle doppelten gestrichen und anschließend alle Nummern gebildet, die sich ergeben, wenn alle zweistelligen Kombinationen von „00“ bis „99“ an die Stämme angehängt werden. Diese Nummern bilden die Auswahlgrundlage, aus der die Stichprobe ausgewählt wird. Die ausgewählten Nummern werden angerufen. Ist eine Nummer noch nicht vergeben oder ist ein Computer- oder Faxmodem angeschlossen oder stellt sich heraus, dass eine Nummer nicht zu einem Privathaushalt gehört, so wird die Nummer gestrichen. Gestrichene Nummern werden durch neue Nummern aus der Auswahlgrundlage ersetzt. Bei den zu privaten Telefonanschlüssen gehörenden Nummern wird versucht, zunächst eine Person im Haushalt zu erreichen. Die Zielperson wird dann in der Regel meist mit dem *„Last-Birthday-Verfahren“* ausgewählt, bei dem die erreichte Person gefragt wird, welche im Haushalt lebende Person der Zielgruppe (z. B. ab einem bestimmten Alter) zuletzt Geburtstag hatte. Diese ist dann zu befragen.

Bei **Personenstichproben für mündliche Interviews** wird das sogenannte **„Random-Route-Verfahren“** angewendet. Ausgangspunkt ist, dass die gesamte

Bundesrepublik Deutschland in künstliche, kleine geographische Einheiten unterteilt ist, in denen jeweils ca. 1500 Menschen wohnen. Für jeden dieser sogenannten **„Sample Points“** gibt es eine Beschreibung der dort befindlichen Straßen mit den dort vorhandenen Hausnummern.

In einem ersten Schritt wird nun eine geschichtete Stichprobe von Sample-Points gezogen, so dass eine ausgewogene Verteilung nach Regionen und Gemeindegrößenklassen gegeben ist. Pro Sample Point wird nun aufgrund der Beschreibung der Sample Points jeweils ein Startpunkt vorgegeben. Ausgehend von diesem Startpunkt wird eine Liste von Namen an Türschildern erstellt. Dabei wird nach festen Regeln vorgegangen. Beispiele für solche Regeln sind:

- Es wird genau definiert, in welche Richtung der Straße gegangen werden soll.
- Bei Mehrfamilienhäusern ist unten links anzufangen, auf der linken Seite nach oben zu gehen und dann rechts wieder nach unten.
- Gewerbebetriebe, Arztpraxen, Rechtsanwaltskanzleien etc. sind nicht aufzuführen. Ist jedoch, z. B. in einer Schule, eine Hausmeisterwohnung vorhanden, so ist diese zu erfassen.

Diese Listen werden an das Institut gesandt, das dann z. B. jeden 10. Haushalt auswählt. Diese Haushalte sind zu kontaktieren. Soweit erweist sich das Verfahren als Beispiel für eine zweistufige Auswahl mit der ersten Stufe „Sample Point“ und der zweiten Stufe „Haushalt“.

Für die Bestimmung der zu befragenden Person bei Mehrpersonenhaushalten gibt es im Wesentlichen drei Verfahren:

- Der *„Schwedenschlüssel“* (das Verfahren wurde von einem Deutschen aus Schweden „importiert“) oder auch das „Kish Selection Grid“ (benannt nach dem Statistiker Leslie Kish): Dabei wird eine nach Alter geordnete Liste aller möglichen Zielpersonen des Haushalts erstellt. Zu jedem Interview und zu jeder Haushaltsgröße gibt es eine Zufallszahl, die von 1 bis zur jeweiligen Haushaltsgröße reicht und die zwischen den Interviews gleichverteilt sind. Nehmen wir an, dass ein Haushalt vier solcher Mitglieder hat. Dann gibt es zur Haushaltsgröße 4 eine Zufallszahl, die von 1 bis 4 reicht. Bei einem Viertel der Interviews steht eine 1, bei einem Viertel eine 2 usf. Nehmen wir an, dass zu diesem Interview die Zufallszahl 3 gehört. Dann ist die dritte Person auf der Liste zu befragen.
- Das *Last-Birthday-Verfahren,* das oben bereits bei den Telefonstichproben erläutert wurde.
- Die *Personenkette*: Dabei wird wie beim Schwedenschlüssel (s. o.) bei Mehrpersonenhaushalten eine Liste der Personen erstellt. Weiter ist eine Schrittweite (z. B. 3) festzulegen und ein Startwert von 1 bis zur Schrittweite zufällig zu bestimmen (z. B. 2). Nehmen wir an, der erste kontaktierte Haushalt ist ein 1-Personen-Haushalt. Da gemäß Startpunkt erst die 2. Person zu befragen ist, wird kein Interview durch-

geführt. Der zweite Haushalt sei ein Vierpersonenhaushalt. Diese vier werden an die erste kontaktierte Person nach Alter sortiert zu einer Personenkette angefügt und sind dort die Personen Nr. 2 bis 5. Gemäß Startwert ist nun die zweite Person der Kette (also die 1. Person des Vierpersonenhaushalts) und gemäß Schrittweite die Person Nr. $2+3=5$, also die vierte Person des Vierpersonenhaushalts, zu befragen. Entsprechend wird mit den weiteren Haushalten verfahren.

Die Personenkette hat dabei gegenüber den anderen Verfahren den *Vorteil,* dass Personen aus 1-Personenhaushalten die gleiche Wahrscheinlichkeit haben, in die Stichprobe zu gelangen, wie Personen aus 5-Personenhaushalten. Da beim Schwedenschlüssel und beim Last-Birthday-Verfahren von jedem erreichten Haushalt genau ein Interview durchgeführt wird, haben dort Personen aus 1-Personenhaushalten die fünffache Wahrscheinlichkeit, ausgewählt zu werden im Vergleich zu Personen aus 5-Personen-Haushalten. Dies muss durch Gewichtung wieder ausgeglichen werden, was für die Genauigkeit der Ergebnisse nachteilig ist.

Nachteilig bei der Personenkette ist, dass es möglich ist, dass (wie oben am Beispiel gezeigt) aus einem kontaktierten Haushalt keine Person oder auch mehr als eine Person befragt wird, wobei das erstgenannte Problem die Kosten erhöht, das zweite insofern problematisch sein kann, weil die zweite zu interviewende Person evtl. den Fragebogen schon kennt.

Bis jetzt wurden allgemeine Bevölkerungsumfragen behandelt, bei denen die Grundgesamtheit definiert ist als alle in Deutschland in Privathaushalten lebenden Personen ab einem bestimmten Alter. In der Praxis wichtig sind noch Verfahren, die es erlauben, eine Stichprobe für eine Grundgesamtheit zu ziehen, die nur einen Teil der Wohnbevölkerung umfasst. Solche Teilgruppen können z. B. Hundebesitzer, PKW-Fahrer oder Motorradfahrer sein. Wesentlich für das anzuwendende Verfahren ist, wie groß der Anteil der Zielgruppe in der Bevölkerung ist, die sogenannte „Inzidenz". Bei hoher Inzidenz wird i. A. eine allgemeine Bevölkerungsstichprobe gezogen und gleich zu Beginn des Interviews gefragt (sogenannte *„Screeningfrage"*), ob die Person zur Grundgesamtheit gehört oder nicht. Ist der Anteil dagegen gering, so wird entweder versucht, über „Adressbroker" Adressen der Zielgruppe zu kaufen oder man geht im Schneeballverfahren vor. Dabei sucht man eine solche Person und versucht, von dieser Kontaktdaten für weitere Interviews zu erhalten.

4.3 Repräsentative Befragungen

4.3.1 Grundlagen der Fragenformulierung

4.3.1.1 Einführung

Die Formulierung von Fragen und der Entwurf von Fragebögen für Befragungsverfahren galten über lange Zeit als „Kunstlehre", die vor allem auf Erfahrung beruhte. Nicht

zufällig trägt das über Jahrzehnte einflussreiche Buch von Stanley Payne zur Frageformulierung von 1951 (!) den Titel: „The *Art* of Asking Questions“. Seit den 1980er Jahren hat eine umfassende und theoretisch fundierte Forschung zur Frageformulierung und Fragebogenentwicklung zu entsprechendem Wissen geführt, das den Fragebogenentwurf zumindest teilweise erlernbar macht. Die Bedeutung der Frageformulierung für valide und reliable Untersuchungsergebnisse wird sofort deutlich, wenn man feststellt, wie stark sich selbst geringfügig wirkende Unterschiede von Erhebungsmethoden auf Untersuchungsergebnisse auswirken. In der Literatur (vgl. z. B. Bradburn & Sudman, 1979; Schuman & Presser, 1981) findet sich dazu eine Fülle von Beispielen.

Beispiel

Als **erstes** (schon „klassisches“) **Beispiel** zur Fehlerempfindlichkeit von Befragungen hier das Ergebnis eines Vergleichs von zwei (scheinbar) nur formal unterschiedlichen Antwortskalen, die zu deutlich verschiedenen Ergebnissen führten:

Schwarz et al. (1991) (siehe auch Schwarz, 1999) verwendeten zwei prinzipiell gleichartige numerische Antwortskalen, die von −5 bis +5 bzw. von 0 bis 10 reichten, zur Messung der Lebenszufriedenheit und es ergaben sich deutlich unterschiedliche Ergebnisse. Es zeigte sich bei der erstgenannten Skala ein wesentlich niedrigerer Anteilswert (13 %) der Angaben in der unteren Hälfte der Skala (von −5 bis 0) als in der anderen Skala im entsprechenden Bereich von 0 bis 5 (Anteilswert 34 %), weil anscheinend Auskunftspersonen die negativen Werte in der erstgenannten Skala nicht als *geringe Zufriedenheit,* sondern als *ausgeprägte Unzufriedenheit* interpretierten. Eine eher als Formalie erscheinende Veränderung der Antwortskala hat also zu einem sehr deutlichen (Faktor 2,6!) Ergebnisunterschied geführt.

Für den Nutzer von Untersuchungsergebnissen ähnlich schlecht erkennbar ist der Einfluss der Fragetechnik beim **zweiten** (ebenfalls klassischen) **Beispiel** (Quelle: Bradburn & Sudman, 1979, S. 14 ff.). Hier ging es um die eher schlichte Fragestellung, wie viele Dosen Bier Amerikaner pro Jahr trinken. Dazu wurden bei unterschiedlichen (jeweils repräsentativen) Teilstichproben verschiedene Fragetechniken verwendet:

Lange oder kurze Frage: Die kurze Frage war direkt auf den Bierkonsum gerichtet, die lange Frage war dagegen mit einer Einleitung versehen, in der auf unterschiedliche Situationen, in denen Bier getrunken wird, Bezug genommen wurde.

„Offene“ oder „geschlossene“ Frage: Bei der sog. offenen Frage war die Angabe zum Bierkonsum im letzten Jahr (z. B. „180 Dosen“) direkt einzutragen. Bei der geschlossenen Frage war eine vorgegebene Antwortkategorie (z. B. „Bis zu 50 Dosen“, „51 bis 100 Dosen“, „101 bis 200 Dosen“, „201 bis 300 Dosen“ und „Über 300 Dosen“ zu wählen).

Es ergaben sich die folgenden Ergebnisse (Mittelwerte):

Lange, offene Frage	320 Dosen pro Jahr
Kurze, geschlossene Frage	131 Dosen pro Jahr

D. h. mit der ersten Fragetechnik lag der Schätzwert für den Bierkonsum von Amerikanern fast 2,5-mal so hoch wie mit der zweiten Frageform! Woran kann das liegen? Anscheinend wirken hier zwei Effekte gemeinsam:

Bei der langen Frage („Es gibt öfter mal Situationen, in denen man ein Bier trinkt, z. B. mit Kollegen, bei Feiern oder abends vor dem Fernseher; wie viele Dosen sind das bei Ihnen pro Jahr?“) findet eine Aktivierung der Erinnerung statt und außerdem wird der sozial eher unerwünschte Bierkonsum gewissermaßen gerechtfertigt. Bei der geschlossenen Frage (mit Antwortkategorien) zögern Auskunftspersonen, sich den höchsten Kategorien zuzuordnen, weil sie erkennen, dass sie sich damit zu den Gruppen mit der stärksten Ausprägung des sozial unerwünschten Verhaltens „Bierkonsum“ bekennen. Bei der offenen Frage ist dagegen die Relation des eigenen Bierkonsums zum „üblichen“ Bierkonsum nicht erkennbar. ◀

Nun also zu den Überlegungen zur Formulierung von Fragen. Ausgangspunkt dafür ist natürlich die Festlegung, welche Meinungen, Sachverhalte, Einstellungen etc. überhaupt ermittelt werden sollen. Diese sind durch die Problemdefinition (siehe Abschn. 2.2.1) und insbesondere die Untersuchungshypothesen (siehe Abschn. 2.2.2) schon weitgehend bestimmt. Wenn man beispielsweise die Hypothese überprüfen will, ob mangelnde Zufriedenheit mit dem Service eines Unternehmens zu geringerer Kundenbindung führt, dann muss man eben erheben, wie groß die Zufriedenheit mit dem Service und wie stark die Bindung der Kunden ist.

Was ist in diesem Zusammenhang eigentlich unter einer **„Frage“** zu verstehen? Der Begriff der Frage geht hier sicher über eine bestimmte sprachliche Form, an deren Ende ein Fragezeichen steht, hinaus. Man meint damit vielmehr jegliche Art der Aufforderung, entsprechende Informationen zu geben (Tourangeau et al., 2000, S. 29).

So könnte eine Frage nach dem Alter beispielsweise lauten:

- „Wie alt sind Sie?“ oder
- „Tragen Sie hier bitte Ihr Alter ein: …“

Beispiel

Hier ein weiteres Beispiel für unterschiedliche Frageformen bezüglich der Zufriedenheit mit einem Produkt:

- „Sind Sie mit der Qualität von Produkt XY zufrieden?“ oder
- „Geben Sie bitte an, inwieweit Sie mit der Qualität von Produkt XY zufrieden sind“ oder
- „Ich würde gern wissen, inwieweit Sie mit der Qualität von Produkt XY zufrieden sind.“ ◀

Angesichts der schon angesprochenen Fehlerempfindlichkeit von Befragungen ist eine sorgfältige und verständige Vorgehensweise bei der Entwicklung von Fragen besonders

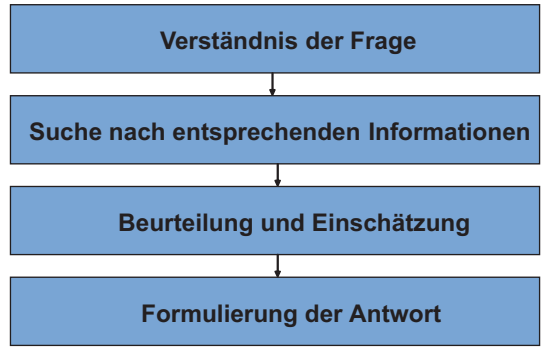


Abb. 4.6 Modell des Antwortprozesses bei Befragungen

wichtig. Typischerweise ist dieser Prozess mit diversen so genannten „Pretests " (siehe Abschn. 4.3.3) verbunden, bei denen die Eignung einzelner Fragen für den Untersuchungszweck überprüft wird. Für eine zweckmäßige (→ Validität) Formulierung von Fragen ist das Verständnis der bei der Auskunftsperson bei der Beantwortung von Fragen ablaufenden Prozesse eine wesentliche Voraussetzung. Dafür hat das in Abb. 4.10 wiedergegebene Modell breite Akzeptanz gefunden (Tourangeau et al., 2000, S. 7 ff.; Groves et al., 2009, S. 218 ff.; Lenzner & Menold, 2015). Am Beginn des Antwortprozesses steht – nicht ganz überraschend – das Verständnis der gestellten Frage. Es folgt die Suche nach Informationen/Erinnerungen im Gedächtnis, die dem erfragten Sachverhalt entsprechen. Diese Informationen werden zu einer entsprechenden Beurteilung/Einschätzung zusammengefasst. Am Ende stehen die Formulierung und Übermittlung der Antwort (Abb. 4.6).

Diese vier Aspekte werden im folgenden Abschnitt eingehender erläutert und durch Beispiele illustriert. Es folgen dann Ausführungen zu einigen weiteren allgemeinen Prinzipien der Frageformulierung. Eine besonders bedeutsame spezielle Fragetechnik, sogenannte Multi-Item-Skalen, bei denen mehrere einzelne Angaben zu einem Messwert (z. B. für eine Einstellung oder die Kundenzufriedenheit) zusammengefasst werden, wird im Abschn. 4.3.2 gesondert behandelt.

4.3.1.2 Grundlegende Anforderungen an Frageformulierungen

Der Überblick über einige „Regeln" bei der Frageformulierung im vorliegenden Abschnitt orientiert sich – wie gesagt – vor allem an den vier gerade skizzierten Elementen des Antwortprozesses.

Der erste Aspekt ist also das **Verständnis der Frage.**

„Eine Survey-Frage sollte von allen Befragten in der vom Fragebogenentwickler intendierten Weise interpretiert werden und möglichst leicht zu verstehen und zu beantworten sein." (Lenzner & Menold, 2015, S. 1). Was gehört nun zum Verständnis einer Frage? Zunächst muss die Auskunftsperson bei der Frage und den damit verbundenen Hinweisen (z. B. zu Antwortmöglichkeiten) zuhören. Sie muss weiterhin die logische Form (den Satzbau) verstehen und daraus entnehmen, welche Angabe gewünscht wird.

Häufig werden dabei auch der Fragebogenkontext und die Antwortkategorien zur Interpretation der Frage herangezogen (Schwarz, 1999). Letztlich ist es notwendig, dass die in der Frage verwendeten Worte bei allen Auskunftspersonen möglichst einheitlich mit den entsprechenden gedanklichen Konzepten in Beziehung gesetzt werden.

Jaccard und Jacoby (2020, S. 403 ff.) haben die Literatur zum Verständnis von Fragen umfassend analysiert und dabei folgende Gruppen von Einflussfaktoren identifiziert:

- *Charakteristika der befragten Personen* mit den einzelnen Aspekten: Lesefähigkeit, kognitive Fähigkeiten, MuttersprachlerIn oder nicht, Aufmerksamkeit, allgemeine Motivation
- *Kontext der Fragen* mit den einzelnen Aspekten: Ablenkung, Motivation durch die Aufgabe, Fragereihenfolge u. a.
- *Vokabular der Fragen* mit den einzelnen Aspekten: Verwendung spezifischer Fachbegriffe, Verwendung von Abkürzungen, Zweideutigkeit von Worten, ungebräuchliche Worte u. a.
- *Grammatikalische Struktur der Fragen* mit den einzelnen Aspekten: Verwendung von Negationen, Komplexität der Sätze u. a.

Ein zentraler Aspekt dabei ist das Problem, ob jede Auskunftsperson die **verwendeten (Fach-) Worte kennt.** Dazu einige Beispiele:

- „Wie hoch ist Ihr Involvement bei …?“
- „Mögen Sie ‚Blanquette de Limoux‘?
- „Wie hoch ist Ihr Annuitätendarlehen?“

Ein zweites Problem, das **präzise und einheitliche Verständnis der Frage,** lässt sich anhand weiterer Beispiele leicht nachvollziehen:

- Haben Sie in letzter Zeit…?“ (Was heißt „in letzter Zeit“? Letzte Woche, letzter Monat, letztes Jahr?)
- Wie hoch ist Ihr Einkommen?“ (brutto/netto? Monatlich/jährlich? persönliches Einkommen oder Familieneinkommen? Arbeitseinkommen oder einschl. Renten, Zinsen etc.?)
- Gehen Sie oft ins Kino?“ (Was ist „oft“? Wöchentlich, zweimal pro Woche, monatlich?)

Regeln für verständliche Fragen

Sudman und Blair (1998, S. 255 f.), Groves u. a. (2009, S. 220 ff.), Dillman et al. (2009, S. 79 ff.) und Lenzner und Menold (2015, S. 1 ff.) formulieren einige Regeln, um Probleme bei der Verständlichkeit von Fragen zu vermeiden bzw. zu reduzieren, von denen hier einige wiedergegeben seien:

- **Spezifisch sein!**

  Beispiel: Anstelle einer Frage *„Haben Sie in letzter Zeit größere Anschaffungen getätigt?“* (Was ist „in letzter Zeit“? Was sind „größere Anschaffungen“?) wäre die spezifischere Formulierung zu verwenden:

  „Sagen Sie mir bitte, ob Sie in den letzten 6 Monaten Möbel oder Elektrogeräte im Wert von mehr als 500 € gekauft haben.“

Möbel für mehr als 500 €	ja/nein
Elektrogeräte für mehr als 500 €	ja/nein

- **Wer, was, wann, wo und wie verdeutlichen!**
  Beispiel: *„Wie hoch war Ihr Brutto-Haushaltseinkommen im Jahr 2023? Bitte berücksichtigen Sie dabei die Einkommen aller Mitglieder Ihres Haushalts einschließlich Renten, Zinsen u. ä.“*
- **Abstrakte und mehrdeutige Begriffe vermeiden!**
  Beispiel: *„Wie wichtig war Ihre Allgemeinbildung für Ihre berufliche Entwicklung?“* Was ist mit Allgemeinbildung gemeint? Gehören dazu nur die „klassischen“ Inhalte wie Geschichte, Literatur, Sprachen etc.? Oder bezieht man auch Kenntnisse über Technik oder Sport ein?
- **Festlegen, wie geantwortet werden soll**!
  Auf die Frage *„Wie zufrieden sind Sie mit Ihrem Auto?“* sind die unterschiedlichsten Antworten denkbar, z. B. „Das war ein Fehlkauf“, „Hervorragend“, „Prima Auto“, „Mittelmäßig“, „Gerade so zufrieden“. Deswegen lautet die Alternative *mit vorgegebenen Antwortmöglichkeiten*:
  „Wie zufrieden sind Sie mit Ihrem Auto?“
  Sehr zufrieden, einigermaßen zufrieden, wenig zufrieden, überhaupt nicht zufrieden?“
- **Doppelte Verneinungen vermeiden!**
  Beispiel: *„Bei ärmeren Familien sollte nicht weniger Beachtung des Umweltschutzes beim Konsum erwartet werden als bei reicheren Familien. Stimmen Sie dieser Aussage zu?“ Ja/nein*
- **Einfache und kurze Sätze verwenden!**
  Jede Frage muss so formuliert werden, dass sie für alle – auch die sprachlich weniger gewandten – Auskunftspersonen voll verständlich ist. Es ist also eine möglichst kurze, grammatikalisch einfache und dem Wortschatz der Alltagssprache angepasste Frageformulierung zu suchen. Das schließt natürlich nicht aus, dass man bei der Befragung spezieller Zielgruppen (z. B. Ärzte, Einkaufsleiter) auch deren möglicherweise hoch entwickelte Fachsprache benutzt.

- **Ausmaße, Häufigkeiten etc. durch Zahlen angeben lassen!**
  Beispiel: *„Wie oft gehen Sie durchschnittlich pro Monat ins Kino?“* „… mal“ an Stelle von *„Gehen Sie oft ins Kino?“*
- **Verdeutlichen, auf wen sich die Frage bezieht!**
  Beispiel: *„Wie oft kaufen Sie Waschmittel?“* Ist bei dieser Frage die Auskunftsperson gemeint oder der Haushalt, in dem diese lebt?

Der zweite Schritt im Antwortprozess ist die **Suche nach entsprechenden Informationen.**

Hier geht es also um die Suche nach spezifischen Gedächtnisinhalten. Diese können sich auf Ereignisse und Erfahrungen in der Vergangenheit (z. B. Urlaubsziel im letzten Jahr oder zuletzt gekaufte Biermarke) beziehen, aber auch auf früher gebildete Einstellungen und Meinungen. Tourangeau et al. (2000, S. 82) stellen die wesentlichen Gründe zusammen, die dazu führen, dass **Gedächtnisinhalte,** auf die in der Frage Bezug genommen wird, nicht verfügbar sind:

- Die relevanten Informationen sind nicht aufgenommen worden. Beispiel: Wer beim Kauf die Marke nicht beachtet, kann später auch nicht angeben, welche Marke er gekauft hat.
- Die Auskunftsperson scheut die Mühe, die mit der Erinnerung an möglicherweise lang zurückliegende Einzelheiten verbunden ist.
- Die Auskunftsperson erinnert sich nicht an das spezifisch erfragte Ereignis (z. B. letzter Einkauf von Kaffee), sondern an allgemeinere entsprechende Informationen (z. B. allgemein bevorzugte Kaffee-Marke).
- Die Auskunftsperson erinnert sich nur an einzelne Bruchstücke der erfragten Information. Beispiel: Wer kann sich schon an die genauen Mengen (Anzahl von Flaschen) und die genauen Preise (€ 0,79 oder € 0,89) beim letzten Einkauf von Erfrischungsgetränken erinnern?
- Die Auskunftsperson erinnert sich falsch. Beispielsweise verwechselt sie Marken oder Einkaufsstätten.

Ebenfalls bei Tourangeau et al. (2000, S. 98) findet sich eine Zusammenstellung von empirischen Ergebnissen zum Erinnerungsvermögen von Auskunftspersonen. Danach zeigt sich – nicht wirklich überraschend –, dass länger zurückliegende Ereignisse schlechter erinnert werden. Die Erinnerung an Ereignisse ist relativ gut, wenn diese in der zeitlichen Nachbarschaft von besonderen Zeitpunkten (z. B. Weihnachten, Semesterbeginn) stattgefunden haben, besonders herausgehoben waren (z. B. Urlaub, Geburtstag) oder wenn diese (emotional) bedeutsam waren (z. B. Heirat, Ortswechsel). Es wird empfohlen, bei entsprechenden Fragestellungen nach Möglichkeit darauf Bezug zu nehmen.

Nun zu einigen Problemen der Suche nach Informationen bei der Auskunftsperson, die in der praktischen Marktforschung häufig auftreten. Ein erstes Problem besteht darin, dass gelegentlich nach Einschätzungen gefragt wird, die entsprechende **Erfahrungen** voraussetzen, welche nicht immer vorhanden sind. Dazu zwei Beispiele:

- Was antwortet der bisher unfallfreie Autofahrer auf die Frage nach der Kulanz bei der Schadensregulierung seiner Kfz-Versicherung?
- Was sagt der passionierte Biertrinker, wenn er gefragt wird, ob er Côtes-du-Rhône-Wein oder Rioja im Geschmack angenehmer findet?

Das zweite Problem bezieht sich auf die **Fähigkeit, sich zu erinnern.** Viele Phänomene, die für Marktforscher interessant sind (z. B. Markenwahl, Mediennutzung), sind für Konsumenten so unwichtig, dass sie sich die entsprechenden Informationen nicht merken und deshalb solche Fragen nicht korrekt beantworten können. Auch dazu zwei Beispiele:

- „Haben Sie in den letzten 6 Monaten mindestens einmal ein Heftpflaster verwendet?“
- „Haben Sie am Mittwoch der vergangenen Woche die ‚Tagesschau' gesehen?“

Drittens geht es hier um **Meinungen** oder **Absichten,** die im Hinblick auf Präferenzen und zukünftiges Kaufverhalten interessant sind, die aber bei den Auskunftspersonen (noch) nicht so ausgeprägt vorhanden sind, dass sie schon hinreichend klar geäußert werden können. Wieder zwei Beispiele:

- „Wie ist Ihre Meinung über den Flug-Komfort bei ‚EasyJet'?“
  (Hat jeder irgendeine Meinung zum Flug-Komfort bei ‚EasyJet'?)
- „Haben Sie die Absicht, in den nächsten vier Wochen einen Haushaltsreiniger zu kaufen?“
  (Kein Mensch macht vier Wochen im Voraus Pläne, Haushaltsreiniger zu kaufen, sondern kauft so etwas, wenn er es gerade braucht.)

Empfehlungen, um die Fähigkeit, sich zu erinnern, zu verbessernHier wieder einige Empfehlungen von Sudman und Blair (1998, S. 260 f.) und Lenzner und Menold (2015, S. 7 ff.) zur Fähigkeit der Auskunftsperson, Antworten zu geben:

- **Angemessenen Zeitrahmen für die Erinnerung wählen!**
  An herausragende Konsum-Entscheidungen (z. B. Studienreisen, Autos, Luxus-Garderobe) erinnert man sich teilweise über Jahre, bei alltäglichen Einkäufen (Lebensmittel, Reinigungsmittel etc.) verblasst die Erinnerung schon nach wenigen Tagen.
  Beispiele: „Wie oft sind Sie in den *letzten 10 Jahren* umgezogen?“ **aber** „Wie oft waren Sie in den *letzten 7 Tagen* in einer Gaststätte?“

- **Gedächtnisstützen geben!**
  Durch Formulierungen wie im folgenden Beispiel kann die Erinnerung aktiviert werden:
  „Es gibt ja verschiedene Anlässe, in eine Gaststätte zu gehen, beim schnellen Hunger oder Durst unterwegs, mit Freunden oder mit Partner/Partnerin zu einem besonderen Anlass. Wie oft haben Sie in den letzten 7 Tagen eine Gaststätte besucht?“
- **Auch die Vorgabe von Antwortkategorien kann die Erinnerung auffrischen!**
  „Welche der folgenden Zeitungen und Zeitschriften haben Sie in den letzten vier Wochen mindestens einmal gelesen?“

• ZEIT	(…)
• Spiegel	(…)
• Süddeutsche	(…)
• Tagesspiegel	(…)

- Zur Heranziehung von Unterlagen (z. B. Rechnungen, Packungen) ermuntern!
- Nicht jeder Konsument weiß z. B., welche Marke von Haushaltsrollen er gekauft hat. Hier erleichtert vielleicht ein Blick in den Küchenschrank die Wahrheitsfindung!
- **Bei der Messung von Kaufabsichten möglichst genau über Produkt, Preis etc. informieren!**
  Je deutlicher die Befragungssituation auf diese Weise an die tatsächliche Kaufsituation angenähert ist, desto realistischer sind die Angaben zu Kaufabsichten.

Nun zum dritten Schritt im Antwortprozess, der Bildung von **Beurteilungen und Einschätzungen.**

Im vorigen Schritt des Modells des Antwortverhaltens ging es ja überwiegend um den Zugriff zu im Gedächtnis gespeicherten Informationen. Jetzt stehen Aspekte der Informationsverarbeitung im Mittelpunkt des Interesses. Dabei geht es nach Tourangeau et al. (2000, S. 8 ff.) in erster Linie um:

- Ziehung von Schlüssen auf der Basis vorhandener Informationen
- Zusammenfassung der vorhandenen Informationen
- Entwicklung einer Einschätzung, insbesondere bei Meinungs- oder Einstellungsfragen

Zunächst also zu den **Schlüssen auf Basis vorhandener Informationen.** Groves et al. (2009, S. 234 ff.) illustrieren diesen Vorgang am Beispiel der Häufigkeit der Einnahme

von Tabletten. Kaum jemand wird das direkt speichern und gewissermaßen einen „Zähler mitlaufen lassen“ („Genau 37 Tabletten seit dem 1. Juni“), wenn er eine Tablette nimmt. Typisch sind eher Schlüsse in folgender Art: „Etwa eine halbe Packung im letzten Jahr“ (gespeicherte Information) → „Eine Packung enthält 60 Tabletten.“ (gespeicherte Information) → „Also etwa 30 Tabletten im Jahr.“ (Schluss). Ein weiteres Beispiel zur Bildung eines Schätzwertes für die Anzahl eingenommener Tabletten: „Etwa alle zwei Wochen“ (gespeicherte Information) → „Also etwa 25 Tabletten im Jahr“ (Schluss).

Für die **Zusammenfassung vorhandener Informationen** sind Beispiele leicht zu identifizieren. So müssen, wenn nach einem Qualitätsvergleich bei zwei Marken gefragt wird, die einzelnen Einschätzungen und Erfahrungen erinnert und dann zueinander in Beziehung gesetzt werden. Bei einer Frage „Wie teuer war Ihre letzte Urlaubsreise insgesamt, wenn Sie an Fahrtkosten, Hotelkosten, Verpflegung und sonstige Kosten denken?“ müssen die einzelnen Kostenfaktoren erinnert und zusammengefasst werden.

Bei der **Entwicklung von Einschätzungen, Meinungen oder Einstellungen** ist der entsprechende kognitive Prozess deutlich aufwendiger, weil die Zusammenfassung einzelner Aspekte mit Bewertungen im Hinblick auf ein zusätzliches Kriterium verbunden ist. Bei einer Frage „Wie waren Sie mit Ihrer letzten Pauschalreise im Hinblick auf Flug, Transferservice, Verpflegung, Hotel und örtliche Reiseleitung zufrieden?“ müssen die einzelnen Aspekte erinnert und dann eine Zufriedenheitseinschätzung gebildet werden. Bei vielen Fragen nach Meinungen oder Einstellungen muss man wohl davon ausgehen, dass die Auskunftspersonen diese nicht gewissermaßen „auf Vorrat“ bilden und laufend aktualisieren, sondern dass diese erst gebildet werden, wenn eine entsprechende Frage gestellt wird.

Empfehlung

Hier wieder eine Empfehlung von Sudman und Blair (1998, S. 260 f.) zur Fähigkeit der Auskunftsperson, Antworten zu geben:

Antwortfähigkeit selbst einschätzen lassen!

Man kann die Auskunftsperson zunächst fragen, ob sie eine entsprechende Meinung hat, das betreffende Produkt schon mal gekauft habe etc. und nur bei positiver Antwort die jeweilige Frage stellen. Beispiel:

„Haben Sie selbst schon Erfahrungen mit der Marke SONY gemacht oder von anderen Leuten über die Marke SONY gehört?“

() Nein	„Weiter zu Frage…“
() Ja	*„Wie ist Ihre Meinung über die Marke SONY?“*

Es ist leicht nachvollziehbar, dass eine Überforderung beim Interview vermieden werden muss. Die Genauigkeit entsprechender Angaben darf auch nicht überschätzt werden. So lassen Untersuchungsergebnisse zur Abrundung bei numerischen Angaben (z. B. Schaeffer & Bradburn, 1989) erkennen, dass – der Alltagserfahrung entsprechend – Zahlen, die durch 5 bzw. 10 teilbar sind, weit überproportional angegeben werden (z. B. weit häufiger 10, 15, 20 als 11, 17, 21). Das weist deutlich darauf hin, dass (verständlicherweise) viele Auskunftspersonen sich nur um eine begrenzte Genauigkeit ihrer Antworten bemühen (Krosnick, 1999, S. 547 f.).

Nun zum letzten Schritt im Antwortprozess, der Formulierung der Antwort
Bei der Formulierung einer Antwort sind hier zwei Aspekte von besonderem Interesse: Die Zuordnung einer Einschätzung zu einer (vorgegebenen) Antwortkategorie und die Überprüfung und gegebenenfalls Modifizierung der Antwort, z. B. im Hinblick auf die soziale Akzeptanz dieser Antwort. Zunächst zur Zuordnung zu Antwortkategorien. Das ist sicher recht einfach, wenn es sich um numerische Angaben handelt. Wenn man weiß, dass man in einem bestimmten Zeitraum z. B. ein bestimmtes Produkt viermal gekauft hat, dann ist die Zuordnung zu einer der Kategorien „Nicht gekauft“, „1–3-mal gekauft“, „4–6-mal gekauft“, „Mehr als 6-mal gekauft“ keine große intellektuelle Herausforderung. Schon schwieriger ist z. B. die Zuordnung eines Qualitätsurteils, das aus den Einschätzungen „funktioniert gut, komplizierte Bedienung, zuverlässig“ besteht, zu einer der Antwortkategorien „Sehr gut“, „Gut“, „Mittelmäßig“, „Schlecht“, „Sehr schlecht“. Einige detaillierte Hinweise zur angemessenen Gestaltung von Antwortkategorien geben z. B. Lenzner und Menold (2019).

Von vorgegebenen Antwortkategorien kann auch ein erheblicher Einfluss auf das Antwortverhalten ausgehen (Schwarz, 1999). Besonders gängig sind sogenannte **Primacy-** und **Recency-Effekte.** Diese beziehen sich darauf, dass Antwortmöglichkeiten, die am Beginn („Primacy“) bzw. am Ende („Recency“) der Liste von Antwortkategorien stehen, oftmals häufiger gewählt werden, als es bei einer anderen Platzierung der Fall wäre (Groves et al., 2009, S. 239 f.). Das ist darauf zurückzuführen, dass sich Auskunftspersonen oftmals für die erste in Frage kommende Kategorie entscheiden, bevor sie alle Antwortmöglichkeiten kennen (Primacy-Effekt) oder darauf, dass die letztgenannten Antwortmöglichkeiten bei der Antwort besser erinnert werden (Recency-Effekt). Aus diesem Grund werden bei computergestützten Befragungen lange Listen in der Regel zufällig rotiert und jeder Befragte eine zufällige Anordnung der Antwortvorgaben angeboten bekommt. In einer ganzen Reihe von Untersuchungen ist auch gezeigt worden, dass die vorgegebenen Antwortkategorien von der Auskunftsperson zur Interpretation der Frage verwendet werden und einen Eindruck von dem üblichen und erwarteten Antwortspektrum vermitteln. Tab. 4.1 zeigt ein entsprechendes Beispiel von Schwarz et al. (1985).

Beispiel

Man erkennt in der folgenden Tabelle leicht, dass bei den links wiedergegebenen Antwortverteilungen 16,2 % der Befragten angeben, mehr als 2 1/2 h pro Tag fernzusehen. Bei den rechts dargestellten Antwortkategorien geben aber 37,5 % der Befragten einen Fernsehkonsum von über 2 1/2 h an. Wie ist dieser Unterschied zu erklären? Bei den links aufgeführten Kategorien ist die Angabe „2 1/2h" der höchste Wert, der in der Wahrnehmung von Auskunftspersonen für besonders hohen Fernsehkonsum steht. Bei den rechts genannten Kategorien liegen dagegen diverse Werte über 2 1/2 h im Rahmen des von den Befragten (durch die Antwortvorgaben) als „normal" wahrgenommenen Spektrums. Nun gilt sehr hoher Fernsehkonsum als sozial eher unerwünscht. Kaum jemand ist stolz darauf, täglich viele Stunden vor dem Fernseher zu sitzen. Das kann dazu führen, dass Auskunftspersonen ihre Angaben zum Fernsehkonsum nach unten korrigieren, wenn sie den Eindruck haben, in einem relativ hohen Bereich zu liegen.

Antwortverteilungen zur täglichen Fernsehdauer von Deutschen bei unterschiedlichen Vorgaben von Antwortkategorien. (Quelle: Schwarz et al., 1985, S. 391)

Antwort-Vorgabe 1	Anteil der Antworten	Antwort-Vorgabe 2	Anteil der Antworten
<½ h	7,4 %	<2 ½ h	62,5 %
½ bis 1 h	17,7 %	2 ½ bis 3 h	23,4 %
1 bis 1 ½ h	26,5 %	3 bis 3 ½ h	7,8 %
1 ½ bis 2 h	14,7 %	3 ½ bis 4 h	4,7 %
2 bis 2 ½ h	17,7 %	4 bis 4 ½ h	1,6 %
>2 ½ h	16,2 %	>4 ½ h	0,0 %
	100 %		100 %

◀

Das Beispiel von Schwarz et al. (1985) leitet über zu dem Gesichtspunkt der **Modifizierung von Antworten,** um bestimmten Anforderungen zu genügen. Hier steht der Aspekt im Mittelpunkt, ob bestimmte Antworten den üblichen **sozialen Normen** entsprechen und ob deswegen das Antwortverhalten der Auskunftspersonen durch Anpassung an diese Normen entsprechend beeinflusst wird. Hier einige drastische Beispiele für Fragen, bei denen dieser Faktor wohl eine Rolle spielt:

- „Haben Sie in den letzten zwei Jahren ein Buch gelesen?"
- „Trinken Sie täglich hochprozentigen Alkohol?"
- „Würden Sie beim Einkauf von Reinigungsmitteln Gesichtspunkte des Umweltschutzes beachten?" (Beim heute hohen Stellenwert des Umweltschutzes bekennen sich viele Leute dazu, deutlich weniger nehmen die meist höheren Preise umweltschonender Produkte tatsächlich in Kauf.)

Regeln von Sudman und Blair

Hier wieder einige auf langer Erfahrung und zahlreichen Untersuchungen basierenden „Regeln“ von Sudman und Blair (1998, S. 263 ff.) zum Problem der (mangelnden) Angaben zu sozial unerwünschtem Verhalten:

- **Neutrales, sachliches Verhalten der Interviewer (persönlich, telefonisch) trainieren!**
  Dadurch soll zumindest eine Verstärkung des Einflusses sozialer Erwünschtheit durch die Person des Interviewers vermieden werden.
- **„Sponsorship-Effekt“ vermeiden!**
  Der Begriff „Sponsorship-Effekt“ bezieht sich darauf, dass in Fällen, in denen der Auftraggeber („Sponsor“) einer Untersuchung erkennbar ist, ein Teil der Auskunftsperson dazu neigt, sich an die vermutlich vom Auftraggeber „gewünschten“ Antworten anzupassen, um eine Disharmonie zu vermeiden und den Interviewer ohne große Probleme „los zu werden“. Deswegen kann es besser sein, wenn der Auftraggeber einer Untersuchung nicht erkennbar wird.
- **In bestimmten Fällen: Offene statt geschlossene Fragen verwenden!**
  Hier kann auf das in Abschn. 4.3.1.1 skizzierte „Bier-Beispiel“ verwiesen werden. Bei geschlossenen Fragen – also Fragen mit vorgesehenen Antwortkategorien – zeigen Auskunftspersonen besondere Zurückhaltung gegenüber extremen Antwortkategorien, wenn es um Fragestellungen geht, bei denen die soziale Akzeptanz von Verhaltensweisen (z. B. Alkoholkonsum, umweltfreundliches Verhalten) eine Rolle spielt.
- **Längere Fragetexte zur Reduktion des „sozialen Stigmas“ verwenden!**
  Beispiel: *„Viele Leute haben heutzutage keine Zeit, Bücher zu lesen. Wie ist das bei Ihnen?“*

4.3.1.3 Weitere allgemeine Prinzipien der Frageformulierung

Grundlegend und allgemein gültig ist natürlich das Prinzip der **Neutralität** jeder Fragestellung. Fragen, bei denen die Attraktivität verschiedener Antwortmöglichkeiten unterschiedlich ist, führen zu entsprechend verzerrten Ergebnissen. Hier ist vor allem an die Auswirkungen suggestiver Formulierung (Plumpes Beispiel: „Sind Sie mit mir der Meinung, dass …?“), an im Hinblick auf positive und negative Meinungen ungleichgewichtige Antwortvorgaben („Sehr gut“, „gut“, „mittel“, „schlecht“) und an die Bindung einzelner Antwortmöglichkeiten an Wertvorstellungen („Sind Sie für die Einführung von Studiengebühren, um die Studienbedingungen zu verbessern?“) zu denken.

In derartigen Fällen werden der Auskunftsperson abweichend von deren eigentlichen Meinungen, Einstellungen etc., die ja ermittelt werden sollen, eine oder mehrere Antwortkategorien nahegelegt. Die Gefahr der Verzerrung der Ergebnisse durch diese Einflüsse ist dann besonders groß, wenn Sachverhalte erfragt werden, zu denen sich der

Befragte erst im Augenblick der Befragung eine Meinung bildet, und nicht auf eine früher gebildete „stabile“ Auffassung zurückgreifen kann.

Empfehlungen zur Fragenformulierung

Hier wieder einige Empfehlungen von Sudman und Blair (1998, S. 265), Dillman et al. (2009, S. 88) und Vomberg und Klarmann (2022, S. 100 f.):

- **Positiv oder negativ besetzte Worte vermeiden!**
  z. B. „Funktionär“, „Gerechtigkeit“, „giftig“
- **Extreme Begriffe („alles“, „immer“, etc.) vermeiden!**
  „Tun die Kraftwerkbetreiber alles, was ihnen möglich ist, für den Umweltschutz?“ Wer tut schon alles, was möglich ist? Deswegen sind hier relativ wenige positive Antworten zu erwarten.
- **Bezugnahme auf Normen vermeiden!**
  „Sind Sie – wie die meisten Menschen – der Ansicht, dass …?“
- **Ausgewogene („balancierte“) Antwortmöglichkeiten geben!**
  z. B. „sehr gut – gut – mittel – schlecht – sehr schlecht.“

Eine Ausnahme von dieser Regel ist die Kundenzufriedenheitsforschung, bei der gelegentlich mit Skalen wie „sehr unzufrieden – eher unzufrieden – eher zufrieden – sehr zufrieden – begeistert“ gearbeitet wird. Der Grund ist, dass der negative Bereich oft nur schwach besetzt ist, weil unzufriedene Kunden eben in der Regel nicht über lange Zeit Kunden sind.

Frage an die Leserin/den Leser (nach Vomberg & Klarmann, 2022, S. 101): Sollte man bei einer Frage nach der Zustimmung von Auskunftspersonen zur Wirtschaftsordnung der Bundesrepublik Deutschland die Formulierung „Marktwirtschaft“ oder „Kapitalismus“ verwenden?

Im Zusammenhang mit der sozialen Erwünschtheit von Antworten ist die Alternative **offene** oder **geschlossene Frage** schon angesprochen worden. Bei offenen Fragen ist die Art und Formulierung der Antwort voll ins Belieben der Auskunftsperson gestellt; bei geschlossenen Fragen sind Antwortkategorien vorgegeben, aus denen eine ausgewählt werden soll. Ausführlichere Diskussionen zur Anwendung von offenen oder geschlossenen Fragen finden sich u. a. bei Bradburn et al. (2004, S. 151 ff.) und Züll, 2015). Geschlossene Fragen haben bestimmte Vor- und Nachteile, die umgekehrt für offene Fragen gelten:

Vorteile geschlossener Fragen

- Einfache Beantwortung (Erleichterung für Auskunftspersonen wegen geringerer kognitiver Anforderungen (Züll, 2015) und dadurch höhere Antwortquote).

- Wenig Probleme bei der Verarbeitung der Angaben (→ Codierung, siehe Abschn. 4.5.3).
- Ermunterung zu Antworten, auf die die Auskunftsperson ohne Vorgaben nicht gekommen wäre.

Nachteile geschlossener Fragen

- Auskunftspersonen können aus der Frage und den Antwortkategorien entnehmen, welche Antworten im üblichen Bereich liegen (Beispiel im Abschn. 4.3.1.1: Antwortkategorie „300 Dosen Bier und mehr“ zeigt, dass diese Antwort für extremen Alkoholkonsum steht).
- Oberflächliches Antwortverhalten durch schnelles, unbedachtes Ankreuzen von Kategorien wird erleichtert.
- Originelles Antwortverhalten (→ Antworten, an die der Gestalter des Fragebogens vorher nicht gedacht hat) wird erschwert.
- Geschlossene Fragen sind nicht geeignet, wenn ein Fragebogen über längere Zeit genutzt wird und sich die Kategorien ändern können, z. B. wenn nach den wichtigsten Problemen gefragt werden soll, die in Deutschland zu lösen sind. Hier sind im Laufe der Zeit Probleme aktuell geworden (z. B. Klimawandel, Energieversorgung), andere haben an Dringlichkeit verloren (z. B. Jugendarbeitslosigkeit).

Da zum Wesen geschlossener Fragen die Vorgabe von zu wählenden **Antwortkategorien** gehört, muss man natürlich auch Überlegungen zu deren Anlage anstellen. Wichtig ist die **Vollständigkeit der Kategorien,** damit sich jede Auskunftsperson irgendwo einordnen kann. Deswegen findet man typischerweise nach unten oder oben offene Kategorien (z. B. „bis 3-mal pro Woche“ oder „80 Jahre und älter“) bzw. eine Kategorie „Sonstiges“ (z. B. „Andere Marke“ oder „Sonstige Gründe“).

Um eine eindeutige Antwortmöglichkeit zu geben, gehört die **Ausschließlichkeit der Kategorien** zum Standard, d. h. dass jede Auskunftsperson sich nur einer Kategorie zuordnen kann (z. B. „Bis 1000 €“, „1001 bis 2000 €“, „2001 bis 3000 €“, „über 3000 €“). In vielen Fällen werden Antwortkategorien wie „weiß nicht“ oder „keine Angabe“ vorgesehen, damit Personen ohne entsprechende Meinung, Kenntnisse etc. sich entsprechend einordnen können. Teilweise wird darauf aber bewusst verzichtet, um den Auskunftspersonen diesen bequemen Ausweg nicht anzubieten und sie zu drängen, „Farbe zu bekennen“.

Ein letzter hier zu behandelnder Aspekt hat nur mittelbar mit der Frageformulierung zu tun, sondern eher mit den Konsequenzen, die man für die **Ergebnisinterpretation** aus der Fehlerempfindlichkeit von Befragungen zieht. Wie schon gezeigt wurde (Abschn. 4.3.1) können geringe Unterschiede bei Frageformulierungen zu deutlichen Ergebnisunterschieden führen. Deswegen soll die Fragebogenentwicklung mit besonderer Sorgfalt unter Beachtung vorliegender Erfahrungen (siehe z. B. Sudman & Blair, 1998; Bradburn et al., 2004) und mit Anwendung mehrerer Pretests (Vorab-Überprüfung von Fragebogenentwürfen bei kleinen Stichproben; siehe Abschn. 4.3.3 erfolgen. Ein

systematischer Weg zur Entwicklung valider und reliabler Messinstrumente bei Befragungen wird im Abschn. 4.3.2 im Zusammenhang mit der Entwicklung von Multi-Item-Skalen aufgezeigt. Dennoch verbleibt oft Unsicherheit hinsichtlich der Aussagekraft von Befragungsergebnissen. Wenn z. B. 38 % der Befragten gesagt haben, dass sie ein bestimmtes neues Produkt kaufen würden, wie viele Konsumenten kaufen es dann wirklich? Wenn 64 % der Kunden eines Unternehmens angeben, dass sie mit den Leistungen des Unternehmens zufrieden sind, was sagt das aus? Sind 64 % viel oder wenig? Wegen derartiger Probleme spielen bei der Interpretation von Befragungsergebnissen häufig *Vergleiche im Zeitablauf* oder *Vergleiche zwischen unterschiedlichen Gruppen von Befragten* eine wichtige Rolle (siehe folgende Beispiele).

Beispiele zur Ergebnisinterpretation

- Wenn 38 % der Konsumenten angeben, ein neues Produkt kaufen zu wollen und der Vergleichswert aus einer früher durchgeführten entsprechenden Untersuchung für ein inzwischen erfolgreich eingeführtes Produkt nur bei 30 % lag, dann spricht vieles dafür, dass die jetzt anstehende Produkteinführung erfolgreich wird.
- Wenn 64 % der eigenen Kunden zufrieden sind, dieser Wert bei Konkurrenzunternehmen aber zwischen 70 und 80 % liegt, dann hat man offenkundig ein Defizit.
- Wenn die eigene Rate zufriedener Kunden bei 64 % liegt, dieser Wert aber ein Jahr zuvor nur unter 50 % lag, dann haben inzwischen ergriffene Maßnahmen zur Steigerung der Kundenzufriedenheit offenbar einigen Erfolg gehabt. ◀

Die Grundidee, die hinter derartigen Vergleichen steht, besteht darin, dass systematische Verzerrungen der Ergebnisse von Befragungen durch Mängel bei der Frageformulierung sich gewissermaßen beim Vergleich (teilweise) neutralisieren. Wenn also z. B. die Zahl der tatsächlichen Käufer von neuen Produkten bei einer bestimmten Befragungstechnik eher überschätzt wird, dann müsste dieser Effekt bei allen Anwendungen dieser Befragungstechnik auftreten und man könnte zumindest aus dem Vergleich Schlüsse ziehen. Das setzt allerdings voraus, dass durchgehend eine identische Erhebungsmethode verwendet wird. Anderenfalls wäre nicht klar, ob Ergebnisunterschiede auf die veränderte Methode oder auf in der Realität gegebene Unterschiede zurückzuführen sind. Weiterhin wird unterstellt, dass Ergebnisverzerrungen in der einen oder anderen Richtung bei allen befragten Teilgruppen und im Zeitablauf etwa gleichmäßig wirken.

4.3.1.4 Key Informant Problem und Common Method Bias

In Verbindung mit wesentlichen Teilgebieten der Marketingforschung der letzten Jahrzehnte haben zwei besondere methodische Probleme entsprechende Beachtung gefunden, nämlich das Key Informant Problem und der Common Method Bias. Im Hinblick auf Untersuchungen zum Kaufverhalten im Business-to-Business-Bereich (siehe z. B. Weiber & Kleinaltenkamp, 2013) stellte sich immer das Problem, dass typischerweise mehrere Personen in einem Unternehmen an Beschaffungsprozessen beteiligt sind (→ Buying Center). Daraus ergaben sich die Probleme, *welche Person(en)* im Unter-

nehmen zu befragen sind, um Aufschluss über diese Prozesse zu erhalten und *welche Aussagekraft* die Angaben solcher „*Key Informants*" haben.

Im Marketingbereich findet man häufig Untersuchungen zum Einfluss bestimmter Einflussfaktoren (z. B. Werbung oder Produktqualität) auf gewisse Erfolgsgrößen (z. B. Markenbindung oder Kundenzufriedenheit). Ein Kritikpunkt an diesen Untersuchungen besteht darin, dass oft sowohl die Daten für die unabhängigen als auch die abhängigen Variablen aus denselben Quellen (z. B. von denselben Auskunftspersonen) stammen, dass also ein **„Common Method Bias"** aufgetreten sein könnte. „Von einem solchen Bias spricht man, wenn die beobachteten Zusammenhänge zwischen unabhängigen und abhängigen Variablen nicht allein auf tatsächliche Zusammenhänge zurückzuführen sind, sondern auch darauf, dass für beide Variablen dieselbe Informationsquelle verwendet wurde." (Homburg, 2007, S. 45).

Als **„Key Informants"** bezeichnet man Auskunftspersonen, die weniger über sich selbst Auskunft geben, sondern eher über die Organisation (z. B. das Unternehmen), der sie angehören, beispielsweise über Umsätze, Struktur, Abläufe oder Entscheidungsprozesse. Hurrle und Kieser (2005) äußern sich auf Basis einer Übersicht über einschlägige empirische Untersuchungen äußerst skeptisch über die Fähigkeit von Key Informants, valide Angaben zu relativ komplexen und/oder abstrakten Konstrukten (z. B. Unternehmenskultur, Ressourcen) zu machen (Key Informant Bias). Selbst hinsichtlich der Fähigkeit zu korrekten Angaben über relativ einfach erscheinende Merkmale (z. B. Umsätze oder bestimmte Qualitätsmerkmale) gibt es Zweifel und einige Studien zeigten einen erheblichen Anteil – teilweise grober – Fehler bei solchen Schätzungen. Damit stellt sich natürlich die Frage, unter welchen Bedingungen die Angaben von Key Informants hinreichend reliabel und valide sind. Homburg et al. (2012a) haben in einer groß angelegten empirischen Untersuchung u. a. folgende – gewissermaßen günstige – Bedingungen für die Aussagekraft entsprechender Daten ermittelt:

- Genauigkeit der Angaben von Key Informants ist eher gegeben, wenn sich diese auf die Gegenwart (also nicht auf länger zurückliegende oder zukünftige Vorgänge) beziehen, wenn diese objektivierbaren Tatbeständen (z. B. eher die Beschäftigtenzahl als Angaben zur Unternehmenskultur) gelten und herausgehobene Entscheidungen und Ereignisse an Stelle von Routine-Vorgängen betrachtet werden.
- Angaben von Personen mit großem Erfahrungsschatz, also Personen mit hoher hierarchischer Stellung und langer Beschäftigungsdauer, haben tendenziell höhere Reliabilität.

Damit deuten sich schon Gründe für die in der Literatur (z. B. Hurrle & Kieser, 2005) kritisierte eingeschränkte Aussagekraft solcher Angaben an. Homburg et al. (2012b) heben in diesem Zusammenhang u. a. die folgenden Gesichtspunkte hervor:

- *Eigene Interessen und funktionsspezifische Sichtweisen* der Key Informants. Manche Führungskräfte neigen wohl zur positiven Selbstdarstellung oder zur besonderen Hervorhebung von Erfolgen in ihrem eigenen Verantwortungsbereich.
- *Begrenzte Kompetenz* der Key Informants. Haben sie tatsächlich die notwendigen Informationen und Erfahrungen?
- Messung *abstrakter Konstrukte,* die den Key Informants nicht vertraut sind. Wenn man beispielsweise die Innovations- oder Marktorientierung eines Unternehmens durch Befragung von Key Informants messen will, dann sind das Variable, die vielfach aus der Sicht dieser Auskunftspersonen nicht eindeutig und klar definiert sind und mit denen sie sich in ihrer laufenden Arbeit kaum beschäftigen.

Hintergrundinformation

Vomberg und Klarmann (2022, S. 80) schätzen die angemessene *Auswahl von Key Informants* als wichtigsten Aspekt für die Reduzierung eines Key Informant Bias ein:

> „Das wichtigste und effektivste Hilfsmittel zur Verringerung von Problemen des Key Informant Bias ist die sorgfältige Auswahl der Key Informants. Der/die ForscherIn muss die Auswahl der kontaktierten Key Informants sorgfältig auf das Untersuchungsziel ausrichten. Typischerweise werden Key Informants nicht zufällig ausgewählt und nicht als repräsentativ für eine Grundgesamtheit angesehen. Eher wählen die Forschenden Auskunftspersonen aus, die über spezielle Qualifikationen verfügen, wie z.B. ihre Position im Unternehmen oder ihr Wissen."

Der sogenannte **„Common Method Bias"** (siehe dazu Podsakoff et al., 2003; Temme et al., 2009; Vomberg & Klarmann, 2022) bezieht sich also darauf, dass bei der Messung von unabhängigen und abhängigen Variablen bei derselben Informationsquelle mit der Abstimmung der verschiedenen Angaben im Hinblick auf Konsistenz zu rechnen ist. Beispielsweise könnte ein Produktmanager, der nach den Kosten und dem Erfolg bestimmter Marketing-Aktivitäten befragt wird, dazu neigen, hier konsistent wirkende Angaben zu machen. Die Analyse entsprechender Daten würde dann zu einer Überschätzung des Zusammenhanges von unabhängigen und abhängigen Variablen führen. Für eine ausführliche Darstellung möglicher Gründe für einen Common Method Bias sei auf die drei o. g. Artikel verwiesen.

Beispiel

Podsakoff et al. (2003, S. 879) erläutern das Problem des Common Method Bias mit einem kleinen Beispiel:

> „Beispielsweise sei angenommen, dass ein Forscher interessiert ist, eine vermutete Beziehung zwischen den Konstrukten A und B zu untersuchen. Auf der Grundlage theoretischer Überlegungen erwartet man, dass die Messungen von Konstrukt A und die Messungen von Konstrukt B korreliert sind. Wenn die Messungen von Konstrukt A und die Messungen von Konstrukt B Gemeinsamkeiten bei den Messmethoden haben, können diese Methoden einen systematischen Einfluss auf die beobachtete Korrelation zwischen den Mes-

sungen ausüben. Also stellt ein Common Method Bias zumindest teilweise eine alternative Erklärungsmöglichkeit für die beobachtete Korrelation zwischen den Messungen dar." ◀

Vomberg und Klarmann (2022, S. 77 f.) geben folgende *Empfehlungen* für die Vermeidung bzw. Einschränkung eines Common Method Bias:

- *Verwendung unterschiedlicher Informationsquellen;* z. B. Befragung *unterschiedlicher* Manager hinsichtlich des eingesetzten Werbebudgets bzw. der entsprechenden (daraus resultierenden ?) Entwicklung des Marktanteils
- *Zeitliche Trennung des Erhebungszeitpunkts verschiedener Variabler;* die gedankliche Abstimmung von Angaben einer Auskunftsperson wird eingeschränkt, wenn die Erinnerung an früher gemachte Angaben reduziert wird.
- *Erhebung von abhängigen und unabhängigen Variablen an verschiedenen Stellen im Interview;* z. B. zuerst die abhängigen Variablen, (deutlich) später die unabhängigen.
- *Verwendung unterschiedlicher Arten von Antwortkategorien bei verschiedenen Fragen;* z. B. eine 7er Skala bei einer unabhängigen Variablen und eine offene Frage bei der zugehörigen abhängigen Variablen. Dadurch wird die einfache Übernahme einer Antwort erschwert.

Wenn man bei der Datenerhebung mit einem Key Informant Problem oder einem Common Method Bias rechnen muss, sind natürlich besondere Bemühungen zur Validierung erforderlich. Dabei steht – analog zur Grundidee der im Abschn. 3.1 gekennzeichneten **Triangulation** – die Verwendung von Angaben mehrerer Informanten im Mittelpunkt. Im Hinblick auf entsprechende Ansätze und Methoden sei insbesondere auf Homburg und Klarmann (2009), Homburg et al. (2012b) und Homburg et al. (2012a) verwiesen.

4.3.2 Entwicklung von Multi-Item-Skalen

4.3.2.1 Einführung: Single- versus Multi-Item-Skalen

Im vorigen Abschnitt sind im Zusammenhang mit Befragungsverfahren die Entwicklung von Messinstrumenten und die dabei zu beachtenden Grundsätze schon angesprochen worden. Im vorliegenden Abschnitt soll dieses Problem etwas systematischer behandelt werden und die Darstellung der Entwicklung von Multi-Item-Skalen, also einer speziellen (aber besonders wichtigen!) Befragungstechnik, erfolgen.

In Abschn. 2.3 wurde die *Reliabilität* als (notwendige, nicht hinreichende) Voraussetzung der Validität gekennzeichnet. Selbst eine (scheinbar) valide Messung, die mit Zufallsfehlern behaftet ist, würde einem „wahren" Wert nicht entsprechen. Andererseits ist eine verlässliche Messung mit geringer *Validität* wohl ebenso nutzlos. Bei der Entwicklung von Messinstrumenten kommt es also darauf an, diese beiden Fehlerarten zu minimieren. Wenn man sicherstellen kann, dass keinerlei systematische oder zufällige Fehler ein Untersuchungsergebnis maßgeblich beeinflussen, dann hat dieses Ergebnis

offenbar Aussagekraft für die interessierenden Phänomene der Realität. Deshalb sei hier an die Kennzeichnung der Validierung als *Ausschluss alternativer Erklärungsmöglichkeiten* für ein Untersuchungsergebnis erinnert.

Wenn man Reliabilität und Validität als zentrale Anforderungen an Messinstrumente charakterisiert, dann stellt sich die Frage, wie geprüft werden kann, ob ein Messinstrument diesen Anforderungen genügt. Eine solche Prüfungsmöglichkeit hat natürlich zentrale Bedeutung für die Entwicklung von Erhebungsmethoden. Die Gegenüberstellung von Untersuchungsergebnissen und „wahren Werten" zur Prüfung der **Validität** einer Messung scheidet im Regelfall aus, da ja der sogenannte „wahre Wert" nicht bekannt ist und erst durch die Untersuchung geschätzt werden soll. Auch der die **Reliabilität** kennzeichnende Aspekt der Unabhängigkeit der Ergebnisse von zufälligen Einflüssen beim einzelnen Messvorgang lässt sich in der Forschungspraxis nicht leicht umsetzen. Eine auf diesem Ansatz basierende Prüfung der Reliabilität einer Messung müsste darauf hinauslaufen, dass der gleiche Messvorgang zu verschiedenen Zeitpunkten zum gleichen (zumindest sehr ähnlichen) Ergebnis führt (→ Test-Retest-Reliabilität).

Neben die Schwierigkeiten, die Datenerhebung für eine Untersuchung – zumindest für Teile davon – mehrfach durchführen zu müssen, tritt das Problem, dass man bei dieser Art der *Reliabilitätsüberprüfung* die Konstanz der zu messenden Phänomene im Zeitablauf unterstellen muss. Für die praktische Anwendung in der empirischen Marketingforschung werden wegen der genannten Probleme andere Hilfsmittel zur Validitäts- und Reliabilitätsprüfung von Messinstrumenten empfohlen (siehe folgende Abschnitte).

Der **Entwicklungsprozess von Messinstrumenten** umfasst deren Entwurf sowie deren Korrektur und Verfeinerung auf der Basis der Ergebnisse entsprechender Prüfungen. Die wesentlichen Schritte sollen im Folgenden skizziert werden. Die Vorgehensweise beruht hauptsächlich auf einem Vorschlag von Churchill (1979) und den weiterführenden Ausführungen von Netemeyer et al. (2003). Wegen des typischerweise hohen Entwicklungsaufwandes solcher Messinstrumente ist man in der Marketingforschung – wie schon deutlich früher in der Psychologie – dazu übergegangen, für häufig relevante Konzepte (z. B. Einstellung oder Involvement) standardisierte Messinstrumente zu entwickeln. Ähnlich wie ein Psychologe, der zur Messung der Intelligenz eines Probanden ein entsprechendes Standard-Messverfahren verwendet, kann der Marktforscher bestimmte – bewährte und geprüfte – Fragetechniken oder eben Multi-Item-Skalen bei entsprechenden Untersuchungsgegenständen immer wieder einsetzen. Eine häufig genutzte umfassende Sammlung solcher Messinstrumente ist das von Bearden et al. (2011) herausgegebene „Handbook of Marketing Scales"; einen konzentrierten Überblick zu 20 besonders gängigen Marketing-Skalen (z. B. Einstellungen zu Werbung oder Produkten, Preis-Empfindlichkeit oder Involvement) gibt Bruner (2013).

Die Entwicklung und Verwendung standardisierter Messinstrumente bringt wesentliche Vorteile (siehe Nunnally & Bernstein, 1994):

- Größere Objektivität der Ergebnisse, weil diese nicht von individuell bestimmten Messverfahren abhängen

- Wirtschaftlichkeit, weil der aufwendige Prozess der Entwicklung und Prüfung einer Skala nur einmal durchgeführt werden muss
- Bessere Kommunikation von Ergebnissen durch Bezugnahme auf in der Fachwelt bekannte und akzeptierte Messverfahren
- Bessere Möglichkeiten zur Durchführung von Replikationsstudien (siehe Abschn. 2.2.3)

Hier ist die Einschränkung anzubringen, dass die während des Entwicklungsprozesses einzusetzenden Hilfsmittel vor allem auf sogenannte **Multi-Item-Skalen** sinnvoll angewandt werden können. Man versteht hierunter Erhebungstechniken, bei denen der gesuchte Messwert nicht nur auf einer einzelnen Angabe einer Auskunftsperson beruht, sondern durch die Zusammenfügung der Angaben bezüglich einer größeren Zahl von Fragen („Items“) zustande kommt. Dazu werden in den meisten Fällen sogenannte „Ratingskalen“ (siehe Abschn. 7.2) verwendet, bei denen die Antwortmöglichkeiten zu den verschiedenen Items numerisch abgestuft (z. B. 1 bis 7 oder −2 bis +2) sind, womit eine akzeptable Annäherung an das Messniveau der Intervallskalierung erreicht werden soll. Solche Multi-Item-Skalen haben seit der Publikation der einflussreichen Artikel von Jacoby (1978) und Churchill (1979) zumindest in der wissenschaftlichen Marketingforschung eine dominierende Stellung. Der Entwicklungsprozess von Skalen in der im Folgenden dargestellten Form ist allerdings nur für *reflektive Messungen* (siehe Abschn. 4.3.2.3) relevant, die auch besonders häufig angewandt werden. Die Beschränkung auf Multi-Item-Skalen ist nicht allzu gravierend, da diese generell zur Messung komplexerer Phänomene empfohlen werden.

Hintergrundinformation

Hulland et al. (2018) kommen auf Basis einer entsprechenden empirischen Untersuchung zu einer Einschätzung der *Relevanz von Multi-Item-Skalen* (zumindest) für die wissenschaftliche Marketingforschung:

Die Autoren haben in drei weltweit führenden Marketing-Zeitschriften (Journal of Marketing, Journal of Marketing Research, Journal of the Academy of Marketing Science) für den Zeitraum von 2006 bis 2015 insgesamt 1561 publizierte wissenschaftliche Artikel ermittelt, von denen 522 (33,4 %) empirische Untersuchungen unter Verwendung von *Befragungen* waren. Von diesen 522 Artikeln haben 97,5 % (!) Multi-Item-Skalen verwendet. Daneben zeigte sich, dass über 90 % der für die jeweilige Untersuchung zentralen Variablen mit Multi-Item-Skalen gemessen worden waren.

Als Gründe für die Bevorzugung von Multi-Item-Skalen sind vor allem zu nennen (Nunnally & Bernstein, 1994, S. 66 f.):

- Mehrere Items sind eher als ein einzelnes geeignet, den verschiedenen Facetten eines zu messenden Konzepts (z. B. Einstellung zu einer Marke) gerecht zu werden (Baumgartner & Homburg, 1996). Man geht also von einem höheren Informationsgehalt von Messungen auf Basis von Multi-Item-Skalen aus.

- Multi-Item-Skalen ergeben feiner differenzierte Messwerte als Single-Item-Skalen, die meist auch besser an ein höheres Messniveau (→ Intervallskala, siehe Abschn. 7.2) angenähert sind.
- Wegen der geringeren Abhängigkeit des ermittelten Messwerts von der Reaktion auf ein *einzelnes* Item ist die Reliabilität von Multi-Item-Skalen tendenziell höher als die von Single-Item-Skalen.

Beispiel

Hier ein einfaches Beispiel für eine Multi-Item-Skala. Die Einschätzung einer Autowerkstatt (positiv – negativ) seitens der Kunden soll gemessen werden. Dazu werden insgesamt fünf Fragen formuliert:

„Das Personal der Werkstatt ist stets so freundlich und zuvorkommend, wie ich es erwarte."

(1)	(2)	(3)	(4)	(5)
Stimme voll zu	Stimme teilweise zu	Teils/teils	Stimme eher nicht zu	Stimme überhaupt nicht zu

„Die Ausstattung der Werkstatt entspricht meinen Erwartungen."

(1)	(2)	(3)	(4)	(5)
Stimme voll zu	Stimme teilweise zu	Teils/teils	Stimme eher nicht zu	Stimme überhaupt nicht zu

„Termine werden pünktlich eingehalten."

(1)	(2)	(3)	(4)	(5)
Stimme voll zu	Stimme teilweise zu	Teils/teils	Stimme eher nicht zu	Stimme überhaupt nicht zu

„Das Personal wirkt kompetent."

(1)	(2)	(3)	(4)	(5)
Stimme voll zu	Stimme teilweise zu	Teils/teils	Stimme eher nicht zu	Stimme überhaupt nicht zu

„Das Personal wirkt vertrauenswürdig."

(1)	(2)	(3)	(4)	(5)
Stimme voll zu	Stimme teilweise zu	Teils/teils	Stimme eher nicht zu	Stimme überhaupt nicht zu

Wie verläuft hier der Messvorgang? Die Auskunftsperson gibt ihren Zustimmungsgrad zu den fünf Einzelfragen an. Je nach Angabe wird der einzelnen Frage ein Zahlenwert (1: sehr positiv bis 5: sehr negativ) zugeordnet. Wenn man diese Zahlen-

werte einfach aufaddiert, erhält man einen Gesamtwert (zwischen 5 und 25), der als Indikator für die Einschätzung der jeweiligen Auskunftsperson hinsichtlich der Autowerkstatt interpretiert wird. Ein niedriger Zahlenwert steht hierbei für eine eher positive Einschätzung, ein hoher Zahlenwert für eine eher negative Einschätzung. Hier ist die Gesamt-Einschätzung also nicht durch eine Frage, sondern durch die *Zusammenfassung* von mehreren einzelnen Angaben („Multi-Items") ermittelt worden. Es sei hinzugefügt, dass das verwendete (sehr schlichte) Beispiel natürlich den noch zu erörternden Anforderungen an solche Messungen nicht voll entspricht, sondern nur dazu dient, einen ersten Eindruck von Multi-Item-Skalen zu vermitteln. ◀

Allerdings wird die *generelle* Überlegenheit von Multi-Item-Skalen auch wieder in Frage gestellt. Naturgemäß erfordern Multi-Item-Skalen, deren Beantwortung auch recht monoton sein kann, mehr Geduld und Anstrengung beim Interview als Single-Item-Skalen, was wiederum zu höheren Abbruchraten und zu höheren Kosten der Datenerhebung führt (Fuchs & Diamantopoulos, 2009). Nicht zuletzt deswegen ist das Interesse an den Möglichkeiten des Einsatzes von **Single-Item-Skalen** wieder gewachsen. Fuchs und Diamantopoulos (2009) haben diesen Aspekt umfassend diskutiert und kommen zu dem Ergebnis, dass der Einsatz von Single-Item-Skalen vor allem unter den folgenden Bedingungen in Frage kommt:

- *Messung „konkreter" Konzepte (bzw. Konstrukte)*. Die Unterscheidung von konkreten und abstrakten Konzepten geht auf Rossiter (2002) und Bergkvist und Rossiter (2007) zurück. Konkrete Konzepte werden von Auskunftspersonen einheitlich und eindeutig verstanden, z. B. „Alter" oder „Kaufhäufigkeit". Dagegen ist es bei abstrakten Konzepten, die weniger einheitlich verstanden werden (z. B. „Unternehmenskultur" oder „Brand Coolness", siehe Abschn. 4.3.2.4), zweckmäßig, den Bedeutungsgehalt des Konzepts durch eine größere Zahl von Items zu kennzeichnen und zu verdeutlichen. Bei mehrdimensionalen Konstrukten ist natürlich in der Regel eine Single-Item-Skala in der Regel nicht angemessen, weil sie der Mehrdimensionalität nicht gerecht werden kann.
- *Redundanz der Items bei vorhandenen Multi-Item-Skalen*. Gelegentlich findet man Messungen, bei denen die verschiedenen Items formal und inhaltlich sehr ähnlich sind. Das führt zwar zu (scheinbar) günstigen Ergebnissen bei der Reliabilitätsprüfung (siehe Abschn. 4.3.2.5), aber eben auch zu hoher Redundanz der verschiedenen Einzel-Messungen. Dem entsprechend empfehlen Diamantopoulos et al. (2012) in ihrer umfassenden Untersuchung bei hochgradig homogenen und semantisch redundanten Items eine Single-Item-Skala als Alternative.
- *Die jeweiligen Forschungsziele erfordern keine Multi-Item-Skalen*. Wenn keine detaillierten Analysen zu dem hauptsächlich interessierenden Konzept vorgesehen sind oder wenn das Konzept nicht den Kernbereich der Untersuchung betrifft, sind oft Single-Item-Skalen ausreichend (Hulland et al., 2018).

- *Anwendung bei heterogenen und/oder sehr großen Stichproben*. Wenn man z. B. eine (etwas pauschale) Single-Item-Skala für die Messung von „Kundenzufriedenheit“ verwendet, so ist diese breiter einsetzbar als eine Multi-Item-Skala, deren Items sich auf Spezifika bestimmter Produkte oder Kundengruppen beziehen. Daneben hat bei sehr großen Stichproben das eingangs genannte Argument des geringeren Untersuchungsaufwandes bei Single-Item-Skalen besonderes Gewicht.

Andererseits haben Sarstedt und Wilczynski (2009) in ihrer Studie eine Überlegenheit von Multi-Item-Skalen im Hinblick auf Reliabilität und Aspekte der Validität festgestellt. Auch Diamantopoulos et al. (2012) kommen zu dem Ergebnis, dass in vielen Fällen Single-Item-Skalen schlechtere Messeigenschaften haben als Multi-Item-Skalen und dass erstere nur unter bestimmten Bedingungen verwendet werden sollten. Die Einschätzungen sind also etwas uneinheitlich, jedenfalls nicht mehr ganz einheitlich im Hinblick auf eine generelle Bevorzugung von Multi-Item-Skalen.

Beispiel

Eine zentrale Idee bei der Anwendung von *Multi-Item-Skalen* wird vielleicht durch ein Beispiel aus dem Hochschulbereich plausibel. Üblicherweise werden in Studiengängen der Wissensstand, das Verständnis und Anwendungsfähigkeit der Studierenden bezüglich des Inhalts einer Lehrveranstaltung u. a. durch Klausuren gemessen. Was wäre davon zu halten, wenn eine Klausur zu einem Kurs „Marktforschung“ nur aus einer einzigen Frage (mit 10 min Bearbeitungsdauer), z. B. „Welche Relevanz hat die Generalisierbarkeit von Ergebnissen für explorative Untersuchungen?“, bestünde? Würden Sie das Ergebnis einer solchen Klausur akzeptieren? Wären damit Ihr Leistungsstand zu Wissen, Verständnis und Anwendungsfähigkeit hinsichtlich der Marktforschung (und die entsprechende Zensur) gut gemessen und begründet? Würde der Zufall beim Ergebnis eine zu große Rolle spielen? Man würde wohl eine größere Zahl von Fragen, die unterschiedliche Fähigkeiten und unterschiedliche Themen betreffen, für deutlich angemessener halten. Analog dazu kann man sich die Grundidee einer Multi-Item-Skala vorstellen. ◀

Im folgenden Abschnitt sollen zunächst Multi-Item-Skalen und insbesondere deren gängigste Form, die Likert-Skalen, genauer gekennzeichnet werden. Dabei erfolgt aus Gründen der Klarheit und Einfachheit eine Beschränkung auf *eindimensionale Skalen,* d. h. auf Skalen, die durch ein einziges Kontinuum (niedrig bis hoch) gekennzeichnet sind. Es folgen in den weiteren Abschnitten Erläuterungen zu den verschiedenen Schritten der Skalenentwicklung einschließlich der Überprüfung von Reliabilität und Validität dieser zu entwickelnden Skalen.

4.3.2.2 Arten von Multi-Item-Skalen

Das zentrale Kennzeichen von Multi-Item-Skalen besteht – wie schon erwähnt – darin, dass sich der Messwert für ein Konzept (bzw. Konstrukt) durch die Zusammenfassung

der Angaben einer Auskunftsperson zu einer gewissen Zahl von Einzelfragen (Items) ergibt. In der Marktforschung häufig genannte und angewandte Formen von Multi-Item-Skalen sind Likert-Skalen und semantische Differenziale, auf die im Folgenden eingegangen wird. Die in früheren Zeiten stärker beachteten Thurstone-Skalen (siehe dazu z. B. Hoyle et al., 2002, S. 167 f.) werden wegen ihrer besonders aufwendigen Entwicklungsprozedur heute kaum noch verwendet und deshalb hier nicht näher betrachtet.

Besonders stark verbreitet in Wissenschaft und Praxis sind Likert-Skalen (benannt nach dem bedeutenden Sozialforscher Rensis Likert). Deren häufige Anwendung liegt wohl vor allem daran, dass Likert-Skalen auf der einen Seite sehr *vielfältige Anwendungsbereiche* bieten und auf der anderen Seite nur einen relativ (!) *begrenzten Untersuchungsaufwand* erfordern. Deswegen sei diese Art von Multi-Item-Skalen als erste dargestellt. Die **Kennzeichen** einer **Likert-Skala** lassen sich folgendermaßen charakterisieren:

- Die Auskunftspersonen bekommen einige (häufig ca. 10) Aussagen (z. B. „Die Marke XY garantiert besonders hohe Qualität“) zu einem Untersuchungsgegenstand (z. B. Einstellung zur Marke XY) vorgelegt. Dabei werden meist positive und negative Aussagen gemischt.
- Meist werden fünf abgestufte Antwortmöglichkeiten (Starke Zustimmung, Zustimmung, Unentschieden, Ablehnung, Starke Ablehnung) vorgesehen. Gelegentlich findet man auch drei- oder siebenfache Abstufungen. Selbstverständlich muss die Ausgewogenheit (Zustimmung – Ablehnung) der Antwortmöglichkeiten gesichert sein.
- Den Antwortmöglichkeiten werden Zahlenwerte (bei 5er Skalen −2 bis +2 oder 1 bis 5) zugeordnet. Bei der Mischung von positiven und negativen Aussagen müssen diese Zuordnungen von Zahlenwerten so angelegt sein, dass – je nach „Richtung“ der einzelnen Items – die Zahlenwerte mit den Ausprägungen des zu messenden Konstrukts korrespondieren. Beispielsweise sollten bei „positiven“ Items die unterschiedlichen Zustimmungsgrade von 1 bis 5 codiert werden und bei „negativen“ Items von 5 bis 1.
- Die (numerisch codierten) Antworten zu den einzelnen Items der Skala werden additiv zu einem Messwert für das interessierende Konstrukt zusammengefasst.

Der ganze Prozess der Suche nach geeigneten Items sowie der Beurteilung und Auswahl dieser Items nach geeigneten Kriterien kann recht aufwendig sein. Dieser Prozess ist Gegenstand der folgenden Abschnitte. Allgemein wird davon ausgegangen, dass Likert-Skalen hinreichend gut den Anforderungen einer **Intervallskalierung** (siehe Abschn. 7.2) entsprechen (Hoyle et al., 2002, S. 176 f.). Dafür ist aber Voraussetzung, dass die Datenerhebung durch numerische Angaben (→ Ratingskala) bei den verschiedenen Antwortmöglichkeiten und graphische Hilfsmittel dieser Anforderung (→ Interpretierbarkeit der Abstände zwischen den Messwerten) entspricht (siehe Abschn. 7.2).

Die Codierung und Zusammenfassung der einzelnen Werte zu einem (Gesamt-) Messwert für die Einstellung erfolgt in der schon beschriebenen Weise.

Beispiel

Hier ein (hypothetisches) Beispiel zur Messung der Einstellung zu einem Einzelhandelsgeschäft mit Hilfe einer Likert-Skala mit den folgenden Items:

„Das Geschäft hat eine einladende Atmosphäre."
„Das Verkaufspersonal ist kompetent."
„Der Service ist zu langsam."
„Das Geschäft bietet eine große Auswahl an."
„Die Einrichtung wirkt eher altmodisch."
„Die meisten VerkäuferInnen sind eher unfreundlich."

Man beachte, dass positive und negative Aussagen über das Geschäft gemischt sind. Der Grad der jeweiligen Zustimmung zu den Items wird mit Hilfe der folgenden Antwortkategorien gemessen:

„Stimme absolut nicht zu	(1) bzw. (5)"
„Stimme nicht zu	(2) bzw. (4)"
„Unentschieden	(3)
„Stimme zu	(4) bzw. (2)"
„Stimme vollkommen zu	(5) bzw. (1)"

Die Codierung und Zusammenfassung der einzelnen Werte zu einem (Gesamt-) Messwert für die Einstellung erfolgt in der schon beschriebenen Weise. ◀

Semantische Differenziale sind ebenfalls eine in der Marktforschung sehr verbreitete Messmethode, die weitgehend den Merkmalen von Multi-Item-Skalen entspricht, aber nicht immer zur Zusammenfassung der einzelnen Angaben zu einem (Gesamt-) Messwert führt. Die Grundidee besteht darin, im Hinblick auf den Untersuchungsgegenstand (z. B. Image einer Marke, Beurteilung des Verkaufspersonals) eine Reihe von gegensätzlichen Adjektiv-Paaren zu formulieren und auf entsprechenden Ratingskalen die Auskunftsperson angeben zu lassen, welches der jeweils zwei Adjektive deren Meinung am ehesten entspricht.

Über alle Auskunftspersonen (oder Teilgruppen davon) wird für jedes Adjektiv-Paar – unter der manchmal etwas optimistischen Annahme, dass Intervallskalierung (siehe Abschn. 7.2) vorliegt – ein Mittelwert errechnet. Eine zusammenfassende Darstellung der Ergebnisse in einer Form, wie sie in Abb. 4.7 wiedergegeben ist, lässt erkennen, woher der Begriff „semantisches Differenzial" stammt. Ein häufig zitiertes Beispiel zur Verwendung eines semantischen Differenzials, bei dem die einzelnen Angaben zu einem Messwert für ein Konzept zusammengefasst werden, ist die Skala zur Messung von Involvement von Zaichkowsky (1985).

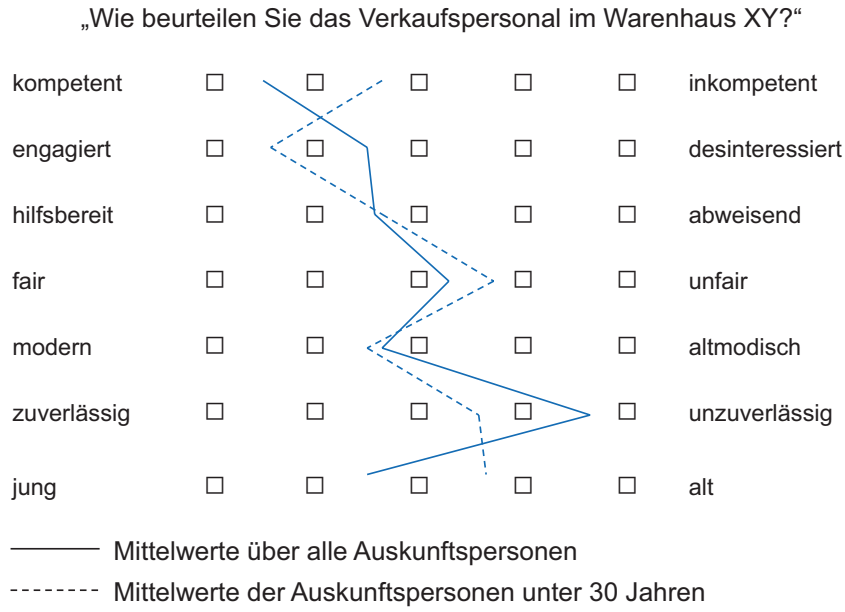


Abb. 4.7 Beispiel eines semantischen Differenzials zur Beurteilung von Verkaufspersonal

Semantische Differenziale sind in der praktischen Marktforschung recht verbreitet, wohl weil sie relativ leicht zu entwickeln und zu interpretieren sind (Iacobucci & Churchill, 2010, S. 242). Man kann auch eine einheitliche Skala (gleiche Adjektiv-Paare) auf unterschiedliche Gegenstände anwenden und diese damit vergleichbar machen.

4.3.2.3 Formative versus reflektive Messungen

In neuerer Zeit hat die Unterscheidung von so genannten formativen und reflektiven Indikatoren große Beachtung gefunden und ist wegen ihrer Bedeutung für die Anwendung von Strukturgleichungsmodellen (siehe Kap. 9) stark diskutiert worden (vgl. z. B. Albers & Hildebrandt, 2006). Der letztgenannte Aspekt würde den Rahmen dieses einführenden Lehrbuchs deutlich sprengen. Hier mag eine Kennzeichnung des Unterschiedes zwischen beiden Arten von Messungen und die Erläuterung daraus resultierender Konsequenzen für die Entwicklung von Messinstrumenten genügen. Die Unterscheidung „formativ versus reflektiv“ bezieht sich auf einen grundlegenden gedanklichen Unterschied zwischen diesen Arten von Konzepten und das Zustandekommen geeigneter Messungen.

Zunächst zu **formativen** Messungen. Der Begriff „formativ“ ist gleichbedeutend mit „gestaltend“. Das ist hier so zu verstehen, dass sich das (gedankliche) Konzept aus mehreren Komponenten zusammensetzt, dass es also gewissermaßen durch diese „gestaltet“ wird. Ein klassisches Beispiel dafür nennen Döring und Bortz (2016, S. 462 f.) mit dem *sozialen Status*. Dieser wird oft als Resultierende aus (Aus-) Bildungsniveau, Berufsprestige und Einkommen/Vermögen angesehen. Das sind also Komponenten des sozia-

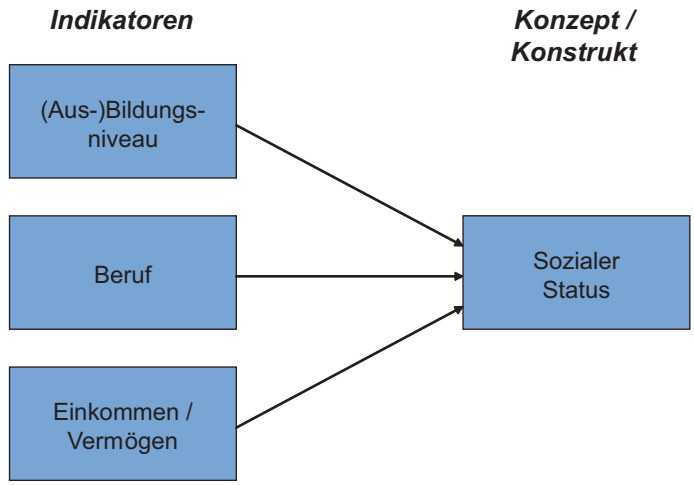


Abb. 4.8 Beispiel einer formativen Messung

len Status. Wenn eine davon sich verändert, dann verändert sich auch der soziale Status. Wenn man eine davon bei der Messung nicht angemessen berücksichtigt, dann wäre das so definierte Konzept „sozialer Status" nicht adäquat (valide) gemessen. In diesem Sinne ist das Konzept das *Ergebnis* der verschiedenen berücksichtigten Indikatoren. Abb. 4.8 illustriert diese Grundidee mit Hilfe des genannten Beispiels. Zu Einzelheiten bezüglich der Entwicklung von formativen Messmethoden sei insbesondere auf Diamantopoulos et al. (2008) verwiesen.

Genau umgekehrt ist die Sichtweise bei **reflektiven Messungen.** Der Begriff „reflektiv" steht hier für „widerspiegeln". Damit ist gemeint, dass sich das Konzept (z. B. die Einstellung zu einem Produkt) auf eine Vielzahl beobachtbarer Indikatoren (beispielsweise eine größere Zahl von Aussagen zu dem Produkt) auswirkt. Entsprechende Messungen werden meist so vorgenommen, dass eine begrenzte Zahl dieser möglichen Indikatoren verwendet wird, die „gute" Eigenschaften im Hinblick auf Reliabilität und Validität haben. Das verweist schon auf die Domain Sampling Theorie, auf die im folgenden Abschnitt kurz eingegangen wird. Diese bezieht sich auf die Grundidee, dass die bei einer Messung verwendeten Items eine *Stichprobe* aus einer *Vielzahl möglicher Items* zur Messung des jeweils interessierenden Konzepts darstellen. Bei *reflektiven* Messungen ist es also möglich und sinnvoll, aus vielen denkbaren Items die geeignetsten auszuwählen. Dagegen würde eine entsprechende Auswahl bei *formativen* Messungen dazu führen, dass relevante Facetten des interessierenden Konzepts unberücksichtigt blieben.

Dieser Ansatz wird in Abb. 4.9 durch ein Beispiel illustriert, das sich auf Konsumentenverwirrtheit (vgl. z. B. Walsh, 2002) bezieht. Ein solches Konzept kann das Antwortverhalten bei einer Vielzahl entsprechender Items, die hier als Indikatoren für Konsumentenverwirrtheit dienen, beeinflussen. In dem Beispiel kann man erkennen,

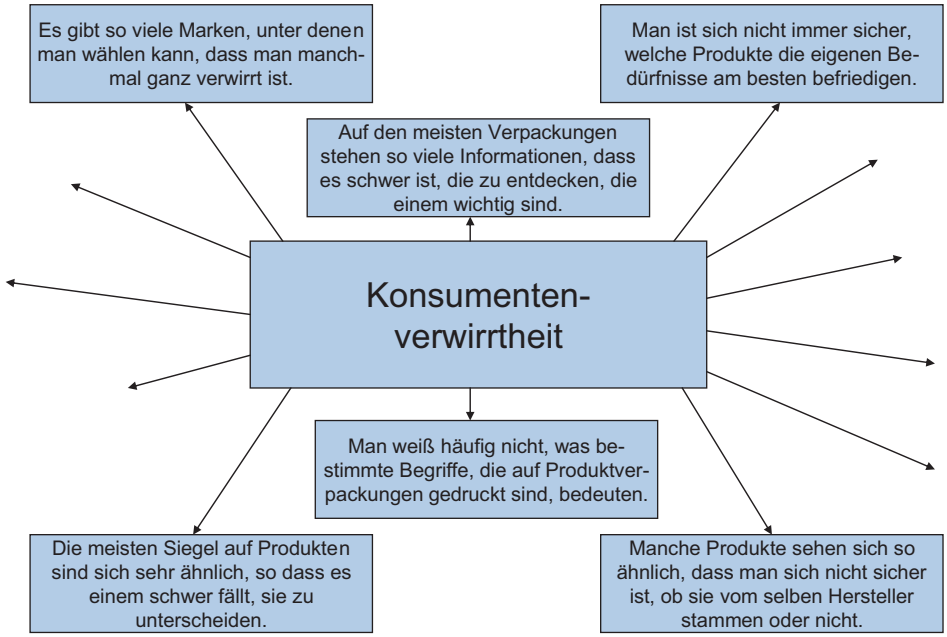


Abb. 4.9 Beispiel einer reflektiven Messung. (Nach Walsh, 2002 und Langer et al., 2008)

dass aus der großen Zahl unterschiedlicher möglicher Items gewissermaßen eine Stichprobe von Items gezogen wurde (in der Abbildung dargestellt), die Gegenstand der entsprechenden Multi-Item-Skala zur Konsumentenverwirrtheit ist. Diese ausgewählten Items wären prinzipiell auch durch andere denkbare Items (in der Abbildung durch „ins Leere" gehende Pfeile angedeutet) ersetzbar.

Hintergrundinformation

Diamantopoulos (2008, S. 1201) kennzeichnet prägnant den Unterschied von reflektiven und formativen Messungen:

> „Die konventionellen Messmethoden in der Marketing- und BWL-Forschung basieren auf reflektiven Messungen, wobei angenommen wird, dass die beobachteten Messwerte (d. h. Indikatoren) die Unterschiede bei den latenten Konstrukten reflektieren. Deswegen wird angenommen, dass die kausale Wirkung vom Konstrukt auf die Indikatoren verläuft und deswegen bei Veränderungen des Konstrukts Wirkungen auf die Indikatoren erwartet werden, die zu der Multi-Item-Skala gehören."

> „Ein alternativer Mess-Ansatz, der neuerdings wachsende Aufmerksamkeit gefunden hat, verwendet formative Indikatoren. Bei diesem Ansatz wird angenommen, dass Veränderungen bei den Indikatoren (in der anderen Richtung) Änderungen der Konstrukte verursachen. Mit anderen Worten *formen* (bzw. bestimmen) die Indikatoren das Konstrukt und dieses wird als (typischerweise lineare) Kombination seiner Indikatoren dargestellt ..."

Tab. 4.1 Unterschiede zwischen formativer und reflektiver Perspektive. (In Anlehnung an Jarvis et al., 2003, S. 203)

Formative Perspektive	Reflektive Perspektive
Kausalität von Indikatoren zum Konzept	Kausalität vom Konzept zu Indikatoren
Indikatoren definieren Merkmale des Konzepts	Indikatoren sind Manifestationen des Konzepts
Veränderungen bei den Indikatoren verursachen Veränderungen des Konzepts	Veränderungen bei den Indikatoren verändern das Konzept nicht
Indikatoren sind typischerweise nicht austauschbar	Indikatoren sind austauschbar
Indikatoren müssen nicht inhaltlich homogen sein	Indikatoren müssen inhaltlich homogen (→Konzept) sein
Verzicht auf einen Indikator kann das Konzept inhaltlich verändern	Verzicht auf einen Indikator verändert das Konzept inhaltlich nicht

Die Entscheidung für eine formative oder eine reflektive Perspektive bei der Messung ist nicht immer ganz eindeutig und leicht. Jarvis et al. (2003) haben dafür eine Reihe von Kriterien vorgeschlagen, von denen einige, die auch zur Illustration des Unterschiedes von formativer und reflektiver Perspektive dienen können, in Tab. 4.1 wiedergegeben sind.

Welche Konsequenzen haben die Unterschiede zwischen formativen und reflektiven Konzepten nun für die *Entwicklung von Multi-Item-Skalen?* Die zentralen Aspekte lassen sich aus der vorstehend skizzierten Unterscheidung ableiten und sollen hier in Anlehnung an Albers und Hildebrandt (2006) und Eberl (2006) gekennzeichnet werden:

Indikatoren (Items) sind bei formativen Konzepten *nicht austauschbar*, während sie bei reflektiven Konzepten nach Messeigenschaften (→ Reliabilität, Validität) *ausgewählt* werden können.

Bei formativen Messungen müssen die Indikatoren nicht untereinander korreliert sein. Diese können vielmehr unabhängig voneinander sein und sich ergänzen. Bei reflektiven Messungen, die sich in prinzipiell gleichartigen Indikatoren niederschlagen, geht man dagegen von Korreliertheit zwischen den Indikatoren aus (siehe dazu die Ausführungen zu Cronbach's α im Abschn. 4.3.2.5).

WICHTIG: Die Ausführungen in den *folgenden Abschnitten* 4.3.2.5 und 4.3.2.6 beziehen sich deswegen *nur* auf *reflektive Messungen*.

4.3.2.4 Definition der zu messenden Konzepte und Sammlung der Items

Im Abschn. 2.2.2 ist skizziert worden, dass sich die Operationalisierungsphase, in der u. a. die einzusetzenden Messinstrumente festgelegt werden, direkt an die Konzeptualisierung anschließt. Ausgangspunkt ist (und muss sein) die exakte **Definition** der zu messenden Konzepte. Diese Forderung ist keineswegs trivial. Vielmehr beobachtet

man in der empirischen Marketingforschung teilweise erhebliche Uneinheitlichkeit der in verschiedenen Untersuchungen verwendeten Definitionen gleicher oder ähnlicher Konzepte, nicht selten fehlt sogar die explizite Angabe von Definitionen.

Beispiel

Beispielsweise fanden Warren et al. (2019) bei ihrer Literaturanalyse zum Begriff **„Coolness“** (für eine größere Studie zur „Brand Coolness“) 70 (!) verschiedene Definitionen. Messverfahren, die auf unterschiedlichen Definitionen aufbauen, können nur zu uneinheitlichen bis widersprüchlichen Ergebnissen führen.

Hier eine Übungsfrage für die Leserin bzw. den Leser: Der Begriff „cool“ bzw. „Coolness“ wird – insbesondere von jüngeren Menschen – geradezu inflationär gebraucht. Was verstehen Sie unter diesem Begriff? Versuchen Sie bitte, eine Definition dafür gedanklich zu entwickeln, bevor Sie die folgende Definition von Warren und Campbell (2014, S. 544) lesen.

Warren et al. (2019, S. 37) verwenden diese Definition und kennzeichnen Coolness als „eine *subjektive* und *dynamische,* sozial konstruierte *positive* Eigenschaft, die mit kulturellen Objekten (Menschen, Marken, Produkten etc.) verbunden wird, die als angemessen *unabhängig* angesehen werden.“

Warren et al. (2019, S. 37) erläutern an gleicher Stelle die vier wesentlichen Merkmale von Coolness:

- „Coolness ist subjektiv. Marken sind nur cool (oder uncool!) in dem Maße, wie KonsumentInnen diese entsprechend einschätzen.“
- „Coolness hat eine positive Wertigkeit. Die meisten Wörterbücher beschreiben ‚cool‘ als einen Begriff, um Zustimmung, Bewunderung und Akzeptanz auszudrücken.“
- „Ein drittes Merkmal hilft, ‚cool‘ von ‚wünschenswert‘ zu unterscheiden: Unabhängigkeit. Unabhängigkeit besteht darin, dass man willens und fähig ist, seinen eigenen Weg zu gehen und sich nicht an die Erwartungen und Wünsche anderer anzupassen.“
- „Coolness ist dynamisch. Die Marken, die heute cool sind, können morgen nicht mehr cool sein.“ ◀

Mit einer Definition wird der Inhalt eines gedanklichen Konzepts (z. B. Kaufabsicht, Bekanntheitsgrad) formuliert und schriftlich festgehalten. Im Wesentlichen wird der Inhalt eines zu definierenden Begriffs („Definiendum“) mit Hilfe anderer (bereits bekannter) Begriffe und entsprechender Formulierungen („Definiens“) möglichst genau charakterisiert. In der Regel basiert die Entwicklung einer Definition u. a. auf der Auswertung der einschlägigen Literatur unter Berücksichtigung der darin enthaltenen Definitionsversuche, theoretischen Überlegungen und praktischen Erfahrungen. Mit einer Definition ist häufig die Abgrenzung zu anderen („benachbarten“) Konzepten verbunden. Beispielsweise wäre bei einer Definition des Konzepts „Markenpräferenz“ eine

Abgrenzung vom Konzept „Kaufabsicht“ vorzunehmen (was keine ganz leichte Aufgabe ist).

Hintergrundinformation
Shelby Hunt (1987, S. 209) kennzeichnet Wesen von und Anforderungen an Definitionen:

> „Definitionen sind ‚Regeln zum Ersetzen‘ (….). Mit einer Definition meint man also, dass ein Wort oder eine Gruppe von Worten (Definiens) äquivalent mit dem zu definierenden Wort (Definiendum) sein soll. Gute Definitionen zeigen Inklusivität, Exklusivität, Unterscheidbarkeit, Klarheit, Kommunizierbarkeit, Konsistenz und Knappheit.“

Mit „Inklusivität“ ist hier gemeint, dass die Phänomene, die dem Definiendum zugerechnet werden, bei der Definition eingeschlossen sind. Andererseits bezieht sich die „Exklusivität“ auf die Abgrenzung gegenüber anderen Phänomenen.

Auf der Basis der einschlägigen Literatur geben Eisend und Kuß (2023, S. 146) einige Empfehlungen für die Formulierung von Definitionen. Diese sollen

- das jeweilige Konzept möglichst prägnant und umfassend charakterisieren sowie von anderen (ähnlichen) Konzepten klar abgrenzen;
- möglichst einfache, eindeutige und klare Begriffe verwenden;
- knapp formuliert sein;
- mit anderen Definitionen innerhalb des Fachgebiets und bisheriger Forschung nach Möglichkeit verträglich sein.

Die exakte (und explizite) Definition der zu messenden Konzepte ist nicht nur unverzichtbare **Grundlage für jede Validitätsprüfung,** weil eben sonst kein Maßstab existiert, anhand dessen zu beurteilen wäre, ob tatsächlich das gemessen wurde, was gemessen werden sollte. Sie bestimmt auch den **Inhalt der** in einer Skala zu verwendenden **Items.** Im Interesse der Vergleichbarkeit von Untersuchungsergebnissen ist im Zweifel der Anpassung an früher verwendete Definitionen der Vorzug gegenüber neuen Definitionen zu geben (vgl. Churchill, 1979, S. 67).

Die Umsetzung (Operationalisierung) eines theoretischen Konzepts in ein adäquates Messinstrument beginnt mit der **Sammlung von Items.** Das Grund-Erfordernis dabei (siehe Abschn. 2.2.2) besteht darin, dass die verwendeten Items dem interessierenden Konzept (und nur diesem!) möglichst gut entsprechen sollen. Items, die eher einem anderen als dem jeweiligen Konzept zuzuordnen sind, müssen frühzeitig eliminiert werden, da eine Messung, bei der sie eine Rolle spielen, eben keine reine Messung des interessierenden Konzepts mehr wäre. Wenn eine Skala diesen Anforderungen genügt, dann entspricht sie den Kriterien der Inhaltsvalidität (siehe Abschn. 4.3.2.6).

Die Wege zur Gewinnung von Items sind unterschiedlich und werden in der Regel parallel begangen:

- Logische und/oder kreative Ableitung aus der Definition eines Konzepts
- Sammlung in früheren Untersuchungen verwendeter Items
- Auswertung von Literatur, in der das interessierende Konzept beleuchtet wird
- Experten-Gespräche
- Qualitative Vorstudien (z. B. Tiefeninterviews, Gruppendiskussionen) bei Angehörigen der für die Untersuchung relevanten Zielgruppe (siehe Kap. 3)

4.3.2.5 Überprüfung der Reliabilität

Sobald Items für die zu entwickelnde Multi-Item-Skala vorliegen, können Überprüfungen der Messeigenschaften von Entwürfen einer solchen Skala vorgenommen werden, um schrittweise festzustellen, welche Items für die letztendlich zu verwendende Skala geeignet sind. Im Mittelpunkt steht dabei die Überprüfung von Reliabilität und Validität der Skala (siehe dazu auch Abschn. 2.3), wobei der erstgenannte Aspekt im vorliegenden Abschnitt und die Validität im folgenden Abschnitt erörtert werden.

Hintergrundinformation

Das Wesen der Reliabilität wird von Schermelleh-Engel und Werner (2007, S. 114) kurz und klar charakterisiert (die Autorinnen kommen aus der Psychologie und sprechen deshalb von „Testverfahren“ an Stelle von „Messverfahren“):

> „Unter Reliabilität wird die Genauigkeit einer Messung verstanden. Ein Testverfahren ist perfekt reliabel, wenn die damit erhaltenen Testwerte frei von zufälligen Messfehlern sind. Das Testverfahren ist umso weniger reliabel, je größer die Einflüsse von zufälligen Messfehlern sind.“

Zunächst wird an die Überlegung angeknüpft, dass sich Reliabilität auf die Unabhängigkeit der Messwerte von den Besonderheiten und Zufälligkeiten eines einzelnen Messvorgangs bezieht. Direkt daran knüpft die Grundidee der so genannten **Test-Retest-Reliabilität** an. Die Bezeichnung lässt schon erahnen, dass es um die Wiederholung einer Messung in einem angemessenen zeitlichen Abstand geht. Als Maßzahl für die Reliabilität in diesem Sinne würde man die Korrelation (siehe Abschn. 7.3.2) der beiden Messungen verwenden. Diese Art der Reliabilitätsüberprüfung setzt natürlich voraus, dass sich das zu messende Konstrukt in der Zwischenzeit nicht verändert hat. Anderenfalls wäre ja eine geringe Korrelation nicht durch mangelnde Reliabilität, sondern durch diese Veränderung begründet.

Eine Reliabilitätsprüfung durch Wiederholung eines Messvorgangs und Vergleich der Ergebnisse wäre sehr aufwendig und auch in methodischer Hinsicht problematisch, u. a. dadurch, dass eine Vormessung das Ergebnis einer Nachmessung beeinflussen kann. In Verbindung damit steht das Problem, dass bei einer zweiten Messung ähnliche Werte auch dadurch zu Stande kommen können, dass die Auskunftspersonen sich bei der zweiten Befragung an ihr Antwortverhalten bei der ersten Befragung erinnern und konsistent antworten möchten.

Den genannten Problemen der Bestimmung der Test-Retest-Reliabilität versucht man beim Ansatz der **Parallel-Test-Reliabilität** dadurch zu entgehen, dass man zum gleichen Zeitpunkt (d. h. im gleichen Fragebogen) eine Vergleichsmessung mit einem entsprechenden Messinstrument durchführt. Beide Messungen sollen bei gegebener Reliabilität hoch korreliert sein. Beispielsweise könnte man zwei verschiedene (aber äquivalente) Likert-Skalen zur Messung desselben Konstrukts anwenden und dann die entsprechenden Ergebnisse korrelieren. Die Schwierigkeit besteht natürlich darin, zwei äquivalente Messinstrumente zu finden bzw. zu entwickeln. Abgesehen davon werden auf diese Weise die Interviews länger und manchmal auch zu eintönig.

Der wohl gängigste Ansatz zur Reliabilitätsüberprüfung ist direkt auf Multi-Item-Skalen bezogen. Dabei wird so vorgegangen, dass nicht alle Item-Werte einer Auskunftsperson durch Addition zu einem Gesamtwert zusammengefügt werden. Vielmehr teilt man die Gesamtheit der Items in zwei Hälften und erhält durch additive Verknüpfung innerhalb der beiden Gruppen dann zwei Messwerte. Man kommt auf diese Weise in einem Messvorgang zu zwei sehr ähnlichen Messinstrumenten (mit gleich strukturierten, aber unterschiedlich formulierten Items) für ein Konzept. Die Reliabilität einer Messmethode müsste sich in einem hohen Korrelationskoeffizienten für die beiden Teil-Skalen niederschlagen.

Die Basis für diesen Ansatz ist die **Domain Sampling Theorie** (Nunnally & Bernstein, 1994, S. 216 ff.), die davon ausgeht, dass jede Menge von in einer Skala verwendeten Items eine Stichprobe aus einer großen Menge von (alle Facetten des interessierenden Konzepts abdeckenden) Items ist. Wenn man in der oben erwähnten Weise zwei Teil-Skalen bildet, so hat man damit zwei Stichproben von Items aus einer (natürlich unbekannten) Grundgesamtheit von Items gezogen, die zu äquivalenten Ergebnissen führen müssten. Wenn dies bei hinreichend großer Zahl von Items nicht der Fall ist, dann sind offenbar andere Einflüsse bei der Messung wirksam, die Reliabilität ist also gering. Der Grundgedanke dieser Vorgehensweise schlägt sich in der Bezeichnung **Split-Half-Reliabilität** (Nunnally & Bernstein, 1994, S. 232 f.) nieder.

Hintergrundinformation

David de Vaus (2002, S. 19) erläutert die Grundidee der Überprüfung von Reliabilität auf der Basis der internen Konsistenz einer Multi-Item-Skala:

> „Wenn Items, die dazu da sind, dasselbe zu Grunde liegende Konzept zu messen, in konsistenter Weise beantwortet werden, wird die Menge von Items als reliabel angesehen. Mit anderen Worten: Reliabilität wird dadurch bestimmt, dass geprüft wird, wie konsistent verschiedene Items dasselbe Konzept darstellen, nicht dadurch, dass man betrachtet, mit welcher Konsistenz dieselben Items im Zeitablauf beantwortet werden. Alle Maßzahlen der internen Konsistenz kennzeichnen die Reliabilität durch Koeffizienten, die zwischen 0 und 1 liegen."

Nun kann die Aufteilung einer Menge von Items in zwei Hälften in unterschiedlicher Weise erfolgen und damit zu nicht eindeutigen Reliabilitätsindikatoren führen. Dieses Problem wird dadurch behoben, dass man üblicherweise den Reliabilitätskoeffizienten

Cronbach's α verwendet, der dem Mittelwert der Korrelationskoeffizienten aller möglichen Kombinationen von Skalenhälften entspricht (vgl. Cronbach, 1951; Peter, 1979). Dieser α-Koeffizient ist somit ein Maß für die interne Konsistenz einer Skala, weil ja die Übereinstimmung (→ Korrelation) der einzelnen Items gemessen wird. Er kann herangezogen werden, um bei der Skalenentwicklung aus der Menge der anfangs vorhandenen Items die weniger geeigneten zu eliminieren. Vergleiche des a -Wertes einer Skala mit den a -Werten für die (fast) gleichen Skalen, bei denen jeweils eines der Items nicht enthalten ist, zeigen an, inwieweit die betreffenden Items geeignet sind, die Reliabilität der Skala zu erhöhen oder zu verringern.

Ein zweiter Indikator für die Nützlichkeit eines Items in einer Skala ist die Korrelation dieses Items mit dem aus den restlichen Items gebildeten Gesamtwert der Skala (siehe z. B. McIver & Carmines, 1981, S. 31 ff.). Eine geringe Korrelation weist darauf hin, dass ein Item die Reliabilität eines Messinstruments eher verringert und/oder dass es nicht hinreichend dem zu messenden Konzept entspricht und insofern die Validität der Messung beeinträchtigt.

4.3.2.6 Überprüfung der Validität

Im Mittelpunkt der Entwicklung und Prüfung von Messinstrumenten steht die Betrachtung der Validität. Mit der Validität steht und fällt die Qualität eines Messinstruments und damit der ganzen Untersuchung, in der dieses verwendet wird. Man geht davon aus, dass sich Unterschiede bei einem Konzept in den Ergebnissen entsprechender (valider) Messungen widerspiegeln (Borsboom et al., 2004). Da man die Validität einer Messung eben nicht durch den Vergleich des Messwerts mit dem typischerweise ja unbekannten wahren Wert des interessierenden Konzepts (bzw. Konstrukts; deswegen wird in der Literatur oft von **Konstruktvalidität** gesprochen) ermitteln kann, bedient man sich (gewissermaßen hilfsweise) verschiedener Kriterien, um festzustellen, ob das entwickelte bzw. in der Entwicklung befindliche Messinstrument den unterschiedlichen Facetten der Validität entspricht. Wenn nicht, sind entsprechende Veränderungen des Messinstruments nötig. Im Folgenden werden dazu folgende Aspekte bzw. Arten der Validität skizziert:

1. Inhaltsvalidität
2. Kriterienvalidität
3. Konvergenzvalidität
4. Diskriminanzvalidität

Wenn ein Messinstrument sich bei diesen Arten der Validitätsüberprüfung bewährt, dann spricht das dafür, dass dieses Instrument tatsächlich misst, was es messen soll, und man kann auf Basis der resultierenden Untersuchungsergebnisse praktische Entscheidungen treffen und/oder wissenschaftliche Schlüsse ziehen. „Mit jeder weiteren Überprüfung wachsen die Belege bezüglich der Validität weiter." (Strauss & Smith, 2009, S. 6) In diesem Sinne ist Validierung das Ergebnis eines Prozesses, kein Einzelergebnis. Abb. 4.10 gibt einen entsprechenden Überblick.

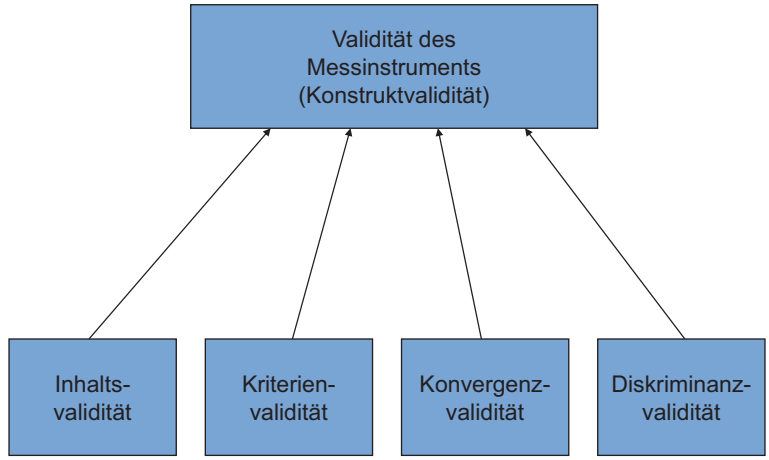


Abb. 4.10 Kriterien für die Validitätsprüfung. (Nach Viswanathan, 2005, S. 65)

Zunächst zur **Inhaltsvalidität** („content validity"). Diese bezieht sich auf die (häufig von Experten beurteilte) Eignung und Vollständigkeit des Messinstruments im Hinblick auf das zu messende Konzept. Hier geht es also darum, dass sich die wesentlichen Aspekte dieses Konzepts in den Skalen-Items widerspiegeln. Aus der Definition des Konzepts müssen also die relevanten Inhalte abgeleitet und in Items „übersetzt" werden. Die Überprüfung erfolgt typischerweise durch Experten, die die logische Eignung der Items hinsichtlich des definierten Konzepts beurteilen („face validity"). Dabei kommt es nicht zuletzt darauf an, dass jedes der verwendeten Items den Ansprüchen der Inhaltsvalidität genügt. Rossiter (2011, S. 14) kritisiert in diesem Zusammenhang eine gelegentlich etwas nachlässige Forschungspraxis, bei der zwar die Skala insgesamt, aber eben nicht alle einzelnen Items sorgfältig geprüft werden. Typischerweise erfolgt die Überprüfung der Inhaltsvalidität vor den weiteren empirischen Tests (s. u.) einer Skala.

Beispiel

David de Vaus (2002, S. 28) gibt ein Beispiel zur Inhaltsvalidität:

„Die Feststellung der Inhaltsvalidität beinhaltet die Überprüfung, in welchem Maße in das Messinstrument die verschiedenen Aspekte des Konzepts einfließen. Beispielsweise wäre ein Messverfahren, das dazu dient, den allgemeinen Gesundheitszustand zu messen, und das darauf begrenzt ist, den Blutdruck zu messen, dem Konzept „Gesundheit" nicht angemessen, zumindest nicht nach dem üblichen Verständnis. Gesundheit wird meist als ein breiteres und komplexeres Phänomen angesehen. Andere Aspekte der physischen Gesundheit und ebenso – beispielsweise – des psychischen Wohlbefindens wären normalerweise Bestandteil eines validen Messverfahrens für Gesundheit." ◀

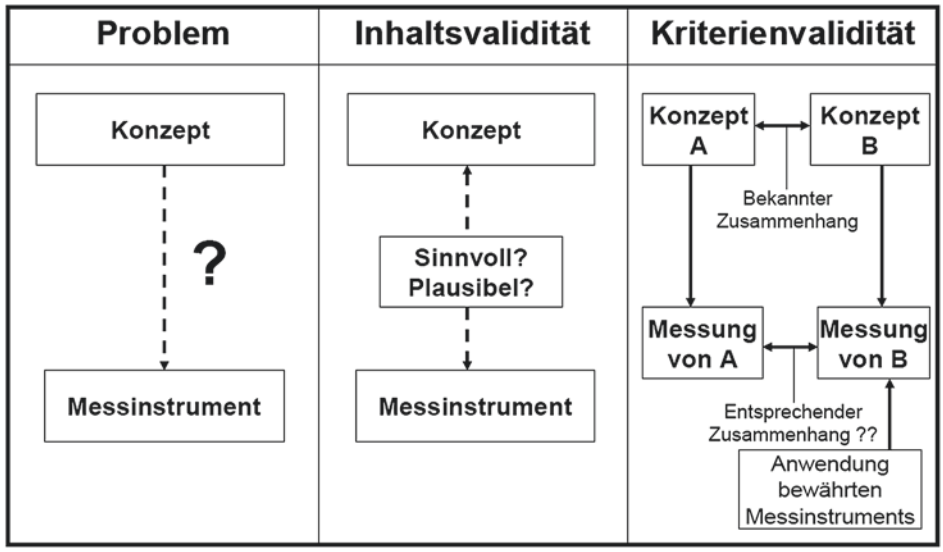


Abb. 4.11 Logik der Prüfung von Inhalts- und Kriterienvalidität

Deutlich konkreter sind die Möglichkeiten zur Überprüfung der **Kriterienvalidität.** Was ist damit gemeint? Kriterienvalidität bezieht sich darauf, dass die Ergebnisse einer Messung einer bekannten („etablierten") Beziehung zu Messungen anderer Konzepte entsprechen. Beispielsweise ist in der Konsumentenforschung seit langem bekannt, dass Einstellungen und Verhalten in einer (nicht deterministischen) positiven Beziehung stehen. Wenn man eine Skala zur Messung von Einstellungen zu einer Marke entwickelt, dann müssten diese Werte mit Messungen der Kaufhäufigkeit dieser Marke positiv korreliert sein. Anderenfalls wäre an der Validität der Einstellungsskala zu zweifeln (vgl. Hildebrandt, 1984). Abb. 4.11 illustriert die Grundideen der Prüfung von Inhalts- und Kriterienvalidität.

In der Literatur wird teilweise noch danach unterschieden, ob das betreffende Kriterium gleichzeitig („concurrent validity") oder zu einem späteren Zeitpunkt („predictive validity") gemessen wird. Einen Spezialfall der Kriterienvalidität bezeichnet Spector (1994, S. 277) als **„Known-Groups-Validity"** und meint damit, dass im Hinblick auf die Messwerte der entwickelten Skala bei bestimmten Gruppen von Befragten (Validität unterstellt) *unterschiedliche Ergebnisse* auftreten müssten. Die Kriterienvariable ist also nicht kontinuierlich, sondern kategorial. Z. B. würde man bei einer Skala, mit der die Einstellung zu klassischer Musik gemessen werden soll, erwarten, dass bei einer Gruppe von Befragten im Alter über 40 Jahren mit hoher Schulbildung positivere Werte zu Stande kommen als bei einer Gruppe im Alter unter 18 Jahren mit geringer Schulbildung. Wenn das nicht so wäre, würde man wohl an der Validität der Skala zweifeln.

Zentrale Bedeutung für die Validitätsüberprüfung haben *Konvergenzvalidität* und *Diskriminanzvalidität*. Die entsprechenden Grundideen sollen hier kurz charakterisiert wer-

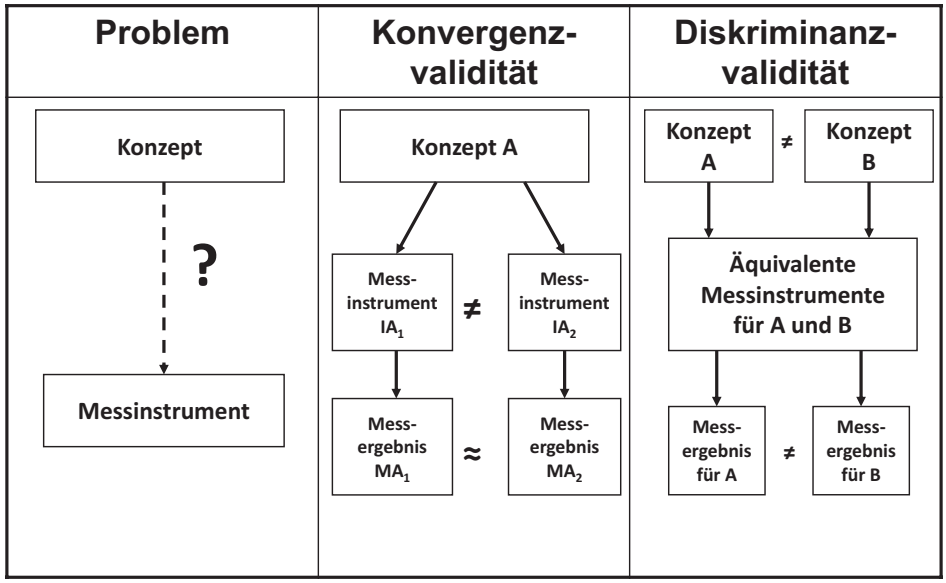


Abb. 4.12 Logik der Prüfung von Konvergenz- und Diskriminanzvalidität

den (vgl. Peter, 1981). Beide werden häufig im Zusammenhang betrachtet, was sich auch im Ansatz der Multitrait-Multimethod-Matrix (vgl. Campbell & Fiske, 1959) niederschlägt, auf den am Ende dieses Abschnitts eingegangen wird.

Zunächst zur **Konvergenzvalidität:** Wenn das gleiche Konzept mit zwei (möglichst) verschiedenen Messinstrumenten gemessen wird, so sollen die Ergebnisse sehr ähnlich sein (konvergieren), sofern diese Instrumente valide sind. Beide Instrumente sollen möglichst wenig methodische Gemeinsamkeiten haben, da sonst die Ähnlichkeit der Messwerte ein Artefakt sein könnte, das durch eben diese Gemeinsamkeiten verursacht wurde. Wenn also zwei sehr *unähnliche Messverfahren* angewandt auf das *gleiche Konzept* zu konvergierenden Ergebnissen führen, dann sind diese Ergebnisse offenbar unabhängig vom Erhebungsverfahren, stellen also wohl kein Methoden-Artefakt dar. Hier sei auf die Analogie zum Stichwort „Triangulation“ hingewiesen, das im Abschn. 3.1 eine Rolle spielte. Was macht dagegen die **Diskriminanzvalidität** aus? Wenn man mit dem gleichen Typ von Messinstrumenten (z. B. Likert-Skalen) verschiedene (nicht zusammenhängende) Konzepte misst, dann sollen die Ergebnisse nicht korreliert sein. Ansonsten würden die Messwerte ja weniger die Unterschiedlichkeit der Konzepte wiedergeben, sondern eher auf systematische Einflüsse der Messmethoden zurückzuführen sein, was natürlich das Vertrauen in deren Validität schwinden ließe. Beispielsweise sollten die Ergebnisse von Messungen der Einstellungen zur Marke VW und zum Urlaubsziel Mallorca, beide gemessen mit Likert-Skalen, nur schwach korreliert sein. Mit *gleichartigen Messverfahren* angewandt auf *verschiedene Konzepte* soll man also die Messwerte für diese Konzepte unterscheiden (→ diskriminieren) können. Abb. 4.12 illustriert auch hier die Grundideen beider Ansätze.

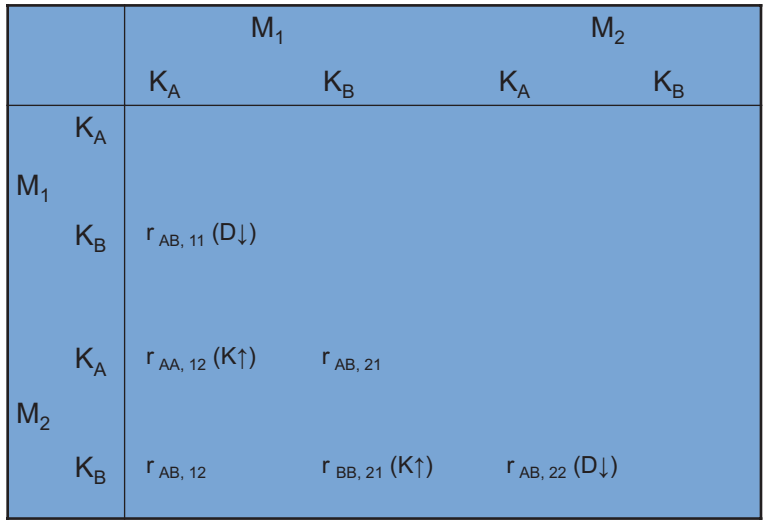


Abb. 4.13 Beispiel einer einfachen Multitrait-Multimethod-Matrix

Eine übersichtliche Darstellung der zur Prüfung von Konvergenz- und Diskriminanzvalidität notwendigen Korrelationskoeffizienten, die sogenannte Multitrait-Multimethod-Matrix (Multimerkmals-Multimethoden-Matrix), geht auf Campbell und Fiske (1959) zurück. In Abb. 4.13 findet sich eine schematische Darstellung des einfachsten Falls einer solchen Matrix mit zwei Konzepten, die jeweils mit Hilfe zweier Untersuchungsmethoden (z. B. Likert-Skala und einfache Ratingskala) gemessen werden. In der Abbildung ist durch die Buchstaben K bzw. D schon eingetragen, welche Korrelationskoeffizienten im Hinblick auf Konvergenz- und Diskriminanzvalidität besonders beachtlich sind. Die daneben eingezeichneten Pfeile deuten an, ob die Korrelationen hier hoch oder niedrig sein sollten.

In der in Abb. 4.13 dargestellten Matrix sind die beiden verwendeten Messmethoden mit M_1 und M_2 gekennzeichnet; K_A und K_B stehen für zwei verschiedene Konzepte/Merkmale A und B. In den Tabellenfeldern stehen Korrelationskoeffizienten r, die auf den durch die Anwendung der beiden Messmethoden auf die zwei Konzepte gewonnenen Daten basieren.

Im Hinblick auf die Kriterien der Konvergenz- und Diskriminanzvalidität sind an die Werte die Korrelationskoeffizienten bestimmte Forderungen zu stellen: Die Koeffizienten $r_{AA,12}$ und $r_{BB,21}$ geben an, wie stark die mit unterschiedlichen Methoden gemessenen Werte für das gleiche Konzept (A bzw. B) korrelieren. Bei Vorliegen von Konvergenzvalidität müssten sich hier hohe Werte ergeben. Auf jeden Fall müssen die Werte deutlich höher sein als die für die Prüfung der Diskriminanzvalidität herangezogenen Korrelationskoeffizienten.

Die Koeffizienten $r_{AB,11}$ und $r_{AB,22}$ zeigen die Korrelationen von Messwerten für verschiedene Konzepte, die durch gleichartige Methoden zustande gekommen sind.

Wenn keine Beziehung zwischen den Konzepten besteht und die entsprechenden Messinstrumente die Konzepte korrekt wiedergeben (Diskriminanzvalidität), dann müssten die Korrelationskoeffizienten sehr gering sein (im Idealfall: Null).

Unbefriedigende Ergebnisse dieser Prüfphase führen zu einer Rückkoppelung im Prozess der Entwicklung des Messinstruments, was typischerweise bedeutet, dass der Prozess der Generierung von Items wieder aufgenommen wird und/oder dass die Angemessenheit der für das Konzept verwendeten Definition in Frage gestellt werden muss.

Beispiel

Ein geradezu „klassisches" Beispiel zur Überprüfung von Konvergenz- und Diskriminanzvalidität mit Hilfe der Multitrait-Multimethod-Matrix stammt aus dem einflussreichen Aufsatz von Churchill (1979) zur Skalenentwicklung im Marketing.

Es wurden dabei die drei Konzepte Job-Zufriedenheit, Rollenkonflikt und Rollenunklarheit bei Verkäufern mit Hilfe von jeweils zwei Methoden, der Likert-Skala und der Thermometer-Skala (Rating-Skala in Form eines Thermometers) gemessen. Einzelne charakteristische Teile der daraus entstehenden Multitrait-Multimethod-Matrix, die nachstehend wiedergegeben ist, sind mit den Ziffern 1 bis 4 gekennzeichnet.

Nun zur Interpretation:

Die „Reliabilitätsdiagonale", die mit (1) gekennzeichnet ist, enthält die Korrelationen von alternativen Formen einer Likert-Skala, die im Abstand von zwei Wochen eingesetzt wurde. Diese Werte sind befriedigend groß.

Maßgeblich für die Konvergenzvalidität ist die „Validitätsdiagonale" (3). Hier werden also die Messwerte für jeweils gleiche Konzepte – gemessen mit unterschiedlichen Methoden – in Beziehung gesetzt. Die Korrelationskoeffizienten sind positiv und hinreichend groß, jedenfalls signifikant von 0 verschieden.

Die Analyse hinsichtlich der Diskriminanzvalidität ist etwas komplizierter. Zunächst sollten die Korrelationen in der Validitätsdiagonalen (3) zumindest größer sein als die Korrelationen für verschiedene Konzepte und verschiedene Methoden (4), was hier eindeutig der Fall ist. Direkt ableitbar aus der Grundidee der Diskriminanzvalidität ist die Anforderung, dass die Koeffizienten in der Validitätsdiagonalen (3) (gleiches Konzept, verschiedene Methoden) höher sein sollten als Korrelationskoeffizienten, die für unterschiedliche Konzepte und gleiche Methoden (2) stehen, was hier gegeben ist (Abb. 4.14).

Für eine ausführlichere Darstellung eines Beispiels einer Multitrait-Multimethod-Matrix sei auf Döring und Bortz (2016, S. 472 ff.) verwiesen. ◄

Die Bedeutung der Reliabilität und Validität von Messungen für die Aussagekraft von empirischen Untersuchungen ist im vorliegenden Buch immer wieder betont worden. In diesem Abschnitt sind anhand der Entwicklung von Multi-Item-Skalen für Befragungen einige praktisch einsetzbare Kriterien für die Überprüfung von Reliabilität und Validität vorgestellt worden. Diese Kriterien sind (hoffentlich) auf der Basis eines gewissen

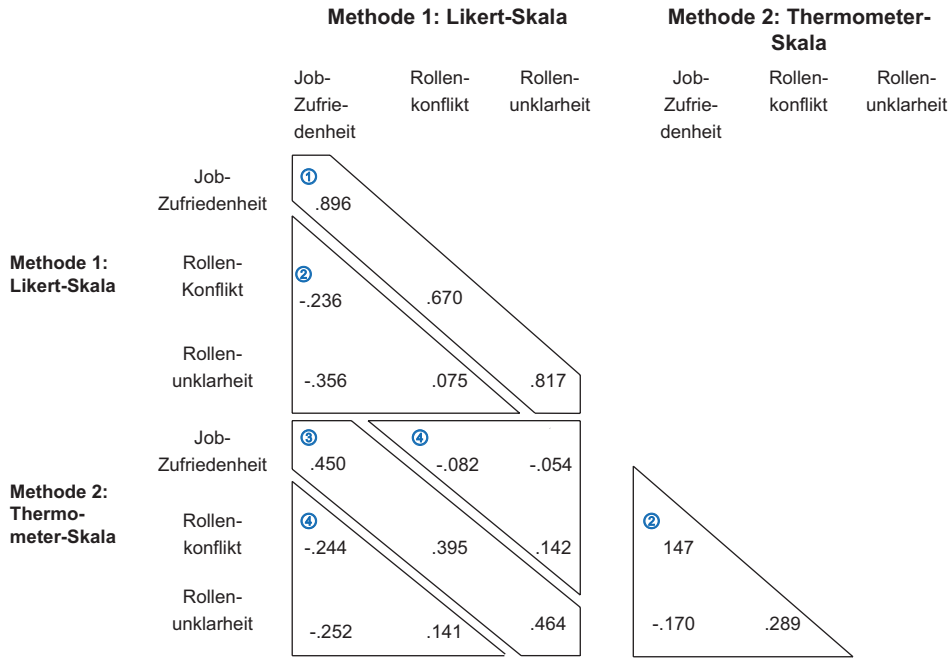


Abb. 4.14 Beispiel einer MTMM-Matrix von Churchill, 1979

methodischen Grundverständnisses relativ leicht nachvollziehbar. Inzwischen gibt es weitergehende – und auch wesentlich anspruchsvollere – Ansätze zur Validierung auf der Basis von Strukturgleichungsmodellen (siehe Abschn. 9.6). Homburg und Giering (1996, S. 8) sprechen hier von „Reliabilitäts- und Validitätskriterien der zweiten Generation". Zu Einzelheiten sei auf den zitierten Aufsatz von Homburg und Giering (1996) sowie auf Hildebrandt (1984), Netemeyer et al. (2003) und Weiber und Mühlhaus (2010) verwiesen.

4.3.3 Entwicklung von Fragebögen

In den vorigen Abschnitten ist (mit gutem Grund) relativ ausführlich die Formulierung von Fragen einschließlich der Entwicklung von Multi-Item-Skalen (als wichtiger und anspruchsvoller Form der Frageformulierung) erörtert worden. Im vorliegenden Abschnitt geht es jetzt vor allem um die Zusammenstellung der einzelnen Fragen zu einem Fragebogen, der natürlich heute nicht mehr an die physische Form eines Drucks auf Papier gebunden ist, sondern auch elektronische Speicherung bei computergestützten Telefon-Befragungen oder bei Online-Befragungen erlaubt.

Zunächst zu der Alternative, ob in einem Fragebogen nur Fragen zu einem Untersuchungsgegenstand **(Einthemen-Umfrage)** oder zu mehreren Problemkreisen **(Mehrthemen-Umfrage)** enthalten sind. Für die letztgenannte Form wird auch der (einigermaßen) anschauliche Begriff **„Omnibus-Befragung“** verwendet, natürlich nicht, weil die Interviews in öffentlichen Verkehrsmitteln durchgeführt werden, sondern weil analog zu einem Omnibus (lat. „Wagen für alle“), in dem mehrere Personen transportiert werden, hier Fragen zu unterschiedlichen Themen in einem einzigen Fragebogen übermittelt werden. Omnibus-Befragungen finden typischerweise auf Initiative von kommerziellen Marktforschungsinstituten (aber auch von sozialwissenschaftlichen Institutionen, siehe z. B. die ALLBUS-Umfragen des Mannheimer Gesis-Instituts, www.gesis.org) statt, die Interessenten anbieten, sich mit einigen Einzelfragen daran zu beteiligen. Durch die Themenmischung bietet die Omnibus-Befragung methodische Vorteile insofern, als sie für die Auskunftsperson abwechslungsreicher ist und keine Ausrichtung auf ein Antwortverhalten erlaubt, das auf einen erkennbaren Auftraggeber der Untersuchung ausgerichtet ist („Sponsorship-Effekt“).

Das zentrale Problem der Fragebogenentwicklung besteht in der Festlegung der Reihenfolge, in der die Fragen gestellt werden. „Es kann sein, dass Umfrageteilnehmer ihre Meinungen erst während der Beantwortung des Fragebogens bilden. D. h. dass Antworten auf Fragen am Beginn Informationen aktivieren können, die Antworten bei später folgenden Teilen des Fragebogens beeinflussen können.“ (Vomberg & Klarmann, 2022, S. 101) Manche dieser **Reihenfolge-Effekte** sind leicht erkennbar und nachvollziehbar, andere werden erst bei Veränderungen der Reihenfolge und Vergleich der jeweiligen Ergebnisse offenbar (siehe das folgende Beispiel). So kann man sich leicht vorstellen, dass beispielsweise bei einer Frage nach den Gründen, ein bestimmtes Auto zu kaufen, die Nennung einer Antwortkategorie „Hohe Sicherheit“ besonders häufig erfolgt, wenn zuvor z. B. die Frage „Haben Sie in letzter Zeit einen Werbespot gesehen, in dem das Sicherheitspaket des neuen VW Golf gezeigt wurde?“ gestellt worden ist. Zur Korrektur solcher Effekte versucht man „Gegengewichte“ („counterbalancing“) zu bilden, indem man bei einem Teil der Fragebögen (z. B. 50 %) die Fragereihenfolge umkehrt (Jacoby, 2013, S. 763).

Beispiel

Zur Illustration von Reihenfolge-Problemen hier ein Beispiel, bei dem sogar die Reihenfolge der Nennung von Antwortvorgaben zu deutlichen Ergebnisunterschieden führte. Das Beispiel stammt aus einer Untersuchung des Survey Research Center der University of Michigan aus dem Jahre 1979 (vgl. Schuman & Presser, 1981, S. 70). Es wurde bei zwei Teilstichproben ermittelt, inwieweit sich die Reihenfolge, in der Antwortmöglichkeiten genannt werden, auf die Antwortverteilung auswirken.

Die Erhebung, aus der das hier interessierende Teilergebnis im Folgenden wiedergegeben wird, bezog sich u. a. auf die Notwendigkeit staatlicher Wohnungspolitik in den USA.

„Manche Leute glauben, dass sich die Bundesregierung um angemessene Wohnungsversorgung für jeden kümmern sollte, während andere Leute glauben, dass sich jeder selbst seine Wohnung besorgen sollte. Was kommt Ihrer Meinung am nächsten?“

1	Regierung	44,6 %
2	Jeder selbst	55,4 %
		100 % ($n=327$)

„Manche Leute glauben, dass sich jeder selbst seine Wohnung besorgen sollte, während andere Leute glauben, dass sich die Bundesregierung um angemessene Wohnungsversorgung für jeden kümmern sollte. Was kommt Ihrer Meinung am nächsten?“

2	Regierung	29,5 %
1	Jeder selbst	70,5 %
		100 % ($n=329$)

Die nur scheinbar belanglose Änderung der Reihenfolge, in der die verschiedenen Standpunkte im Fragetext erscheinen, bewirkt also eine (signifikante) Veränderung der Antwortverteilung um etwa 15 %. Woran kann das liegen? Schuman und Presser (S. 68) vermuten, dass beide genannten Alternativen für die Auskunftspersonen attraktiv sind und dass – bei bisher nicht festgelegter Meinung – die erstgenannte gedanklich akzeptiert wird, bevor die zweite vorgelesen wird. ◀

Ausgehend vom relativ neuen Verständnis der kognitiven Prozesse bei einer Befragung (siehe Abschn. 4.3.1.2) sollen hier sogenannte **Kontext-Effekte** kurz erörtert werden. Man versteht darunter die Beeinflussung des Antwortverhaltens bei einer Frage durch andere im Fragebogen enthaltene Fragen. Der typische Fall ist der gerichtete Kontext-Effekt, bei dem eine Frage später folgende andere Fragen beeinflusst. Sudman et al. (1996, S. 81) heben hervor, dass jede Frage in einem Kontext steht, dass also entsprechende Wirkungen nicht völlig ausgeschlossen werden können.

Die wohl wichtigsten (Moore, 2002) gerichteten Effekte sind der Konsistenz- und der Kontrast-Effekt, die vor allem bei Einstellungsfragen auftreten. Beim **Konsistenz-Effekt** kann man aus der Bezeichnung schon erahnen, was damit gemeint ist: Eine frühere Frage hat Einfluss auf eine später folgende, wenn die Auskunftsperson versucht, ihr Antwortverhalten bei der späteren Frage so zu gestalten, dass ihr Verhalten insgesamt als konsistent erscheint. Die Antwort wird also in Richtung auf die bisherigen Antworten aus dem Kontext verändert. Dagegen bestehen **Kontrast-Effekte** darin, dass frühere Fragen bei späteren zu Antworten führen, die sich stärker als normal voneinander abheben. Beispielsweise kann es bei aufeinander folgenden Einstellungsmessungen hinsichtlich

zweier Produkte dazu kommen, dass die Messwerte sich deutlicher als ohne diesen Effekt unterscheiden, weil die Auskunftsperson einen Vergleich vornimmt und den Unterschied stärker betont (Sudman et al., 1996, S. 100 ff.).

Die Wirkung von Kontexten lässt sich auch auf das in Abb. 4.6 dargestellte Vier-Stufen-Modell des Befragungsprozesses beziehen. Im Folgenden werden diese Wirkungen überblicksartig dargestellt (vgl. Sudman et al., 1996, S. 83 ff.):

- **Stufe 1** „Verständnis der Frage“: Bisher schon im Fragebogen verwendete Begriffe erleichtern und prägen deren Verständnis in einer folgenden Frage. Wenn beispielsweise zuvor Probleme des technischen Fortschritts angesprochen wurden, dann wird später der Fortschritts-Begriff entsprechend verstanden. Auf spezielle Fachbegriffe, die schon eingeführt wurden, kann bei späteren Fragen Bezug genommen werden.
- **Stufe 2** „Suche nach entsprechenden Informationen“: Wenn schon Fragen zum gleichen Thema gestellt wurden, dann ist die Erinnerung an den jeweiligen Gegenstand schon aktiviert und der Zugriff zu gespeicherten Informationen erleichtert.
- **Stufe 3** „Beurteilungen und Einschätzungen“: Hier sind die schon angesprochenen Konsistenz- und Kontrast-Effekte einzuordnen. Diese sind aber nur wirksam, wenn die Erinnerung an frühere Fragen und Antworten zum jeweiligen Zeitpunkt noch vorhanden ist.
- **Stufe 4** „Formulierung der Antwort“: Vorhergehende Fragen können die Beachtung von Aspekten der sozialen Erwünschtheit von Angaben oder der Selbstdarstellung verstärken.

Ein wesentliches Mittel zur Beeinflussung von **Kontext-Effekten** ist die räumliche Anordnung der Fragen im Fragebogen und damit ihr zeitlicher Abstand im Interview. Bei enger „Nachbarschaft“ werden die positiven Kontext-Effekte bei Stufe 1 und 2 stärker wirksam, aber auch die störenden Wirkungen auf Stufe 3 und 4 (und umgekehrt). Allerdings kann eine Steuerung der Anordnung von Fragen nur wirksam erfolgen, wenn die Interviews persönlich oder telefonisch durchgeführt werden (siehe dazu Abschn. 4.3.4). Bei schriftlichen Befragungen ist es üblich, dass die Auskunftspersonen den Fragebogen vor der Beantwortung durchsehen, so dass die Steuerung von Reihenfolge und Abstand von Fragen wenig Wirkung hat. Bei Online-Befragungen gibt es unterschiedliche Varianten, die eine vorherige Betrachtung der Fragen erlauben oder nicht.

Hintergrundinformation

Hinsichtlich der üblicherweise empfohlenen Fragereihenfolge seien hier einige allgemeine Regeln aus der einschlägigen Literatur (Döring & Bortz, 2016, S. 405 ff.; Noelle-Neumann & Petersen, 2000, S. 120 ff.; Sudman & Blair, 1998, S. 285 ff.; Vomberg & Klarmann, 2022, S. 101 f.) zusammengestellt:

1. Zu Beginn des Fragebogens Kontakt zur Auskunftsperson herstellen mit einigen leicht beantwortbaren und Interesse weckenden Fragen, den so genannten „**Eisbrecher-Fragen**“. Beispiel: „Zunächst eine Frage zu Ihrem letzten Urlaub. Heute werden ja viele Urlaubsreisen zu

außereuropäischen Zielen angeboten. Wie stehen Sie dazu? Fahren Sie lieber weit weg oder bleiben Sie lieber in Europa?“
Fahre lieber weit weg ()
Bleibe lieber in Europa ()

2. **Fragen zu persönlichen Merkmalen** wie Alter, Einkommen, Schulbildung etc. sollten nach Möglichkeit *am Ende des Fragebogens* platziert werden. Diese Fragen sind zwar für zahlreiche Auswertungen wichtig, werden aber von vielen Auskunftspersonen als Eindringen in ihre Intimsphäre wahrgenommen. Zu Beginn eines Interviews könnten solche Fragen Misstrauen wecken, während am Ende schon etwas Vertrauen entstanden ist und eher erklärt werden kann, dass diese Fragen zur (anonymen) Auswertung der sonstigen Angaben benötigt werden.
3. Zur Unterstützung des Gedächtnisses der Auskunftsperson und zur Erleichterung der Beantwortung sollte eine **logische Reihenfolge der Fragen** eingehalten werden. Beispiel:
„Wo waren Sie vor zwei Jahren im Urlaub?“
„Wo waren Sie im letzten Jahr im Urlaub?“
„Wo verbringen Sie in diesem Jahr Ihren Urlaub?“
4. **Inhaltlich zusammengehörende Fragen** sollten zusammengefasst werden, um gedankliche Sprünge zu vermeiden und die Beantwortung zu erleichtern. Wenn man allerdings deutliche Reihenfolge-Effekte befürchten muss, dann kann man auch „Puffer-Fragen“ einfügen, durch die die Auskunftsperson vom Inhalt einer Frage wieder abgelenkt wird, bevor eine möglicherweise dadurch beeinflusste Frage gestellt wird.
5. Innerhalb eines Themas **vom Allgemeinen zum Speziellen.** Wenn also zur Zufriedenheit mit einer Urlaubsreise gefragt wird, dann sollte zunächst die Zufriedenheit mit dem Urlaub allgemein und dann die Zufriedenheit mit einzelnen Aspekten wie Flug, Hotel etc. abgefragt werden. Es kann aber sein, dass das zunächst geäußerte Gesamturteil die Ergebnisse bei den einzelnen Aspekten beeinflusst. Bei umgekehrter Vorgehensweise kann es dagegen sein, dass die jeweils abgefragten Einzel-Aspekte das Gesamturteil stärker als sonst beeinflussen (Vomberg & Klarmann, 2022, S. 102).
6. Beim **Wechsel eines Themas** sollten **Übergänge** zwischen den entsprechenden Fragebogen-Abschnitten hergestellt werden, um die Auskunftsperson durch den Fragebogen zu leiten. Beispiel:
„Nun folgen einige Fragen zum Bereich …“
7. Am Ende des Fragebogens kann man den Auskunftspersonen ermöglichen, ein **Feedback** zu geben. Ganz am Schluss stehen der Dank an die Auskunftspersonen und eine Verabschiedung.

Für die Fragenreihenfolge hat ein spezieller Fragentyp eine besondere Bedeutung, die so genannten **„Filterfragen“**. Was ist damit gemeint? Der Name bezieht sich darauf, dass damit die Teilmenge von Auskunftspersonen herausgefiltert werden soll, für die eine folgende Frage bzw. ein folgender Fragebogenteil zutrifft. Die anderen Befragten werden gewissermaßen an diesem Teil „vorbei geleitet“. Beispiel:

- Sind Sie Raucher?“
  - Nein ()→ Bitte weiter mit Frage XX
  - Ja () ↓
    „Welche Zigarettenmarke bevorzugen Sie?“

Ein Problem bei einer großen Zahl solcher Filterfragen in einem Fragebogen entsteht dadurch, dass die Auskunftsperson nach einiger Zeit merkt, wie sie bei diesen Fragen

antworten muss, um eine Reihe folgender Fragen zu umgehen und das Interview abzukürzen.

Die **formale/optische Gestaltung von Fragebögen** kann in diesem einführenden Lehrbuch nicht umfassend behandelt werden. Dazu muss vor allem auf die einschlägige Spezialliteratur (Noelle-Neumann & Petersen, 2000; Sudman & Blair, 1998; Bradburn et al., 2004; Dillman et al., 2009) verwiesen werden. Einige gängige und einfache **Regeln** mögen genügen:

- Fragebogen als handliche kleine Broschüre gestalten.
- Große, klare Schrifttypen verwenden.
- Fragen übersichtlich anordnen.
- Fragen nicht über mehrere Seiten hinziehen.
- Optische Hilfsmittel (Pfeile, Hervorhebungen etc.) verwenden.
- Alle Fragen nummerieren.
- Anweisungen für die Beantwortung (z. B. „Bitte nur eine Antwort ankreuzen“) deutlich machen.

Angesichts der zahlreichen Probleme und Fehlermöglichkeiten bei der Formulierung von Fragen und bei der Entwicklung von Fragebögen gilt heute die Durchführung von so genannten **„Pretests“** vor dem Einsatz des Fragebogens bei einer größeren Stichprobe als Standard (Kaase, 1999, S. 49). Man versteht unter einem *Pretest* die Erprobung eines Fragebogens unter Bedingungen, die möglichst weitgehend der Untersuchungssituation entsprechen. Üblich ist hier eine Anzahl von Auskunftspersonen, die etwa zwischen 20 und 50 liegt. Sudman und Blair (1998, S. 301) nennen drei Aufgaben von Pretests, wobei die erstgenannte herausgehobene Bedeutung hat:

- **Identifizierung von Unklarheiten, Fehlern, Missverständnissen** etc. bei Frageformulierungen, Antwortkategorien und Erläuterungen zum Fragebogen.
- Realitätsnahe Abschätzung der **Interviewdauer** für Planung des Interviewereinsatzes.
- Feststellung, ob die Antworten auf die verschiedenen Fragen **Varianz** haben. Fragen, die von praktisch allen Auskunftspersonen einheitlich beantwortet werden (z. B. „Sehen Sie gelegentlich fern?“), bringen keine Information und können eliminiert werden.

Hintergrundinformation

Die American Association for Public Opinion Research fasst in ihren „Codes of Ethics“ (zitiert nach Kaase, 1999, S. 133) die Bedeutung von Pretests zusammen:

> „Qualitativ hochwertige Umfragen sehen grundsätzlich ein angemessenes finanzielles und zeitliches Budget zum Pretesten von Fragebogen und Feldarbeit vor. Pretests sind der einzige Weg, um herauszufinden, ob alles „funktioniert“, insbesondere dann, wenn in einer Umfrage neue Techniken oder neue Fragebatterien zum Einsatz kommen sollen. Weil es kaum möglich ist, alle potenziellen Missverständnisse oder Verzerrungen der verschiedenen

> Fragen und Verfahren vorherzusehen, ist es für eine gut geplante Umfrage existenziell, Vorkehrungen für Pretests zu treffen. Alle Fragen sollten vorgetestet werden, um sicher zu stellen, dass sie von den Befragten verstanden werden, dass sie von den Interviewern ordentlich abgearbeitet werden können und dass sie die Antwortbereitschaft nicht negativ beeinflussen."

In der Analyse empirischer Studien, die im Journal of the Academy of Marketing Science publiziert worden sind, durch Hulland et al. (2018) ließen nur 58,4 % der Studien erkennen, dass Pretests durchgeführt worden waren. Vor diesem Hintergrund betonen Hulland et al. ebenfalls die Notwendigkeit von Pretests.

Zu den üblichen Methoden bei **Pretests** gehören (Groves et al., 2009, S. 259 ff.):

- **Experten-Gespräche:** Spezialisten für den Untersuchungsgegenstand oder für Befragungstechniken beurteilen die Fragen hinsichtlich inhaltlicher Angemessenheit bzw. Verständlichkeit, Eindeutigkeit, Beantwortbarkeit, Antworthemmnissen und Leichtigkeit der Handhabung.
- **Gruppendiskussionen:** Diskussion der Fragebogen-Entwickler mit Angehörigen der Zielgruppe über Inhalt des Fragebogens und entsprechenden Sprachgebrauch.
- **Feld-Pretests:** Relativ kleine Zahl von Probe-Interviews mit (nicht unbedingt repräsentativ) ausgewählten Angehörigen der Zielgruppe mit dem entworfenen Fragebogen und anschließenden Auswertungsgesprächen mit diesen Auskunftspersonen.
- **Split-Ballot-Experimente:** Verwendung unterschiedlicher Frageformen in (jeweils repräsentativen) Teilstichproben und Schlussweise von Ergebnisunterschieden auf Wirkungen der verschiedenen Frageformen (→ experimentelles Design).
- **„Kognitive Interviews":** Es geht hier um eine spezielle Form von Interviews zur Analyse der kognitiven Prozesse (z. B. Verständnis, Erinnerung), die bei der Auskunftsperson während des Interviews ablaufen. Dazu werden u. a. verbale Protokolle („Methode des lauten Denkens") angefertigt, wobei während des Befragungsprozesses von der Auskunftsperson möglichst alle dabei verwendeten Informationen, Erinnerungen, Schlussfolgerungen etc. laut ausgesprochen und vom Forscher aufgezeichnet und analysiert werden.

Im Hinblick auf Einzelheiten zur Durchführung von Pretests sei hier auf Groves et al. (2009, S. 259 ff.), Madans et al. (2011) und auf Presser et al. (2004) verwiesen.

Empfehlung zur Fragebogenentwicklung

In den bisherigen Abschnitten sind verschiedene Schritte bei der Entwicklung eines Fragebogens dargestellt worden. Wegen der großen Fehlerempfindlichkeit von Befragungen erfordert dieser Prozess besondere Sorgfalt und mehrere Überprüfungen. Als Maßstab dafür, aber auch als Illustration des typischen Aufwandes bei der Fragebogenentwicklung, sei hier in Anlehnung an Bradburn et al. (2004,

S. 315 f.) eine entsprechende Empfehlung in insgesamt 17 (!) Schritten wiedergegeben:

1. Bestimmung der bei der Umfrage zu erhebenden Informationen.
2. Suche nach entsprechenden bereits in früheren Umfragen verwendeten Fragen.
3. Entwurf (Formulierung) neuer Fragen bzw. Überarbeitung früher verwendeter Fragen.
4. Festlegung der Fragereihenfolge.
5. Entwurf der äußeren Gestaltung des Fragebogens.
6. Entwurf von Codierungsregeln (numerische Verschlüsselung der Antworten für die Computer-Eingabe, siehe Abschn. 5.3).
7. Erster Pretest (bei Kollegen, Bekannten etc.).
8. Überarbeitung des Fragebogens aufgrund der Pretest-Ergebnisse und erneuter (kleiner) Pretest.
9. Vorbereitung von Interviewer-Anweisungen zur Durchführung der Interviews.
10. Pretest bei 20 bis 50 Personen aus der Zielgruppe der Befragung.
11. Sammlung und Auswertung von Kommentaren der Interviewer und der Auskunftspersonen zum Fragebogen.
12. Eliminierung von Fragen ohne Varianz der Antworten oder mit geringer Validität.
13. Überarbeitung von Fragen, bei denen es Probleme gibt.
14. Erneuter Pretest.
15. Erarbeitung der Endfassung der Interviewer-Anweisungen.
16. Beobachtung von Problemen bei der Interviewer-Schulung und während der Anlaufphase der Umfrage sowie gegebenenfalls Vornahme entsprechender Korrekturen.
17. Auswertung von Interviewer-Kommentaren und sonstigen Erfahrungen nach der Untersuchung zur Verwendung bei künftigen Untersuchungen.

Hier sei hervorgehoben, dass die vorstehende Empfehlung zur Fragebogenentwicklung insgesamt vier (!) Pretests vorsieht.

4.3.4 Kommunikationsformen bei Befragungen

4.3.4.1 Überblick

In der Markt- und Sozialforschung haben sich inzwischen vier Hauptformen der Kommunikation mit Auskunftspersonen etabliert, die hier charakterisiert und im Folgenden jeweils kurz diskutiert werden. Es werden dabei unterschieden:

- **Persönliche bzw. mündliche Befragung (Face-to-Face Interviews)**
  Beim persönlichen Interview wird die Auskunftsperson in der Regel in ihrer Wohnung oder am Arbeitsplatz vom Interviewer aufgesucht (manchmal auch auf der Straße oder auf Messen angesprochen) oder in ein spezielles Studio gebeten und auf der Grundlage eines gedruckten oder im Laptop gespeicherten Fragebogens befragt.
  Ist der Fragebogen nur in Papierform vorhanden, so spricht man auch von *PAPI-Interviews,* wobei die Abkürzung für „Paper and Pencil Interviews" steht. Heute werden die weitaus meisten persönlichen Interviews mit dem Computer durchgeführt. Diese werden auch als *CAPI-Interviews* bezeichnet. Dabei steht die Abkürzung für Computer Assisted Personal Interview.
- **Schriftliche Befragung (Mail)**
  Eine schriftliche Befragung vollzieht sich meist so, dass der Auskunftsperson auf postalischem Wege der Fragebogen zugesandt wird, den diese dann ausfüllen und zurückschicken soll.
- **Telefonische Befragung**
  Bei dieser Art der Befragung übermittelt ein Interviewer per Telefon Fragen an die Auskunftsperson, die er in der Regel von einem Computer-Bildschirm (*CATI*/Computer Assisted Telephone Interview) abliest, nimmt die Antworten auf und macht sofort die entsprechenden Eingaben am Computer.
- **Online-Befragung**
  Bei Online-Befragungen (die auch als *CAWI*-Interviews bezeichnet werden, wobei das Kürzel für Computer Assisted Web Interview steht) wird ein Fragebogen, der auf einem Server gespeichert ist, über das Internet ausgefüllt oder es wird ein Fragebogen per E-Mail an die Auskunftsperson verschickt, beantwortet und dann wieder per E-Mail zurückgesandt.
  Eine Sonderform stellt die Befragung mit mobilen Apps dar, bei der die Fragen auf dem Mobiltelefon beantwortet werden. Dies bietet sich insbesondere an, wenn Aktivitäten außer Haus bewertet werden sollen, wie z. B. die Zufriedenheit mit Hotelübernachtungen.

Nach Kaase (1999, S. 46) lassen sich diese Kommunikationsformen repräsentativ angelegter Befragungen in der in Abb. 4.15 dargestellten Weise untergliedern, wobei der vorhandene bzw. nicht vorhandene Einsatz von Interviewern im Vordergrund steht, was wegen den damit verbundenen Kosten sinnvoll ist.

In den folgenden Abschnitten sollen die hier vorstehend gekennzeichneten Formen der Befragung jeweils kurz diskutiert, ihre spezifischen Probleme genannt und Lösungsansätze für einige dieser Probleme aufgezeigt werden. Dabei wird jeweils anhand der folgenden wichtigen Kriterien vorgegangen:

- Repräsentanz
- Qualität und Umfang der zu erhebenden Daten
- Organisatorischer, zeitlicher und finanzieller Aufwand

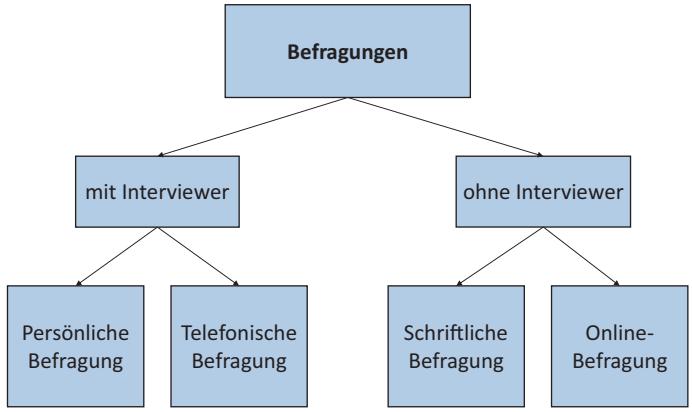


Abb. 4.15 Kommunikationsformen von Befragungen. (Nach Kaase, 1999, S. 46)

Im Zusammenhang mit der **Repräsentanz** geht es zunächst darum, ob das wichtigste Prinzip bei der Auswahl einer repräsentativen Stichprobe, nämlich die Berechenbarkeit der Chance für alle Elemente der Grundgesamtheit in die Stichprobe zu gelangen (siehe Abschn. 4.2), durch die Kommunikationsform der Befragung beeinflusst wird. Weiterhin ist das Problem der mangelnden Stichprobenausschöpfung dadurch, dass ein Teil der in der Stichprobe befindlichen Zielpersonen den Fragebogen nicht erhält (z. B. wegen einer Adressenänderung) oder die Antwort verweigert, zu erörtern. Wenn man die Stichprobe in die Gruppe der Antwortenden und derer, die nicht antworten (Verweigerer u. a.), unterteilt, kann es oftmals sein, dass sich diese beiden Gruppen hinsichtlich für die Untersuchung relevanter Merkmale unterscheiden. Mangelnde Repräsentanz der Ergebnisse beeinträchtigt direkt deren Generalisierbarkeit (siehe Abschn. 2.2.3).

Beispiel

Beispielsweise müsste man damit rechnen, dass bei einer Untersuchung über soziale Kontakte die Ergebnisse bei mangelhafter Stichprobenausschöpfung dadurch verzerrt werden, dass kontaktarme Leute besonders häufig Auskünfte verweigern und damit unterrepräsentiert sind. Dieses Problem ist durch eine Vergrößerung der Stichprobe natürlich nicht zu lösen. Allein eine möglichst weitgehende Ausschöpfung der Stichprobe kann die Repräsentanz einer solchen Umfrage verbessern. ◀

Ein weiteres Repräsentanzproblem bei Befragungen besteht darin zu gewährleisten, dass die für die Stichprobe ausgewählte Person den Fragebogen auch selbst beantwortet **(Identitätsproblem)**. Analog zur Problematik geringerer Stichproben-Ausschöpfung können die Ergebnisse einer Umfrage verzerrt werden, wenn die in der Stichprobe

ausgewählte nicht mit der antwortenden Person identisch ist (beispielsweise, weil der Interviewer an Stelle der eigentlich zu befragenden Person, die er nicht angetroffen hat, deren Ehepartner befragt), da damit die Zufälligkeit der Auswahl der Erhebungselemente nicht mehr gegeben wäre.

Bei der Erörterung der Beeinflussung der **Qualität und des Umfangs der zu erhebenden Daten** durch die Kommunikationsform der Befragung steht die Frage im Vordergrund, in welchem Maße das Instrumentarium der Frageformulierung (z. B. Verwendung von Vorlagen, Abbildungen oder Videos) und des Fragebogenaufbaus (z. B. Steuerung der Fragenreihenfolge) einsetzbar ist. Wichtig ist weiter, wie lang die maximale Interviewdauer ist, die sich mit einer Methode erzielen lässt. Weiterhin sind in diesem Zusammenhang auch mögliche Beeinflussungen des Antwortverhaltens, die mit der Form der Befragung zusammenhängen, zu diskutieren.

Für die Praxis entscheidend ist auch der **organisatorische, zeitliche und finanzielle Aufwand,** der mit einer Methode verbunden ist. Ein Teil des mit einer Umfrage verbundenen Aufwandes – wie z. B. der für die Fragebogenerstellung, die Datenanalyse und die Berichterstattung – ist von der Kommunikationsart weitgehend unabhängig und braucht deswegen nicht besonders beachtet zu werden. Deutliche Unterschiede zeigen sich vor allem beim Erhebungsaufwand pro Interview und in geringerem Maße bei den mit der Stichprobenziehung zusammenhängenden Tätigkeiten. Dies hat selbstverständlich auch Einfluss auf die Felddauer, also auf die Zeitspanne, die ein Institut üblicherweise braucht, um die Daten zu erheben.

Der Aufwand wird dabei vor allem dadurch beeinflusst, ob Interviewer eingesetzt werden müssen und ob diese reisen müssen. Der Aufbau, die Schulung und Pflege eines Interviewerstabs sind sehr aufwendig und teuer. Einen Anhaltspunkt für die verschiedenen Kosten gibt die ESOMAR Pricing Study (ESOMAR, 2018). Dabei wurde ein Standardprojekt zugrunde gelegt, nämlich eine 15-minütige repräsentative Befragung von 500 regelmäßigen Nutzern von Schokolade. Dabei ergaben sich für Deutschland folgende Preise (ESOMAR, 2018, S. 36).

- Persönlich (CAPI): 23.543 €
- Telefonisch (CATI): 19.213 €
- Online: 10.192 €

Die telefonische Befragung war also rund 18 % billiger als die persönliche Befragung, die online Befragung sogar 57 % preisgünstiger.

4.3.4.2 Persönliche/mündliche Befragung

Bei der persönlichen Befragung muss zunächst festgelegt werden, wo befragt werden soll. **Auf der Straße oder in einem Geschäft** lassen sich nur kurze Interviews bis ca. 8 min Dauer realisieren. Außerdem können bevölkerungsrepräsentative Stichproben so nur sehr eingeschränkt realisiert werden. Ebenso ist bei Befragungen **im Studio** die

Repräsentativität deutlich eingeschränkt, auch wenn zur besseren geografischen Verteilung mehrere Studios angemietet werden. Studiointerviews werden vor allem angewendet, wenn das Befragungsmaterial nicht zum Haushalt transportiert werden kann, z. B. wenn verschiedene mögliche Neuprodukte miteinander verglichen werden sollen. Studiobefragungen erlauben ähnlich lange Interviews, wie Befragungen in der Wohnung der Befragten (sogen. **Inhome-Befragungen**). Inhome Befragungen bieten darüber hinaus durch die Anwendung des Random-Route-Verfahrens (vgl. Abschn. 4.2.3) ein hohes Maß an möglicher Repräsentativität.

Die Repräsentativität wird jedoch auch bei der persönlichen Befragung durch die sinkende Ausschöpfung beeinträchtigt, die bei längeren Interviews heute oft unter 40 % liegt. So lag die Ausschöpfungsquote bei der mit hoher Qualität durchgeführten Umfrage „Allbus“ im Jahr 2018 bei 32,4 % (vgl. Gesis, 2018). Erfahrungen der Institute zeigen, dass vor allem Angehörige der höheren sozialen Schichten besonders häufig persönliche Interviews zu Hause ablehnen, wohl auch aus Angst vor Kriminalität.

Das **Identitätsproblem** stellt sich bei mündlichen Befragungen im Zusammenhang mit der Qualität des Interviewerstabs. Bei absolut zuverlässigen Interviewern, die so lange Wiederholungsbesuche machen, bis sie eine vorgegebene Zielperson tatsächlich antreffen, zeigt sich dieses Problem natürlich nicht. In der Praxis wird man aber damit rechnen müssen, dass mancher Interviewer der Versuchung nicht widerstehen kann, eine andere Person aus dem gleichen Haushalt oder eine sonstige „ähnliche“ Person zu befragen, um einen Wiederholungsbesuch zu vermeiden. Das kann so weit gehen, dass Interviews vollständig gefälscht werden. Zur Sicherung der Identität von zu befragender und antwortender Person können entsprechende Interviewerkontrollen durch telefonische Nachbefragungen vorgenommen werden.

Der entscheidende Vorteil der mündlichen Umfrage liegt in der **Qualität der erhobenen Daten.** Diese Form der Befragung gestattet den Einsatz des gesamten Instrumentariums der Frageformulierung und der Fragebogengestaltung, da die Befragungssituation vom Interviewer den Anweisungen des Untersuchungsleiters entsprechend gestaltet werden kann. Dadurch kann die Einhaltung einer für den Untersuchungszweck notwendigen Fragenreihenfolge garantiert werden. Es können bei der Befragung Vorlagen (Bilder, Töne, kurze Filmausschnitte und Texte oder auch Produkte) verwendet werden. Der Interviewer kann bei komplexen Fragen Hilfen geben und gegebenenfalls auch die Informationen, die das Interview liefert, durch eigene Beobachtungen ergänzen (z. B. „Person wohnt in freistehendem Einfamilienhaus“).

Ein weiterer wichtiger Vorteil ist auch die **Interviewlänge,** die mit dieser Befragungsform erzielt werden kann. Durch die vielfältigen Möglichkeiten der Fragebogengestaltung kann man eine mündliche Befragung recht abwechslungsreich anlegen. In Verbindung damit ermöglicht es die motivierende Anwesenheit eines Interviewers bzw. einer Interviewerin, Befragungsdauern von bis zu ca. 50 min zu realisieren.

Die Anwesenheit eines Interviewers oder einer Interviewerin birgt aber auch die Gefahr, dass dadurch Verzerrungen entstehen. Dieser sogen. **Interviewer-Bias** kann einerseits dadurch entstehen, dass der Interviewer durch seine Persönlichkeit (extremes Bei-

spiel: Farbiger führt Interviews über Rassendiskriminierung durch) oder sein Verhalten (z. B. Auftreten, persönliche Bemerkungen) das Antwortverhalten der Auskunftsperson beeinflusst. Andererseits kann – insbesondere dann, wenn der Interviewer selbst eine prononcierte Meinung zum Untersuchungsgegenstand hat – das Phänomen der selektiven Wahrnehmung auftreten. Es handelt sich dabei um eine meist unbewusste Tendenz des Interviewers, die gegebenen Antworten in Richtung auf seine eigenen Erwartungen verfälscht aufzunehmen. Mittel zur Erreichung eines möglichst neutralen Verhaltens des Interviewers sind vor allem in der weitgehenden Standardisierung des Fragebogens, in der klaren personellen Trennung von Untersuchungsanlage und Interviewdurchführung und in der zweckmäßigen Interviewerauswahl und -schulung zu sehen.

Persönliche Interviews werden heute meist nicht mehr mit Papierfragebögen, sondern mit Computerunterstützung als **CAPI-Interviews** („CAPI“ steht für Computer Assisted Personal Interviewing) durchgeführt. Bei einem CAPI-Interview erscheinen die Fragen und die Antwortkategorien auf dem Bildschirm eines Laptop- oder Tablet-Computers und werden vom Interviewer der Auskunftsperson vorgelesen oder auch gezeigt. Deren Antwort wird vom Interviewer über eine Tastatur oder mit einem Stift auf dem berührungsempfindlichen Display sofort in den Rechner eingegeben und dieser präsentiert die nächste Frage auf dem Bildschirm, wobei die Einhaltung der manchmal recht komplizierten Verzweigungslogik („Filterfragen“) automatisch gesteuert wird.

Nachteilig ist, dass der Fragebogen entsprechend programmiert werden muss und dass die Interviewer mit Laptop- oder Tablet-Computer ausgestattet werden müssen. Dem stehen gravierende Vorteile von CAPI-Interviews im Vergleich zur herkömmlichen persönlichen Befragung mit Papierfragebögen (auch **PAPI** genannt, für Paper and Pencil Interview) gegenüber:

- Durch die direkte Eingabe der Antworten in den Rechner entfällt der häufig zeitaufwendige Prozess der Codierung (siehe Abschn. 4.5.3) und Fehlerkontrolle.
- Bei richtig programmierten Fragebögen sind Filterfehler während des Interviews ausgeschlossen.
- Unmittelbar nach der Eingabe einer Antwort können Fehlerkontrollen und gegebenenfalls Korrekturen vorgenommen werden. In Verbindung mit dem Wegfall von Fehlermöglichkeiten bei gesonderter Codierung und Eingabe erhöht dies die Qualität der Daten beträchtlich.
- CAPI-Systeme erlauben eine weitgehende Individualisierung der Befragung. Das bezieht sich nicht nur auf am Computer mögliche komplizierte Filteranweisungen, indem z. B. Informationen, die zu Beginn eingegeben wurden, im weiteren Verlauf des Interviews verwendet werden können. Beispielsweise kann zu Beginn nach einer präferierten Marke gefragt werden und dieser Markenname dann in alle Fragetexte, die sich darauf beziehen, vom Computer eingesetzt werden.
- Besonders wichtig ist, dass bei der Vorlage von langen Listen die Reihenfolge der Antwortvorgaben zufällig erfolgen kann (Randomisierung), wodurch der Einfluss der Reihenfolge neutralisiert wird.

Nachteil der persönlichen Befragung sind vor allem die hohen Kosten (s. oben) sowie der hohe Zeitbedarf. Vor allem diese Nachteile sind es, welche dazu führen, dass die persönliche Befragung, die bis 1995 die am meisten genutzte Interviewform war, immer seltener durchgeführt wird. Laut ADM waren 2018 nur etwa 23 % aller von den Mitgliedsinstituten durchgeführte Interviews persönliche Interviews. Zwanzig Jahre zuvor waren es noch 39 % (vgl. ADM, 2020).

4.3.4.3 Schriftliche Befragung

Neben der Zusendung des standardisierten Fragebogens durch die Post **(Mail)** gibt es auch noch die Möglichkeit, dass der Fragebogen verteilt und zurückgesandt bzw. wieder eingesammelt wird. Die letztere Möglichkeit findet z. B. bei Veranstaltungen (auch bei Lehrveranstaltungen der Universitäten und Hochschulen) oder nach Besuchen z. B. von Museen oder auch Hotels, Anwendung. Dies soll hier aber nicht weiter thematisiert werden.

Zusammen mit dem Anschreiben wird stets ein erklärendes und motivierendes **Begleitschreiben** sowie ein Freiumschlag für den Rückversand des ausgefüllten Fragebogens versendet. Das Begleitschreiben muss über den Verantwortlichen der Befragung sowie die Einhaltung des Datenschutzes informieren. Auch sollte dargestellt werden, warum das Ausfüllen des Fragebogens wichtig ist. Wichtig ist auch noch, einen spätesten Rücksendetermin zu nennen, der ausreichend, aber auch nicht zu viel Zeit zum Ausfüllen lässt, also etwa eine Woche.

Der **Fragebogen** selbst muss einfach und klar sein. Die Möglichkeit einer Rückfrage bei einer Interviewerin entfällt ja. Es ist auch davon auszugehen, dass der Fragebogen erst ganz durchgelesen und dann ausgefüllt wird. Dadurch ist es z. B. ausgeschlossen, erst nach der ungestützten Markenbekanntheit zu fragen (z. B. welche Automarken kennen Sie?) und anschließend nach der gestützten Markenbekanntheit, bei der Marken vorgegeben werden und nach deren Bekanntheit gefragt wird. Wichtig ist auch eine ansprechende Gestaltung des Fragebogens, die dem Befragten Wertigkeit signalisiert. Die Dauer der Befragung ist auf etwa 8 min begrenzt. Bilder können begrenzt eingesetzt werden, andere Stimuli wie Videos, Töne oder Produkte sind nicht möglich.

Empfehlung für schriftliche Befragungen

Dillman et al. (2009, S. 234 ff.) geben u. a. die folgenden Empfehlungen für die Gestaltung schriftlicher Befragungen mit dem Ziel einer hohen Rücklaufquote:

- Alle Kontakte zur Auskunftsperson möglichst persönlich gestalten.
- Gutschein für eine Belohnung mit dem Fragebogen versenden.
- Bei mehreren Kontakten (Anschreiben, Mahnung etc.) unterschiedliche Gestaltungsformen wählen.
- Sorgfältige zeitliche Abstimmung der verschiedenen Kontaktversuche.
- Sinnvolle Wahl des Versandzeitpunkts.
- Verwechslungsmöglichkeit der Befragungsunterlagen mit Werbebriefen vermeiden.

Wichtigste Voraussetzung dieser Methode ist das Vorhandensein von Adressen der Personen der Grundgesamtheit. Dies erklärt, warum die Methode vor allem zur Messung der Kundenzufriedenheit in den Branchen eingesetzt wird, in denen die Adressen der Kunden bekannt sind, wie z. B. Banken, Autohäuser oder Telekommunikation.

Die **Repräsentanz** einer schriftlichen Befragung wird sehr stark von der **Rücksendequote** bestimmt. Diese hängt ganz wesentlich vom Thema ab. Sie kann von 1 % z. B. bei einer Befragung eines Telekommunikationsanbieters bis über 60 % bei einer Kundenzufriedenheitsstudie eines Krankenhauses gehen. Zu ihrer Steigerung empfiehlt es sich, kurz nach dem Rücksendetermin an die Zielpersonen, die noch nicht zurückgeschickt haben, ein Erinnerungsschreiben zu senden.

Die Rücksendequote bestimmt letztlich auch über die **Kosten** der Methode im Vergleich zu anderen Verfahren. Müssen im Fall einer 1 %igen Ausschöpfung 100.000 Fragebögen versendet werden, um 1000 ausgefüllte Fragebögen zu erhalten, so kann eine telefonische oder gar persönliche Befragung günstiger sein. Bei üblichen Rücksendequoten von 10 bis 20 % ist die schriftliche Befragung jedoch kostengünstig.

Ein weiteres Problem für die Repräsentanz einer postalischen Befragung besteht darin, dass nicht garantiert ist, dass die für die Stichprobe ausgewählte Person den Fragebogen auch tatsächlich selbst ausfüllt **(Identitätsproblem)**.

Insgesamt hat die schriftliche Befragung so erhebliche Nachteile, dass sie in den letzten Jahren sehr an Bedeutung verloren hat und vor allem durch die Online-Umfrage ersetzt wurde. Nur noch ca. 5 % aller Interviews werden schriftlich durchgeführt (vgl. ADM, 2020).

4.3.4.4 Telefonische Befragung

Die telefonische Befragung installierte sich in den 1970er Jahren in Folge der damals schnell steigenden Ausstattung der Haushalte mit Festnetztelefonen. Diese **Telefondichte,** die in den 2000er Jahren weit über 90 % lag, ist inzwischen jedoch rückläufig, weil immer mehr Haushalte ausschließlich mobil telefonisch erreichbar sind. Zum Jahresbeginn 2019 waren nur noch 86,4 % aller Haushalte per Festnetztelefon erreichbar. Der Rückgang im Vergleich zum Vorjahr betrug 4,5 Prozentpunkte (vgl. Statistisches Bundesamt, 2019a, S. 12),

Positiv für die **Repräsentanz** der Telefonbefragung ist dagegen, dass mit dem Gabler-Häder-Verfahren (vgl. Abschn. 4.2.3) eine Methode zur Verfügung steht, mit der sehr gut eine Zufallsauswahl realisiert werden kann. Auch sind einfach wiederholte Kontaktversuche möglich, wenn eine Zielperson nicht erreicht wurde. Viele Telefonstudios sind zu diesem Zweck mit **Autodialing** ausgestattet. Dabei werden in einem Computer die anzurufenden Telefonnummern gespeichert. Der Computer ruft an. Hebt niemand ab, so wird der Anrufversuch später wiederholt. Wird abgehoben, so wird der Anruf an einen freie Interviewpatz durchgestellt.

Nachteilig ist dagegen, dass die **Ausschöpfung** der telefonischen Befragung deutlich geringer ist als bei der persönlichen Befragung und in der Regel bei unter 20 % liegt. Dies liegt vor allem daran, dass die Interviewer nur eingeschränkt zur Teilnahme

motivieren können. Das für die Repräsentanz einer Untersuchung ebenfalls bedeutsame Identitätsproblem stellt sich weniger als bei der mündlichen Befragung, da durch die Möglichkeit des Mithörens der Interviews durch die Studioleitung eine größere Kontrolle erreichbar ist.

Die **Qualität und Menge der erhebbaren Daten** ist bei der telefonischen Befragung geringer als bei persönlichen Interviews. Durch den Computereinsatz (**CATI** = Computer Assisted Telephone Interviewing) sind zwar auch komplexe Filterführungen möglich. Die geringere Motivation durch die Interviewer führt aber dazu, dass die Dauer eines Telefoninterviews 20 min nicht übersteigen sollte. Auch komplexe Fragen oder lange Auswahllisten, für die bei persönlichen Interviews ein Hilfsblatt übergeben werden kann, müssen vermieden werden. Schließlich ist es auch nicht möglich, mit Abbildungen, Produktproben oder Videos zu arbeiten. Positiv ist dagegen zu werten, dass durch die Verringerung des Kontakts zwischen Interviewer und Interviewtem auch die Gefahr des Interviewereinflusses geringer ist. Schließlich ist die Kontrolle der Interviewer besonders gut möglich, da die Studioleitung jederzeit mithören kann.

Die **Kosten** der telefonischen Befragung sind geringer als bei der persönlichen Befragung, da die Reisezeiten und -kosten für die Interviewer entfallen (s. oben). Dies führt auch dazu, dass die Ergebnisse deutlich **schneller** als bei persönlichen Interviews zur Verfügung stehen. Die Kosten sind jedoch noch immer deutlich höher als bei Online-Umfragen.

Insgesamt hat die telefonische Befragung so viele Vorteile, dass sie aktuell nach der Onlinebefragung die am zweitmeisten genutzte Methode der Datenerhebung ist, allerdings mit großem und wachsendem Abstand (2018: 28 % Telefon vs. 40 % Online, vgl. ADM, 2020).

Eine Sonderform der telefonischen Befragung ist die **Befragung per Mobiltelefon,** die insbesondere deswegen zunimmt, weil immer mehr Menschen nur oder vor allem mobil erreichbar sind. Dabei ist darauf zu achten, dass dem Befragten keine Kosten entstehen (z. B. durch Roaming, wenn er sich außerhalb seines Heimatlandes aufhält). Auch lassen sich in der Regel nur kürzere Interviews erzielen. Inzwischen wird versucht, durch eine Kombination von per Festnetz mit per Mobilfunk durchgeführten Interviews eine bessere Repräsentanz sicherzustellen (sogenannte *„Dual-Frame-Ansätze"*). Die dabei entstehenden Probleme und Lösungsmöglichkeiten beleuchtet ein von der Marktforschungsorganisation ADM herausgegebener Forschungsbericht (vgl. ADM, 2012).

4.3.4.5 Online-Befragung

Seit etwa Mitte der 90er Jahre haben sich durch die starke Ausbreitung der Internet-Nutzung in den hoch entwickelten Ländern auch die Möglichkeiten zur Nutzung dieses Mediums für Befragungen wesentlich verbessert.

Aktuell sind ca. 87 % aller Haushalte in Deutschland mit stationärem Internet erreichbar. Damit ist die Ausstattung mit Internetzugang inzwischen besser als die Ausstattung mit Festnetztelefon. Das Kriterium der **Repräsentanz** ist von daher mit nur kleinen Einschränkungen gegeben. Zwar sind insbesondere ältere potenzielle Auskunftspersonen

und Angehörige niedriger sozialer Schichten mit Online-Befragungen schlecht erreichbar. Da diese Zielgruppen aber weniger im Fokus des Marketings sind, wird dies oft in Kauf genommen. Man spricht dann auch davon, dass die Stichprobe repräsentativ für die Online-Bevölkerung ist. Wenn das nicht akzeptabel ist, dann kann dieser Mangel durch eine anders (z. B. telefonisch) erhobene Zufallsstichprobe behoben werden (sogenannte „Mixed Mode-Surveys“).

Die Repräsentanz von Online-Stichproben wird jedoch dadurch deutlich eingeschränkt, dass es kein dem Gabler-Häder- oder dem Random-Route-Verfahren vergleichbare Methode zu Erzeugung von zufälligen Online-Stichproben gibt. Schon weil es kein Verzeichnis der Email-Adressen gibt und weil viele Menschen mehrere Email-Adressen haben, ist es nicht möglich, eine Auswahlgrundlage für Zufallsstichproben zu schaffen. Von daher sind auch **keine Zufallsstichproben** für allgemeine Bevölkerungsumfragen wie bei telefonischen oder persönlichen Befragungen möglich.

Für die Stichprobenziehung bei allgemeinen Bevölkerungsumfragen kommen insbesondere **Online-Panels** zum Einsatz. 77 % aller Online-Interviews wurden 2018 auf diese Weise rekrutiert (vgl. zu diesem und den folgenden Anteilen ADM, 2020). Online-Panels sind zum Teil sehr große Pools von Personen (bis mehrere Millionen), die ihre Bereitschaft bereit erklärt haben, sich befragen zu lassen. Sie werden in der Regel von großen Marktforschungsunternehmen aufgebaut und betrieben, Anbieter können aber auch andere Firmen wie z. B. Betreiber Sozialer Medien sein. Die wesentlichen soziodemografischen Merkmale der teilnehmenden Personen werden zu Beginn der Mitarbeit erhoben und gespeichert, so dass sich sehr einfach eine für die Online-Bevölkerung repräsentative Stichprobe ziehen lässt. Bei einer Befragung erhalten die ausgewählten Personen eine Mail mit einem Link zum Fragebogen. Dieser führt zu einem Fragebogen, der online ausgefüllt wird. Die Beantwortung des Fragebogens wird in der Regel vergütet. In Online-Panels lassen sich Interviewdauern von bis zu 30 min erzielen, wenn die Vergütung entsprechend hoch ist.

Eine weitere Methode besteht darin, dass auf einer Website ein Fenster geöffnet wird, das zur Teilnahme an einer Befragung auffordert. Dieses **River-Sampling** wird vor allem genutzt, wenn die Benutzer einer Website kurz (maximal ca. 8 min) zu befragen sind, so z. B. wenn die Benutzerfreundlichkeit eines Konfigurators auf der Website eines Automobilherstellers beurteilt werden soll. Etwa 6 % der Online-Umfragen wurden mit diesem Verfahren durchgeführt.

Eine dritte Möglichkeit (14 % aller Online-Umfragen) wird eingesetzt, wenn **Email-Listen der Grundgesamtheit** vorliegen. Dies ist beispielsweise bei Mitarbeiterbefragungen oder bei der Befragung von Studierenden der Fall. In diesem Fall wird häufig auch eine Befragung der Grundgesamtheit angestrebt (z. B. bei Befragungen zur Mitarbeiterzufriedenheit) oder es wird eine Stichprobe gezogen, was sehr einfach nach dem im Abschn. 4.2.3 dargestellten Verfahren möglich ist. Hier ersetzt heute die Online-Befragung meist die früher häufig durchgeführte schriftliche Befragung.

Schließlich gibt es noch einen sehr kleinen Teil (ca. 2 %) der Online-Umfragen, der **Offline rekrutiert** wurde. Dabei kann die Anwerbung der Interviewten z. B. telefonisch

erfolgen. Denjenigen, die zum Interview bereit sind, wird dann ein Link zu einem Online-Fragebogen zugesandt. Auf diese Weise können die Möglichkeiten der Telefonstichprobe zur Bildung einer Zufallsauswahl mit den Möglichkeiten der Online-Befragung wie z. B. die Verwendung von Bildern, Videos etc. kombiniert werden. Nachteilig sind bei dieser Methode die hohen Kosten und der hohe organisatorische und zeitliche Aufwand.

Damit die Befragten das Interview auch vollständig durchführen, ist es wichtig, dieses ansprechend zu gestalten. Detaillierte Empfehlungen zur Gestaltung von Online-Befragungen findet man in den „Standards zur Qualitätssicherung für Online-Befragungen" des „Arbeitskreises Deutscher Markt- und Sozialforschungsinstitute" (www.adm-ev.de). Die Empfehlungen von Dillman et al. (2009, S. 271 ff.) für Online-Befragungen entsprechen sinngemäß weitgehend den Empfehlungen dieser Autoren für schriftliche Befragungen (s. o.).

Hinsichtlich der **Qualität der Daten** bieten Web-Befragungen gute Bedingungen. Die Möglichkeiten zur Steuerung von Verzweigungen des Fragebogens entsprechen denen von mündlicher und telefonischer computergestützter Befragung. Darüber hinaus können auch Bilder, Video-Sequenzen oder akustische Signale (z. B. Melodien) verwendet werden. Ferner erlaubt die Web-Befragung auch einige verdeckte Arten der Datenerhebung, wie z. B. die Messung der Reaktionszeiten bei den einzelnen Antworten. Befragte, welche den Fragebogen einfach schnell durchklicken, lassen sich dadurch feststellen, dass die Dauer für die Beantwortung der Fragen sehr kurz ist. Allerdings wird es immer wichtiger, die Fragebögen so zu gestalten, dass sie auch am Mobiltelefon beantwortet werden können, da immer mehr Panelteilnehmer die Fragebögen mobil beantworten wollen. Dies schränkt die Gestaltungsmöglichkeiten ein.

Ein weiterer Vorteil von Online-Befragungen besteht in der kurzen **Untersuchungsdauer,** ähnlich wie bei telefonischen Befragungen. Die **Kosten der Untersuchung** sind in der Regel deutlich geringer als bei persönlicher und auch bei telefonischer Befragung, da weder Interviewer-Honorare noch Reise- oder Studiokosten anfallen. Es lassen sich sogar – nach angemessener sprachlicher Anpassung der Online-Fragebögen – Befragungen mit weltweit verstreuten Zielgruppen durchführen, was bei den anderen Methoden kaum möglich ist.

Inzwischen ist die Realisierung „elektronischer Fragebögen" durch entsprechende Standard-Software (z. B. „Unipark" oder „SurveyMonkey") mit begrenztem technischem Know-how und zu überschaubaren Kosten möglich geworden (siehe z. B. www.unipark.info oder www.surveymonkey.de). Für die Durchführung wissenschaftlicher Befragungen an Hochschulen bietet SoSci Survey (www.socisurvey.de) ein kostenfreies Instrument zur Erstellung von einfachen Online-Fragebögen an.

Allerdings spricht die besondere Bedeutung von ansprechend gestalteten und abwechslungsreichen Fragebögen gerade für die Online-Forschung eher für eine Programmierung der Fragebögen durch Profis. Denn eine wichtige Motivation für Panelteilnehmer ist der Spaß am Ausfüllen von Fragebögen (Comley, 2007), der naturgemäß bei schlechten, eintönigen und repetitiv gestalteten Fragebögen leidet.

Dabei sollten jedoch die *Nachteile* von Online-Befragungen nicht übersehen werden. Zunächst gibt es keinerlei Kontrolle, wer den Fragebogen ausfüllt. Weiter sind manche Online-Panelisten darauf aus, möglichst viele Vergütungen zu erhalten. Sie melden sich bei mehreren Panelbetreibern. Werden für eine Studie mehrere Panels kombiniert – was insbesondere bei kleinen Zielgruppen häufig geschieht – so kann es sein, dass die gleiche Person den Fragebogen mehrmals ausfüllt. Doch diese Gefahr besteht auch, wenn nur ein Panel genutzt wird, da sich manche Panelteilnehmer mehrmals unter verschiedenen Identitäten anmelden. Es ist auch möglich, dass eine Person die Screeningfragen, mit denen festgestellt wird, ob die Person zur Zielgruppe gehört, bewusst falsch beantwortet, um interviewt zu werden. Dies führt dann natürlich zu verzerrten Daten.

Gut geführte Online-Panels achten daher darauf, dass zwischen zwei Befragungen des gleichen Teilnehmers eine bestimmte Zeit (z. B. sechs Wochen) verstreicht. Doch auch dies kann zu Einschränkungen der Repräsentanz führen. Wurde beispielsweise kurz vor der Befragung eine andere Befragung unter den Fahrern eines bestimmten PKW-Modells durchgeführt, dann befinden sich in der Befragung u. U. keine oder nur sehr wenige Fahrer dieses Modells. Damit können sich trotz guter soziodemografischer Repräsentativität einer Onlinestichprobe nur schwer kontrollierbare Verzerrungen ergeben. Schließlich lassen sich auch in Online-Panels trotz der Tatsache, dass sich die Panelteilnehmer grundsätzlich bereit erklärt haben, Fragebögen auszufüllen, häufig nur geringe Responseraten von 25 bis 30 % erzielen.

Trotzdem überwiegen die Vorteile der Methode, insbesondere die Schnelligkeit, die geringen Kosten und die großen Möglichkeiten der Interviewgestaltung. Diese haben dazu geführt, dass Online-Befragungen in den letzten Jahren zur weitaus wichtigsten Befragungsform überhaupt aufgestiegen sind.

Für die Plattform Amazon Mechanical Turk bestehen mehrere Studien, welche die Validität von Analysen auf Basis prozessgenerierter Daten der Plattform mit anderen webbasierten Designs vergleichen. Mit Blick auf die externe Validität zeigen Levay et al. (2016) und Buhrmester et al. (2011) dass sich die Charakteristika zufällig gezogener Stichproben aus dem Datenbestand von Amazon Mechanical Turk nicht wesentlich von den Charakteristika, welche die US-Bevölkerung repräsentieren, unterscheiden. Die Validität von Untersuchungsergebnissen kann außerdem anhand ihrer Reproduzierbarkeit überprüft werden. Horton (2011) und Paolacci et al. (2010) gelingt es, Ergebnisse von bereits in der Vergangenheit durchgeführten Experimenten mit plattformbasierten Experimenten zu reproduzieren und zeigen, dass bei einer Befragung die Messgenauigkeit einer plattformbasierten Stichprobe mindestens genauso gut ist, wie diejenige einer Stichprobe aus Teilnehmern eines Laborexperiments an einer US-amerikanischen Universität oder einer Stichprobe, die sich aus dem Inserat in einem Internet-Diskussionsforum ergibt.

Mobile Befragungen am Smartphone oder Tablet-Computer waren lange eine Sonderform. Inzwischen ist es jedoch so, dass Online-Fragebögen zunehmend so gestaltet sind, dass sie sowohl am PC als auch am Mobiltelefon beantwortet werden können. Denn dadurch lassen sich insbesondere auch mobile Zielgruppen erreichen, die mit

anderen Methoden nur noch schwer erreichbar sind und es steigt die Antwortbereitschaft, da von den Befragten sonst unproduktive Zeit, z. B. in der U-Bahn genutzt werden kann.

Daneben spielen rein mobil durchgeführte Befragungen eine steigende Rolle. Diese wurden nicht zuletzt durch die schnell gestiegene Ausstattung mit Smartphones möglich. 81,6 % aller Haushalte waren 2019 mit mindestens einem Smartphone ausgestattet (Statistisches Bundesamt, 2019b). Mobile Befragungen haben potenziell eine Reihe von *spezifischen Vorteilen:*

- Zusammen mit der Feststellung des Aufenthaltsortes des Befragten (entweder über Funkzellenortung oder genauer über GPS) lassen sich Befragungen abhängig vom Ort generieren wie z. B. die Zufriedenheit mit einem Freizeitpark nach einstündigem Aufenthalt dort.
- Es können mobile Online-Panels aufgebaut werden, mit deren Mitgliedern vereinbart wird, dass sie unmittelbar nach vorher definierten Ereignissen (z. B. Besuch eines Restaurants) von sich aus eine Kurzbefragung initiieren. Ein Vorteil einer solchen Befragung gegenüber der traditionellen Methode von zu Hause aus ist, dass das Erlebte noch frisch im Gedächtnis ist.

Dem stehen jedoch auch *Nachteile* entgegen:

- Die Ausstattung mit Smartphones ist noch nicht so umfassend, dass für alle Zielgruppen hinreichend repräsentative Stichproben möglich sind.
- Mobile Endgeräte arbeiten mit unterschiedlichen Betriebssystemen, deren wichtigste derzeit Android von Google und IOS von Apple sind. Das bedeutet, dass der Fragebogen mehrmals programmiert werden muss.
- Darüber hinaus haben mobile Endgeräte unterschiedliche Bildschirmgrößen, welche eine Anpassung des Fragebogens erforderlich machen können.

Insgesamt spielen solche Befragungen noch eine begrenzte, aber mit der zunehmenden Penetration internetfähiger mobiler Endgeräte und der Verlagerung der Internetnutzung von zu Hause auf außer Haus eine schnell steigende Rolle. 2018 wurden 8 % aller Interviews mobil durchgeführt (vgl. ESOMAR, 2019, S. 154).

4.3.4.6 Zusammenfassung

Gerade zu den unterschiedlichen Kommunikationsformen bei Befragungen sind im Lauf der letzten Jahrzehnte umfangreiche Erfahrungen gesammelt und zahlreiche Studien durchgeführt worden. Hier ist es nicht leicht, eine entsprechende Übersicht zu gewinnen und zu behalten. Es kommt hinzu, dass manche Entwicklungen in Technik und Gesellschaft die Anwendungsbedingungen für diese Kommunikationsformen im Zeitablauf verändern. Hier seien nur folgende Gesichtspunkte genannt:

- Steigende *Telefondichte* ermöglichte repräsentativ angelegte telefonische Umfragen.

	Persönliche / mündliche Befragung	Schriftliche Befragung	Telefonische Befragung	Online-Befragung
Repräsentanz	+	-	+	o
Qualität der Daten	+	o	o	+
Aufwand	-	+	o	+
Untersuchungsdauer	-	-	+	+

Abb. 4.16 Stärken (+) und Schwächen (−) verschiedener Kommunikationsformen bei Befragungen

- *Ausbreitung von Computern* und deren gesunkene Kosten ermöglichten computergestützte Interviews (z. B. CATI).
- *Ausweitung der Internet-Nutzung* erlaubt bei immer mehr Zielgruppen (annähernd) repräsentative Online-Befragungen.
- Zunehmendes *Desinteresse* und *Misstrauen* gegenüber Umfragen führt zu erhöhten Verweigerungsraten.

Trotz aller Komplexität der Materie und Vorsicht gegenüber Vereinfachungen der methodischen Probleme sei hier der Versuch unternommen, wichtige Stärken und Schwächen der verschiedenen Kommunikationsformen von Befragungen in einer einfachen Tabelle zusammenzufassen. Dazu werden in der Abb. 4.16 die jeweiligen Stärken jeweils durch ein „+“, die Schwächen durch ein „−“ gekennzeichnet. Eine „0“ ist eingetragen, wenn keine eindeutigen und ausgeprägten Stärken und Schwächen festgestellt werden können. Als Kriterien werden die Gesichtspunkte verwendet, die die Diskussion in den vorangehenden Abschnitten bestimmt haben. Bei der konkreten Auswahl einer Erhebungsmethode für eine Problemstellung ist insbesondere die Qualität noch weiter aufzufächern. Aspekte wie mögliche Dauer des Interviews, ob Stimuli wie Produkte, Töne, Bilder oder Videos notwendig sind, die mögliche Komplexität der Fragefolgen, die mögliche Kontrolle des Interviewvorgangs – das alles sind Aspekte, die zu berücksichtigen sind.

Durch die Abb. 4.17 soll noch ein Eindruck von der Bedeutung und der Entwicklung der verschiedenen Kommunikationsformen für die Marktforschungspraxis vermittelt werden. Man erkennt daran u. a., dass die Online-Befragung inzwischen die mit großem

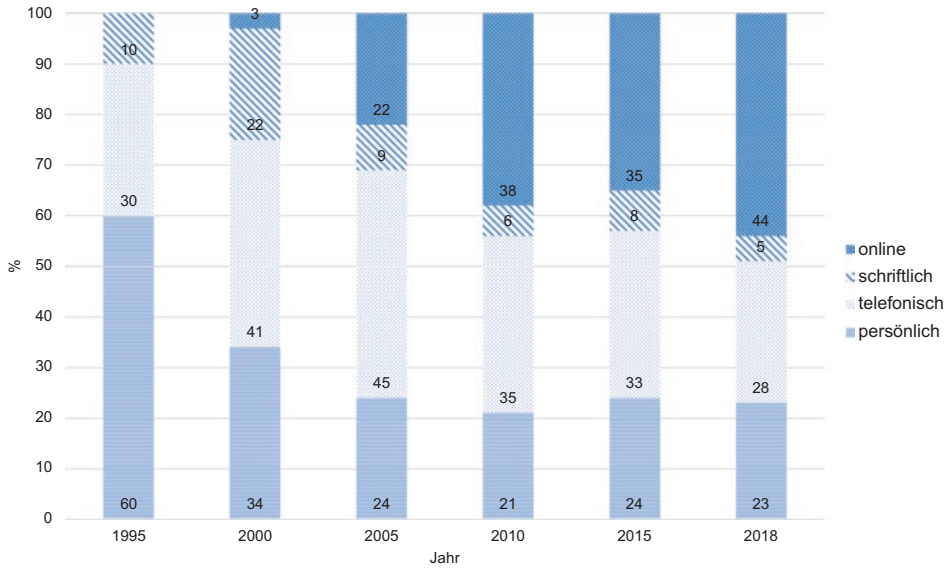


Abb. 4.17 Quantitative Interviews nach Befragungsart in Deutschland. (Quelle: ADM, 2020)

Abstand meist genutzte Erhebungsmethode ist, gefolgt von der telefonischen Befragung. Die persönliche Befragung bleibt stabil, die schriftliche Befragung führt nur noch ein Nischendasein.

Die Daten beruhen auf der Erhebung des Arbeitskreises Deutscher Marktforschungsinstitute (ADM) bei seinen Mitgliedsinstituten. Insgesamt ist der Online-Anteil in der Realität wohl noch größer, da wegen der für persönliche und telefonische Interviews vom ADM zur Verfügung gestellten Infrastruktur vor allem solche Institute Mitglied sind, welche diese Befragungsarten einsetzen. Institute, die nur Online befragen, sind dagegen häufiger nicht Mitglied beim ADM.

4.4 Beobachtungsverfahren

4.4.1 Kennzeichnung von Beobachtungen

Die Beobachtung ist eine Technik der Datenerhebung, die auf eine Kommunikation zwischen Erhebendem und Auskunftspersonen durch Fragen und Antworten verzichtet. Dabei wird so verfahren, dass die zu untersuchenden Gegebenheiten und Verhaltensweisen direkt erfasst werden. Zur Beobachtung gehören die Auswahl, die Aufzeichnung und die Codierung von Verhaltensweisen und anderen interessierenden Phänomenen (z. B. Wege im Supermarkt, Ablauf von Verkaufsgesprächen).

▶ **Definition** Döring und Bortz (2016, S. 324) **definieren Beobachtungen** folgendermaßen:

> „Unter einer wissenschaftlichen Beobachtung (‚scientific observation') versteht man die zielgerichtete, systematische und regelgeleitete Erfassung, Dokumentation und Interpretation von Merkmalen, Ereignissen oder Verhaltensweisen mithilfe menschlicher Sinnesorgane und/oder technischer Sensoren zum Zeitpunkt ihres Auftretens."

Auch Beobachtungsverfahren unterliegen natürlich den üblichen *Anforderungen an wissenschaftliche Methodik*. Deswegen ist auch hier die Ausrichtung auf ein definiertes Forschungsziel und die Konzentration auf die entsprechenden Ausschnitte aus der Realität (siehe Abschn. 2.2.2) erforderlich. Systematische Planung, Durchführung und Dokumentation einer Untersuchung sind ebenso erforderlich wie die sorgfältige Prüfung von Reliabilität, Validität und Generalisierbarkeit der erhobenen Daten (Hoyle et al., 2002, S. 366).

Gegenstand von Beobachtungen können Eigenschaften und Verhaltensweisen von Personen (z. B. Wege einer Person in einem Supermarkt) und von Gruppen von Personen (z. B. Kommunikationsprozess in einer Gruppe) sein. In diesem Zusammenhang ist hervorzuheben, dass die Untersuchung verbalen Verhaltens (z. B. Ablauf eines Verkaufsgesprächs) durchaus Gegenstand einer Beobachtung sein kann. Die Daten kommen aber nicht durch die Reaktion auf Fragen eines Interviewers zu Stande. Der Prozess der Datenerhebung durch entsprechend geschulte BeobachterInnen ähnelt durchaus dem Antwortprozess bei Befragungen (siehe Abschn. 4.3.1.1), hat aber natürlich seine Spezifika (Jaccard & Jacoby, 2020, S. 423 ff.):

- Zunächst müssen die BeobachterInnen ein klares *Verständnis* des Verhaltens bzw. der Situation, der die Beobachtung gilt, entwickeln.
- Im zweiten Schritt muss *eingeschätzt* werden, inwieweit das Verhalten bzw. die Situation bestimmten (theoretischen) Vorgaben oder Kategorien entspricht (z. B. „längeres Verweilen", „aggressive Rückfrage" oder „Kaufentscheidung für Produkt XY").
- Letztlich muss diese Einschätzung *kommuniziert* werden, z. B. durch eine verbale Angabe im Erhebungsbogen oder durch Ankreuzen auf einer Ratingskala.

Natürlich gibt es auch die Möglichkeit, für bestimmte Untersuchungen Befragung und Beobachtung zu kombinieren. Aus praktischen Gründen muss man sich bei der Beobachtung auf Sachverhalte beschränken, die hinreichend oft auftreten und nicht zu lange dauern. Beispielsweise wird man kaum die Ursachen für tödliche Verkehrsunfälle durch Beobachtung des fließenden Verkehrs, wo glücklicherweise nur selten Unfälle zu sehen sind, klären oder den Einfluss der schulischen Erziehung auf das Kommunikationsverhalten von Erwachsenen durch die langjährige Beobachtung der Entwicklung von ausgewählten Personen überprüfen können. In diesen Fällen wäre der Einsatz von Beobachtungsverfahren zu aufwendig und/oder zu langwierig.

Beispiel

Hier einige Beispiele für Gegenstände der Beobachtung in der Marktforschung u. a. nach Zikmund (1997, S. 251):

- **Informationsverhalten,** z. B. Clickstream-Analyse bei der Internet-Nutzung (siehe Abschn. 4.4.3)
- **Physische Bewegung,** z. B. Wege von Konsumenten in einem Supermarkt
- **Verbales Verhalten,** z. B. Kommentare von Passagieren, die am Schalter einer Fluggesellschaft warten, oder Inhalt von Verkaufsgesprächen
- **Andere Ausdrucksformen**, z. B. Gesichtsausdruck, Stimmfrequenz, Körpersprache
- **Räumliche Beziehungen,** z. B. Standort von Betrachtern bei Displays, Schaufenstern, Plakaten etc.
- **Abläufe,** z. B. Wartezeit und Verzehrdauer bei McDonald's, Blickverlaufsregistrierung
- **Physische Objekte,** z. B. im Haushalt vorhandene Markenartikel
- **Verbale und bildliche Inhalte,** z. B. von Anzeigen oder Protokolle von Verkaufsgesprächen
- **„Spuren" von Verhalten,** z. B. Vorräte in Haushalten, die auf frühere Käufe schließen lassen, oder Abnutzung von Fußbodenbelägen in verschiedenen Zonen eines Supermarkts
- **Tatsächliche Einkäufe,** aufgezeichnet durch Scanner-Daten (z. B. zur Analyse des Kaufverbunds, also der Kombinationen gleichzeitig gekaufter Artikel) ◀

Im Vergleich zu den weitaus häufiger angewandten Befragungsverfahren haben Beobachtungen einige Vor- und Nachteile. Zunächst zu den **Vorteilen:**

- Vermeidung einiger Validitätsprobleme der Befragung (z. B. begrenztes Erinnerungsvermögen, bewusst verzerrte Angaben, mangelnde Auskunftsfähigkeit)
- Unabhängigkeit von Bereitschaft und Fähigkeit der beobachteten Personen zur Verbalisierung der Angaben
- Bei Beobachtungen kann man *reales Verhalten* erfassen, nicht nur Aussagen über (angebliches) früheres oder beabsichtigtes Verhalten. Beobachtungen lassen sich auch in realen Kaufsituationen durchführen.
- Unreflektiertes und daher kaum verbalisierbares Verhalten (z. B. bei Impulskäufen oder das Blickverhalten bei der Betrachtung von Anzeigen) kann erfasst werden.
- Bei Befragungen ist mit Ergebnisverzerrungen durch Einflüsse der Frageformulierung und des Interviewers zu rechnen (→ Reaktivität). Bei Beobachtungsverfahren können derartige Verzerrungen vermieden werden.

Andererseits sind Beobachtungsverfahren natürlich auch nicht frei von Fehlermöglichkeiten. Im Vergleich mit Befragungsverfahren sind sie vor allem mit dem gravierenden Nachteil behaftet, bei weitem nicht so breit einsetzbar zu sein wie diese. Daneben haben sie im Vergleich zu Befragungen auch die folgenden **Nachteile:**

- Gründe für beobachtete Verhaltensweisen sind in der Regel nicht erkennbar. Deswegen entsteht hinsichtlich der Erklärung von Untersuchungsergebnissen ein Defizit.
- Die Datenerhebung und die Aussagemöglichkeiten sind weitgehend auf den Beobachtungszeitpunkt begrenzt.
- Die Untersuchungsgegenstände sind auf Bereiche begrenzt, die in beschränkter Zeit beobachtbar sind. So wäre es kaum möglich, die langwierige Entstehung und Festigung einer Markenbindung in einem entsprechend langen Beobachtungszeitraum zu verfolgen.
- Die Ziehung von Zufallsstichproben bereitet besondere Probleme, weil die Ergebnisse oftmals vom gewählten Untersuchungszeitpunkt abhängen (siehe auch Abschn. 4.2),
- Die Datenerhebung wird durch die begrenzten Fähigkeiten der Beobachter eingeschränkt, in kurzer Zeit eine Vielzahl von Verhaltensweisen korrekt wahrzunehmen, bestimmten Kategorien zuzuordnen und zu erfassen.

4.4.2 Auswahlprobleme und Gestaltungsmöglichkeiten bei Beobachtungen

Bei der Beobachtung stellen sich dem Forscher meist komplexere **Auswahlprobleme** als bei der Befragung. Zunächst müssen die für den Untersuchungsgegenstand als relevant erscheinenden *Merkmale*, die in die Erhebung einbezogen werden sollen, ausgewählt werden. Analog zur Befragung, bei der festgelegt werden muss, mit welchen konkreten Fragestellungen welche Teilaspekte des Untersuchungsthemas geklärt werden können, muss bei der Beobachtung entschieden werden, welche Einzelmerkmale beobachtet werden sollen, da es in der Regel unmöglich ist, alle beobachtbaren Tatbestände zu erfassen. Hier stellt sich das Problem der Validität einer Messung: Kann man von dem beobachteten Merkmal tatsächlich auf den untersuchten Sachverhalt schließen; kann man beispielsweise daraus, dass in einem Haushalt ein bestimmtes Produkt vorrätig ist, darauf schließen, dass dieses Produkt dort auch regelmäßig verwendet wird?

Darüber hinaus sind

- *Beobachtungseinheiten* (z. B. Personen, Geschäfte) und
- *Erhebungszeiträume/-punkte* (wegen der Abhängigkeit der Ergebnisse von der Erhebungszeit)
- auszuwählen.

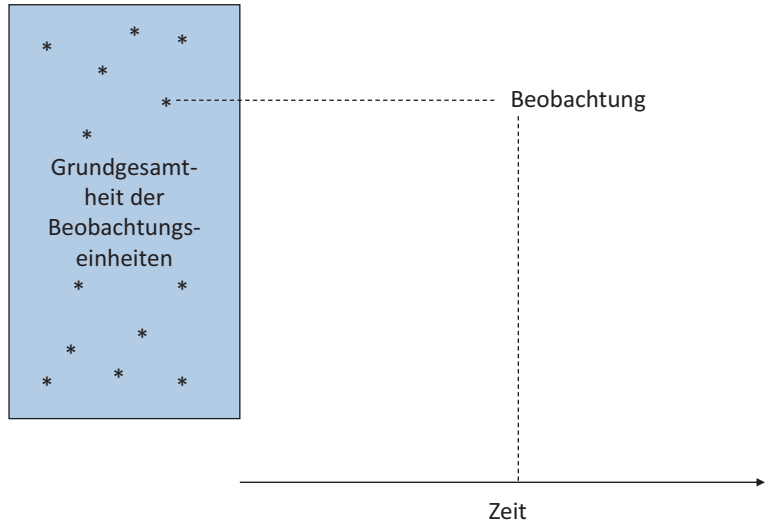


Abb. 4.18 Auswahl von Beobachtungseinheiten und -zeitpunkten

Zunächst zu den **Beobachtungseinheiten.** Während sich eine Befragung immer an eine Zielperson richtet, können hier auch Personengruppen und Sachen untersucht werden. Dabei sind die Auswahlprobleme aber komplexer als bei der relativ übersichtlichen Ziehung einer Personenstichprobe, da sowohl die Definition der Grundgesamtheit als auch die Abgrenzung der Untersuchungseinheiten Schwierigkeiten bereiten können (Beispiel: Untersuchung von Produkt-Empfehlungen im Kollegenkreis).

Mit Hilfe von Befragungsverfahren kann man durch vergangenheits- oder zukunftsbezogene Fragen auch Aussagen machen, die vom Befragungszeitpunkt (weitgehend) unabhängig sind. Dagegen ist das Ergebnis einer Beobachtung in der Regel auf die Erhebungszeit bezogen. Es stellt sich also zusätzlich das Problem, diese *Zeit* festzulegen. Dafür sind Anfang und Ende des Beobachtungszeitraumes, die Zahl der Beobachtungsintervalle und die Dauer der Einzelbeobachtung zu bestimmen. Abb. 4.18 illustriert das zweifache Auswahlproblem.

Im Hinblick auf **Gestaltungsmöglichkeiten von Beobachtungen** werden einige wesentliche Alternativen unterschieden, die im Folgenden kurz umrissen seien.

Standardisierte und nicht-standardisierte Beobachtung Die standardisierte Beobachtung ist durch vorherige Festlegung der zu untersuchenden Einzel-Merkmale und der entsprechenden Erfassungskategorien gekennzeichnet, während sich die nicht-standardisierte Beobachtung als eher impressionistische Informationssammlung darstellt. Letztere ist hauptsächlich für explorative Vorstudien einsetzbar, während die standardisierte Beobachtung in der Regel eine weitgehende Konkretisierung des Untersuchungsgegenstandes voraussetzt. Die eindeutige Zuordnung von beobachteten Tatbeständen zu festgelegten Erfassungskategorien erfordert auch eine angemessene Schulung der für die

Beobachtung eingesetzten Personen oder aber den Einsatz standardisierter technischer Erfassungsmethoden. Für deskriptive Untersuchungen kommt fast ausschließlich die standardisierte Beobachtung in Betracht.

Teilnehmende und nicht-teilnehmende Beobachtung Bei der teilnehmenden Beobachtung ist der Beobachter selbst in den untersuchten Prozess einbezogen. Bei einer Untersuchung über das Verhalten von Verkäufern kann sich das z. B. so vollziehen, dass der Beobachter selbst im Verkauf tätig wird und in dieser Rolle versucht, seine Beobachtungen vorzunehmen. Berühmte Beispiele teilnehmender Beobachtung finden sich auch in der kulturanthropologischen Forschung, wo die Rolle des Beobachters oftmals nicht getarnt sein konnte wie das bei anderen Untersuchungen (trotz forschungsethischer Bedenken) in Hinblick auf unverzerrte Ergebnisse oft als zweckmäßig erscheint. Die teilnehmende Beobachtung hat gegenüber der nicht-teilnehmenden den Vorteil eines engen und tiefen Kontaktes zur Untersuchungseinheit (z. B. einer Gruppe), bringt aber Probleme hinsichtlich der Genauigkeit der Ergebnisse mit sich, die durch die beschränkte Aufnahmekapazität des Beobachters, seine Doppelrolle beim beobachtenden Prozess (Teilnahme und Distanziertheit) und durch den Einfluss des Beobachters auf den zu untersuchenden Vorgang verursacht sein können.

Offene und getarnte Beobachtung Die offene Beobachtung ist für die zu beobachtende Person oder Gruppe erkennbar. Diese Form ist zwar aus forschungsethischen Gründen zu bevorzugen, führt aber häufig zu dem Problem, dass das Bewusstsein, beobachtet zu werden, eine unerwünschte Verhaltensänderung der betroffenen Personen mit sich bringt. So könnte beispielsweise eine Untersuchung über den Einfluss von Verkaufsgesprächen auf den Verkaufserfolg eines Produktes daran scheitern, dass die beobachteten Verkäufer für den Untersuchungszeitraum von ihrem üblichen Verhalten abweichen und sich an bestimmte Normen, die sie für erwünscht halten, anpassen. Die Mittel der Tarnung von Beobachtungen sind recht vielfältig. Sie können die Person des Beobachters betreffen, indem z. B. bei Kaufverhaltensuntersuchungen in einem Supermarkt der Beobachter durch eine dem sonstigen Personal entsprechende Arbeitskleidung getarnt wird. Andererseits können auch technische Hilfsmittel wie z. B. versteckte Kameras oder einseitig durchsichtige Spiegel zur Verdeckung von Beobachtungen dienen.

Feld- und Labor-Beobachtung Wird die Situation, in der die Beobachtung durchgeführt wird, vom Forscher geschaffen oder beeinflusst, so spricht man von einer Labor-Beobachtung, im anderen Falle – bei einer unbeeinflussten Situation – von einer Feld-Beobachtung. Wenn für die Beobachtung spezielle technische Geräte (z. B. Hautgalvanometer zur Messung von emotionalen Reaktionen) notwendig sind, ist man in der Regel auf Labor-Untersuchungen angewiesen. Meist ist bei einer Labor-Beobachtung eine Tarnung nicht oder nur eingeschränkt möglich. Ein Beispiel für eine aktuelle Entwicklung bei Labor-Beobachtungen ist das sogenannte „Virtual Shopping“, bei dem die Einkaufssituation in einer Verkaufsstätte sehr realistisch am Bildschirm simuliert wird und das

Verhalten von Versuchspersonen bei ihren (simulierten) Einkäufen sehr umfassend und detailliert aufgezeichnet werden kann (vgl. Burke, 1996).

Bei den vorstehenden Unterscheidungen sind implizit zwei Aspekte angesprochen worden, die vor allem für Beobachtungen weitreichende Bedeutung haben: Die Fragen der Aufdringlichkeit und der Reaktivität von Messungen. Der Gesichtspunkt der **Aufdringlichkeit** bezieht sich darauf, dass die Testperson *bemerkt*, dass eine Beobachtung stattfindet. Bei der **Reaktivität** liegt der Fokus eher auf der *Beeinflussung* des Verhaltens der beobachteten Personen. Als Beispiel für eine Datenerhebung, die nicht (in diesem Sinne) aufdringlich und deswegen auch nicht reaktiv ist, sei die Aufzeichnung (und spätere Auswertung) von **Scanner-Daten** genannt. Dabei werden die auf fast allen erhältlichen Produkten angebrachten Strichcodes an der Kasse automatisch erfasst und erlauben eine genaue und umfassende Auswertung aller Einkaufsvorgänge. Hier bemerken die betroffenen Konsumenten nicht, dass Daten erhoben werden. Das ohnehin vor dieser Datenerhebung an der Scanner-Kasse sich vollziehende Einkaufsverhalten kann also davon nicht beeinflusst werden. Dagegen ist die Beobachtung der Gehirnaktivität im Magnetresonanztomographen eine extrem aufdringliche Beobachtungsform: Die untersuchten Personen müssen sich in einer engen Röhre, die zudem noch laute Geräusche von sich gibt, fixieren lassen.

Für die **Datenerfassung bei Beobachtungen** gibt es im Allgemeinen die Wege des Einsatzes von Beobachtern und die Verwendung von technischen Geräten. Bei der Auswahl der Personen, die den Erhebungsvorgang einer Beobachtung durchführen, gilt analog zur Interviewerauswahl, bei der die Interviewer möglichst wenig über die Ziele einer Untersuchung wissen sollen, das Prinzip der Trennung von Untersuchungsanlage und Datenerhebung, um (z. B. im Hinblick auf selektive Wahrnehmung) unverzerrte Ergebnisse zu erhalten. Zur Erfassung der beobachteten Vorgänge und Merkmale gibt es mehrere Möglichkeiten. Die erste besteht in der parallel laufenden oder nachträglichen Protokollierung. Hier sind oft recht enge Grenzen der Aufnahme- und Wiedergabefähigkeit des Beobachters gesetzt. Eine Vereinfachung besteht darin, ein übersichtliches und eindeutiges Kategoriensystem mit entsprechender Software-Unterstützung (Hoyle et al., 2002, S. 376) zu verwenden. Weiterhin bestehen technische Möglichkeiten (Video-Aufzeichnung), die eine fast vollständige Aufzeichnung des beobachteten Vorganges zur nachträglichen Protokollierung und deswegen eine durch die Aufnahmekapazität der beobachteten Person weniger behinderte Analyse erlauben. Bestimmte Beobachtungen lassen sich nur mit Hilfe spezieller technischer Geräte (z. B. Augenkameras) vornehmen; einige Beispiele dafür werden im Abschn. 4.4.4 kurz vorgestellt.

4.4.3 Clickstream-Analyse zur Beobachtung der Internet-Nutzung

Die schnell fortschreitende Ausbreitung der Internet-Nutzung – nicht zuletzt im kommerziellen Bereich – ist allgemein bekannt und bedarf keiner Erläuterung. Im Abschn. 4.3.4.5 ist bereits die Nutzung des Internets für Befragungen angesprochen wor-

den. Hier geht es vor allem um die Beobachtung des Verhaltens von Internet-Nutzern bei der Informationssuche, beim Kontakt mit Werbung und beim Einkauf. So haben inzwischen (2019) 94 % der deutschen Haushalte einen Internet-Zugang und 84 % der Internet-NutzerInnen haben im Jahre 2019 auch Waren oder Dienstleistungen über das Internet eingekauft (Quelle: Statistisches Bundesamt; www.destatis.de). Angesichts der Größe entsprechender Märkte und der Bedeutung dieses Kommunikationsmediums verwundert es natürlich nicht, dass in Praxis und Wissenschaft entsprechendes Interesse an der Erhebung von Daten über das Verhalten von Internet-Nutzern entstanden ist.

Nun kann man leicht nachvollziehen, dass eine Datenerhebung über Befragungen im Hinblick auf Internet-Nutzung wohl wenig erfolgversprechend wäre. Wer ist schon willens und in der Lage, sich Einzelheiten seines Suchverhaltens (z. B. Adressen, Reihenfolge und Betrachtungsdauer der besuchten Websites) zu merken und darüber bei einer Befragung korrekt zu berichten (siehe dazu auch Abschn. 4.3.1.2)? Dagegen bietet die Kommunikation über einen Computer ideale Voraussetzungen, um im Sinne der in Abschn. 4.4.1 skizzierten Überlegungen entsprechende Beobachtungen vorzunehmen, aufzuzeichnen und zu analysieren. Im Mittelpunkt stehen dabei so sogenannte **Clickstreams.** Man versteht darunter den Weg, den eine Internet-Nutzerin – eben mittels diverser „Mausklicks" – auf einer oder mehreren Websites nimmt (Bucklin et al., 2002). Typische Arten von erhobenen Daten sind dabei u. a. (Bucklin & Sismeiro, 2009):

- Namen der besuchten Websites und der betrachteten Seiten
- Reihenfolge des Zugriffs
- Verweildauer auf Websites oder einzelnen Unter-Seiten
- Angeklickte Werbebanner
- Abgegebene Angebote und gekaufte Produkte

Mit Blick auf **Arten von Datenquellen** unterscheidet man hauptsächlich zwischen auf eine Website bezogenen („site-centric") und auf die Nutzer bezogenen („user-centric") Quellen. Beide Arten von Quellen können zur wechselseitigen Validierung verwendet werden. Hinsichtlich der erstgenannten Art, also einer bestimmten Website, werden dazu beim entsprechenden Server alle Eingaben und Abrufe der einzelnen Nutzer detailliert aufgezeichnet, was Analysen bei einer großen Zahl von Nutzern erlaubt, aber nur bezogen auf die jeweilige Website mit ihren Unter-Seiten. Beim Zugriff auf Cookies ist auch das Verhalten von Nutzern im Hinblick auf eine Website über einen längeren Zeitraum grundsätzlich beobachtbar.

Umfassendere Informationen über das Suchverhalten eines Nutzers (user-centric) bekäme man, wenn man die beim Internet-Provider darüber gespeicherten Daten auswerten könnte, was aber aus leicht nachvollziehbaren Gründen des Datenschutzes in der Regel nicht möglich ist. In der Praxis werden für derartige Untersuchungszwecke spezielle **Internetnutzungspanels** (siehe dazu Kap. 5) verwendet. Die an einem solchen Panel teilnehmenden Personen gestatten es den Panel-Anbietern, die verschiedenen Daten (s. o.) über alle ihre Zugriffe zu den verschiedensten Websites aufzuzeichnen

und zu analysieren. Internationale Anbieter solcher Dienstleistungen und Informationen sind z. B. ComScore (www.comscore.com) und Nielsen NetRatings (www.nielsen-netratings.com). In der Regel sind besondere Bemühungen und Anreize (z. B. Prämien und Gewinnspiele) nötig, um TeilnehmerInnen für eine dauerhafte Beteiligung zu gewinnen. Bucklin und Sismeiro (2009) verweisen aber auch auf einige Einschränkungen der Aussagekraft dieser Daten. So kann es sein, dass trotz großer Teilnehmerzahlen bei einem Panel die Zahl der Nutzer spezieller Websites dennoch gering und die Aussagekraft entsprechender Daten begrenzt ist. Wenn ein Nutzer mehrere Computer nutzt (z. B. daheim und am Arbeitsplatz) oder wenn ein Computer von mehreren Personen genutzt wird, dann ist der Rückschluss auf die Internet-Zugriffe *einer Person* problematisch. Besonders die Erfassung des Surfverhaltens am Arbeitsplatz ist in der Regel nicht möglich, da eine solche Software von den Arbeitgebern i. d. R. nicht akzeptiert wird.

Ebenfalls von Bucklin und Sismeiro (2009) stammt ein Überblick über wichtige Anwendungsgebiete von Clickstream-Analysen sowie eine Literaturübersicht zu einschlägigen Studien. Diese Autoren unterscheiden dabei die folgenden Anwendungsbereiche:

- **Suchverhalten im Internet:** Auswahl von Websites; Suche auf einer Website; Design von Websites; Lernprozesse bei der Internet-Nutzung
- **Online-Werbung:** Kontakthäufigkeiten; Klickraten bei Bannerwerbung; Werbung mit Suchmaschinen
- **Online-Shopping und E-Commerce:** Prognose von Kaufverhalten; Entstehung von Consideration Sets; Verhalten bei Online-Auktionen; Electronic Word of Mouth

Hintergrundinformation

Bucklin et al. (2002, S. 256) kennzeichnen Relevanz und Perspektive der Clickstream-Analyse:

> „Die Ausbreitung des Internets und die detaillierten Informationen zum Nutzungsverlauf durch Clickstream-Daten eröffnen einen riesigen und vielversprechenden Bereich von Forschungsmöglichkeiten für die empirische Analyse von Kaufentscheidungen. Scannerdaten führten zu einer Forschungsrichtung der Modellierung von Kaufentscheidungen, die zwanzig Jahre nach ihrem Beginn immer noch Beiträge zum Erkenntniszuwachs leistet. Jedoch waren diese Daten begrenzt auf die relativ starre Auswahlsituation in Supermärkten und verpackte Konsumgüter des laufenden Bedarfs. Andererseits werden Clickstream-Daten den Forschern das Fenster zum Auswahlprozess vor dem Kauf öffnen, die Untersuchung vieler Arten von Waren und Dienstleistungen über verpackte Konsumgüter hinaus ermöglichen und Analysen des Wahlverhaltens in dynamischen Situationen erlauben, in denen die Anbieter Anreize speziell auf einzelne Kunden ausrichten können."

4.4.4 Implizite Methoden und Consumer Neuroscience

Für zahlreiche Phänomene des Konsumentenverhaltens, die mit Hilfe der Marktforschung untersucht werden sollen, sind weder Befragungen noch Beobachtungen, die

durch bloßen Augenschein erfolgen, geeignete Erhebungsmethoden. Hier wird es notwendig, Methoden der Marktforschung einzusetzen, die darauf abzielen, unbewusste, automatische oder schwer verbalisierbare Reaktionen, Meinungen und Präferenzen von Verbrauchern zu erfassen; sogenannte „implizite Forschungsmethoden".

Dies kann vor allem durch den Einsatz von neurologischen Messungen, Blickverfolgung oder physiologischen Reaktionen wie Hautleitfähigkeit erreicht werden. Die Nutzung moderner Technologien wie Eye-Tracking, EEG (Elektroenzephalographie) und Facial Coding (Kodierung zur Beschreibung von Gesichtsausdrücken) gehört zu den impliziten Methoden und ermöglicht den Einsatz dieser oft erst in einem für die Marktforschung relevanten Ausmaß. Solche Technologien ermöglichen eine genauere Erfassung von physiologischen und neurologischen Reaktionen. Implizite Methoden fokussieren sich oft auf die Erfassung von Emotionen und unbewussten Assoziationen, die bei traditionellen Umfragen oder Fokusgruppen leicht übersehen werden könnten. Dazu zwei kurze Beispiele aus dem Bereich der Kommunikationstests (siehe hierzu auch Abschn. 6.6):

- Für Untersuchungen zur Gestaltung von Werbemitteln (z. B. Anzeigen) ist es von größtem Interesse zu wissen, welche Teile einer Anzeige überhaupt beachtet werden und in welcher Reihenfolge und Intensität (z. B. Dauer des Blickkontakts) dies geschieht. Auch solche Daten lassen sich durch Befragungen und einfache Beobachtungen nicht gewinnen, weil eben niemand diese Vorgänge selbst hinreichend genau wahrnimmt und erinnert. Daher hat sich seit einigen Jahrzehnten über verschiedene technische Entwicklungsstufen die Methode der Blickregistrierung (Eye Tracking) mit speziellen Kameras zur Aufzeichnung des Blickverlaufs durchgesetzt (siehe Kroeber-Riel & Gröppel-Klein, 2019, S. 287 ff.). Ein entsprechendes Anwendungsbeispiel ist im Abschn. 6.6 des vorliegenden Buches dargestellt.
- Um die emotionale Reaktion von Verbrauchern oder Zielgruppen in Bezug auf Produkte, Dienstleistungen oder Marketingkampagnen zu verstehen, werden sogenannte Emotion Analytics-Verfahren eingesetzt. Neben der Auswertung von biometrischen Daten wie Herzfrequenz, Hautleitfähigkeit, Gesichtsausdrücke und Gehirnaktivität kommen dabei verstärkt Methoden der Gesichts- und Mimik-Analyse (sog. Facial Coding) zum Einsatz), über die dann, oftmals mit Hilfe des Einsatzes von KI, auf Emotionen der Probanden geschlossen werden kann. In der Marktforschungspraxis werden diese Verfahren häufig, bspw. zum Test von Werbespots, verwendet. Der Proband sieht sich an einem Bildschirm (z. B. Laptop) einen oder mehrere Werbespots an, während über eine entsprechende Software die emotional aktivierenden Momente identifiziert und interpretiert werden (siehe hierzu Marichalar Quezada et al., 2022). Der dabei verwendete Ansatz des Facial Coding und Tracking geht auf die Arbeiten von Paul Ekman zurück, einem renommierten Psychologen, der sich intensiv mit der Erforschung von Emotionen und nonverbaler Kommunikation beschäftigt hat. Er hat wichtige Beiträge zur Identifizierung und Kategorisierung von Gesichtsausdrücken gemacht, insbesondere im Zusammenhang mit den sechs universellen

Grundemotionen (Freude, Trauer, Angst, Überraschung, Ekel und Wut), um Gesichtsausdrücke zu interpretieren (z. B.: Ekman & Friesen, 1978).

Implizite Methoden versuchen oft, ein ganzheitlicheres Bild zu schaffen, indem sie verschiedene Elemente miteinander kombinieren. Zum Beispiel können Blickverfolgung und emotionale Gesichtsreaktionen zusammen analysiert werden, um ein umfassenderes Verständnis der Konsumentenreaktionen zu gewinnen.

Insgesamt handelt es sich also um einen innovativen Ansatz, um tiefergehende Einblicke in das Verhalten und die Präferenzen der Verbraucher zu erhalten. Gleichzeitig sind solche Verfahren nicht ohne Herausforderungen. Der Einsatz erfordert ein sorgfältiges Design und eine präzise Auswertung, um aussagekräftige Erkenntnisse zu gewinnen. Die Interpretation impliziter Daten ist in der Regel kompliziert, da sie oft mehrdeutige Ergebnisse liefern. Häufig ist deshalb der Einsatz von Fachleuten mit Kenntnissen in Neurowissenschaften, Psychologie und Forschungsmethodik erforderlich, wodurch ein starker Bezug zu einem weiteren innovativen Forschungsansatz, der Consumer Neuroscience, deutlich wird.

Consumer Neuroscience ist ein Teilaspekt der so genannten **Neuroökonomik,** ein Forschungsbereich der Wirtschaftswissenschaften der sich in den letzten ca. 20 Jahren immer stärker entwickelt.

▶ **Definition** Peter Kenning (2014, S. 22) definiert Consumer Neuroscience „als die systematische Integration neurowissenschaftlicher Theorien, Methoden, Konzepte und Erkenntnisse in die Konsumentenverhaltensforschung."

Im Wesentlichen geht es also darum, Erkenntnisse und Methoden der Hirnforschung für Analyse und Verständnis des Konsumentenverhaltens zu nutzen und damit Einsichten zu gewinnen, die (mittelbar) für das Marketing nutzbar sind. Kosslyn (1999) sieht die Schwerpunkte der Hirnforschung bei zwei Arten von Fragestellungen: Ablauf von Informationsverarbeitungsprozessen im menschlichen Hirn und Ursachen/Bedingungen für das Auftreten bestimmter Reaktionen oder Prozesse. Die Beziehung zum Marketing ist hier leicht erkennbar, weil es eben bei der Ausgestaltung der verschiedenen Marketing-Instrumente darum geht, bestimmte Reaktionen bei potenziellen Kunden zu erzielen. Der zweite von Kosslyn genannte Schwerpunkt deutet auch darauf hin, dass bei entsprechenden Untersuchungen experimentelle Designs (siehe Abschn. 6.1) eine wesentliche Rolle spielen, weil typischerweise bestimmte Stimuli (z. B. Anzeigen- oder Packungsentwürfe) eingesetzt werden, um die Reaktion einer Versuchsperson darauf zu messen.

Jeder, der bereits Veröffentlichungen über die Hirnforschung gelesen hat, kann sich leicht vorstellen, dass derartige Untersuchungen sehr spezielle und hoch entwickelte Methoden und Geräte erfordern, die den Rahmen der in den Sozialwissenschaften üblichen Methodenkenntnisse bei weitem sprengen. Nach Kenning (2014, S. 30) „ist es unerlässlich, zumindest Grundkenntnisse im Bereich der Neuroanatomie vorweisen zu können."

Welche Marktforscherin oder welcher BWL-Student hat diese schon? Daneben wird eine besondere Kompetenz verlangt, um Untersuchungen mit Hilfe (teilweise extrem teurer und) komplexer Spezialgeräte durchführen zu können. Als entsprechende Methoden, deren Bezeichnungen schon die Komplexität der verwendeten Geräte und Methoden erahnen lassen, nennt Kenning (2014) u. a.:

- Functional Magnetic Resonance Imaging (fMRI)
- Elektroenzephalographie (EEG)
- Magnetenzephalographie (MEG)
- Nahinfrarotspektroskopie (NIRS)

Beispiel

Ein Beispiel eines Ergebnisses aus dem Gebiet der Consumer Neuroscience soll die Vorgehensweise bei solchen Untersuchungen etwas veranschaulichen. Damit wird gleichzeitig die oben erwähnte experimentelle Vorgehensweise illustriert, bei der unabhängige Variable manipuliert und entsprechende Reaktionen der Versuchspersonen gemessen werden.

Wildner und Jäncke (2010) berichten von einem Experiment, bei dem verschiedene Methoden zur Messung der Markenstärke mit fMRI validiert wurden. Bei der Anlage des Experiments war zu berücksichtigen, dass fMRI die Hirnaktivität mit einer Verzögerung von mehreren Sekunden misst. Denn die Aktivität einer Hirnregion führt zu erhöhtem Sauerstoffverbrauch, was wiederum zum Fluss von sauerstoffgetränktem Blut zu dieser Region führt. Dieser Blutfluss wird durch fMRI sichtbar gemacht. Er erreicht erst ca. 4 bis 10 s nach der Hirnaktivität sein Maximum.

Die Probandinnen (es waren 19 Studentinnen der Universität Zürich) wurden zunächst außerhalb des Scanners zu Schokolade befragt. Dabei wurden drei Marken bestimmt: Die Lieblingsmarke, eine weitere gute Marke und eine gerade noch akzeptable Marke. Weiter wurden drei Methoden zur Bestimmung des Markenwertes angewendet, deren Validität überprüft werden sollten: Als erstes eine Auswahlmethode, bei der die drei Marken mehrmals unter verschiedenen Preiskonstellationen gezeigt wurden und die Person die Marke auswählt, die sie bei der jeweiligen Preissituation kaufen würde. Markenpräferenz zeigt sich hier als Mehrpreisbereitschaft. Die zweite Methode war ein Chipgame, bei der Spielmarken entsprechend der Bevorzugung auf die drei Marken zu verteilen waren. Bei der dritten Methode wurden schließlich die drei Marken mit je 10 Statements z. B. zur Mehrpreisbereitschaft oder zum Markenvertrauen bewertet.

Dann folgte der eigentliche Test im Scanner, der als Glücksspiel angelegt war, bei dem pro Runde jeweils eine Tafel Schokolade gewonnen werden konnte. Bei jeder Runde wurde zunächst die Marke gezeigt, um die es in dieser Runde ging. Dann drehte sich auf dem Bildschirm ein Glücksrad, das nach einigen Sekunden entweder auf „Gewinn“ oder auf „Niete“ stehen blieb. Die anschließende Schwarzphase von 6 s diente der Normalisierung des Blutflusses. Pro Person wurden 40 solcher Durchgänge realisiert.

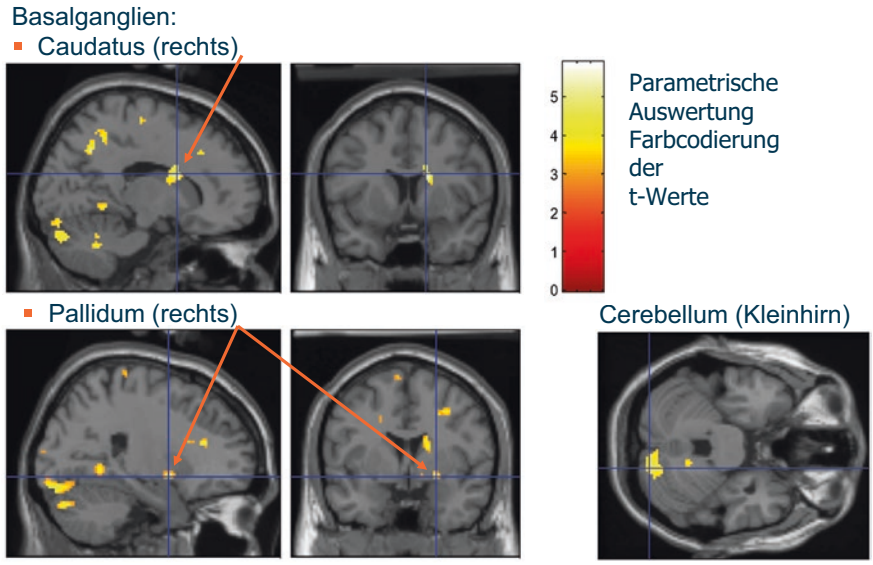


Abb. 4.19 Hirnregionen, die nach dem Gewinn der Schokolade bei Lieblingsmarke signifikant mehr aktiviert waren als bei noch akzeptabler Marke. (Quelle: Wildner & Jäncke, 2010)

Die Abbildung Abb. 4.19 zeigt die Hirnregionen, die bei der bevorzugten Marke nach dem Gewinn deutlich stärker aktiviert waren als bei der gerade noch akzeptablen Marke. Für die Interpretation sind detaillierte Kenntnisse der Hirnanatomie erforderlich. So muss man z. B. wissen, dass das Pallidum bei der Generierung positiver affektiver Reaktionen auf angenehme Reize beteiligt ist und die Aktivierung des Cerebellums für allgemeines Arousal steht.

Im Ergebnis zeigte sich, dass die Ergebnisse aller drei Methoden zur Markenwertbestimmung hoch mit den gemessenen Unterschieden bei der Hirnaktivität korrelieren, dass die Korrelation bei der Auswahlmethode aber am höchsten und beim Chip-Game am geringsten ist. ◀

Hier kann nur ein allererster Eindruck von diesem dynamischen und anspruchsvollen Forschungsgebiet vermittelt werden. Vertiefende Information bietet die umfangreiche Spezialliteratur. Für einen Einstieg eignen sich besonders die Bücher von Kenning (2014) und von Bruhn und Köhler (2010) sowie der Überblick im Artikel von Ariely und Berns (2010).

4.5 Datensammlung und -aufbereitung

4.5.1 Grundlagen

In den bisherigen Teilen dieses Buches sind vor allem Aspekte der Datenerhebung für die Marktforschung behandelt worden. Es wurden in diesem Zusammenhang

hauptsächlich Probleme der Festlegung von Untersuchungsdesigns und der Entwicklung von Messinstrumenten diskutiert. Im Mittelpunkt des vorliegenden Kapitels stehen die Tätigkeiten, die zwischen der Festlegung der Untersuchungsdesigns mit allen methodischen Einzelheiten und der statistischen Datenanalyse stattfinden: die **Sammlung der Daten im „Feld“** und deren **Aufbereitung.** Manche der dabei relevanten Gesichtspunkte sind eher technischer oder handwerklicher Art und beruhen auf Erfahrungen, sind also theoretisch schwer zu vermitteln. Es muss aber daran erinnert werden, dass die Genauigkeit und Gültigkeit von Untersuchungsergebnissen nur so gut ist, wie das schwächste Glied in der Kette der Schritte, die zu diesen Ergebnissen geführt haben.

So kann man sich leicht vorstellen, dass alle Sorgfalt bei der Entwicklung von Fragebögen oder der Stichprobenziehung vergeblich bleiben muss, wenn der Interviewer bei der Datensammlung bestimmte Anweisungen nicht beachtet oder wenn erhebliche Fehler bei der Codierung und Eingabe der Daten in den Computer auftreten. Deswegen sollen hier Aspekte der Datensammlung und -aufbereitung vor allem aus dem Blickwinkel der damit verbundenen **Fehlermöglichkeiten** betrachtet werden. Dabei wird die im vorliegenden Kapitel behandelte besonders gängige Form der Datenerhebung, die repräsentative Befragung, zugrunde gelegt. Viele der darauf bezogenen Überlegungen können auch auf andere Untersuchungsdesigns übertragen werden.

Die Arten von Fehlern, die in der Marktforschung auftreten, sollen anhand des in Abb. 4.20 dargestellten Schemas charakterisiert und abgegrenzt werden. Es werden dort drei **Fehlerarten** aufgeführt: Fehler durch die Auswahl der Auskunftspersonen, durch fehlende oder unzutreffende Angaben über die jeweilige Person und Fehler bei der Durchführung der Befragung. Die Darstellung der Fehlerarten in Form eines Eisbergs ist keineswegs zufällig gewählt worden. Sie soll vielmehr verdeutlichen, dass ein großer Teil der bei derartigen Untersuchungen auftretenden Fehler dem Nutzer der Untersuchungsergebnisse verborgen bleiben. Oft wird in entsprechenden Berichten nur der Stichprobenfehler ausgewiesen, das aber mit manchmal übertriebener Genauigkeit. So haben die zu Beginn des Abschn. 4.3.1.1 vorgestellten Beispiele gezeigt, dass Fehler durch unterschiedliche Frageformulierungen wesentlich gravierender sein können als ein Stichprobenfehler, der mit der (Schein-) Genauigkeit von ein oder zwei Kommastellen angegeben wird. Das ist in der Abb. 4.20 auch durch die „Grenze der Sichtbarkeit“ angedeutet. Einige der in der Abbildung aufgeführten Fehlerarten wurden schon angesprochen. So sind vor allem Messfehler bei den Auskunftspersonen (z. B. durch Erinnerungsmängel oder bewusst verzerrtes Antwortverhalten) schon ausführlich erörtert worden.

Der **Stichprobenfehler** (auch Zufallsfehler genannt) resultiert aus der zufälligen Schwankung von Stichprobenergebnissen um einen „wahren Wert“, der für die Grundgesamtheit gilt. Diese Abweichungen können von Stichprobe zu Stichprobe unterschiedlich sein und sind nie ganz zu vermeiden, da sie mit dem Prinzip der Stichprobenziehung verbunden sind. Durch die Vergrößerung einer Stichprobe kann man Fehler dieser Art allerdings reduzieren. Die Größe des Stichprobenfehlers ist im Gegensatz zu anderen Arten von Fehlern berechenbar, sofern die Auswahl der Stichprobenelemente zufällig erfolgt ist (siehe dazu Abschn. 4.2 und Kap. 8).

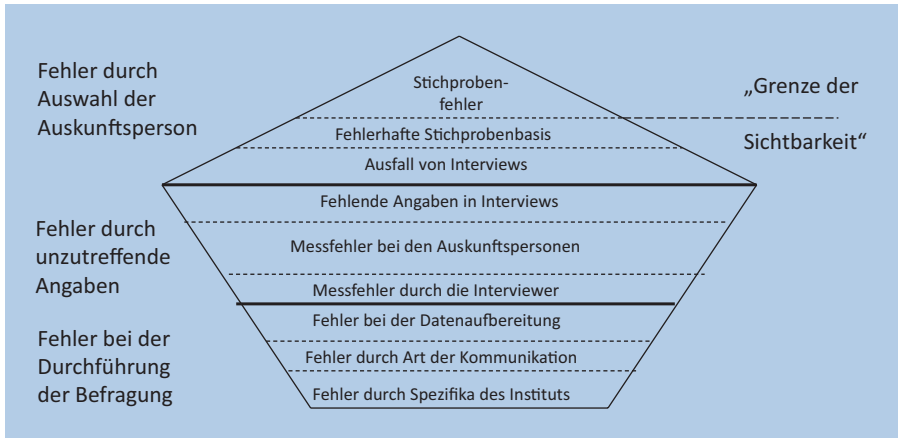


Abb. 4.20 Fehlerarten bei Umfragen. (Nach Weisberg, 2005, S. 19)

Alle anderen Fehler werden zusammen als systematische Fehler bezeichnet. Im Gegensatz zum Stichprobenfahler lässt sich der systematische Fehler nicht berechnen und reduziert sich auch nicht, wenn die Stichprobe vergrößert wird.

Ein Teil des systematischen Fehlers ist eine **fehlerhafte Stichprobenbasis.** Diese besteht darin, dass Verzeichnisse von Personen, Haushalten, Unternehmen etc., die als Basis für die Auswahl der Erhebungseinheiten verwendet werden, die Grundgesamtheit, über die Aussagen gemacht werden sollen, nicht angemessen abdecken. Das wird insbesondere dann zum Problem, wenn sich die Stichprobenbasis systematisch von der Grundgesamtheit unterscheidet. So kann man sich beispielsweise leicht vorstellen, dass bei der Verwendung eines Wählerverzeichnisses als Stichprobenbasis für eine repräsentative Befragung in einer Region die frisch zugezogenen Personen (die offenbar eher mobil sind) noch nicht in dem Verzeichnis enthalten sind während andere Personen, die schon verzogen sind (also eher nicht mehr zu der Grundgesamtheit in der Region lebender Personen gehören), noch im Wählerverzeichnis sind. Allgemein spricht man von einer fehlerhaften Stichprobenbasis, wenn bestimmte Gruppen von Elementen der Grundgesamtheit eine zu geringe oder zu große Wahrscheinlichkeit haben, Elemente der Stichprobe zu werden. Abb. 4.21 illustriert dieses Problem.

Man erkennt deutlich, dass in den natürlich sehr einfachen und fiktiven Beispielen einige Elemente der Grundgesamtheit keine Chance hätten, in die Stichprobe zu kommen („Undercoverage“), und andererseits Elemente, die eigentlich nicht zur Grundgesamtheit gehören, in die Stichprobe kommen können. Für eine ausführliche und praxisnahe Diskussion dieses Problems sei hier auf Groves et al. (2009, S. 54 ff. und 69 ff.) verwiesen.

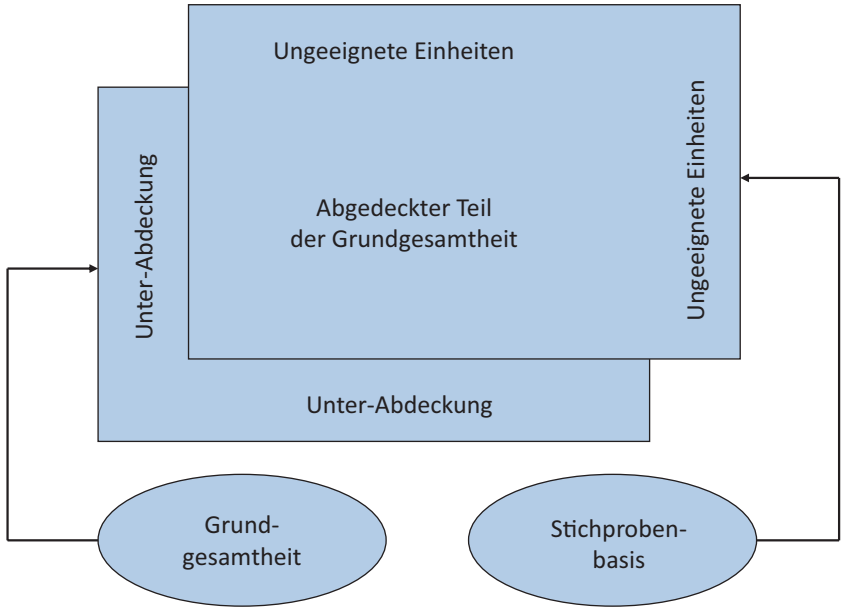


Abb. 4.21 Abdeckung einer Grundgesamtheit durch eine Stichprobenbasis. (Quelle: Groves et al., 2009, S. 55)

Beispiele für fehlerhafte Stichprobenbasen

- Eine Stichprobe aus der Grundgesamtheit „erwachsene Gesamtbevölkerung" wird auf der Basis von Telefonverzeichnissen gezogen. Problem: Haushalte ohne Festnetz-Telefon und Haushalte mit Telefonanschlüssen, die nicht aufgelistet sind, haben keine Chance, ausgewählt zu werden.
- Eine Personen-Stichprobe wird so gebildet, dass aus (mit gleicher Wahrscheinlichkeit zufällig) ausgewählten Haushalten jeweils eine Person befragt wird. Problem: Personen, die allein im Haushalt leben, haben eine größere Chance, befragt zu werden, als Personen, die mit mehreren anderen Menschen in einem Haushalt wohnen. Personen aus kleinen Haushalten wären also überrepräsentiert. Dieser Fehler kann allerdings berechnet und durch entsprechende Gewichtung wieder ausgeglichen werden.
- Eine Stichprobe von Unternehmen einer Branche wird auf der Basis des Mitgliederverzeichnisses des entsprechenden Unternehmensverbandes gezogen. Problem: Unternehmen, die nicht Mitglied des Verbandes sind (z. B. sehr kleine Unternehmen), haben keine Chance, in die Stichprobe zu kommen ◀

Die nächste Art der in Abb. 4.20 dargestellten Fehlerarten ist der **Ausfall von Interviews,** vor allem dadurch, dass Auskunftspersonen nicht erreicht werden oder die Teilnahme an der Untersuchung verweigern. Die Abb. 4.22 gibt – bezogen auf telefonische

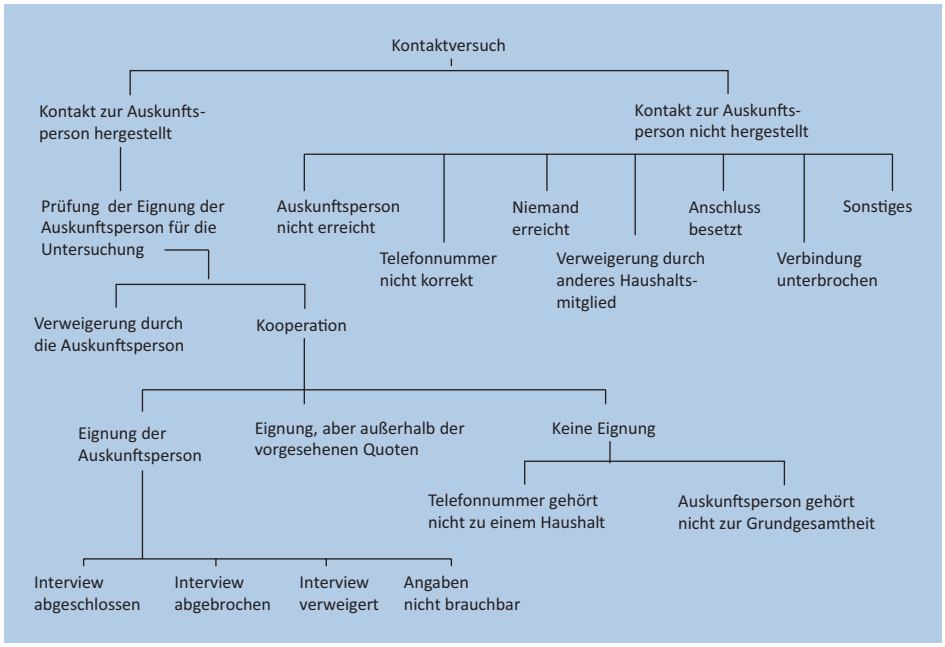


Abb. 4.22 Kontaktversuche bei telefonischen Befragungen. (In Anlehnung an Wisemann & McDonald, 1980, zitiert nach Churchill & Iacobucci, 2002, S. 528)

Befragungen – einen Überblick über die Vielzahl von Situationen, die zum Ausfall von Messungen führen können.

Die Probleme, die sich durch fehlende Angaben von Personen, Haushalten, Unternehmen etc. ergeben, die zur Stichprobe gehören, sind in der an eine Zielscheibe angelehnten Abb. 4.23 illustriert. Man kann sich vorstellen, dass auch die Probleme der Stichprobenausschöpfung damit zu tun haben, wie genau man mit einer durchgeführten Untersuchung die „wahren" Merkmalsverteilungen in einer Grundgesamtheit trifft. Hier ist auch wieder auf die Beziehungen zu dem zentralen Aspekt der **Validität** einer Untersuchung hinzuweisen. Das Problem mangelnder Stichprobenausschöpfung ist recht gravierend, wenn man bedenkt, dass nach einer Studie des Arbeitskreises Deutscher Markt- und Sozialforschungsinstitute (Engel & Schnabel, 2004) die durchschnittliche Antwortrate bei einer Meta-Analyse zahlreicher Umfragen nur bei etwa 48 % lag. Auch eine Analyse von Umfragedaten, die in international führenden Marketing-Zeitschriften publiziert waren, ergab eine Ausschöpfungsrate unter 50 % (Collier & Bienstock, 2007, S. 172).

Im linken Teil der Abb. 4.23 ist die Situation gezeigt, in der eine Stichprobe vollständig ausgeschöpft wird und deswegen keine Verzerrung der Untersuchungsergebnisse entsteht. Die vollständige Stichprobenausschöpfung stellt in der Praxis (nicht nur) der Marktforschung allerdings einen seltenen Ausnahmefall dar. Deshalb ist die Aussage-

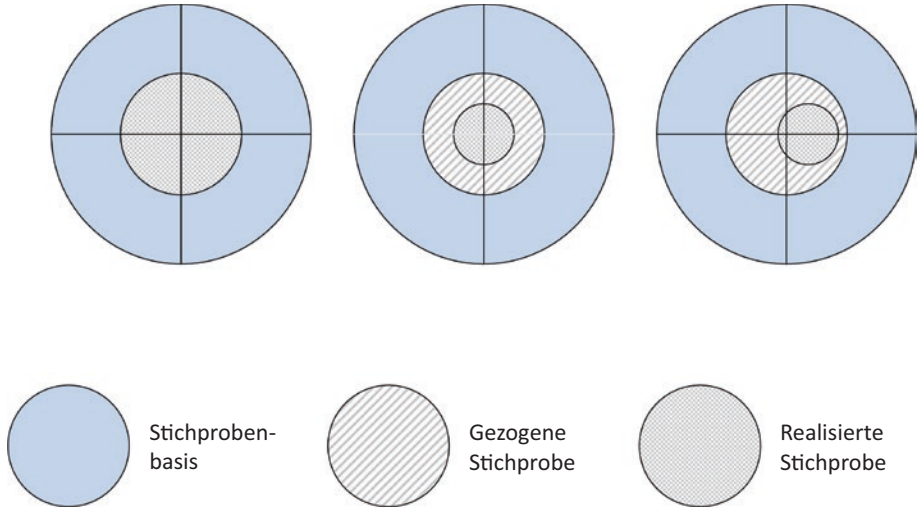


Abb. 4.23 Vollständige und unvollständige Stichprobenausschöpfung und ihre Konsequenzen. (Quelle: Parasuraman, 1986, S. 526)

kraft der Stichprobentheorie, bei der eben von vollständiger Stichprobenausschöpfung ausgegangen wird, für die Genauigkeit bzw. Sicherheit von Ergebnissen zusätzlich begrenzt.

Gängiger ist der im mittleren Teil von Abb. 4.23 dargestellte Fall, bei dem nur bei einem Teil der bei der Stichprobenziehung ausgewählten Elemente Daten erhoben werden können. In der in der Mitte der Abbildung wiedergegebenen Situation wird dadurch keine Verzerrung der Ergebnisse verursacht. Man muss aber eher damit rechnen, dass die Merkmalsverteilung bei den Stichprobenelementen, bei denen keine Daten erhältlich sind, anders als bei den restlichen ist. Beispielsweise kann man sich leicht vorstellen, dass unter den bei einer mündlichen Umfrage schwer erreichbaren Auskunftspersonen der Anteil mobiler, in Beruf und Freizeit besonders aktiver Menschen höher ist als in der Gruppe derer, die bei einem der ersten Kontaktversuche erreicht werden. Im rechten Teil der Abb. 4.23 ist die sich aus derartigen Gründen ergebene Ergebnisverzerrung symbolisch dargestellt. Wesentlich für die Vermeidung derartiger Verzerrungen sind die Hartnäckigkeit und das Geschick des Interviewers im Hinblick auf den Kontakt zur und die Motivierung der ausgewählten Auskunftsperson (siehe dazu auch Abschn. 4.5.2).

Hinsichtlich der Fehler durch unzutreffende Angaben werden *fehlende Angaben in Interviews* durch mangelnde Angaben zu einzelnen Fragen und *Messfehler durch die Interviewer* im folgenden Abschn. 4.5.2 kurz angesprochen. Hinsichtlich von *Messfehlern bei den Auskunftspersonen* ist schon auf die umfassende Diskussion im Abschn. 4.3.1 verwiesen worden. Hinsichtlich der Fehler bei der Durchführung der Befragung sind die spezifischen Stärken und Schwächen verschiedener *Arten der*

Kommunikation bereits im Abschn. 4.3.4 erörtert worden. Um *Fehlermöglichkeiten bei der Datenaufbereitung* geht es im Abschn. 4.5.3.

Bezüglich der *Fehlermöglichkeiten durch Spezifika des Instituts* bezieht sich Weisberg (2005, S. 297 ff.) insbesondere auf sogenannte „Haus-Effekte" und knüpft damit an das aus der politischen Umfrageforschung bekannte Phänomen an, dass bestimmte Institute als eher konservativ oder links-liberal orientiert gelten. Neben solchen Tendenzen können auch von Institut zu Institut verschiedene Praktiken und Gewohnheiten zu systematischen Ergebnisunterschieden führen. Ähnliche Konsequenzen haben interkulturelle Unterschiede (z. B. Bedeutungsgehalt bestimmter Begriffe) bei länderübergreifenden Untersuchungen.

Hintergrundinformation
Herbert Weisberg (2005, S. 299) kennzeichnet „Haus-Effekte" in folgender Weise:

> „Der Begriff „Haus-Effekte" bezieht sich darauf, dass Umfragen zum selben Thema, die von verschiedenen Instituten durchgeführt werden, zu unterschiedlichen Ergebnissen führen können. Natürlich führt der Stichprobenfehler dazu, dass man keine identischen Resultate erwarten sollte, aber die Unterschiede sind manchmal weit größer als man es durch Stichprobenfehler erklären kann. …. Er (der Haus-Effekt, Anm. d. Verf.) kann angesehen werden als eine Residualgröße, die verbleibt, wenn man die Einflüsse unterschiedlicher Frageformulierungen berücksichtigt hat. Haus-Effekte werden hier der Durchführung von Befragungen zugerechnet, weil sich die Ergebnisunterschiede häufig auf Anleitung, Training und Überwachung von Interviewern zurückführen lassen, die außerhalb des Umfrage-Instituts nicht erkennbar sind."

4.5.2 Datensammlung

Wenn in der einschlägigen Literatur bei der Datensammlung auftretende Probleme behandelt werden, so steht dabei meist die Datenerhebung mittels mündlicher Befragung im Vordergrund. Das ist dadurch gerechtfertigt, dass die entsprechenden Fragestellungen bei anderen Formen der Befragung (schriftlich, telefonisch) damit weitgehend abgedeckt sind und dadurch, dass bei anderen Erhebungsverfahren (z. B. bei Beobachtungen) Probleme der Datensammlung oftmals so stark auf die jeweilige Untersuchungssituation bezogen sind, dass eine generelle Behandlung in vertretbarem Umfang kaum möglich ist.

Hinsichtlich **persönlicher Merkmale** (z. B. Alter, Geschlecht, soziale Schicht) wird oftmals eine möglichst große Entsprechung zwischen Interviewer und Auskunftsperson empfohlen, da man auf diese Weise die Bereitschaft, sich an der Untersuchung zu beteiligen und korrekte Auskünfte zu geben, fördern kann. Allerdings lässt sich diese Forderung aus praktischen Gründen höchstens teilweise realisieren. Bezüglich **psychischer Faktoren,** die das Verhalten der Interviewer und damit (indirekt) das Antwortverhalten von Auskunftspersonen beeinflussen (z. B. durch eigene Meinungen der Interviewer), kommt es am ehesten darauf an, durch Schulung die Interviewer zu einem möglichst neutralen Verhalten zu bewegen, um eben diese Wirkungen zu minimieren.

Von den Verhaltensweisen der Interviewer sind hinsichtlich der Genauigkeit bzw. Fehlerhaftigkeit von Untersuchungsergebnissen die Folgenden besonders wichtig:

- **Fehler bei der Fragestellung** Im vorliegenden Kapitel ist schon erörtert worden, wie bedeutsam die Entwicklung von Messinstrumenten für die Qualität von Untersuchungsergebnissen ist und wie empfindlich diese Ergebnisse selbst gegenüber geringfügig wirkenden methodischen Veränderungen sind. Daraus folgt unmittelbar die Notwendigkeit, den Einsatz der Messinstrumente durch die Interviewer genau in der festgelegten Weise vornehmen zu lassen. Hinzu kommt der Gesichtspunkt, dass Angaben von Auskunftspersonen nur vergleichbar sind, wenn sie durch einheitliche Erhebungstechniken zustande gekommen sind. Erstaunlich ist der hohe Anteil von Abweichungen von den vorgegebenen Fragen, die man in einschlägigen speziellen Untersuchungen festgestellt hat. Beispielsweise haben Bradburn und Sudman (1979, S. 27 ff.) in einer solchen Studie 372 Interviews aufgezeichnet und bei etwa 30 % der insgesamt gestellten Fragen Lesefehler aufseiten der Interviewer festgestellt.
- **Fehler bei der Motivierung von Auskunftspersonen** Insbesondere wegen der Gefahr der mangelnden Repräsentativität einer Untersuchung durch nicht erreichte Zielpersonen und/oder abgebrochene bzw. verweigerte Interviews (siehe Abschn. 4.5.1) ist es notwendig, erhebliche Anstrengungen im Hinblick auf eine weitgehende Stichprobenausschöpfung zu unternehmen. Neben den üblichen Erläuterungen der Bedeutung des einzelnen Interviews für den Erfolg der Untersuchung und der Zusicherung der Einhaltung von Regeln des Datenschutzes durch den Interviewer kommt es hier besonders darauf an, durch eine gewisse Zahl von *Kontaktversuchen* zu unterschiedlichen Zeiten möglichst viele der in einer Stichprobe ausgewählten Personen zu erreichen.
- **Fehler hinsichtlich der Vollständigkeit der Angaben** Die gängigste Möglichkeit für eine Auskunftsperson, sich schwierigen oder unangenehmen Fragen zu entziehen ist die „Weiß nicht"-Angabe. Oftmals wird eine solche Möglichkeit bei geschlossenen Fragen angeboten. Weisberg (2005, S. 133 ff.) empfiehlt hier zur Steigerung der Antwortbereitschaft vor allem ein „Nachhaken" des Interviewers und finanzielle Anreize für die Auskunftsperson bei vollständigen Angaben.
- **Fehler bei der Erfassung von Antworten** Angesichts der komplexen Aufgabe des Interviewers, ein lebendig wirkendes und zur Fortsetzung motivierendes Gespräch mit einer Auskunftsperson zu führen und dabei gleichzeitig eine Fülle von Anweisungen zu beachten, wundert es nicht, dass bei der Übertragung von Antworten in den Erhebungsbogen Fehler auftreten können. Hinzu kommt die *Beeinflussung von Wahrnehmungen* von Antworten seitens des Interviewers durch dessen *Erwartungen*. So haben Bradburn und Sudman (1979, S. 51 ff.) in einer Studie Interviewer nach ihren Erwartungen bezüglich der Ergebnisse einer Umfrage befragt. Es zeigte sich bei diversen Variablen (Einkommen, Teilnahme an Glücksspielen, Alkoholkonsum), dass die im Fragebogen eingetragenen Angaben der Auskunftspersonen, die von Interviewern befragt worden waren, die niedrige Ergebnisse erwartet hatten, deutlich

unter denen lagen, die bei der anderen Gruppe von Interviewern zustande gekommen waren.

- **Fälscherproblem** Wenn man an die recht anspruchsvolle, aber mäßig bezahlte Arbeit von InterviewerInnen denkt, so kann man nicht völlig ausschließen, dass sich gelegentlich InterviewerInnen diese durch komplette oder teilweise Fälschungen „erleichtern", indem sie nur wenige oder keine der verlangten Angaben bei den in der Stichprobe enthaltenden Auskunftspersonen erheben, sondern diese nach eigenem Gutdünken am heimischen Schreibtisch eintragen. Eine besondere Spielart von Fälschungen sind sog. „*Filter-Fälschungen*". Diese bestehen darin, dass der (unseriöse) Interviewer bei Filterfragen (siehe Abschn. 4.3.3) – unabhängig von der tatsächlichen Angabe der Auskunftsperson – die Antwortmöglichkeit ankreuzt, die dazu führt, dass eine größere Zahl folgender Fragen übersprungen und das Interview damit abgekürzt wird.

Hier sei in Erinnerung gerufen, dass sich derartige Probleme und Fehlermöglichkeiten bei schriftlichen und Online-Befragungen natürlich nicht stellen. Bei telefonischen Befragungen stellen sie sich nur in deutlich geringerem Maße. Zunächst ist der Interviewereinfluss durch den nur akustischen Kontakt geringer. Daneben gibt es hier wesentlich bessere Möglichkeiten der Interviewerkontrolle und -schulung als bei persönlichen Interviews, weil (ausgewählte) Gespräche aufgezeichnet oder vom Untersuchungsleiter mitgehört werden können. Letztlich entfällt der Haupt-Anreiz für Fälschungen, weil Telefon-Interviewer in der Regel im Zeitlohn und nicht nach der Zahl der durchgeführten Interviews bezahlt werden.

Die wichtigsten Ansatzpunkte zur Verminderung von Fehlern, die durch das Verhalten von Interviewern entstehen, liegen im Bereich der so genannten **„Interviewer**-" oder **„Feld-Organisation"**. Dazu gehören die Bereiche:

- Anwerbung und Auswahl von Interviewern (Anwerbung durch Zeitungsanzeigen oder über das Internet, Auswahl nach Einstellungsgespräch, psychologische Tests und der Durchführung von Probe-Interviews)
- Interviewer-Ausbildung (Schriftliche Ausbildungsunterlagen, Rollenspiele, Übungen mit „präparierten" Auskunftspersonen)
- Interviewer-Anweisungen für die jeweilige Untersuchung (Erläuterungen zum Hintergrund der Untersuchung, Erläuterungen zur Stichprobe und zum Ablauf des Interviews, Motivierung des Interviewers)
- Interviewer-Einsatz (Betreuung durch „Chef-Interviewer", Honorierung, Festlegung der Zahl von Interviews, Terminplanung)
- Interviewer-Kontrolle (Kontrolle hinsichtlich Fälschungen, Einhaltung von Anweisungen, Verweigerungsraten und Termineinhaltung)

Dieser ganze Komplex ist eher durch praktische Erfahrungen als durch theoretisch geprägte Methodik geprägt (Was nicht geringschätzig gemeint ist! Eher im Gegenteil, weil

ja die Sorgfalt und Kreativität der Praktiker in vielen Instituten großen Respekt verdient.). Die Vielzahl relevanter Einzelaspekte lässt sich hier nicht darstellen. Dazu muss auf die Spezial-Literatur verwiesen werden, wobei insbesondere die Bücher von Groves et al. (2009), Noelle-Neumann und Petersen (2000) und Weisberg (2005) zu nennen sind.

Hintergrundinformation
In den „Best Practices" der American Association for Public Opinion Research (www.aapor.org; zitiert nach Kaase, 1999, S. 133) wird die Bedeutung von Interviewer-Auswahl und –Schulung formuliert:

> „Wichtig für vorbildliche Umfragen ist das Beharren auf hohen Standards bei Auswahl und Schulung der Interviewer. Um qualitativ hochwertige Daten zu erheben, müssen die Interviewer für eine telefonische oder persönlich-mündliche Umfrage sorgfältig geschult werden. Dies kann durch direkte Gruppenschulung, durch telefonische Schulung, durch Eigenstudium oder eine Kombination davon erfolgen. Gutes Interviewerverhalten sollte in den Vordergrund gestellt werden, zum Beispiel für die Art und Weise der ersten Kontaktierung von Befragten, für das professionelle Durchführen von Interviews und für das Vermeiden von Beeinflussungen der Befragten. Die Interviewerschulung sollte auch Übungsinterviews vorsehen, um die Interviewer auf die vielfältigen Situationen vorzubereiten, die sie bei ihrer Arbeit wahrscheinlich antreffen werden (…)."

4.5.3 Datenaufbereitung

Bei der Aufbereitung der Daten für die Analyse mit Hilfe verbreiteter Statistik-Software (z. B. SPSS, SAS/JMP oder R) handelt es sich um einen Untersuchungsschritt, der eher technisch geprägt ist. Deshalb muss hier wieder ein knapper Überblick dazu genügen. Ausführliche Hinweise finden sich z. B. bei Fowler (2009, S. 145 ff.); Groves et al. (2009, S. 329 ff.); Jacoby (2013, S. 865 ff.); Karweit und Meyers (1983); Sudman und Blair (1998, S. 413 ff.).

Die Hauptschritte bei der Datenaufbereitung sind

1. die Editierung der vorliegenden Erhebungsbögen,
2. die Codierung der Erhebungsbögen,
3. die Dateneingabe in den Rechner,
4. die Fehlerkontrolle,
5. die Ergänzung fehlender Daten,
6. die Identifizierung und der Umgang mit Ausreißern und
7. gegebenenfalls die Gewichtung bzw. Hochrechnung der Daten.

Diese einzelnen Schritte sollen im Folgenden kurz gekennzeichnet werden. Dabei ist zu beachten, dass einige dieser Schritte bei computergestützten Datenerhebungen (computergestützte telefonische oder persönliche Interviews und Online-Befragungen) entfallen bzw. schon bei der Entwicklung und Implementierung des „Computer-Fragebogens" erfolgen. Dies sind insbesondere die Schritte 1–3.

Editierung der Erhebungsbögen Als Editierung bezeichnet man eine Überprüfung und gegebenenfalls Korrektur der ausgefüllten Fragebögen. Sie sollte möglichst kurzfristig nach der Datenerhebung vorgenommen werden, um Fehler durch Rückfragen bei dem jeweiligen Interviewer klären zu können. Nach Churchill und Iacobucci (2005, S. 407) werden bei der Editierung vor allem folgende Gesichtspunkte geprüft:

- Vollständigkeit der Angabe
  Auslassung einzelner Angaben oder ganzer Teile des Fragebogens)
- Lesbarkeit der Eintragungen
  (Entschlüsselung von Handschriften, Abkürzungen etc.)
- Verständlichkeit der Angaben
- Konsistenz der Angaben
  (Eliminierung/Aufklärung widersprüchlicher Antworten)
- Vergleichbarkeit der Angaben
  (Einheitlichkeit verwendeter Maßeinheiten etc.)

Gelegentlich wird man bei der Editierung auch Anhaltspunkte für eine offenkundig fehlerhafte oder zu wenig sorgfältige Beantwortung eines Fragebogens finden und die entsprechenden Daten dann aus dem Datensatz eliminieren.

Codierung von Erhebungsbögen Unter der *Codierung* versteht man die Übersetzung der im Fragebogen eingetragenen Angaben in zweckmäßig gewählte Symbole, wofür fast immer Zahlen gewählt werden, einschließlich der Zuordnung verbaler Angaben zu Kategorien und damit verbundenen Symbolen/Zahlen. Bei geschlossenen Fragen (siehe Abschn. 4.3.1) entstehen hier wenige Probleme. Dagegen müssen bei offenen Fragen zunächst Kategorien für die unterschiedlichen Arten auftretender Antworten gebildet werden (vgl. Noelle-Neumann & Petersen, 2000, S. 383 ff.). Alle Codierungsregeln für eine Untersuchung werden in einem Codeplan (auch Codebuch genannt) festgelegt, um eine einheitliche Verfahrensweise bei allen an der Datenaufbereitung beteiligten Personen zu gewährleisten.

„Regeln" für Codierung
Sudman und Blair (1998, S. 422 ff.) stellen einige „Regeln" für die Codierung zusammen:

- Die Codes müssen umfassend sein, damit jede mögliche Antwort codierbar ist (gegebenenfalls Codes für „Trifft nicht zu", „Sonstiges", „Keine Angabe" vorsehen; Fragen mit Mehrfach-Antworten auf mehrere Variable verteilen).
- Fehlende Werte einheitlich codieren (z. B. „9" oder „99").
- Alle Antwortmöglichkeiten müssen überschneidungsfrei codiert werden (Beispiel Alter *falsch*: 10–20 Jahre, 20–30 Jahre, … *richtig*: 10–20 Jahre, 21–30 Jahre, …).

- Die Zahlen, die irgendwie geordneten Antwortkategorien zugeordnet werden, sollen dieser Ordnung entsprechen (z. B. „unzufrieden“ = 0, „mittelmäßig zufrieden“ = 1, „sehr zufrieden“ = 2).
- Ähnlich aufgebaute Fragen sollen ähnlich codiert werden.
- Bei mehreren Untersuchungen möglichst einheitliche Codierungen verwenden.

Die Eingabe der Rohdaten erfolgt oft noch über die Tastatur am Bildschirm. Durch Lese- und Tippfehler kann dabei die Qualität der auszuwertenden Daten erheblich beeinträchtigt werden. Bei der Verwendung von Scannern und dafür geeigneter Erhebungsbögen kann die manuelle Dateneingabe vermieden werden. Bei computergestützten und Online-Befragungen ist natürlich in der Regel keine gesonderte Dateneingabe erforderlich.

Beispiel zur Codierung

Frage	Antwortmöglichkeiten	Code
Sind Sie Raucher?	Ja	1
	Nein → Übergehen zu Frage 3	2
	Keine Angabe	3
Wie viele Zigaretten rauchen Sie pro Tag?	Anzahl eintragen	…
	96 und mehr	96
	Rauche keine Zigaretten, sondern Zigarre, Pfeife etc.	97
	Trifft nicht zu (Nichtraucher → Frage 1)	98
	Keine Angabe	99
Wie alt sind Sie?	Unter 18 Jahre	1
	18 bis 30 Jahre	2
	31 bis 50 Jahre	3
	51 bis 65 Jahre	4
	Über 65 Jahre	5
	Keine Angabe	9
Wie groß ist Ihr Interesse an Sport?	Sehr gering	1
	Gering	2
	Mittelmäßig	3
	Groß	4
	Sehr groß	5
	Keine Angabe	9

◀

	Variable				
Fälle	1	Geschlecht	Alter		m
1	X_{11}	0	32	…	X_{1m}
2	X_{21}	0	19	…	X_{2m}
3	X_{31}	1	54	…	X_{3m}
…	…	…	…	…	…
n	X_{n1}	0	42	…	X_{nm}

Abb. 4.24 Beispiel einer Datenmatrix

Nach der Dateneingabe steht im Rechner eine **Datenmatrix** für die weitere Analyse zur Verfügung, deren Spalten die einzelnen Werte der verschiedenen Variablen enthalten und in deren Zeilen die Angaben jeweils einer Auskunftsperson (eines „Falles“) stehen. Jede Position in dieser Datenmatrix ist durch den Codeplan (siehe oben) definiert. Abb. 4.24 zeigt ein einfaches Beispiel einer solchen Datenmatrix. Die beiden möglichen Kategorien der Variablen Geschlecht sind in dem Beispiel mit „0“ bzw. „1“ codiert, wobei es sich hier um eine nominalskalierte Variable handelt (siehe Abschn. 7.2).

Was versteht man nun unter einer **„Variablen“**? Variable stehen für ein Merkmal, das unterschiedliche Ausprägungen haben kann, z. B. kann das soeben genannte Merkmal „Geschlecht“ zwei, u. U. mit der Ausprägung „divers“ auch drei Ausprägungen und das Merkmal „Jahreseinkommen“ nahezu unendlich viele Ausprägungen (von „0 €“ bis „x €“) haben. Wenn man an das im Abschn. 2.2.2 erörterte Grundmodell empirischer Forschung denkt, dann lässt sich der Begriff der Variablen folgendermaßen veranschaulichen: Man kann sich den Prozess der empirischen Forschung so vorstellen, dass Konzepte bzw. Konstrukte wesentliche Elemente theoretischer Überlegungen und Hypothesen sind. Über den Prozess der Operationalisierung werden entsprechende Messmethoden entwickelt und diese bei Messungen angewandt. Die anfallenden Daten werden dann als *Variable* erfasst bzw. gespeichert. Oftmals stehen also Variable für ein bestimmtes Konzept/Konstrukt (z. B. bei der Variablen „Geschlecht“). Es kann aber auch sein, dass mehrere Variable gemeinsam ein Konzept/Konstrukt repräsentieren, z. B. bei der Verwendung von Multi-Item-Skalen (siehe Abschn. 4.3.2). Die Frage, inwieweit eine Variable einem zu messenden Konzept entspricht, ist die schon ausführlich diskutierte Frage der Validität.

Fehlerkontrolle Vor der statistischen Datenanalyse findet eine Fehlerkontrolle bei dem eingegebenen Datensatz statt. Dadurch sollen bisher unentdeckte und bei der Dateneingabe aufgetretene Fehler identifiziert und nach Möglichkeit eliminiert werden. Hauptsächlich über drei Wege versucht man, Fälle zu ermitteln, die fehlerhaft sind oder bei denen zumindest der Verdacht nahe liegt, dass sie fehlerbehaftet sind:

- Prüfung auf Vollständigkeit: Wurden alle Fälle und alle Variablen übermittelt?
- Prüfung, ob bei Variablen *Werte* auftreten, die laut Codeplan *nicht vorgesehen* sind („Tippfehler“ oder auch Fehler im Codeplan)
- Prüfung auf *logische Konsistenz* der Werte (Beispiel: Alter 17 Jahre; Beruf Rentner)
- Ermittlung von *Ausreißern,* also von Werten, die extrem vom sonstigen Wertebereich abweichen. Karweit und Meyers (1983, S. 395) nennen fünf Alternativen im Hinblick auf identifizierte Fehler:
  1. Rückgriff auf den Original-Fragebogen, um festzustellen, ob es sich um einen Übertragungsfehler handelt
  2. Rückfrage bei der Auskunftsperson
  3. Ersatz des fehlerbehafteten Wertes durch einen sinnvoll geschätzten
  4. Eliminierung des fehlerhaften Wertes und Kennzeichnung der entsprechenden Position in der Datenmatrix als „fehlender Wert“
  5. Eliminierung des gesamten Falles

Ausreißerwerte Sind Extremwerte, die sich deutlich von der erhobenen Messwertreihe abheben. Für die Feststellung von Ausreißern existiert kein allgemein gültiges Standardverfahren. Im Falle von symmetrischen und eingipfligen Verteilungen werden Ausreißer allerdings in der Regel über die Standardabweichung identifiziert. In diesem Fall wird ein Wert oftmals als Ausreißer betrachtet, wenn er sich 2.5 (bzw. 3) Standardabweichungen ober- oder unterhalb des Mittelwertes befindet.

Gewichtung von Daten Bei manchen Untersuchungen ist vor der Datenanalyse noch eine *Gewichtung* der Fälle im Datensatz erforderlich. Ziel dabei ist die Korrektur von Verzerrungen in der Stichprobe, die beispielsweise durch eine unzureichende Abdeckung der Grundgesamtheit, durch systematisch ungleich verteilte Antwortverweigerungen oder durch disproportionale Schichtung bei der Stichprobenziehung (siehe Abschn. 4.2) verursacht sein kann. Wenn also eine bestimmte Gruppe (z. B. alleinstehende Frauen über 70 Jahre) in der Stichprobe – gemessen an ihrem Anteil in der Bevölkerung – zu schwach vertreten ist, dann kann man dies durch die Zuordnung eines (relativ hohen Gewichtungsfaktors) korrigieren. Zu Einzelheiten sei wieder auf Groves et al. (2009, S. 347 ff.) verwiesen.

Literatur

ADM. (Hrsg.). (2012). ADM-Forschungsprojekt ‚Dual-Frame-Ansätze' 2011/ 2012, HYPERLINK https://www.adm-ev.de/index.php?eID=tx_nawsecuredl&u=0&file=fileadmin/user_upload/PDFS/ADM_Dual_Frame_Projekt_-_Forschungsbericht.pdf&t=1502419459&hash=2f329caf8f6f4ca2cbe8b18f9f6050fc46663234.

ADM. (2016). Arbeitskreis Deutscher Marktforschungsinstitute, Jahresbericht 2015, HYPERLINK „www.adm-ev.de".

ADM. (Hrsg.). (2020). Die Marktforschung in Zahlen, https://www.adm-ev.de/die-branche/mafo-zahlen/. Zugegriffen: 14. Juni 2020.

Albers, S., & Hildebrandt, L. (2006). Methodische Probleme bei der Erfolgsfaktorenforschung – Messfehler, formative versus reflektive Indikatoren und die Wahl des Strukturgleichungsmodells. *Zeitschrift für betriebswirtschaftliche Forschung, Jg., 58*, 2–33.

Ariely, D., & Berns, G. (2010). Neuromarketing: The hope and hype of neuroimaging in business. *Nature Reviews Neuroscience, 11,* 284–292.

Baumann, H., Schulz, S., & Thiesen, S. (2024). ALLBUS 2021 – Variable Report, Studien-Nr. 5280, GESIS Datenarchiv für Sozialwissenschaften, 2. Auflage. Mannheim

Baumgartner, H., & Homburg, C. (1996). Applications of structural equation modelling in marketing and consumer research: A review. *International Journal of Research in Marketing, 13*, 139–161.

Bearden, W., Netemeyer, R., & Haws, K. (2011). *Handbook of marketing scales – Multi-Item measures for marketing and consumer behavior research* (3. Aufl.). Sage.

Bergkvist, L., & Rossiter, J. (2007). The predictive validity of multiple-item versus single-item measures of the same constructs. *Journal of Marketing Research, 44*, 175–184.

Borsboom, D., Mellenbergh, G., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*, 1061–1071.

Bradburn, N., & Sudman, S. (1979). *Improving interview method and questionnaire design*. Jossey-Bass.

Bradburn, N., Sudman, S., & Wansink, B. (2004). *Asking questions* (Revised). Jossey-Bass.

Bruhn, M., & Köhler, R. (Hrsg.). (2010). *Wie Marken wirken – Impulse aus der Neuroökonomie für die Markenführung*. Vahlen.

Bruner, G. (2013). *Marketing Scales Handbook: Top 20*. GCBII Productions.

Bucklin, R., & Sismeiro, C. (2009). Click here for internet insight: Advances in clickstream data analysis in marketing. *Journal of Interactive Marketing, 23*, 35–48.

Bucklin, R., Lattin, J., Ansari, A., Gupta, S., Bell, D., Coupey, E., Little, J., Mela, C., Montgomery, A., & Steckel, J. (2002). Choice and the internet: From clickstream to research stream. *Marketing Letters, 13*, 245–258.

Buhrmester, M., Kwang, T., & Gosling, S. D. (2011). Amazon's Mechanical Turk a new source of inexpensive, yet high-quality, data? *Perspectives on psychological science, 6*(1), 3–5.

Burke, R. (1996). Der virtuelle Laden – Testmarkt der Zukunft. *Harvard Business Manager, Jg., 18*, 107–117.

Campbell, D., & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56*, 81–105.

Chauduri, A., & Stenger, H. (2005). *Survey sampling: Theory and methods*. Taylor and Francis Group.

Churchill, G. (1979). A paradigm for developing better measures of marketing constructs. *Journal of Marketing Research, 16*, 64–73.

Churchill, G., & Iacobucci, D. (2002). *Marketing research – Methodological foundations* (8. Aufl.). South-Western.

Churchill, G., & Iacobucci, D. (2005). *Marketing research – Methodological foundations* (9. Aufl.). South-Western.
Cochran, W. (1977). *Sampling techniques*. Wiley.
Comley, P. (2007). Online market research. *Market research handbook* (S. 401–419). ESOMAR.
Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*, 297–334.
Collier, J., & Bienstock, C. (2007). An analysis how non-response error is assessed in academic marketing research. *Marketing Theory, 7*, 163–168.
De Vaus, D. (2002). *Analyzing social science data*. Sage.
Diamantopoulos, A. (2008). Formative indicators: Introduction to the special issue. *Journal of Business Research, 61*, 1201–1202.
Diamantopoulos, A., Riefler, P., & Roth, K. (2008). Advancing formative measurement models. *Journal of Business Research, 61*, 1203–1218.
Diamantopoulos, A., Sarstedt, M., Fuchs, C., Wilczynski, P., & Kaiser, S. (2012). Guidelines for choosing between multi-item and single-item scales for construct measurement: A predictive validity perspective. *Journal of the Academy of Marketing Science, 40*, 434–449.
Dillman, D., Smyth, J., & Christian, L. (2009). *Internet, mail, and mixed-mode surveys* (3. Aufl.). Wiley.
Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation* (5. Aufl.). Springer.
Eberl, M. (2006). Formative und reflektive Konstrukte und die Wahl des Strukturgleichungsverfahrens. *Die Betriebswirtschaft, Jg., 66*, 651–668.
Eisend, M., & Kuß, A. (2023). *Grundlagen empirischer Forschung – Zur Methodologie in der Betriebswirtschaftslehre* (3. Aufl.). Springer Gabler.
Ekman, P., & Friesen, W. (1978). *Facial action coding system: A technique for the measurement of facial movement*. Consulting Psychologists Press.
Engel, U., & Schnabel, C. (2004). Markt- und Sozialforschung – Metaanalyse zum Ausschöpfungsgrad. www.adm-ev.de. Zugegriffen: 1. März 2014.
ESOMAR. (2018). *ESOMAR Global Prices Study 2018*. ESOMAR.
ESOMAR. (2019). *Global Market Research 2019 – An ESOMAR Industry Report*. ESOMAR.
Fowler, F. (2009). *Survey research methods* (4. Aufl.). Sage.
Fuchs, C., & Diamantopoulos, A. (2009). Using single-item measures for construct measurement in management research: Conceptual issues and application guidelines. *Die Betriebswirtschaft, 69*, 195–210.
Gesis. (Hrsg.). (2018). Allbus – die allgemeine Bevölkerungsumfrage der Sozialwissenschaften, https://www.gesis.org/allbus/inhalte-suche/studienprofile-1980-bis-2018/2018/. Zugegriffen: 14. Juni 2020.
Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E., & Tourangeau, R. (2009). *Survey Methodology* (2. Aufl.). Wiley.
Günther, M., Vossebein, U., & Wildner, R. (2006). *Marktforschung mit Panels* (2. Aufl.). Gabler.
Hildebrandt, L. (1984). Kausalanalytische Validierung in der Marketingforschung. *Marketing ZFP, 6*, 41–51.
Homburg, C. (2007). Betriebswirtschaftslehre als empirische Wissenschaft – Bestandsaufnahme und Empfehlungen. In G. Schreyögg (Hrsg.), *Zukunft der Betriebswirtschaftslehre, ZfbF-Sonderheft* (Bd. 56, S. 27–60).
Homburg, C., & Giering, A. (1996). Konzeptualisierung und Operationalisierung komplexer Konstrukte – Ein Leitfaden für die Marketingforschung. *Marketing ZFP, 18*, 5–24.
Homburg, C., & Klarmann, M. (2009). Multi-Informant-Designs in der empirischen betriebswirtschaftlichen Forschung. *Die Betriebswirtschaft DBW, 69*, 147–171.

Homburg, C., Klarmann, M., Reimann, M., & Schilke, O. (2012a). What drives key informant accuracy? *Journal of Marketing Research, 49*, 594–608.

Homburg, C., Klarmann, M., & Totzek, D. (2012b). Using multi-informant designs to address key informant and common method bias. In A. Diamantopoulos, W. Fritz, & L. Hildebrandt (Hrsg.), *Quantitative marketing and marketing management* (S. 81–102). Wiesbaden.

Horton, J. J. (2011). The condition of the Turking class: Are online employers fair and honest? *Economics Letters, 111*(1), 10–12.

Hoyle, R., Harris, M., & Judd, C. (2002). *Research methods in social relations* (7. Aufl.). Wadsworth.

Hulland, J., Baumgartner, H., & Smith, K. (2018). Marketing survey research best practices: Evidence and recommendations from a review of JAMS articles. *Journal of the Academy of Marketing Science, 46*, 92–108.

Hunt,S. (1987). Marketing research – Proximate purpose and ultimate value. In R. Belk, G. Zaltman, & R. Bagozzi (Hrsg.), *Marketing Theory* (S. 209–213). American Marketing Association.

Hurrle, B., & Kieser, A. (2005). Sind key informants verlässliche Datenlieferanten? *Die Betriebswirtschaft, 65*, 584–602.

Iacobucci, D., & Churchill, G. (2010). *Marketing research – Methodological foundations* (10. Aufl.). South-Western.

Jaccard, J., & Jacoby, J. (2020). *Theory construction and model building skills* (2. Aufl.). Guilford.

Jacoby, J. (1978). Consumer research – A state of the art review. *Journal of Marketing, 42*, 87–96.

Jacoby, J. (2013). *Trademark surveys – Designing, implementing, and evaluating surveys* (Bd. 1). American Bar Association.

Jarvis, C., MacKenzie, S., & Podsakoff, P. (2003). A critical review of construct indicators and measurement model misspecification in marketing and consumer research. *Journal of Consumer Research, 30*, 199–218.

Kaase, M. (Hrsg.). (1999). *Qualitätskriterien der Umfrageforschung*. Akademie.

Karweit, N., & Meyers, E. (1983). Computers in survey research. In P. Rossi, J. Wright, & A. Anderson (Hrsg.), *Handbook of survey research* (S. 379–414). Academic.

Kenning, P. (2014). *Consumer neuroscience*. Kohlhammer.

Kosslyn, S. (1999). If neuroimaging is the answer, what is the question? *Philosophical Transactions of the Royal Society B Biological Sciences, 354*(Heft 1387), 1283–1294.

Kroeber-Riel, W., & Gröppel-Klein, A. (2019). *Konsumentenverhalten* (11. Aufl.). Vahlen.

Krosnick, J. (1999). Survey research. *Annual Review of Psychology, 50*, 537–567.

Langer, A., Eisend, M., & Kuß, A. (2008). Zu viel des Guten? Zum Einfluss der Anzahl von Ökolabels auf die Konsumentenverwirrtheit. *Marketing – ZFP, 30,* 19–28.

Lenzner, T., & Menold, N. (2015). Frageformulierung. Mannheim, GESIS – Leibniz-Institut für Sozialwissenschaften.

Lenzner, T., & Menold, N. (2019). Slide Set: Question Wording. *GESIS Survey Guidelines*. Mannheim, GESIS – Leibniz Instute for the Social Sciences.

Levay, K. E., Freese, J., & Druckman, J. N. (2016). The demographic and political composition of mechanical turk samples. *SAGE Open, 6*(1), 1–17.

Madans, J., Miller, K., Maitland, A., & Willis, G. (Hrsg.). (2011). *Question evaluation methods – Contributing to the science of data quality*. Wiley.

Marichalar Quezada, R., Bartl, M., & Garrecht, G. (2022). Emotion AI: Neue Formen der Emotionsmessung durch Künstliche Intelligenz. In U. Lichtenthaler (Hrsg.), *Künstliche Intelligenz erfolgreich umsetzen – Praxisbeispiele für integrierte Intelligenz*. Springer Gabler.

McIver, J., & Carmines, E. (1981). *Unidimensional scaling*. Sage.

Moore, D. (2002). Measuring new types of question-order effects: Additive and subtractive. *Public Opinion Quarterly, 66*, 80–91.

Netemeyer, R., Bearden, W., & Sharma, S. (2003). *Scaling procedures – Issues and applications*. Sage.
Nisbett, R., & Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review, 84*, 231–259.
Noelle-Neumann, E., & Petersen, T. (2000). *Alle, nicht jeder* (3. Aufl.). Springer.
Nunnally, J., & Bernstein, I. (1994). *Psychometric theory* (3. Aufl.). McGraw-Hill.
Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on amazon mechanical turk. *Judgement and Decision Making, 5*, 411–419.
Parasuraman, A. (1986). *Marketing research*. Addison-Wesley.
Peter, J. (1979). Reliability – A review of psychometric basics and recent marketing practices. *Journal of Marketing Research, 16*, 6–17.
Peter, J. (1981). Construct validity – A review of basic issues and marketing practices. *Journal of Marketing Research, 18*, 133–145.
Podsakoff, P., MacKenzie, S., Lee, J., & Podsakoff, N. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology, 88*, 879–903.
Presser, S., Rothgeb, J., Couper, M., Lessler, J., Martin, E., Martin, J., & Singer, E. (Hrsg.). (2004). *Methods for testing and evaluating survey questionnaires*. Wileyn.
Rossiter, J. (2002). The C-OAR-SE procedure for scale development in marketing. *International Journal of Research Marketing, 19*, 305–335.
Rossiter, J. (2011). *Measurement for the social sciences – The C-OAR-SE method and why it must replace psychometrics*. Springer.
Sarstedt, M., & Wilczynski, P. (2009). More or less? A comparison of single-item and multi-item measures. *Die Betriebswirtschaft DBW, Jg., 69*, 211–227.
Schaeffer, N., & Bradburn, N. (1989). Respondents behavior in magnitude estimation. *Journal of the American Statistical Association, 84*, 402–413.
Schermelleh-Engel, K., & Werner, C. (2007). Methoden der Reliabilitätsbestimmung. In H. Moosbrugger & A. Kelava (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (S. 113–133). Springer.
Schuman, H., & Presser, S. (1981). *Questions and answers in attitude surveys*. Academic.
Schwarz, N. (1999). Self-reports – How questions shape the answers. *American Psychologist, 54*, 93–105.
Schwarz, N., Hippler, H., Deutsch, B., & Strack, F. (1985). Response categories: Effects on behavioral reports and comparative judgments. *Public Opinion Quarterly, 49*, 388–395.
Schwarz, N., Knäuper, B., Hippler, H., Noelle-Neumann, E., & Clark, F. (1991). Rating scales: Numeric values may change the meaning of scale labels. *Public Opinion Quarterly, 55*, 618–630.
Spector, P. (1994). Summated rating scale construction – An introduction. In M. Lewis-Beck (Hrsg.), *Basic measurement* (S. 229–300). Sage.
Statistisches Bundesamt. (Hrsg.). (2019a). Fachserie 15 Reihe 2 Wirtschaftsrechnungen – Laufende Wirtschaftsrechnungen Ausstattung privater Haushalte mit ausgewählten Gebrauchsgütern. Statistisches Bundesamt (Destatis) Wiesbaden.
Statistisches Bundesamt (Hrsg.) (2019b). Statistisches Jahrbuch - Deutschland und Internationale 2019
Strauss, M., & Smith, G. (2009). Construct validity: Advances in theory and methodology. *Annual Review of Clinical Psychology, 5*, 1–25.
Sudman, S., & Blair, E. (1998). *Marketing research – A problem solving approach*. McGraw-Hill.
Sudman, S., Bradburn, N., & Schwarz, N. (1996). *Thinking about answers – The application of cognitive processes to survey methodology*. Jossey-Bass.

Temme, D., Paulssen, M., & Hildebrandt, L. (2009). Common method variance. *Die Betriebswirtschaft DBW, 69*, 123–146.
Thompson, S. K. (2012). *Sampling, Hoboken*. Wiley.
Tourangeau, R., Rips, L., & Rasinski, K. (2000). *The psychology of survey response*. Cambridge University Press.
Viswanathan, M. (2005). *Measurement error and research design*. Sage.
Vomberg, A., & Klarmann, M. (2022). *Crafting survey research: A systematic process for conducting survey research*. In C. Homburg et al. (Hrsg.), *Handbook of market research* (S. 67–119). Springer Nature.
Walsh, G. (2002). *Konsumentenverwirrtheit als Marketingherausforderung*. Gabler.
Warren, C., & Campbell, M. (2014). What makes things cool? How autonomy influences perceived coolness. *Journal of Consumer Research, 41*, 543–563.
Warren, C., Batra, R., Loureiro, L., & Bagozzi, R. (2019). Brand coolness. *Journal of Marketing, 83*(5), 36–56.
Weiber, R., & Mühlhaus, D. (2010). *Strukturgleichungsmodellierung*. Springer.
Weiber, R., & Kleinaltenkamp, M. (2013). *Business- und Dienstleistungsmarketing*. Kohlhammer.
Weisberg, H. (2005). *The total survey error approach*. University of Chicago Press.
Wildner, R., & Jäncke, L. (2010). Validierung von Messinstrumenten für die Markenstärke mit bildgebenden Verfahren. In M. Bruhn & R. Köhler (Hrsg.), *Wie Marken wirken: Impulse aus der Neuroökonomie für die Markenführung* (S. 93–107). Verlag Franz Vahlen.
Zaichkowsky, J. (1985). Measuring the involvement construct. *Journal of Consumer Research, 12*, 341–352.
Zikmund, W. (1997). Exploring Marketing Research, 6. Aufl., Dryden Press
Züll, C. (2015). *Offene Fragen*. Mannheim, GESIS – Leibniz-Institut für Sozialwissenschaften (GESIS Survey Guidelines).

Längsschnittuntersuchungen

5

Zusammenfassung

Im Mittelpunkt dieses Kapitels stehen sogenannte „Panels“. Das sind Untersuchungen, bei denen bei einer gleichbleibenden Menge von Personen, Unternehmen etc. zu regelmäßigen Zeitpunkten gleichartige Daten (z. B. über gekaufte Marken, Ausgaben für bestimmte Produktarten) auf die immer gleiche Art und Weise erhoben werden. Dadurch werden Analysen zu Veränderungen im Zeitablauf (z. B. Marktanteilswachstum) möglich, die für zahlreiche Unternehmen außerordentlich wichtig sind. Deswegen haben Panels einen sehr großen Anteil am „Markt für Marktforschung“.

5.1 Wesen und Arten von Panels und Wellenbefragungen

Panels sind Instrumente, die auf die Messung von Veränderungen hin optimiert sind. Oft sind es ja die Veränderungen, welche im Marketing Aktionen auslösen. So wird ein Marktanteil von 30 % grundsätzlich anders beurteilt, wenn der Marktanteil im Vorjahr 20 % betrug, als wenn diese Zahl bei 40 % lag.

Um zu gewährleisten, dass Änderungen in den Ergebnissen auch tatsächlich auf Änderungen bei dem untersuchten Phänomen zurückzuführen sind und nicht auf Änderungen aufgrund des Instruments, werden nach Möglichkeit eine Reihe von Parametern gleich gelassen.

H. Kreis et al., *Marktforschung,* https://doi.org/10.1007/978-3-658-44456-3_5

▶ **Definition**

Panels können gekennzeichnet werden als Erhebungsinstrumente, bei denen

- die Erhebungsinhalte,
- die Erhebungszeitpunkte,
- die Stichprobe und
- die Erhebungsmethode

nach Möglichkeit konstant gehalten werden.

Das bedeutet aber auch, dass sogenannte Online- oder Befragungspanels (siehe Abschn. 4.3.4.5) keine Panels im Sinne dieser Definition sind. Solche Panels sind Pools von potenziell zu befragenden Menschen, für die jeweils für eine bestimmte Befragung eine Stichprobe ausgewählt und befragt wird. Hier wechseln also Thema, Stichprobe und Erhebungszeitpunkte.

Die *Vorteile* von Panels lassen sich gut am Beispiel des Haushaltspanels aufzeigen, bei dem regelmäßig die Einkäufe privater Haushalte erfasst werden (vgl. Abschn. 5.3). Zunächst ist klar, dass die Erhebungsinhalte konstant bleiben müssen, wenn Veränderungen gemessen werden sollen. Ein Vergleich der Einkaufsmenge von Colagetränken im Jahr 2022 mit der Einkaufsmenge von Mineralwasser im Jahr 2023 ist sinnlos. Ebenso leicht verständlich ist, dass die Erhebungszeitpunkte gleich bleiben müssen. Wird der Einkauf von Sekt im Dezember mit dem Einkauf von Sekt im Januar verglichen, so wird man einen deutlichen Rückgang feststellen, der aber keinen Zusammenbruch des Marktes abbildet, sondern lediglich die Tatsache widerspiegelt, dass ein großer Anteil der jährlich verbrauchten Sektmenge um Weihnachten/Silvester eingekauft wird. Sinnvoll ist also z. B. ein Vergleich Dezember 2023 mit Dezember 2022 oder Januar 2024 mit Januar 2023.

Weiter wird die Stichprobe nach Möglichkeit konstant gehalten. Dafür gibt es zwei Gründe. Einmal ist es dadurch in der Regel möglich, Veränderungen genauer zu messen. Wird ein Haushalt der Stichprobe durch einen anderen Haushalt ersetzt, so kann eine Veränderung bereits dadurch zustande kommen, weil der bisherige Haushalt – aus welchen Gründen auch immer – andere Präferenzen hat, als der neue Haushalt. Es können sich also Unterschiede zwischen zwei Perioden allein aufgrund der Veränderung der Stichprobe ergeben. Misst man hingegen beim gleichen Haushalt eine Veränderung im Einkaufsverhalten, so ist dies eine tatsächliche Änderung im Markt, die sich auch in den Daten niederschlagen soll. Weiter ist es dadurch möglich, Kaufgeschichten zu erfassen. So wird es den Anbieter eines neuen Getränks sehr interessieren, ob die Käufer seines neuen Getränks vorher ein anderes Getränk des gleichen Herstellers gekauft haben oder ein Getränk der Konkurrenz, weil im ersten Fall keine neuen Marktanteile für den Hersteller durch die Neuprodukteinführung gewonnen werden können, im zweiten Fall aber schon. Solche Aussagen sind aber nur für solche Haushalte möglich, von denen das Einkaufsverhalten in beiden Perioden bekannt ist.

Es ist klar, dass sich die Konstanz der Stichprobe nicht vollständig erreichen lässt. Haushalte verlassen das Haushaltspanel aus verschiedenen Gründen. Dies kann u. a. durch Umzug in ein anderes Land, durch Tod, aber auch dadurch entstehen, dass die weitere Zusammenarbeit für den Haushalt zu mühselig ist und daher vom Haushalt eingestellt wird. Der Ausfall von Panelhaushalten wird als **„Panelsterblichkeit“** bezeichnet. Haushalte, welche die Stichprobe verlassen, werden durch gleich strukturierte Haushalte ersetzt. Die Gesamtheit der Panelteilnehmer, welche in mehreren Perioden (z. B. den 12 Monaten eines Jahres) berichten, wird als die „durchgehende Masse“ für diesen Zeitraum bezeichnet. Dagegen werden alle berichtenden Panelteilnehmer einer Periode (z. B. eines Monats), einschließlich der in dieser Periode neu angeworbenen Panelteilnehmer und der Panelteilnehmer, welche im folgenden Monat ausfallen, als „volle Masse“ bezeichnet. Abb. 5.1 verdeutlicht diese Zusammenhänge.

Es ist klar, dass ein hoher Anteil der durchgehenden Masse ein wichtiges Qualitätsmerkmal eines Panels ist. Bei gut geführten Haushaltspanels beträgt der Anteil der durchgehenden Masse eines Jahres 70 bis 80 %.

Schließlich wird angestrebt, die Erhebungsmethode nach Möglichkeit konstant zu halten. Der Grund ist, dass auch die Änderung der Erhebungsmethode zu einer Änderung der Daten führen kann. So wurde früher im **Handelspanel**, bei dem die Abverkäufe von Handelsgeschäften erfasst werden (vgl. Abschn. 5.4), die Inventurmethode eingesetzt, bei der aus den Einkäufen des Geschäfts sowie aus den Beständen zu Beginn und zum Ende einer Periode die Verkäufe berechnet werden. In den 1980er und 1990er Jahren wurde die Methode auf die Erfassung der Verkäufe durch Scannerkassen umgestellt. Dies führte dazu, dass Diebstahl, der früher als Verkauf erfasst wurde, nunmehr nicht mehr in der Verkaufsstatistik berücksichtigt ist.

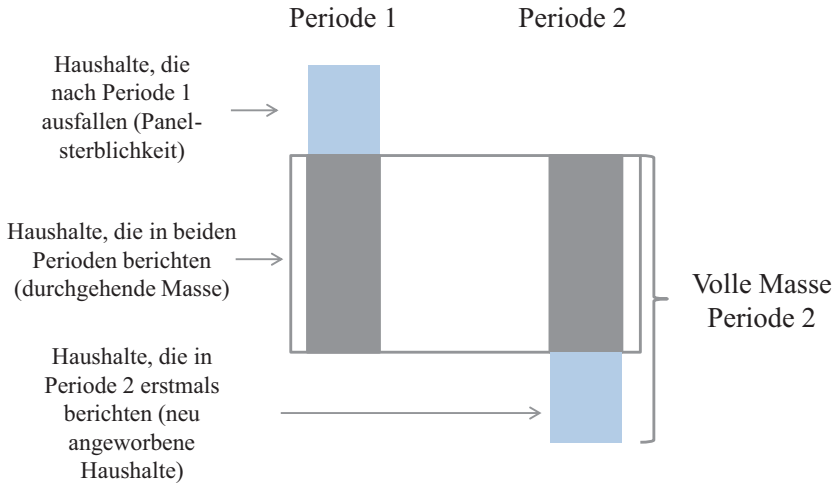


Abb. 5.1 Volle und durchgehende Masse am Beispiel des Haushaltspanels. (Mit freundlicher Genehmigung des GfK e. V.)

Panels lassen sich nach der Art der Panelteilnehmer und den erhobenen Daten unterscheiden. Wichtig sind in der Marktforschung insbesondere folgende Arten von Panels:

- **Verbraucherpanels:** Diese erfassen – anders als der Name es suggeriert – nicht den Verbrauch sondern die Einkäufe von Verbrauchern. Dabei lassen sich unterscheiden:
  - *Großverbraucherpanels,* z. B. Panels von Krankenhäusern oder Kantinen.
  - *Haushaltspanels:* Panels zur Erfassung der Einkäufe von privaten Haushalten.
  - *Individualpanels:* Panels zur Erfassung der Einkäufe von Personen, die in privaten Haushalten leben. Das Individualpanel wird eingesetzt, um solche Einkäufe zu erfassen, die üblicherweise nicht für den Haushalt als Ganzes eingekauft werden, sondern die jede Person üblicherweise für sich einkauft. In der Praxis sind dies Körperpflege, Kosmetik und Süßwaren.
- Der weitaus wichtigste Verbraucherpaneltyp ist das Haushaltspanel.
- **Handelspanels** zur Erfassung der Abverkäufe von Einzelhändlern. Neben den Verkäufen von Geschäften können hier auch Verkäufe von online-Händlern erfasst werden. Dabei gibt es je nach der Art der Geschäfte sehr viele Panels. So werden in einem Lebensmitteleinzelhandelspanel die Verkäufe von Lebensmitteleinzelhandelsgeschäften, in einem Optikerpanel die von Optikern usf. erfasst.
- **Fernsehzuschauerpanels** zur Erfassung des TV-Zuschauererhaltens in privaten Haushalten.
- **Internetpanels** zur Erfassung des Internetsurfverhaltens von Personen in Privathaushalten.
- **Testmarktpanels** zur Bestimmung der Reaktion von Verbrauchern auf Marketingimpulse, wie z. B. neue Produkte oder Werbung. Das wichtigste Beispiel eines solchen Panels ist der Testmarkt GfK BehaviorScan in Haßloch.
- Weitere **Spezialpanels,** wie z. B. Ärztepanels zur Erfassung des Verschreibungsverhaltens von Ärzten oder der Besuche von Pharmareferenten.

Der Aufbau und der Betrieb von Panels erfordern hohe Investitionen und hohe laufende Kosten. Panels werden daher insbesondere von großen Marktforschungsinstituten angeboten, wie A.C. Nielsen, TNS, GfK, IRI und IMS. Panels haben in der Marktforschung eine große Bedeutung. Dies zeigt sich nicht zuletzt dadurch, dass 2022 weltweit etwa 36 % der Marktforschungsumsätze auf Panels entfielen, in Deutschland waren es 41 % (ESOMAR, 2023, S. 164 f.).

Eine Besonderheit der Panelmarktforschung ist, dass häufig die Daten vom jeweiligen Institut auf eigene Rechnung erhoben werden und dann auch dem Institut gehören. Während Umfrageforschung meist im Auftrag eines Kunden stattfindet und die Daten dann auch nur diesem Kunden exklusiv zur Verfügung stehen, können Paneldaten vom Institut an mehrere Kunden verkauft werden. Dies eröffnet den Instituten gute Gewinnmöglichkeiten. Von den Kunden werden Panels in der Regel langfristig bezogen. Dadurch ist die Panelforschung für das Institut auch wesentlich weniger konjunkturabhängig als die Ad Hoc Forschung, was die Panelforschung zusätzlich für die Institute attraktiv macht.

Auch **Wellenbefragungen** sollen Veränderungen abbilden. Es gibt aber einen wesentlichen Unterschied zum Panel. Wellenbefragungen können dadurch gekennzeichnet werden, dass der stets gleiche Erhebungsgegenstand mit der stets gleichen Methode bei *wechselnden* Stichproben erhoben wird. Sie kommen immer dann zum Einsatz, wenn eine feststehende Stichprobe nicht möglich ist (z. B. weil die Mitarbeitsbereitschaft nicht dauerhaft hergestellt werden kann) oder nicht sinnvoll ist, weil z. B. die wiederholte Befragung die Ergebnisse zu sehr verändert. Letzteres gilt für das Haupteinsatzgebiet der Wellenbefragungen, dem Werbetracking, bei dem regelmäßig u. a. nach der Bekanntheit der Werbung gefragt wird (vgl. Abschn. 5.6).

Wellenbefragungen haben gegenüber den Panels den Nachteil, dass sie in der Regel Veränderungen nur ungenauer messen können als Panels. Dieser Verlust an Genauigkeit lässt sich berechnen. Diesbezüglich muss jedoch auf die Spezialliteratur verwiesen werden (vgl. Günther et al., 2019, S 5 ff.)

5.2 Spezielle methodische Probleme der Panelforschung

Aus der Definition von Panels (vgl. Abschn. 5.1) wird deutlich, dass sich für die Panelerfassung nur solche Phänomene eignen, zu deren Erfassung eine *konstante Stichprobe* gebildet werden kann und dies auch sinnvoll ist. Gebildet werden kann eine konstante Stichprobe nur dann, wenn über längere Zeit die Mitarbeitsbereitschaft der Mitglieder eines Panels hergestellt werden kann. Ob dies gelingt, hängt einmal davon ab, wie die Teilnahme belohnt wird. Beim Haushaltspanel kommt hier ein ganzes Paket mit Treueprämien, Geschenken und attraktiven Verlosungen zum Einsatz. Beim Handelspanel sind es dagegen Datenlieferungen (wie erfolgreich ist das betreffende Geschäft im Vergleich zu anderen vergleichbaren Geschäften?) sowie Geldzahlungen, die als Motivation für eine Mitarbeit dienen.

Die Teilnahmebereitschaft hängt aber insbesondere beim Haushaltspanel auch davon ab, wie groß der Aufwand der Teilnahme ist. Hier wurden und werden von den Panelinstituten erhebliche Anstrengungen unternommen, durch den Einsatz entsprechender Geräte diesen Aufwand zu minimieren. Schließlich hängt die Teilnahmebereitschaft auch von der Zielgruppe ab. So lassen sich die Angehörigen der obersten und der untersten Schichten kaum dazu motivieren, über längere Zeit an der Erfassung von Marktforschungsdaten mitzuwirken. Dies führt dazu, dass Haushaltspanels generell eine Tendenz zur Mitte aufweisen. Für manche Warengruppen (z. B. Waschmittel) führt dies kaum zu Verzerrungen. Bei anderen Warengruppen wie z. B. Sekt/Champagner ist jedoch feststellbar, dass sehr preiswerte und sehr teure Produkte nur unterdurchschnittlich erfasst werden.

Des Weiteren ist eine Panelerfassung nur dann sinnvoll, wenn die wiederholte Erhebung bei der gleichen Stichprobe nicht zu einer zu großen Veränderung der Daten führt. Eine solche Verzerrung wird auch als **Paneleffekt** bezeichnet. So ist es nicht sinnvoll, die Kenntnis von Werbung als Panel abzufragen, weil die wiederholte Abfrage zu

einer kontinuierlichen Steigerung der Werbebekanntheit führen würde. Deshalb kommen beim Werbetracking *Wellenbefragungen* zum Einsatz (vgl. Abschn. 5.6). Auch Meinungen oder Überzeugungen sollten nicht in einem Panel abgefragt werden, weil sie durch die mehrfache Beschäftigung mit einem Thema beeinflusst werden können. Damit ist es vor allem das Verhalten, z. B. beim Einkaufen oder bei der Mediennutzung, das panelmäßig erfasst werden kann.

Auch das Einkaufsverhalten von Haushalten kann durch die Abfrage im Haushaltspanel beeinflusst werden, schon weil das Produkt bei den meisten Erfassungsarten nochmals in die Hand genommen und der Preis dazu eingegeben wird. Es lässt sich zeigen, dass Panelhaushalte tendenziell preissensibler reagieren als der Durchschnitt. Dennoch ist diese Verzerrung so gering, dass die Vorteile der Panelerfassung gegenüber den Nachteilen überwiegen.

Eine wesentliche Rolle spielt die Coverage eines Panels. Dabei wird unter **Coverage** der Teil des Marktes verstanden, der durch das Panel erfasst wird. So wird vom Fernsehzuschauerpanel der AGF/GfK in Deutschland der TV-Konsum am Hauptwohnsitz erfasst. Nicht erfasst wird u. a. das Fernsehzuschauerverhalten in Ferienwohnungen, auf Campingplätzen, in Hotels oder Gaststätten.

Eine besondere Bedeutung hat die Coverage bei Handels- und Haushaltspanels für die täglichen Verbrauchsgüter, weil hier der gleiche Markt einmal von der Verkaufs- und einmal von der Einkaufsseite betrachtet wird. Das *Haushaltspanel* erfasst die Einkäufe der privaten Haushalte mit Hauptwohnsitz in Deutschland. Nicht erfasst werden alle sonstigen Einkäufe, u. a. die Einkäufe von Touristen, Asylbewerbern ohne dauerhafte Aufenthaltsgenehmigung, Altenheimen, Krankenhäusern, Büros, Kantinen, Kasernen, Justizvollzugsanstalten und der Gastronomie. Die Einkäufe der privaten Haushalte werden aber grundsätzlich unabhängig davon erfasst, wo sie getätigt wurden. Das *Handelspanel* dagegen erfasst die Kanäle des Handelspanelsegments, also z. B. Rewe, Edeka, netto oder real. Darüber hinaus gibt es aber auch Kanäle, die nicht erfasst werden, weil sie eine Zusammenarbeit mit den Marktforschungsinstituten ablehnen. Dies sind insbesondere die Discounter Aldi, Lidl und Norma. Darüber hinaus gibt es noch Handelsgeschäfte, deren Daten nicht erhoben werden, weil ihre Erfassung zu aufwendig wäre, wie z. B. Flughafenshops oder der Verkauf auf Schiffen.

Die dadurch entstehende Situation verdeutlicht die Abb. 5.2, bei der der Gesamtmarkt durch das große Rechteck, der durch das Handelspanel erfasste Teil durch eine waagrechte und der durch das Haushaltspanel erfasste Teil durch eine senkrechte Schraffur gekennzeichnet ist. Der Gesamtmarkt zerfällt demnach in vier Bereiche, die am Beispiel einer Packung Kaffee verdeutlicht werden sollen. Zunächst gibt es links oben einen Bereich, der von beiden Panels erfasst wird. Der Einkauf einer Packung Kaffee bei Rewe durch einen Privathaushalt ist so ein Beispiel. Rechts daneben ist ein Bereich, der vom Haushaltspanel erfasst wird, nicht jedoch vom Handelspanel, z. B. der Einkauf einer Packung Kaffee durch einen Privathaushalt bei Aldi. Links unten ist ein Bereich, der zwar im Handelspanel erfasst wird, nicht jedoch im Haushaltspanel und der im Beispiel als Einkauf einer Packung Kaffee durch eine Gastwirtschaft bei Edeka beschrieben werden

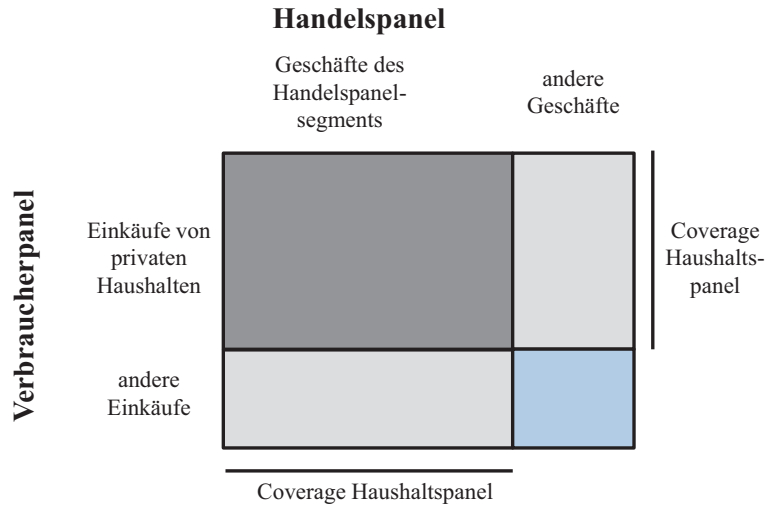


Abb. 5.2 Coverage von Haushalts- und Handelspanel für tägliche Verbrauchsgüter. (Mit freundlicher Genehmigung der GfK SE.)

kann. Schließlich gibt es rechts unten einen Bereich des Marktes, der von keinem der beiden Panels erfasst wird, z. B. der Einkauf durch ein Büro bei Aldi.

Für die Berichterstattung hat dies erhebliche Folgen. Bleibt man beim Beispiel Kaffee, so sind die in beiden Panels ausgewiesenen Mengen zu niedrig. Noch wichtiger ist jedoch, dass die Marktanteile sich erheblich unterscheiden. Im Handelspanel finden sich keine Handelsmarken von Aldi, Lidl oder Norma. Entsprechend sind die Marktanteile der in den Geschäften des Handelspanelsegments verkauften Herstellermarken (z. B. Jacobs, Dallmayr oder Melitta) größer.

Die Höhe der Coverage hängt dabei von verschiedenen Faktoren ab, die im Folgenden anhand des Haushaltspanels erläutert werden sollen. So hat Shampoo eine höhere Coverage als Seife, da Seife anders als Shampoo auch in Firmen, öffentlichen Einrichtungen und in der Gastronomie genutzt wird. Die Coverage kann weiter von der Packungsgröße abhängen. So werden die Großpackungen bei Haushaltsreinigern vor allem im professionellen Einsatz verwendet und haben daher eine geringere Coverage als die üblichen Haushaltspackungen. Bei Schokoriegeln haben die Single-Packungen eine geringere Coverage als die Multipackungen, da die Einer-Schokoriegel oft unterwegs verzehrt werden und daher ihre Erfassung vergessen wird. Die Coverage wird vor allem dann zum Thema zwischen dem Panelkunden und dem Panelanbieter, wenn sie sich ändert, weil dann die Veränderungsrate der Abverkäufe des Herstellers und die im Panel gemessene Veränderungsrate auseinanderklaffen. Dies ist zum Beispiel dann der Fall, wenn der Konsum aufgrund geringerer Arbeitslosigkeit von zuhause auf Kantinen verlagert wird und dadurch die Coverage sinkt.

5.3 Verbraucherpanelforschung

Die Verbraucherpanelforschung soll am Beispiel des besonders wichtigen Haushaltspanels erläutert werden. **Haushaltspanels** zur Erfassung verpackter Verbrauchsgüter wurden in Deutschland von der GfK in Nürnberg und von Nielsen IQ in Frankfurt angeboten. Mit der Übernahme der GfK-Gruppe durch Nielsen IQ wurde das Haushaltspanel der GfK im Januar 2024 an die Firma You Gov verkauft und wird seitdem von dort angeboten. Das umsatzstärkere Haushaltspanel von You Gov hat 30.000 Haushalte, das von Nielsen IQ 20.000 Haushalte.

Für die weitere Erläuterung wird das 30.000er Haushaltspanel von You Gov, früher GfK zugrunde gelegt. Die *Grundgesamtheit* des Haushaltspanels wurde bereits im vorhergehenden Abschnitt erläutert. Die *Stichprobe* ist eine Quotenstichprobe, weil eine Zufallsstichprobe schon aufgrund der zu hohen Anteile der Personen, die nicht zur Mitarbeit bereit sind, nicht möglich ist. Der Anteil der Verweigerer beträgt etwa 95 %. Quotenmerkmale sind hierbei die Region, die Haushaltsgröße, die Zahl der Kinder unter 15 Jahre, das Alter der haushaltsführenden Person sowie die Berufsgruppe des Hauptverdieners bzw. der Hauptverdienerin. Die Stichprobe ist im Wesentlichen proportional. Lediglich Einpersonenhaushalte sind unterrepräsentiert, da sie einmal schwer zu rekrutieren sind, aber auch besonders wenig zur Erfassung der Märkte beitragen. Diese Disproportionalität wird mit der Hochrechnung ausgeglichen.

Die Erhebung im 30.000er Haushaltspanel geschieht bei den Haushalten, die einen Internetanschluss haben, mithilfe eines Lesegeräts, das etwas größer als ein USB-Stick ist. Mit dem Gerät werden alle EAN-Codes der gekauften Artikel erfasst und auf dem Stick gespeichert. Danach wird der Stick über seine USB-Schnittstelle mit einem Rechner verbunden. Der Panelteilnehmer wird zunächst gefragt, wann und in welcher Einkaufsstätte der Einkauf erfolgte. In einem weiteren Schritt werden über das Internet dann die Artikelbeschreibungen der gescannten Artikel aus einer von der GfK gepflegten Artikelstammdatei geholt und dem Panelteilnehmer gezeigt. Die Artikel, die er bearbeiten möchte, werden markiert. Anschließend wird für jeden der markierten Artikel nach der Zahl der gekauften Stück und dem Preis sowie danach, ob der Artikel zum Aktions- oder Normalpreis gekauft wurde gefragt. Abb. 5.3 zeigt diese Erfassung.

Bei Haushalten ohne Internetzugang (ca. 15 % der Panelhaushalte) kommt ein Handgerät zum Einsatz, das eine Leseeinrichtung für Barcodes sowie eine Tastatur enthält. Mithilfe eines Codebuchs wird die Einkaufsstätte erfasst. Dann werden die Artikel jeweils gescannt und Stückzahl, Preis und ob zum Normalpreis oder zum Aktionspreis gekauft wurde, über Menü und die Gerätetastatur eingegeben. Das Gerät speichert diese Informationen und überträgt sie über die Telefonleitung, wenn es in seine Dockingstation gesteckt wird. Die Dockingstation lädt auch die Batterien des Geräts wieder auf.

Für die Berichterstattung werden die erfassten Daten auf die Grundgesamtheit aller privaten Haushalte hochgerechnet. Sieht man von der Disproportionalität der Einpersonenhaushalte oder durch den Ausfall von Panelhaushalten ab, so hat jeder Haushalt

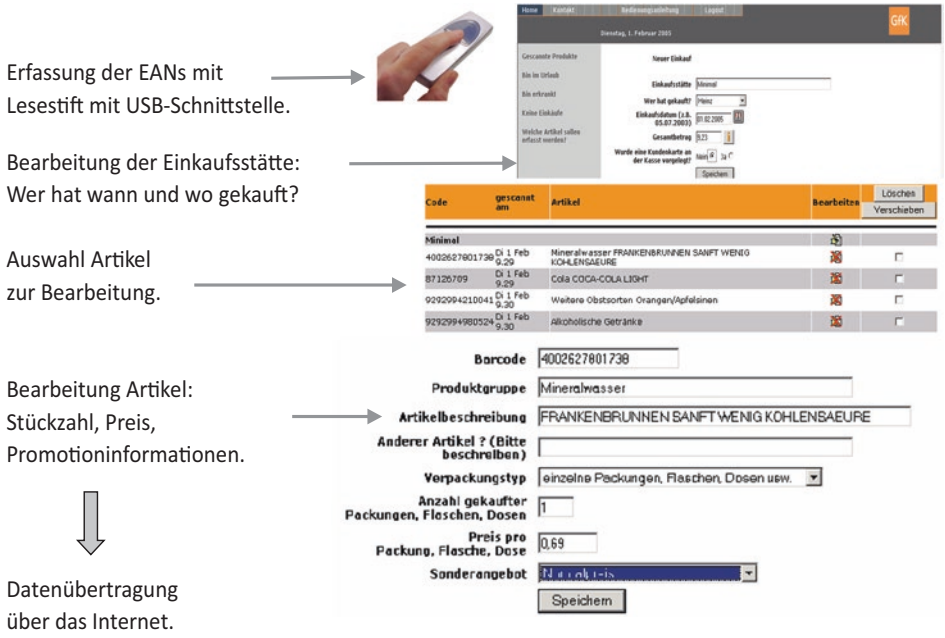


Abb. 5.3 Erfassung von Haushaltspaneldaten. (Mit freundlicher Genehmigung der GfK SE.)

den gleichen Hochrechnungsfaktor. Bei einer Panelstichprobe von 30.000 Haushalten und 40,22 Mio. Haushalten in der Grundgesamtheit ergibt sich ein Hochrechnungsfaktor von 1340. Das bedeutet, dass die Einkäufe eines Panelhaushalts für durchschnittlich 1340 Haushalte der Grundgesamtheit stehen.

Abb. 5.4 zeigt beispielhaft als Ergebnis die Entwicklung der Anteile der Geschäftstypen für verpackte Güter des täglichen Bedarfs. Sehr deutlich ist zu sehen, dass es den LEH Food Vollsortimentern (dabei steht „LEH“ für Lebensmitteleinzelhandel; der Geschäftstyp „Food Vollsortimenter“ umfasst die Supermärkte und kleineren Verbrauchermärkte) durchgehend gelungen ist, Marktanteile auszubauen. Dies ist vor allem auf die erfolgreichen Maßnahmen von Edeka und Rewe zurückzuführen. Weiter ist zu sehen, dass auch die Drogeriemärkte kontinuierlich gewonnen haben. Vor allem DM Werner und Rossmann haben es geschafft, die durch die Pleite von Schlecker entstandenen Lücken schnell wieder zu füllen. Dagegen haben die SB-Warenhäuser, also die großen Verbrauchermärkte auf der grünen Wiese, kontinuierlich verloren. Sie leiden unter der abnehmenden Bereitschaft der Verbraucher, für den Einkauf auch längere Wege in Kauf zu nehmen. Schließlich gelang es den Discountern, die vor 2012 jahrzehntelang gewachsen sind, 2016 ihren Rückgang zu stoppen und 2017 wieder zu gewinnen.

Neben den Informationen zu Einkauf Menge und Wert sind vor allem die käuferbezogenen Maßzahlen relevant. Dabei sind *„Käufer“* definiert als alle Haushalte, die ein

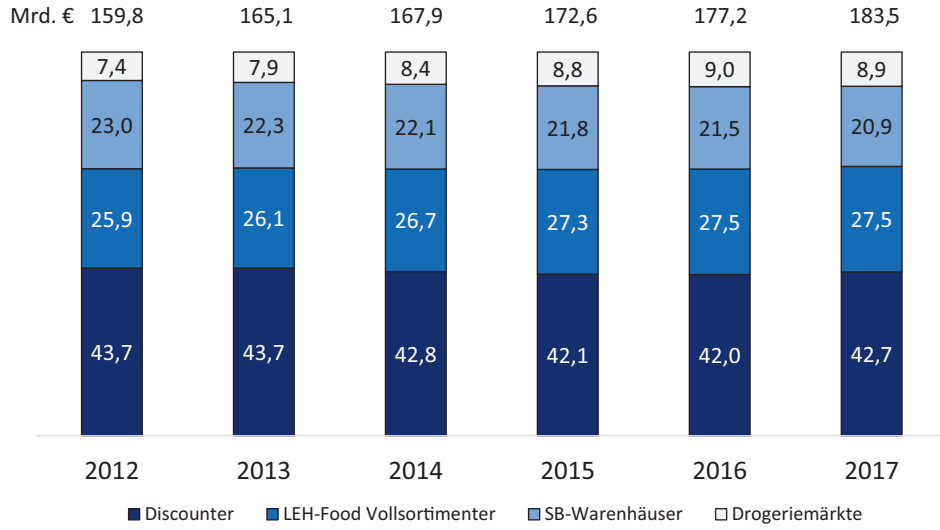
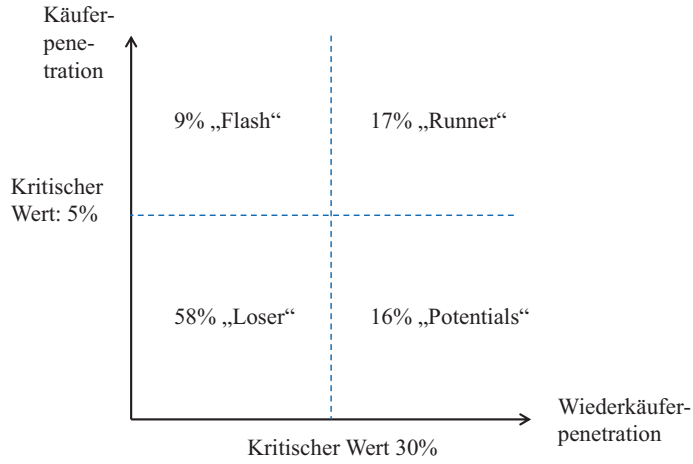


Abb. 5.4 Marktanteile Wert für verpackte tägliche Verbrauchsgüter 2012–2017. (Quelle: GfK Haushaltspanel. Mit freundlicher Genehmigung der GfK SE.)

Produkt in einem bestimmten Zeitraum mindestens einmal kaufen. Eng damit verwandt ist der *Käuferkreis*. Dieser ist definiert als die Zahl der Käufer definiert durch die Zahl aller Haushalte. Er drückt also aus, welcher Anteil aller Haushalte ein Produktangebot probiert hat. Für Warengruppen mit nur eingeschränkter Zielgruppe (z. B. Tiernahrung) interessant ist auch die *Penetration* oder auch *Käuferpenetration,* welche die Zahl der Käufer eines Produkts dividiert durch die Zahl der Warengruppenkäufer. Die Penetration drückt also aus, welcher Anteil der Warengruppenkäufer ein bestimmtes Produkt nutzt.

Die Höhe der Zahl der Käufer drückt aus, ob ein Angebot bekannt, erhältlich und als attraktiv empfunden wird. Ist diese Zahl zu gering, so müssen zunächst die Bekanntheit bzw. der Werbedruck und die Distribution überprüft werden. Ist beides ausreichend vorhanden, so ist die Attraktivität des Produktversprechens, in zweiter Linie auch der Preis des Produkts zu überprüfen.

Vor allem für den langfristigen Produkterfolg wichtig sind auch die Wiederkäufer eines Produkts. *Wiederkäufer* sind definiert als Haushalte, welche ein Produkt in einem bestimmten Zeitraum mindestens zweimal kaufen. Abgeleitet davon ist zunächst die *Wiederkaufsrate,* die berechnet wird, indem die Zahl der Wiederkäufer eines Produkts durch die Zahl der Käufer des Produkts dividiert wird. Diese Zahl drückt also aus, welcher Anteil der Käufer das Produkt wiederkauft. Bei der damit eng verwandten *Wiederkäuferpenetration* wird dagegen die Zahl der Wiederkäufer dividiert durch die Zahl der Käufer, welche nach dem Erstkauf des Produkts in der Warengruppe mindestens noch einmal gekauft haben, egal ob dieses Produkt oder ein anderes Produkt. Die hinter der Berechnung der Wiederkäuferpenetration stehende Logik wird klar, wenn man sich ver-



Die Prozentzahlen beziehen sich auf den Anteil von 265 Neuprodukten der Jahre 2003 und 2004, die 2005 von der GfK untersucht wurden.

Abb. 5.5 Innovationsforschung mit Käuferfacts

deutlicht, dass Wiederkäufe ein Ausdruck der Zufriedenheit mit einem Produkt sind. Von den Haushalten, die das Produkt zwar gekauft haben, danach aber nicht wieder in der Warengruppe gekauft haben, ist nicht bekannt, ob sie zufrieden waren. Die Zufriedenheit kann sich erst im nächsten Kaufakt in der Warengruppe zeigen. Kaufen sie wieder das Produkt, so hat sie dieses wohl überzeugt, ansonsten greifen sie eher zu einem Konkurrenzprodukt.

Ist die Zahl der Wiederkäufer zu gering, so liegt das meist am schlechten Preis-Leistungsverhältnis oder aber daran, dass das Produktversprechen nicht erfüllt wurde. Der Wiederkauf kann gesteigert werden, indem die Produktqualität verbessert wird, der Preis gesenkt wird oder aber indem die Aussagen der Werbung besser auf die tatsächliche Leistungsfähigkeit des Produkts abgestimmt werden.

Abb. 5.5 zeigt die Zusammenschau von Käuferpenetration und Wiederkäuferpenetration in einem Portfolio, das aufgrund einer Untersuchung der GfK von 265 Neuprodukteinführungen beruht.

- 58 % der untersuchten Neuprodukteinführungen erreicht weder bei der Penetration noch bei der Wiederkäuferpenetration die kritischen Werte und müssen deshalb als in keiner Weise erfolgreich bezeichnet werden.
- 9 % der untersuchten Neuprodukteinführungen schafften es zwar, den kritischen Wert bei den Käufern zu erreichen oder zu übertreffen, haben aber einen geringen Wiederkauf. Solche Produkte können anfangs oft schöne Verkaufserfolge erzielen. Wurden jedoch alle potenziellen Erstkäufer erreicht, dann fallen die Abverkäufe aufgrund der fehlenden Wiederkäufer. Wegen dieses kurzfristigen Erfolgs werden sie als „Flash“

bezeichnet. Solche Produkte haben oft ein attraktives Produktversprechen, enttäuschen aber ihre Käufer, weil sie dieses nicht einlösen können.

- Weitere 16 % der untersuchten Neuprodukteinführungen erreichen nur wenige Käufer, können diese aber überzeugen, sodass sie einen hohen Wiederkauf erzielen. Bei diesen Produkten wurde häufig zu sehr am Marketing gespart, sodass das Produkt zu wenig bekannt und/oder zu wenig erhältlich ist. Bei entsprechenden Anstrengungen haben sie sehr wohl das Potenzial zum Erfolg. Sie werden daher als „Potentials" bezeichnet. Die große Gefahr für solche Produkte ist, dass sie von einem Konkurrenten kopiert und die Kopien mit großen Marketingbudgets eingeführt werden.
- Schließlich konnten 17 % aller untersuchten Neuprodukteinführungen genügend Käufer und genügend Wiederkäufer erreichen. Diese auf Anhieb und dauerhaft erfolgreichen Produkte werden als „Runner" bezeichnet.

5.4 Handelspanelforschung

Während das *Verbraucherpanel* die Frage beantwortet, welche *Konsumenten* wie viel einkaufen und damit insbesondere für das Marketing interessant sind, liefert das **Handelspanel** die Information, welche *Geschäfte* die untersuchten Produkte führen und wie viel sie davon unter welchen Bedingungen verkaufen. Handelspanels liefern daher besonders für den Vertrieb unverzichtbare Informationen.

Ein Kennzeichen der Handelspanels ist, dass in ihrer Stichprobe wesentlich größere Mengen der untersuchten Warengruppen bewegt werden. Während in einem Handelspanel meist 1 % oder mehr der gesamten Verkaufsmenge in der Stichprobe enthalten ist, ist es bei Verbraucherpanels häufig weniger als 1 Promille. Dies macht das Handelspanel besonders geeignet für die Marktbeobachtung von seltener gekauften Produkten wie z. B. Mobiltelefonen, Fotoapparaten oder Fernsehern. Zwar können auch diese Produkte in Verbraucherpanels beobachtet werden, deren Berichterstattung liefert aber nur zusätzliche Informationen (z. B. zu welchen Anteilen die Besitzer eines Fernsehers einer bestimmten Marke beim erneuten Kauf eines Fernsehgeräts wieder ein Gerät der gleichen Marke wählen). Die Basisberichterstattung über die Entwicklung von Marktanteilen und Preisen kommt hier vom Handelspanel. Dieser Vorteil des Handelspanels der im Vergleich zum Verbraucherpanel größeren beobachteten Mengen spielt bei den täglichen Verbrauchsgütern aus zwei Gründen eine geringere Rolle: Einmal ist die Zahl der Kaufakte z. B. bei Joghurt oder Waschmittel wesentlich größer als bei Fernsehgeräten oder Computern. Dann ist aber auch die Produktvielfalt bei technischen Geräten wesentlich größer als bei täglichen Verbrauchsgütern.

Auch Handelspanels werden vor allem von großen Instituten angeboten. Für den Bereich der täglichen Verbrauchsgüter ist das Panel von A.C. Nielsen Marktführer. Darüber hinaus bietet in Deutschland IRI ein Handelspanel an. Für den Bereich der Gebrauchsgüter ist es die GfK, die verschiedene Handelspanels anbietet, z. B. für Unterhaltungselektronik, Foto, Informationstechnologie, Elektrogroß- und -kleingeräte für den Haushalt oder auch die Sortimente von Bau- und Gartenmärkten.

Die *Grundgesamtheit* eines Handelspanels wird in der Regel definiert durch Umsatzanteile für bestimmte Sortimente und Verkaufsfläche. So definiert A.C. Nielsen Verbrauchermärkte wie folgt: Diese sind „Einzelhandelsgeschäfte mit mindestens 1000 m^2 Verkaufsfläche, die ein breites Sortiment des Lebensmittel- und Nichtlebensmittelbereichs in Selbstbedienung anbieten." (Nielsen, 2013, S. 10).

Die *Stichprobe* eines Handelspanels wird durch Handelsgeschäfte gebildet. Da die Zusammenarbeit mit Marktforschungsinstituten in der Regel mit der Zentrale der Handelsorganisation vereinbart wird, ist es in der Regel so, dass vom Institut Wünsche für die Stichprobe geäußert werden und von der Zentrale des Handelsunternehmens Stichprobengeschäfte zugeteilt werden. Noch mehr als beim Verbraucherpanel ist also auch beim Handelspanel eine Zufallsstichprobe nicht möglich. Dabei wird die Stichprobe mehrdimensional geschichtet, wie am Beispiel des Lebensmitteleinzelhandelspanels gezeigt werden soll. Schichtungskriterien sind:

- Das Gebiet, wobei sich die sogenannten *Nielsen-Gebiete* als Standard etabliert haben (vgl. Nielsen, 2013, S. 20).
- Der Geschäftstyp, wobei unterschieden wird zwischen Discountern, Verbrauchermärkten, Supermärkten und Drogeriemärkten (Nielsen, 2013, S. 10).
- Die Organisationsform, also das Handelsunternehmen bzw. eine Zusammenfassung von Handelsunternehmen, also z. B. Edeka oder Rewe.

Ein weiterer wichtiger Unterschied zwischen Handels- und Verbraucherpanel betrifft die *Proportionalität der Stichprobe.* Während Verbraucherpanelstichproben im Wesentlichen proportional sind, sind Handelspanelstichproben disproportional. Das bedeutet, dass der Anteil der großen Geschäfte in der Stichprobe größer ist als in der Grundgesamtheit. Das ist insofern sinnvoll, da große Geschäfte wesentlich mehr Menge bewegen als kleine Geschäfte und daher für eine zutreffende Markterfassung die Umsätze der großen Geschäfte genauer geschätzt werden müssen. Diese gewollte Schiefe wird bei der *Hochrechnung* wieder ausgeglichen.

Die *Erhebung* im Handelspanel fand bis in den Anfang der 90er Jahre des vorigen Jahrhunderts zu einem großen Teil noch manuell statt. Dabei wurden die Bestände gezählt und die Einkäufe seit der letzten Zählung durch Belegerfassung bestimmt. Über die Inventurgleichung:

Verkauf = Bestand Vorperiode + Einkauf – Bestand aktuelle Periode

wurden die Verkäufe errechnet. Da dies sehr aufwendig war, wurde nur zweimonatlich berichtet.

Heute werden die Daten nur noch ausnahmsweise manuell erfasst. Bei der großen Mehrheit der Geschäfte werden die Daten aus den Warenwirtschaftssystemen und/oder den Scannerkassen der Geschäfte an das Institut übermittelt. Dadurch konnte der Berichtsrhythmus von zweimonatlich auf wöchentlich verkürzt werden.

Neben den Preisen und den Verkäufen sowie den daraus abgeleiteten Größen wie Marktanteile Menge oder Wert ist es vor allem die Distribution, die im Handelspanel

wichtig ist. Dabei wird unterschieden zwischen der numerischen und der gewichteten Distribution. Dabei drückt die *numerische* Distribution den Anteil der Geschäfte aus, der einen Artikel führt. Die *gewichtete* Distribution drückt dagegen aus, in welchem Anteil des Warengruppenumsatzes ein Produkt vertreten ist.

Bei der Analyse der Einführung von Neuprodukten ist es insbesondere relevant, wie sich die Distribution und die Verkäufe pro führendem Geschäft entwickeln. Geringe Distribution deutet auf eine mangelnde Attraktivität für die Händler oder auf eine geringe Leistung des Außendienstes hin. Sind dagegen die Verkäufe pro führendem Geschäft zu gering, dann liegt dies meist an mangelnder Bekanntheit oder Attraktivität für den Verbraucher (vgl. Wildner, 2007).

Beispiel zur numerischen und gewichteten Distribution

Zur Erläuterung der Distribution gehen wir von 5 Geschäften aus, die alle den gleichen Hochrechnungsfaktor haben und einer Warengruppe, die aus den drei Produkten A, B und C besteht. In einer Periode sind die Verkäufe der Geschäfte bezüglich der drei Produkte wie folgt:

	Geschäfte					Total
Produkte	1	2	3	4	5	
A	3000	4000	1000	6000	–	14.000
B	1000	–	2000	2000	3000	8000
C	–	8000	–	5000	–	13.000
Σ Warengruppe	4000	12.000	3000	13.000	3000	35.000

Produkt B hat danach eine numerische Distribution von 80 %, weil es in vier von fünf Geschäften verkauft wurde. Die gewichtete Distribution beträgt dagegen 66 %, weil die Geschäfte, die B verkauft haben, zusammen einen Warengruppeumsatz von 23.000 € erzielen von insgesamt 35.000 €. Die gewichtete Distribution ist also kleiner als die numerische Distribution, was ein Ausdruck der Tatsache ist, dass B eher in kleineren Geschäften verkauft wird.

Produkt C hat nur 40 % numerische Distribution, aber 71 % gewichtete Distribution. Die gewichtete Distribution ist also größer als die numerische Distribution, was darauf hinweist, dass C vor allem in den großen Geschäften verkauft wird. ◀

5.5 Fernsehzuschauerpanels und Internetnutzungspanels

Fernsehzuschauerpanels dienen der Erfassung des Sehverhaltens von Personen in privaten Haushalten. Dies dient den Fernsehanstalten einmal als Information zur Optimierung des Programms. Die sogenannten „Einschaltquoten“ (das ist der Anteil der tatsächlichen Haushaltsnutzung eines Fernsehprogramms oder einer Sendung an der theoretisch

möglichen Nutzung) sind ein wichtiger Gradmesser für den Erfolg einer Sendung. Dies dient aber auch als Beleg für die Leistungsfähigkeit des Fernsehens als Werbeträger. Die Zahl und die Art der durch die Werbung erreichten Haushalte und Menschen bestimmt letztlich auch den Preis, der für eine Werbeeinschaltung zu bezahlen ist.

Die Organisation der Fernsehforschung unterscheidet sich deutlich von den Handels- und Verbraucherpanels. Handels- und Verbraucherpaneldaten werden von den Instituten erhoben und dann an die interessierten Kunden verkauft. Dabei bleiben die Daten Eigentum der Institute, die sie auch nach Belieben auswerten können. Bei der Fernsehforschung ist es jedoch so, dass sich die Sender gemeinsam die AGF Videoforschung GmbH gegründet haben (vgl. www.agf.de). Die AGF Videoforschung schreibt den Auftrag zur Fernsehforschung aus und kauft dem durchführenden Institut – seit 1985 ist das die GfK-Gruppe in Nürnberg – die Daten ab. Die AGF bestimmt auch, welche Daten ausgewertet werden. So legt die AGF beispielsweise fest, dass für Werbeblöcke die Daten nicht weiter zeitlich differenziert werden. Es gibt demnach keine Reichweiten für den Beginn oder das Ende eines Werbeblocks, sondern nur für den Werbeblock insgesamt.

Durch diese Konstruktion sind Fernsehforschungsdaten eine Währung in dem Sinne, dass die Marktteilnehmer sie in der Regel als die besten verfügbaren Daten anerkennen und mit ihnen arbeiten (vgl. Smith, 2007).

Die Stichprobe der Fernsehforschung in Deutschland beträgt 5.400 berichtende Haushalte. Voraussetzung für die Teilnahme ist, dass der Haushalt mindestens ein stationär betriebenes Fernsehgerät besitzt und der/die Haupteinkommensbezieher/in die deutsche oder eine andere EU-Staatsbürgerschaft hat. Die Erfassung erfolgt durch eine Messtechnologie, die automatisch den eingeschalteten Kanal erfasst. Auch das zeitversetzte Sehen einer Sendung, das z. B. dann entsteht, wenn eine Sendung erst auf einem Festplattenrecorder aufgenommen und später angesehen wird, wird erfasst. Die *Erfassung* erfolgt an allen stationär betriebenen Fernsehgeräten am Hauptwohnsitz des Haushalts. Der Fernsehkonsum in einem Ferienhaus oder in einem Wohnwagen wird also nicht erhoben.

Die Haushalte erhalten auch eine Fernbedienung, auf der sich für jede Person des Haushalts eine Taste befindet. Die Haushaltsmitglieder werden gebeten, sich immer dann anzumelden, wenn sie fernsehen. Auch für Gäste sind Personenanmeldetasten vorgesehen. Dadurch ist es möglich, auch personenbezogene Reichweiten und Einschaltquoten auszuweisen. Die Daten werden an das Institut übertragen und dort ausgewertet. Beispiele für Auswertungen finden sich unter https://www.agf.de/daten/.

Internetnutzungspanels dienen der Feststellung, auf welchen Websites Verbraucher surfen. Dazu ist es erforderlich, eine spezielle Software zu installieren, welche das Surfverhalten erfasst und an das Institut übermittelt. Besonders interessant sind dabei die Zahl der Seitenaufrufe und die Verweildauer auf den Seiten (siehe hierzu auch Abschn. 4.4.3). Derzeit funktioniert dies nur befriedigend auf privaten Computern einschließlich Tabletcomputer und Mobiltelefone. Bei Rechnern am Arbeitsplatz verhindert die Systemadministration der Firma meist die Installation der Erhebungssoftware.

Besondere Schwierigkeiten gibt es, die Kontakte mit der Onlinewerbung zu ermitteln, da diese anders als bei den klassischen Medien Print, Radio oder Fernsehen nicht fest mit dem Medium verbunden wird, sondern in der Regel flexibel aufgrund von Zielgruppendefinitionen eingeschaltet wird. Werbekontakte können daher nur dann ermittelt werden, wenn die Werbung vorab speziell verpixelt wurde (vgl. Nielsen, 2013, S. 65).

Internetnutzungspanels werden derzeit von Nielsen und der GfK angeboten. Während das ältere NetRatings Panel von Nielsen Marktführer ist, hat die GfK 13.000 der 30.000 Haushalte des Haushaltspanels für die Mitarbeit im Online-Panel gewinnen können. Dadurch ist es möglich, das Kaufverhalten für tägliche Verbrauchsgüter mit dem Internetverhalten zu verbinden.

Einen Überblick über Ergebnisse von Paneldatenanalysen zum Kaufverhalten von Konsumenten im Internet liefern Lohse et al. (2000).

5.6 Wellenbefragungen

Wellenbefragungen erheben bei wechselnden Stichproben die stets gleichen Inhalte zu stets wiederkehrenden Zeitpunkten. Sie werden besonders beim Werbetracking eingesetzt, da die wiederholte Befragung in einem Panel das Ergebnis (z. B. die Werbeerinnerung) zu stark beeinflussen würde. Werbetrackings dienen dazu, zu ermitteln, wann eine laufende Werbekampagne verändert bzw. ersetzt werden soll.

Beim **Werbetracking** werden regelmäßig ab ca. 300 Personen der Zielgruppe (also z. B. Käufer der Warengruppe) online zur Aufmerksamkeitswirkung sowie zur Kommunikationsleistung der Werbung befragt. Bei der Aufmerksamkeitswirkung sind wichtig die ungestützte (z. B. „Können Sie sich erinnern, in letzter Zeit Werbung für PKWs gesehen zu haben?“) und die gestützte Werbeerinnerung (z. B. „Haben Sie in letzter Zeit Werbung für Volkswagen gesehen?“), die erinnerten Inhalte sowie die Zuordnung von einzelnen Teilen der Werbung (z. B. Slogans) zu Marken. Bei der Kommunikationsleistung wird erhoben, wie die Werbung die Wahrnehmung und die Bevorzugung der beworbenen Marke verändert. Hier sind u. a. Image, Markenpräferenz und Kaufabsicht wichtig.

Für die Auswertung werden die Werbeausgaben, die der Werbetrackingkunde zur Verfügung stellt, mit den Werten zur Erinnerungs- und Kommunikationsleistung korreliert. Dabei zeigt sich, ob und wie die Leistungswerte auf die Erhöhung oder Reduzierung der Werbeausgaben reagieren. Wichtig sind dabei das Niveau, die Reagibilität und der Trend bei den Leistungswerten.

Dies soll anhand der Kennzahl der gestützten Werbeerinnerung für vier verschiedene Muster erläutert werden (vgl. Abb. 5.6):

- Im *Beispiel 1* ist deutlich zu sehen, dass die Kurve der Erinnerungswerte auf die Erhöhung der Werbeausgaben oder Spendings reagiert. Gehen die Werbeausgaben zurück, dann geht zwar auch die Erinnerung zurück, sie fällt aber nicht auf das ursprüngliche Niveau zurück, d. h. der Trend ist positiv. Diese Kampagne sollte also noch weitergefahren werden.

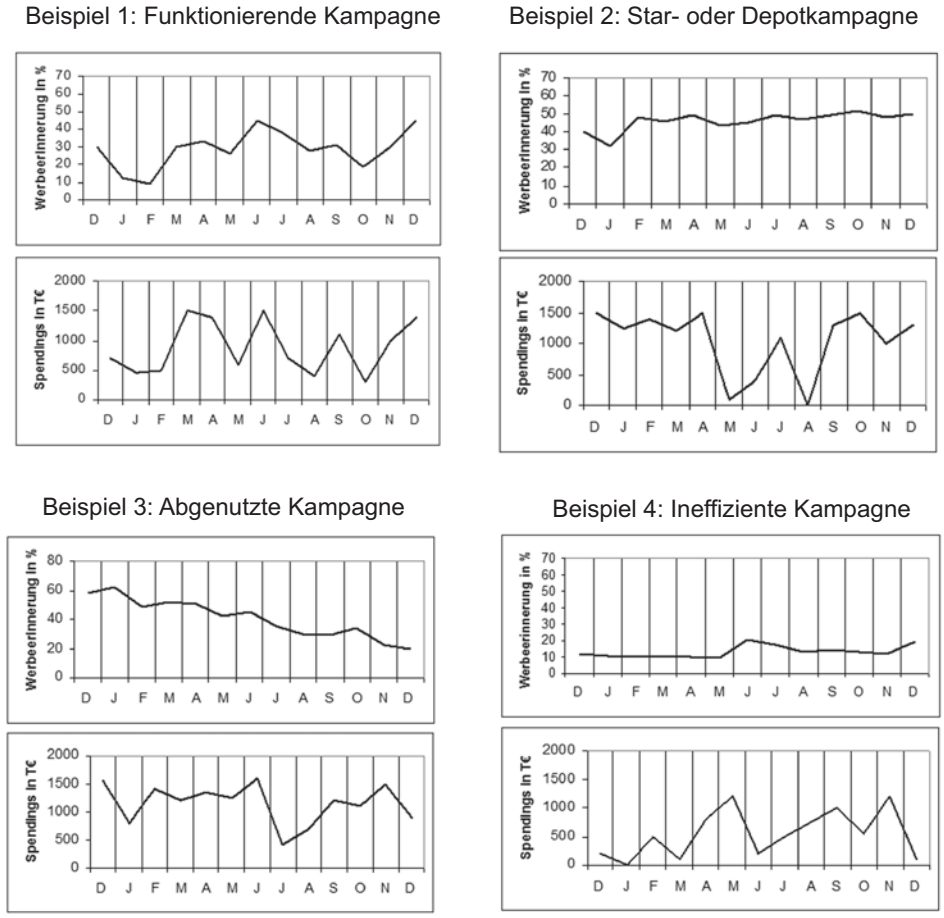


Abb. 5.6 Beispiele für Auswertungen des Werbetrackings

- *Beispiel 2* zeigt den eher seltenen Fall einer sehr erfolgreichen Kampagne, die sich so fest eingeprägt hat, dass sie auch dann erinnert wird, wenn die Werbung zurückgeht. Dies ist einerseits positiv zu werten, weil es ermöglicht, die Werbung zurückzufahren. Andererseits wird es sehr schwer sein, eine neue Kampagne in den Köpfen zu verankern, wenn dies notwendig werden sollte.
- *Beispiel 3* zeigt eine Kampagne, die in der Vergangenheit durchaus erfolgreich war und ein hohes Erinnerungsniveau erzielen konnte, die nun aber abgenutzt ist und auch bei hohen Spendings keine Steigerungen mehr erzielen kann. Diese Kampagne sollte schnell modifiziert bzw. ersetzt werden.
- Schließlich zeigt *Beispiel 4* eine Kampagne, die von Anfang an erfolglos war und gar nicht hätte geschaltet werden sollen. Auch hier ist eine schnelle Änderung erforderlich.

Literatur

ESOMAR. (2023). *Global Market Research 2023*. ESOMAR.

Günther, M., Vossebein, U., & Wildner, R. (2019). *Marktforschung mit Panels* (3. Aufl.). Gabler

Lohse, G. L., Bellman, S., & Johnson, E. J. (2000). Consumer buying behavior on the internet: Findings from panel data. *Journal of Interactive Marketing, 14*(1), 15–29.

Nielsen. (2013). *Deutschland 2013 – Handel, Verbraucher, Werbung*. Nielsen.

Smith, D. (2007). The role and changing nature of marketing intelligence. In ESOMAR (Hrsg.), *Market research handbook* (5. Aufl., S. 3–36). Wiley.

Wildner, R. (2007). Launch and monitoring of in-market performance. In ESOMAR (Hrsg.), *Market research handbook* (5. Aufl., S. 199–215). Wiley.

Experimentelle Untersuchungen und Markttests

6

Zusammenfassung

Experimente haben für unterschiedlichste wissenschaftliche Disziplinen zentrale Bedeutung und eine lange Tradition. Die Grundidee von Experimenten in der Marktforschung besteht darin, dass man nicht nur gegebene Merkmale (z. B. Einkommen, Mediennutzung, Markenpräferenzen) misst, sondern dass man „Manipulationen" an so genannten „unabhängigen Variablen" vornimmt und beobachtet, welche Auswirkungen diese Manipulationen auf so genannte „abhängige Variable" haben. Beispielsweise kann man sich vorstellen, dass man in verschiedenen Supermärkten die Preise eines Produkts unterschiedlich verändert („manipuliert") und dann analysiert, wie sich diese Preisänderungen auf die jeweiligen Absatzmengen auswirken (→ Preis-Absatz-Funktion). Dieses Untersuchungsdesign erlaubt Aussagen über Ursache-Wirkungs-Beziehungen, stellt aber auch besondere Anforderungen an die Durchführung entsprechender Untersuchungen, auf die im 6. Kapitel eingegangen wird. Zahlreiche in der Praxis gängige Produkttests, Kommunikationstests etc. beruhen auf den Prinzipien experimenteller Designs.

6.1 Experimentelle Designs

Wie bereits in Abschn. 2.4.2.2 dargestellt, versteht man (nicht nur) in der Marktforschung unter einem **Experiment** eine Vorgehensweise, bei der eine oder mehrere der sogenannten unabhängigen Variablen derart manipuliert werden, dass die entsprechenden Auswirkungen auf abhängige Variable beobachtet werden können. Experimente spielen in der Marketingforschung insbesondere bei Kausal-Untersuchungen (siehe Abschn. 2.4) eine zentrale Rolle. Generell gelten Experimente als eine der wichtigsten Forschungsmethoden für unterschiedlichste Disziplinen und werden deshalb auch

H. Kreis et al., *Marktforschung,* https://doi.org/10.1007/978-3-658-44456-3_6

in der wissenschaftstheoretischen Literatur entsprechend beachtet (siehe z. B. Arabatzis, 2008; Feest & Steinle, 2016). Es geht dabei darum festzustellen, ob bestimmte (unabhängige) Variable *tatsächlich* der Grund (bzw. die Ursache) für Veränderungen anderer (abhängiger) Variabler (bzw. der Wirkungen) sind. Daneben spielen Experimente für die Marketing-Praxis eine bedeutsame Rolle, wenn es gilt, die Wirksamkeit von Marketing-Maßnahmen, die im Markt zuvor noch nicht eingesetzt wurden, abzuschätzen und zu prognostizieren. Zahlreiche in der Praxis angewandte Produkt-, Preis- und Werbe-Tests (Koschate, 2008, S. 119) dienen diesem Zweck, indem Experimente oder Vorformen davon durchgeführt werden (siehe dazu die Abschn. 6.4 bis 6.8).

Typisch für Experimente ist die gewissermaßen isolierte Betrachtung der interessierenden Variablen. Man will hier nicht die Vielzahl von z. B. auf eine Kaufentscheidung einwirkenden Faktoren betrachten, sondern fokussiert die Untersuchung beispielsweise auf den Einfluss der Werbung auf eine Kaufentscheidung. Deswegen findet man bei experimentellen Untersuchungen häufig eine gewisse Künstlichkeit der Untersuchungssituation, die durch Konstanthaltung bzw. Ausschluss von anderen Einflussfaktoren (→ „Ausschluss alternativer Erklärungsmöglichkeiten“) begründet ist.

Beispiel

Chalmers (1999, S. 28) illustriert das für Experimente typische Bestreben der isolierten Betrachtung der relevanten Variablen an einem einfachen Beispiel:

> „Viele Arten von Prozessen wirken in unserer Umwelt gleichzeitig und sie überlagern und beeinflussen sich wechselseitig in komplizierter Weise. Ein herabfallendes Blatt ist gleichzeitig der Schwerkraft, dem Luftwiderstand, der Kraft des Windes und ein wenig einem Verrottungsprozess ausgesetzt. Es ist nicht möglich, diese verschiedenen Prozesse zu verstehen, wenn man die typischen Abläufe in natürlicher Umgebung sorgfältig beobachtet. Die Beobachtung fallender Blätter führt nicht zu Galileo's Fallgesetzen. Die Lehre, die daraus zu ziehen ist, ist ziemlich klar. Um Daten zu erhalten, die für die Identifizierung und Beschreibung der verschiedenen in der Natur ablaufenden Prozesse relevant sind, ist es im Allgemeinen notwendig zu intervenieren, um den untersuchten Prozess zu isolieren und die Wirkungen anderer Prozesse zu eliminieren. Kurz gesagt: Es ist notwendig, Experimente durchzuführen.“ ◀

Die grundlegende Idee von – hier zunächst sehr vereinfachend dargestellten – Experimenten sei anhand des in Abb. 6.1 dargestellten Beispiels illustriert. Es geht dabei um die Fragestellung, ob eine erhöhte Kundenzufriedenheit zu höherer Markentreue führt. Das wird hier dadurch überprüft, dass man bei einer Teilmenge der Kunden in einem bestimmten Verkaufsgebiet die Kundenzufriedenheit durch besondere Anstrengungen bei Produktqualität, Service, Lieferzeiten etc. beeinflusst *(„manipuliert“)* und dann beobachtet, ob sich die Markentreue bei diesen Kunden in der erwarteten Weise verändert. Dahinter steht der simple Gedanke, dass bei *Vorliegen eines Kausalzusammenhanges* von zwei Merkmalen eine Veränderung eines Merkmals (des vermuteten Grundes) eine entsprechende Veränderung des anderen Merkmals (des Effekts) zur Folge haben müsste.

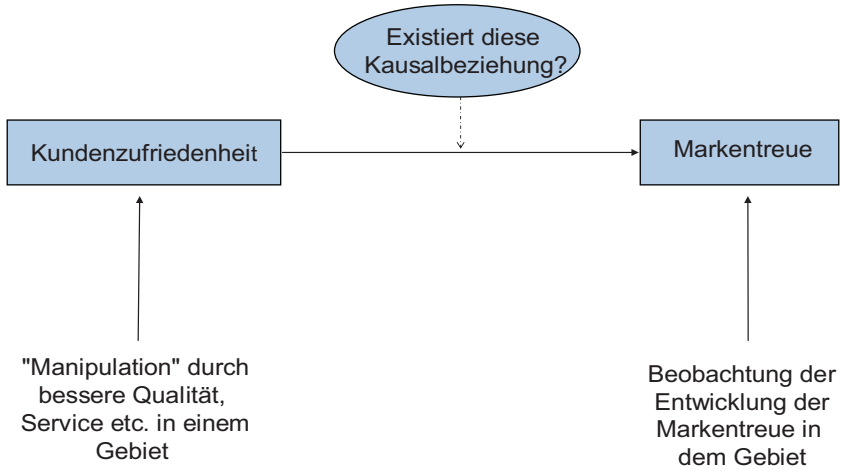


Abb. 6.1 Beispiel für die Vorform eines Experiments in der Marktforschung

In diesem sehr einfachen Beispiel erkennt man schon, dass zu Experimenten normalerweise die *Manipulation der unabhängigen Variablen* und die *Messung der abhängigen Variablen* gehören. Zunächst sei kurz erläutert, wie diese Grundidee bei der Anlage experimenteller Untersuchungen in der Marktforschung realisiert wird. Für die **Manipulation unabhängiger Variabler** muss natürlich zunächst festgelegt werden, welche das sein sollen, was sich im Wesentlichen aus der zu untersuchenden praktischen oder theoretischen Fragestellung ergibt.

Beispiel

Zum Beispiel erfordert die Untersuchung der Auswirkungen von Preisänderungen auf Absatzmengen eben entsprechende Manipulationen von Angebotspreisen (in verschiedenen Läden oder Verkaufsgebieten). „Bei Preisexperimenten werden in realen oder der Realität nachgestellten Kaufsituationen alternative Preise vorgegeben und die Wirkung auf Absatz oder Marktanteil anhand des Verhaltens der Testkäufer erfasst." (Simon & Fassnacht, 2016, S. 144). Wenn man das für unterschiedliche Angebotspreise tut und die sich ergebenden Mengenänderungen bei Verkäufen erfasst, dann erhält man entsprechend viele Messpunkte, um eine Preis-Absatz-Funktion schätzen zu können. Hier geht es also nicht darum, Kausalität zu überprüfen, sondern um Messwerte für die die Feststellung eines funktionalen Zusammenhangs (wie das übrigens in den Naturwissenschaften oft getan wird). ◀

Wie kann man aber psychische Merkmale von Konsumentinnen und Konsumenten (z. B. Wissen, Motivation, Interesse an Werbung) manipulieren, um deren Wirkun-

gen zu untersuchen? Dazu bedarf es im experimentellen Design eines Schrittes, in dem man die Versuchspersonen so beeinflusst, dass die gewünschten Zustände bei den verschiedenen Untersuchungsgruppen erreicht werden. Beispielsweise könnte man versuchen, unterschiedliches Interesse an Werbebotschaften durch eine einleitende entsprechende Aufforderung und/oder durch materielle Anreize zu erreichen. Hier handelt es sich um eine Art der **Operationalisierung** (siehe auch Abschn. 2.2.2), bei der bestimmte Ausprägungen eines interessierenden Konzepts (hier: „Interesse an Werbung“) durch bestimmte Maßnahmen bei den Versuchspersonen erreicht werden sollen. Aronson et al. (1998, S. 111) sprechen deshalb in diesem Zusammenhang von der „Konstruktion der unabhängigen Variablen“. Um zu überprüfen, ob diese Manipulation tatsächlich gelungen ist, werden oftmals sogenannte **„manipulation checks“** durchgeführt. Dabei werden entsprechende Messungen der (manipulierten) unabhängigen Variablen vorgenommen, die (nicht zuletzt den Auftraggebern einer Untersuchung bzw. den Gutachtern bei einer Publikation) zeigen sollen, dass die verschiedenen Versuchs- und Kontrollgruppen (s. u.) eines Experiments sich im Hinblick auf diese Variable tatsächlich in der beabsichtigten Weise systematisch voneinander unterscheiden.

Bei einem experimentellen Design wird also gewissermaßen jede Ausprägung einer unabhängigen Variablen (z. B. „geringes Interesse an Werbung“ vs. „großes Interesse an Werbung“) durch eine entsprechende Gruppe von Versuchspersonen repräsentiert. Der Vergleich zwischen diesen Gruppen im Hinblick auf die interessierende abhängige Variable (z. B. „Erinnerung an eine Werbebotschaft“) erlaubt dann Schlüsse hinsichtlich eines Kausalzusammenhanges, z. B. „Interesse“ → „Erinnerung“. Allerdings zeigt sich hier schon, dass eine größere Zahl von unabhängigen Variablen mit mehr oder weniger verschiedenen Ausprägungen auch eine entsprechende Zahl von Untersuchungsgruppen erfordert (siehe Abb. 6.7 und 6.8), was wiederum zu einer schnell wachsenden Komplexität des Untersuchungsdesigns und großem Untersuchungsaufwand führt. Das gilt insbesondere, wenn auch Interaktionseffekte zwischen verschiedenen Variablen untersucht werden sollen (siehe dazu das Beispiel am Ende des vorliegenden Abschnitts).

Nun zur **Messung der abhängigen Variablen.** Auch die Festlegung der Variablen, deren Veränderung durch die Untersuchung bei der unabhängigen Variablen erklärt werden soll, ergibt sich naturgemäß aus der jeweiligen praktischen oder theoretischen Fragestellung. Generell werden hier Messungen durch Befragungen (z. B. im Hinblick auf Einstellungen, Wissen, Absichten) und Beobachtungen (z. B. von Wahl- oder Informationsverhalten) unterschieden (Aronson et al., 1998). Für die Entwicklung entsprechender Messinstrumente gelten die Grundsätze, wie sie im Abschn. 4.3 diskutiert wurden (siehe dazu auch Abb. 6.11).

Beispiel

Die vorstehend skizzierten Gesichtspunkte der Manipulation und Messung von Variablen seien durch ein einfaches Beispiel illustriert. Es geht um die (eher praktische) Fragestellung, ob „Interesse an Smartphones“ eine Ursache für eine höhere „Preisbereitschaft bei Smartphones“ ist.

Zunächst müsste man wohl definieren, was mit „Interesse“ bzw. „Preisbereitschaft“ wohl gemeint ist. Keine einfachen Fragen. Die Leserin bzw. der Leser dieses Lehrbuchs sei ermuntert, sich an einer solchen Definition (schriftlich formuliert!) zu versuchen.

Hier wird beim „Interesse“ von einer Aufmerksamkeit und Neigung gegenüber einer Sache ausgegangen und bei der „Preisbereitschaft“ von dem Betrag, den ein Kunde in einer bestimmten Marktsituation als Preis eines Produkts *gerade noch* akzeptieren würde. Für eine experimentelle Untersuchung müsste man also zunächst das Interesse der Versuchspersonen an Smartphones manipulieren und dann für die verschiedenen Gruppen (geringes oder großes Interesse) die Preisbereitschaften messen und feststellen, ob sich der erwartete Unterschied (größere Preisbereitschaft bei größerem Interesse) ergibt.

Wie könnte man das Interesse der Versuchspersonen manipulieren? Evtl. könnte man der Gruppe, bei der das Interesse gesteigert werden soll, zu Beginn eine kurze Präsentation anbieten, in der besondere Vorteile von Smartphones in Notsituationen, hinsichtlich eigener Fitness-Programme und bei der Überwachung der eigenen Finanzen anschaulich und eindrucksvoll dargestellt werden. Vielleicht könnte man die Wirkung solch einer Präsentation noch dadurch steigern, dass man im Anschluss eine Art „Quiz“ zu diesem Thema durchführt, bei dem die Versuchspersonen Geldbeträge gewinnen können. Bevor man eine Veränderung der Preisbereitschaften misst, müsste man zunächst feststellen bzw. messen, ob die Manipulation in der einen (Versuchs-) Gruppe tatsächlich zu einem erhöhten Interesse geführt hat (manipulation check).

Anschließend müsste die Messung der Preisbereitschaft erfolgen. Auch das ist keine einfache Aufgabe. Eine direkte Frage („Wie viel € würden Sie für ein Smartphone maximal bezahlen?“) dürfte problematisch sein, weil viele Versuchspersonen gewissermaßen „taktisch“ antworten, da sie wissen, dass niedrige Angaben die künftige Preisgestaltung des Anbieters in der gewünschten Richtung beeinflussen können. Hier bedürfte es also spezieller Messmethoden für die Preisbereitschaft (siehe z. B. Voeth & Niederauer, 2008). ◄

Wenn man unabhängige Variable manipuliert und abhängige Variable gemessen hat und wenn die Ergebnisse den Erwartungen entsprechen, hat man dann schon einen *Kausalzusammenhang* bestätigt? Dazu sei hier auf Überlegungen im Abschn. 2.4.1 zurückgegriffen. Dort waren die folgenden vier Charakteristika eines Kausalzusammenhanges identifiziert worden:

1. Theoretische Begründung des Zusammenhanges
2. Gemeinsame Variation von „Grund“ und „Effekt“
3. Veränderung des „Grundes“ geht der Veränderung des „Effekts“ voraus
4. Ausschluss alternativer Erklärungsmöglichkeiten für den beobachteten Zusammenhang

Inwieweit wird man nun mit dem bisher skizzierten (sehr einfachen) Untersuchungsansatz diesen Charakteristika gerecht? Der erste Gesichtspunkt *(theoretische Begründung; einfacher gesagt: Der vermutete Zusammenhang muss sinnvoll sein.)* ist typischerweise eine de facto notwendige Voraussetzung für eine solche Untersuchung. Wie könnte man eine Untersuchung mit der Manipulation von (unabhängigen) Variablen und der anschließenden Messung abhängiger Variablen anlegen, ohne dass man (durch theoretische Überlegungen oder eine praktische Problemstellung) bestimmt hat, welche Variablen dabei zu berücksichtigen sind? Die Festlegung der in die Untersuchung einzubeziehenden Variablen ergibt sich also weitgehend aus der jeweiligen theoretischen Fragestellung bzw. der praktischen Problemstellung.

Die Gesichtspunkte (2) und (3) ergeben sich gewissermaßen automatisch aus dem experimentellen Design. Wenn man einerseits unterschiedliche Ausprägungen einer unabhängigen Variablen festgelegt hat und andererseits die Werte der entsprechenden abhängigen Variablen misst, dann kann man sofort oder mit relativ einfachen statistischen Methoden feststellen, ob diese in der erwarteten Weise *gemeinsam variieren.* Beispielsweise könnte man vermuten, dass geringe (bzw. intensive) Werbung eine der Ursachen für geringe (bzw. hohe) Marktanteile ist, dann müssten in Verkaufsgebieten mit geringer (bzw. intensiver) Werbung („Grund“) die Marktanteile („Effekt“) auch eher gering (bzw. eher hoch) sein, also in diesem Sinne gemeinsam variieren. Die *zeitliche Abfolge* von Veränderung des „Grundes“ und Veränderung des „Effekts“ ist bei Experimenten ebenfalls durch den Untersuchungsablauf gegeben, weil eben immer die unabhängige Variable („Grund“) manipuliert wird, *bevor* der „Effekt“ gemessen wird. Dem entsprechend gibt es (oft, aber nicht immer; siehe Abb. 6.2 bis 6.6) eine „Vormessung“ vor der Wirkung der unabhängigen Variablen und eine „Nachmessung“ nach dem Einsatz dieser Variablen.

Deutlich schwieriger ist der *Ausschluss alternativer Erklärungsmöglichkeiten* für einen Zusammenhang, womit ja sichergestellt werden soll, dass eine Veränderung des „Effekts“ tatsächlich auf den vermuteten „Grund“ zurückzuführen sein wird und nicht auf andere Einflüsse. Weil man in der Forschungspraxis sicher nicht *sämtliche denkbaren* alternativen Erklärungsmöglichkeiten ausschließen kann, beschränkt man sich hier auf eine überschaubare Anzahl *plausibler* alternativer Erklärungen. Das folgende Beispiel (Quelle: Jacoby, 2013, S. 223 ff.) mit den Abb. 6.2 bis 6.8 ist hauptsächlich auf typische Schritte beim Ausschluss alternativer Erklärungsmöglichkeiten für ein Untersuchungsergebnis ausgerichtet. Wenn man (im kaum realisierbaren Idealfall) *alle* alternativen Erklärungen ausschließen kann, dann kann das Ergebnis eben *nur* durch die (manipulierte) unabhängige Variable verursacht sein.

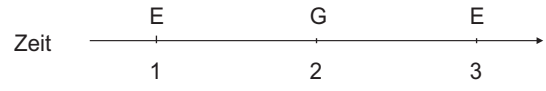


Abb. 6.2 Design 1. (Nach Jacoby, 2013, S. 230)

Wichtig

Jaccard und Becker (2002, S. 248) formulieren einen wesentlichen Grundsatz für die Feststellung von Kausal-Zusammenhängen:

> „Die Möglichkeit, einen Kausal-Zusammenhang zwischen zwei Variablen festzustellen, hängt vom jeweiligen Untersuchungsdesign ab, nicht von den statistischen Methoden, die zur Analyse der erhobenen Daten verwendet werden."

In den verschiedenen Schritten des Beispiels von Jacoby (2013) werden für eine hypothetische Fragestellung (hier: Zusammenhang von Werbung und Kaufabsichten) verschiedene in Frage kommende experimentelle Designs vorgestellt und jeweils mögliche alternative Erklärungsmöglichkeiten erläutert, die dann zu einer Weiterentwicklung dieser Designs zum Ausschluss der jeweiligen Alternativen führen. Dafür ist eine bestimmte Terminologie bzw. Darstellungsform zweckmäßig, die zunächst erläutert sei. Eine Kausal-Hypothese bezieht sich darauf, dass ein vermuteter **G**rund (eine unabhängige Variable, ein „Stimulus"), der mit **G** bezeichnet wird, einen vermuteten **E**ffekt (eine abhängige Variable, eine „Reaktion"), der als **E** bezeichnet wird, zur Folge hat. Im Rahmen experimenteller Untersuchungen bezeichnet man die Gruppe von Versuchspersonen, die dem „vermuteten Grund" ausgesetzt waren, als Experimentgruppe, als Testgruppe oder auch als **Versuchsgruppe** (v). Die Vergleichsgruppe von Versuchspersonen, die dem „vermuteten Grund" *nicht* ausgesetzt waren, nennt man **Kontrollgruppe** (k).

Als Beispiel für die Darstellung der verschiedenen folgenden experimentellen Designs wird also eine Untersuchung der Auswirkungen der Konfrontation von Personen mit Werbung (unabhängige Variable) auf deren Kaufabsichten (abhängige Variable) unterstellt. Alle Designs sind geeignet, drei der vier oben genannten Bedingungen für die Annahme eines Kausal-Zusammenhangs zu überprüfen. Mit zunehmend komplexeren Designs versucht man, der vierten Bedingung – *Ausschluss alternativer Erklärungsmöglichkeiten* – immer besser zu genügen.

Beim *Design 1* (Abb. 6.2) sieht man, dass zu zwei Zeitpunkten ($t=1$; $t=3$) die Ausprägung der abhängigen Variablen (E) „Kaufabsichten" gemessen wird. Zum Zeitpunkt $t=2$ wird die unabhängige Variable (G) „Konfrontation mit Werbung" wirksam. Wenn danach (zum Zeitpunkt $t=3$) die Kaufabsichten bei den Versuchspersonen sich deutlich von denen zum Zeitpunkt $t=1$ unterscheiden, dann vermutet man, dass die Werbung zum Zeitpunkt $t=2$ die Ursache dafür sein könnte. Kann man aber alle alternativen Erklärungsmöglichkeiten für die beobachteten Werte ausschließen? Es könnte ja sein, dass die Veränderung der Kaufabsichten zwischen $t=1$ und $t=3$ nicht auf die Werbung, sondern auf einen anderen Einflussfaktor – z. B. eine allgemeine Geschmacks- oder Einkommensänderung – zurückzuführen ist.

Mit *Design 2* (Abb. 6.3) versucht man, im Hinblick auf diese alternative Erklärungsmöglichkeit eine Kontrolle vorzunehmen, indem man die Versuchspersonen in zwei Gruppen (Versuchsgruppe und Kontrollgruppe) aufteilt. Bei Design 2 wird also nur die Versuchsgruppe mit der Werbung konfrontiert, bei der Kontrollgruppe wird nur eine Ver-

gleichsmessung dergestalt vorgenommen, dass geprüft wird, ob sich (ohne Einfluss der Werbung) eine Veränderung der Kaufabsichten zwischen den Zeitpunkten $t=1$ und $t=3$ ergeben hat. Wenn die entsprechenden Unterschiede der Kaufabsichten (vorher – nachher) in der Versuchsgruppe deutlich größer sind, so spricht das für die Hypothese, dass die Werbung die Ursache dieser Veränderung ist.

Die **Bildung von Versuchs- und Kontrollgruppen** ist eine der wichtigsten und gängigsten Arten des Ausschlusses alternativer Erklärungsmöglichkeiten. Die Grundidee ist ganz einfach: Die Gesamtheit der Versuchspersonen wird in Gruppen aufgeteilt, von denen nur eine/einige der Wirkung der unabhängigen Variablen ausgesetzt werden (Versuchsgruppen). Bei den anderen Gruppen (Kontrollgruppen) ist dies nicht der Fall und die Messung der abhängigen Variablen erfolgt ohne den Einfluss der unabhängigen Variablen. Deutliche (signifikante) Unterschiede der Werte der abhängigen Variablen in Versuchs- und Kontrollgruppen weisen darauf hin, dass die unabhängige Variable die Ursache für die Unterschiede war.

Warum sind für diese Schlüsse in der Regel solche Vergleiche erforderlich? Es könnte ja sein, dass die Veränderung der abhängigen Variablen in Wirklichkeit nicht durch die unabhängige Variable verursacht wurde, sondern durch andere Einflüsse, z. B. könnte eine Veränderung des Marktanteils eines Produkts nicht durch die vermutete Ursache „Intensivierung der Werbung“ begründet sein, sondern durch Ausscheiden von Konkurrenten, Preiserhöhungen von Konkurrenten, Geschmackstrends etc. Wenn man nicht nur beobachtet, wie sich der Marktanteil im Zusammenhang mit einer Intensivierung der Werbung in bestimmten Regionen (Versuchsgruppe) entwickelt hat, sondern auch beobachtet hat, wie sich der Marktanteil in Gebieten entwickelte, in denen die Werbung unverändert blieb (Kontrollgruppe), dann hat man einen Maßstab, um im Vergleich zu beurteilen, ob sich die Marktanteilsänderung in der „Versuchsgruppe“ deutlich von der Marktanteilsentwicklung in den Regionen mit unveränderter Werbung („Kontrollgruppe“) unterschied. Wenn das der Fall ist, dann hätte man die alternativen Erklärungsmöglichkeiten, die sich auf externe Einflüsse (z. B. Ausscheiden von Konkurrenten) beziehen, damit ausgeschlossen. Es sei denn, es gibt andere Einwände, die dann zum Design 3 führen.

Beispiel

Jack Jacoby (2013, S. 219 f.) illustriert mit einem einfachen Beispiel die Bedeutung von Versuchs- und Kontrollgruppen für die Aussagekraft experimenteller Untersuchungen:

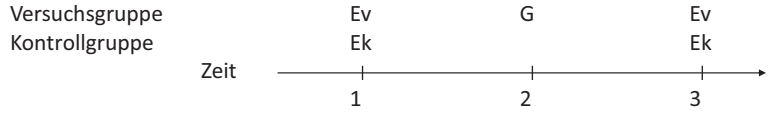


Abb. 6.3 Design 2. (Nach Jacoby, 2013, S. 231)

„Die Leser sind mit dem Begriff *„Placebo“* – wie er in der medizinischen Forschung verwendet wird – vertraut …. . Angenommen ein Pharma-Hersteller hat ein neues Medikament X entwickelt, das die Schmerzen von Krebspatienten mindern soll. Nachdem sie Medikament X genommen haben gaben 35 % einer Zufallsstichprobe von Krebspatienten an, dass es ihre Schmerzen reduziert hat. Bedeutet das, dass Medikament X bei der Verminderung von Schmerzen von mehr als einem Drittel der Patienten wirksam war? Vielleicht, vielleicht auch nicht.

Angenommen, dass 35 % einer vergleichbaren Gruppe von Krebspatienten ebenfalls sagen, ihre Schmerzen seien verringert nachdem sie eine Placebo-Pille genommen haben. Ist die Aussage, dass die Bestandteile von Medikament X eine Verminderung der Schmerzen von Krebspatienten bewirken, gerechtfertigt? Wohl nicht.

Man betrachte jetzt ein anderes Szenario. Angenommen *65 %* von denen, die *Medikament X* genommen haben, sagen, dass es ihre Schmerzen gemindert hat, während *35 %* einer vergleichbaren Gruppe von Krebspatienten, die das *Placebo* genommen hatten, angeben, dass dieses ihre Schmerzen verringert hat. Der Unterschied von 30 % zwischen den beiden Gruppen ist beachtlich und wird als Beleg dafür angesehen, dass Medikament X tatsächlich eine Verminderung der Schmerzen verursacht, zumindest bei einigen Patienten.

Was mag der Grund dafür sein, dass Leute sagten, Medikament X verringere ihre Schmerzen, wenn das nicht der Fall war? Plausible alternative Erklärungen für die Ergebnisse sind u. a. (1) der Einfluss positiven Denkens, (2) der Einfluss positiven Glaubens (Glauben an die Wirkung des Medikaments), (3) nicht undankbar sein wollen für die Teilnahme am Test des Medikaments, (4) nicht wünschen, als „Meckerer“ zu erscheinen oder als jemand mit positiver Einstellung wahrgenommen werden wollen, (5) tatsächliche Wahrnehmung geringerer Schmerzen während des Interviews und Zurückführung dessen auf Medikament X, (6) Wunsch, dem Interviewer zu „gefallen“, indem man die „richtige“ Antwort gibt, usw., usw.“ ◀

Gibt es aber auch für ein Untersuchungsergebnis auf Basis von Design 2 alternative Erklärungsmöglichkeiten? Vielleicht waren Versuchs- und Kontrollgruppe deutlich unterschiedlich zusammengesetzt (z. B. im Hinblick auf Affinität zum beworbenen Produkt oder auf Offenheit gegenüber Werbung), was den Unterschied bei den Effekten verursacht haben könnte.

Mithilfe von *Design 3* (Abb. 6.4) wird also versucht, die Möglichkeit *systematischer Unterschiedlichkeit* von Versuchs- und Kontrollgruppen dadurch auszuschließen, dass die Zuordnung der einzelnen Versuchspersonen zu Versuchs- und Kontrollgruppen *zufällig* erfolgt. Man spricht dann von **Randomisierung** (in den Abbildungen gekennzeichnet durch ein „R“). Für Experimente mit Versuchs- und Kontrollgruppen ist es also zentral, die *Zuordnung* der Versuchspersonen zu den Gruppen nach dem *Zufallsprinzip* vorzunehmen, damit systematische Unterschiede bei der Zusammensetzung der Gruppen vermieden werden. Damit ist *im Idealfall* sichergestellt, dass es zwischen den Gruppen im Hinblick auf alle Merkmale *keine systematischen* Unterschiede gibt. Ansonsten wüsste man ja nicht, ob ein Ergebnisunterschied bei Versuchs- und Kontrollgruppen auf die Wirkung der unabhängigen Variablen oder auf die Unterschiedlichkeit der Gruppenzusammensetzung zurückzuführen ist. In der Realität wird es meist – insbesondere bei

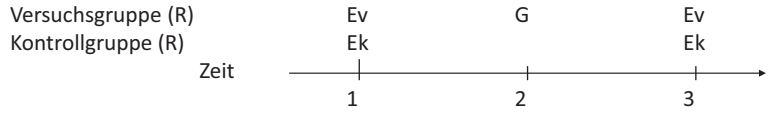


Abb. 6.4 Design 3. (Nach Jacoby, 2013, S. 233)

kleinen Gruppengrößen – dennoch gewisse Unterschiede geben; die Wahrscheinlichkeit für das Auftreten *gravierender* Unterschiede bleibt aber in der Regel gering. Nun ist es allerdings so, dass bei sozialwissenschaftlichen Untersuchungen sehr viele psychische, situative, physische oder soziale Einflussfaktoren die Ergebnisse beeinflussen *könnten.* Angesichts dieser großen Anzahl kann es gut sein, dass trotz geringer Wahrscheinlichkeiten für *einzelne große* Unterschiede zwischen Versuchs- und Kontrollgruppen leicht sein, dass sich in einer experimentellen Untersuchungen bei einzelnen der zahlreichen *möglichen* Einflussfaktoren deutliche Unterschiede zeigen, die wiederum alternative Erklärungsmöglichkeiten für einen Einfluss auf die abhängige Variable sein könnten, was man ja mit einem experimentellen Design gerade ausschließen wollte (Worrall, 2002; Crasnow, 2019).

Die Schlussweise bei der Interpretation der Messwerte bei Design 3 entspricht der bei Design 2. Die hier skizzierte *zufällige Zuordnung* von Versuchspersonen zu Gruppen ist deutlich zu unterscheiden von der *Zufallsauswahl* von Versuchspersonen. Letztere wird insbesondere bei deskriptiven Untersuchungen (siehe Abschn. 2.4) angewandt; bei Experimenten am ehesten dann, wenn man besonderen Wert auf externe Validität legt (siehe Abschn. 6.2), um eine möglichst weitgehende *Generalisierbarkeit* der Ergebnisse zu fördern. Manchmal ist die Realisierung einer zufälligen Zuordnung zu Gruppen nicht möglich. Dann wird gelegentlich versucht, Gruppen mithilfe des sogenannten **„Matching“** so zu bilden, dass diese im Hinblick auf eine Reihe von Merkmalen (z. B. Alter, Geschlecht, Konsumverhalten) möglichst ähnlich sind (siehe dazu z. B. Jaccard & Becker, 2002, S. 246; Shadish et al., 2002, S. 118 ff.).

Beim Vergleich der Ergebnisse von Gruppen, kommen Aspekte statistischer Signifikanz hinzu, die für die darauf aufbauenden Schlussweisen bedeutsam sind. Die elementarste Frage gilt den Gründen für die Verwendung von Daten aus Gruppen an Stelle der Daten von Einzelpersonen. Immerhin werden ja im naturwissenschaftlichen Bereich oft nur einzelne oder sehr wenige Beobachtungen durchgeführt, z. B. hinsichtlich der Wirkung einer Chemikalie auf einen bestimmten Stoff. Bei sozialwissenschaftlichen Untersuchungen würde man dagegen eine solche Vorgehensweise kaum akzeptieren, weil bei Menschen (Konsumentinnen, Patienten etc.) beobachtete Wirkungen so unterschiedlich sein können, dass man von einer einzelnen Beobachtung kaum auf eine *generelle* Wirkung der unabhängigen Variablen schließen kann. Diverse *individuelle* Merkmale (z. B. kognitive Fähigkeiten, Motivation, physische Merkmale wie Hunger oder Durst, bisherige Erfahrungen) können den Zusammenhang von unabhängigen und abhängigen Variablen beeinflussen (z. B. verstärken, behindern, verhindern), sodass Einzelfälle für

generalisierende Ergebnisse zu wenig aussagekräftig wären. Also bildet man eben Gruppen von Versuchspersonen, deren Ergebnisse bei hinreichender Größe (→ Stichprobe) naturgemäß nicht so stark von individuellen Besonderheiten beeinflusst werden und eher *systematische* Zusammenhänge oder Unterschiede bei Variablen widerspiegeln. Sofern mehrere Messwerte solch einer Gruppe vorliegen, kann man leicht feststellen, in welchem Wertebereich diese typischerweise liegen. Wenn die Unterschiede zwischen Versuchs- und Kontrollgruppen deutlich (→ „signifikant“) über diesen Schwankungsbereich hinausgehen, dann spricht das dafür, dass ein *systematischer* Einfluss der unabhängigen Variablen vorliegt. Derartige Fragestellungen analysiert man mithilfe der **Varianzanalyse,** auf die im Abschn. 9.2 noch genauer eingegangen wird. Bei der Varianzanalyse beruht nämlich die Entscheidung über Signifikanz bzw. Nicht-Signifikanz der Messwerte von Gruppen auf einem Vergleich der Schwankungen (→ Varianz) *innerhalb* der Gruppen mit den Unterschieden *zwischen* den Gruppen.

Die Leserin bzw. den Leser wird es nicht überraschen, wenn im Hinblick auf Design 3 erneut die Frage nach anderen Erklärungsmöglichkeiten unterschiedlicher Kaufabsichten in Versuchs- und Kontrollgruppe zum Zeitpunkt $t=3$ aufgeworfen wird. Könnte es sein, dass die Wirkung der unabhängigen Variablen „Kontakt zur Werbung“ dadurch verstärkt wird, dass die Angehörigen der Versuchsgruppe bei der Vormessung (zum Zeitpunkt $t=1$) der Kaufabsichten auf das betreffende Produkt aufmerksam gemacht wurden und die später eingesetzte Werbung intensiver als sonst wahrnehmen (Konditionierung)? Vielleicht hat die Vormessung auch dazu geführt, dass in der Kontrollgruppe erneutes Nachdenken über die Kaufabsichten stattfindet und zu veränderten Messwerten zum Zeitpunkt $t=3$ führt.

Mit dem *Design 4* (Abb. 6.5) soll versucht werden, eine entsprechende Kontrolle vorzunehmen. Durch Vergleich der Werte von Ev1 mit Ev2 bzw. von Ek1 mit Ek2 am Zeitpunkt $t=3$ kann man jetzt prüfen, ob die Vormessung den befürchteten Einfluss hatte. Dieser würde sich nur bestätigen, wenn sich bei den genannten Vergleichen deutliche Abweichungen ergeben. Im anderen Fall (ohne Konditionierung durch die Vormessung) könnte man jetzt am Zeitpunkt $t=3$ Ev1 und Ev2 auf der einen Seite und Ek1 und Ek2 auf der anderen Seite gegenüberstellen, um die Wirkung der Werbung auf Kaufabsichten abzuschätzen.

Weitere Probleme bei der Festlegung eines experimentellen Designs können mit der Bestimmung der Zeitpunkte für die verschiedenen Messungen verbunden sein. Es könnte

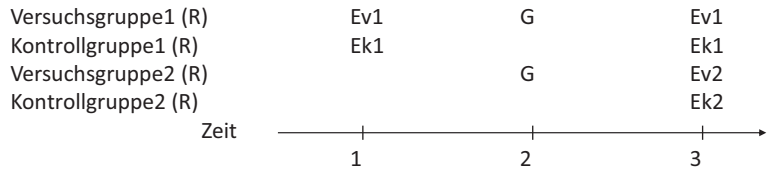


Abb. 6.5 Design 4. (Nach Jacoby, 2013, S. 234)

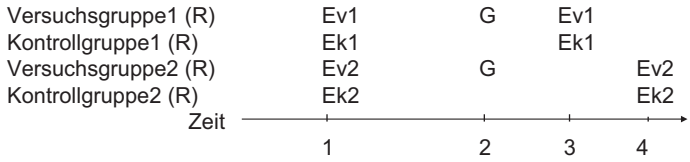


Abb. 6.6 Design 5

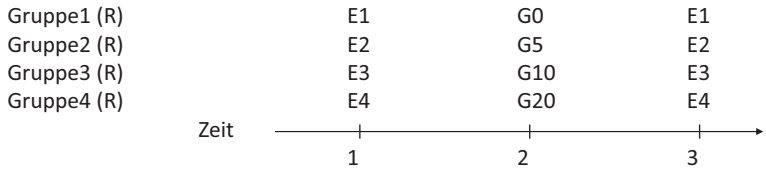


Abb. 6.7 Design 6

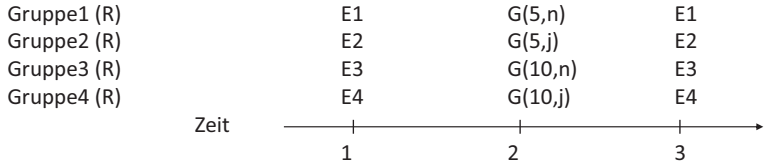


Abb. 6.8 Design 7

ja sein, dass sich zum bisher für die Nachmessung vorgesehenen Zeitpunkt $t=3$ die Wirkung der unabhängigen Variablen noch nicht entwickelt hat oder dass diese schon wieder erloschen ist. Mit Design 5 (Abb. 6.6) wird versucht, eine solche Problemstellung zu berücksichtigen. Wenn beispielsweise der Zeitpunkt $t=3$ einen Tag und $t=4$ eine Woche nach der Konfrontation der Versuchspersonen mit Werbung liegt, könnte man daran kurz- und längerfristige Wirkungen messen.

Ein anderer Aspekt ist der, ob die unabhängige Variable (hier: Kontakt zur Werbung) – wie bisher angenommen – nur zwei Ausprägungen (Kontakt erfolgt – nicht erfolgt) haben kann oder ob nicht unterschiedliche Intensitäten dafür eher typisch sind. Das Design 6 (Abb. 6.7) ist so angelegt, dass die Wirkungen von 0, 5, 10 oder 20 Kontakten zur Werbung gemessen werden können.

Ein höherer Komplexitätsgrad für experimentelle Designs ergibt sich, wenn **Interaktionen** zwischen mehreren unabhängigen Variablen berücksichtigt werden müssen. Beispielsweise ist es ja durchaus plausibel, dass die Wirkung von Werbung dadurch beeinflusst werden kann, dass gleichzeitig eine Preissenkung stattfindet, weil in diesem Fall die Werbung eine für Konsumenten relevante Information übermitteln kann. Im Design 7 (Abb. 6.8) ist dargestellt, wie das Zusammenwirken der Werbung (5 Kontakte oder 10 Kontakte) mit einer Preissenkung (ja – nein) überprüft werden kann.

Eine solche Anordnung, bei der zwei unabhängige Variable mit jeweils zwei Ausprägungen gleichzeitig untersucht werden, bezeichnet man als 2×2-faktorielles Design. Generell spricht man von einem **faktoriellen Design**, wenn dabei „zwei oder mehr unabhängige Variable (genannt Faktoren), jede mit mindestens zwei Ausprägungen" (Shadish et al., 2002, S. 263) verwendet werden. Diese erlauben es auch, das Zusammenwirken bzw. Interaktionen von unabhängigen Variablen zu untersuchen. Allerdings nimmt dabei die Komplexität des Designs rasch zu. So führt beispielsweise ein Design mit drei Faktoren mit 2 bzw. 3. bzw. 4 Ausprägungen (ein $2 \times 3 \times 4$-faktorielles Design) schon zu 24 Gruppen von Versuchspersonen, die jeweils mit unterschiedlichen Kombinationen von Ausprägungen der drei unabhängigen Variablen konfrontiert werden müssten.

Faktorielle Designs erlauben also differenzierte Aussagen über die Wirkungen der unabhängigen Variablen. Man kann daraus entnehmen,

- wie Kombinationen von Merkmalen der unabhängigen Variablen sich auf die abhängige Variable auswirken **(Interaktionseffekte)**;
- wie die Wirkungen der einzelnen unabhängigen Variablen unabhängig von der Wirkung weiterer Variabler (**Haupteffekte**) sind.

Man erkennt leicht, dass bei faktoriellen Designs mit wachsender Zahl von Variablen bzw. wachsender Zahl von Ausprägungen pro Variable die Zahl der Versuchsgruppen und damit Komplexität und Aufwand der Untersuchung stark ansteigen. Deswegen findet man in der psychologischen Methoden-Literatur (z. B. Shadish et al., 2002), wo Experimente eine zentrale Rolle spielen, diverse Ansätze zur Reduktion des Untersuchungsaufwandes („reduzierte Designs"). Bei vielen Fragestellungen muss man aber doch von Interaktionen zwischen mehreren unabhängigen Variablen ausgehen, wie im Beispiel der Abb. 6.9 dargestellt. Man spricht in diesem Zusammenhang auch von einem **Moderator.** D. h. der Effekt einer unabhängigen Variablen auf eine abhängige Variable wird durch eine zweite unabhängige Variable beeinflusst („moderiert"). Der Einfluss der unabhängigen auf die abhängige Variable fällt also stärker oder schwächer aus, je nach Wirkungsweise des Moderators. „Ein Moderator ist eine qualitative oder quantitative Variable, die die Richtung und/oder Stärke einer Beziehung zwischen einer unabhängigen und einer abhängigen Variablen beeinflusst" (Baron & Kenny, 1986, S. 1174). So könnte man sich leicht vorstellen, dass intensivierte Werbung in Verbindung mit einer Preissenkung größere Wirkung haben kann als ohne Preissenkung (siehe Beispiel unten), weil die Wirkung von Werbung größer sein könnte, wenn die Werbung eine Botschaft über einen besonderen Preisvorteil übermittelt. Bei einer solchen gegenseitigen Verstärkung (oder Abschwächung) des Wirkens unabhängiger Variabler spricht man eben von **Interaktionseffekten.**

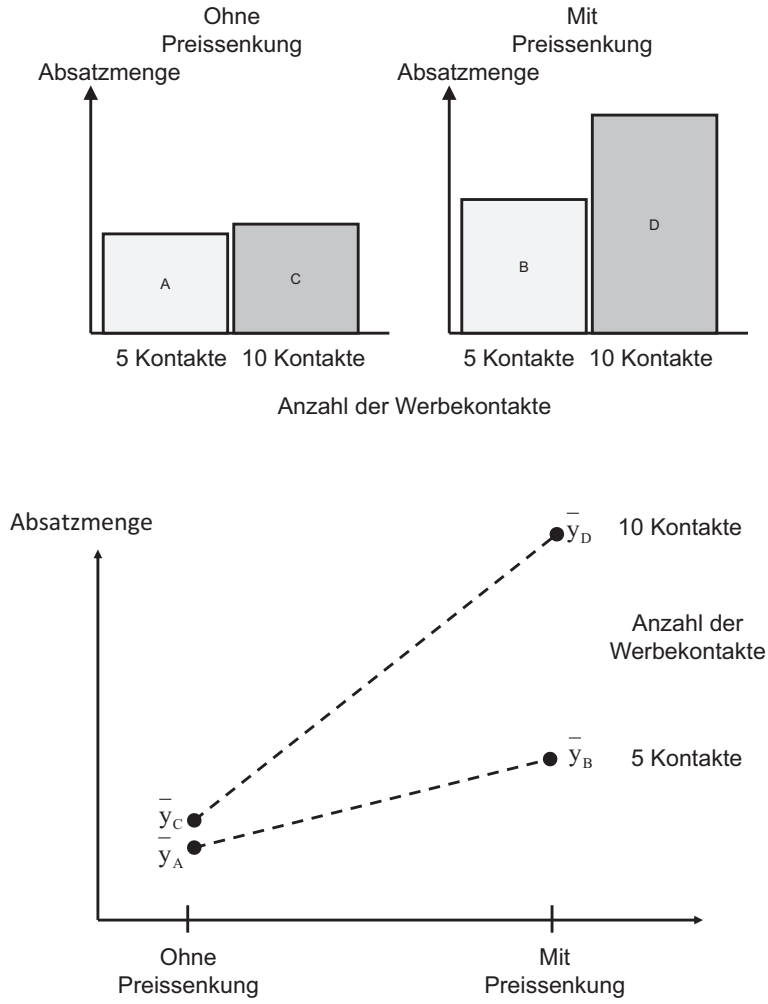


Abb. 6.9 Beispiel zu Interaktionswirkungen

Hintergrundinformation
Kerlinger und Lee (2000, S. 352) kennzeichnen das Wesen von Interaktionseffekten folgendermaßen:

> „Interaktion ist das Zusammenwirken von zwei oder mehr unabhängigen Variablen bei ihrem Einfluss auf die abhängige Variable. Genauer gesagt bedeutet Interaktion, dass die Wirkung einer unabhängigen Variablen auf die abhängige Variable von der Ausprägung einer weiteren unabhängigen Variablen abhängt".

Zur Illustrierung von Interaktionseffekten werden graphische Darstellungsformen, insbesondere Balkendiagramme oder **Interaktionsdiagramme,** verwendet, die anhand des

einfachen Beispiels aus Abb. 6.9 erklärt werden sollen. In dem (natürlich fiktiven) Beispiel geht es um die Wirkungen einer Preissenkung und verschiedener Werbe-Intensität auf Absatzmengen. Es gibt in dem Beispiel zwei verschiedene Preise (mit oder ohne Preissenkung) und zwei Intensitätsgrade der Werbung (5 oder 10 Kontakte). Daraus ergeben sich folgende vier Kombinationen von Merkmalen der beiden unabhängigen Variablen, die in einer größeren Zahl abgegrenzter Verkaufsgebiete angewandt wurden:

a) Ohne Preissenkung, 5 Werbekontakte
b) Mit Preissenkung, 5 Werbekontakte
c) Ohne Preissenkung, 10 Werbekontakte
d) Mit Preissenkung, 10 Werbekontakte

Die Mittelwerte, die sich für diese Merkmalskombinationen in den entsprechenden Verkaufsgebieten ergeben haben, werden in den beiden Grafiken in der Abb. 6.9 dargestellt. Im Balkendiagramm sind die Mittelwerte durch die Höhe der Balken gekennzeichnet, im darunter dargestellten Interaktionsdiagramm werden die Mittelwerte durch die Endpunkte der Linien dargestellt.

Wenn man die in der Abbildung wiedergegebenen Ergebnisse betrachtet, erkennt man deutlich, dass die Mittelwerte der Absatzmengen mit einer Preissenkung höher liegen als ohne Preissenkung. Weiterhin liegen die Absatzmengen bei starker Werbung (nicht ganz überraschend) höher als bei mittlerer. Der Interaktionseffekt, der sich hier durch das Zusammenwirken von Preissenkung und Intensität der Werbung ergibt, wird insofern erkennbar, als die intensivere Werbung ohne Preissenkung nur einen geringen, in Verbindung mit der Preissenkung aber einen starken Zuwachs bei der Absatzmenge bringt.

Von Moderatoren deutlich abzugrenzen sind sogenannte **Mediatoren.** Diese bezeichnen indirekte Beziehungen zwischen Variablen (Jaccard & Jacoby, 2020; Baron & Kenny, 1986). In der Abb. 6.10 ist ein entsprechendes Beispiel dargestellt, in dem man erkennt, dass die Wirkung von Kundenzufriedenheit auf den wirtschaftlichen Erfolg eines Unternehmens auch so betrachtet werden kann, dass ein indirekter Zusammenhang existiert, weil die Wirkung „über“ die Variable Kundenbindung erfolgt. Eine direkte Beziehung in der einen theoretischen Sichtweise kann durchaus eine indirekte Beziehung in einer anderen Perspektive sein.

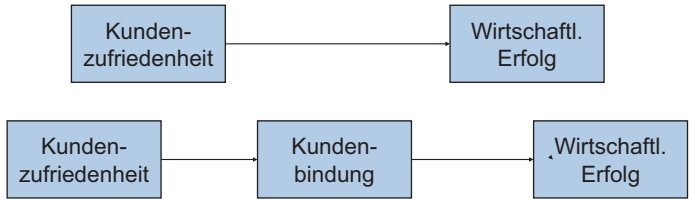


Abb. 6.10 Beispiel für einen Mediator bei einer indirekten Kausalbeziehung

Die vorgestellten Designs sollen nur einen ersten Eindruck von der Vielfalt entsprechender Möglichkeiten und den bei deren Wahl angestellten Überlegungen vermitteln. Es sollte dabei deutlich geworden sein, dass die zentrale Idee des Ausschlusses alternativer Erklärungsmöglichkeiten (→ Validität) für einen beobachteten Zusammenhang die Gestaltung eines experimentellen Designs maßgeblich bestimmt. Für weiterführende Darstellungen muss auf die Spezial-Literatur (z. B. Shadish et al., 2002) verwiesen werden.

Bei den Überlegungen im vorliegenden Abschn. 6.1 ist von einer Art der Manipulation der unabhängigen Variablen ausgegangen worden, bei der jede Versuchsperson (allgemeiner: jede Untersuchungseinheit) mit nur einer Ausprägung der unabhängigen Variablen (zünftig ausgedrückt: einem „treatment“) konfrontiert wird. Die verschiedenen Werte der unabhängigen Variablen sind gewissermaßen auf verschiedene Gruppen aufgeteilt. Bei einem Wirkungstest von drei unterschiedlichen Werbespots würde man also drei Gruppen von Versuchspersonen bilden, jeder Gruppe nur einen der Werbespots vorführen und dann in jeder Gruppe die Wirkung messen. Man spricht in diesem Fall von einem **„Between Subjects Design“** bzw. **„Independent Groups Design“**.

Eine alternative Vorgehensweise kann im vorliegenden einführenden Lehrbuch nicht ausführlich behandelt werden; ein kurzer Hinweis muss genügen: Man könnte auch jede Versuchsperson jedem „treatment“, also jeder Ausprägung der unabhängigen Variablen, aussetzen. Im obigen Werbe-Beispiel würden also die Versuchspersonen nacheinander alle drei Spots zu sehen bekommen und dann jeweils die Wirkung bei dieser Person gemessen werden. Ein solches Vorgehen wird als **„Within Subjects Design“** oder **„Repeated Measures Design“** bezeichnet.

Beim erstgenannten Ansatz ist natürlich die Belastung für die Versuchspersonen geringer, aber auch die Menge erhobener Daten pro Person. Dagegen erbringt der zweite Ansatz relativ viele Daten pro Proband und es entsteht kein Problem möglicherweise eingeschränkter Vergleichbarkeit verschiedener Gruppen. Es kann aber sein, dass die verschiedenen bei einer Person vorgenommenen Messungen sich untereinander beeinflussen (z. B. durch Lern- oder Ausstrahlungseffekte). Dem versucht man zu begegnen, indem man die Reihenfolge der Messungen bei den verschiedenen Personen variiert („counterbalancing“).

6.2 Interne und externe Validität von Experimenten

In den vorangegangenen Teilen dieses Lehrbuchs ist die Bedeutung von Reliabilität und Validität einer Untersuchung hinsichtlich der Aussagekraft ihrer Ergebnisse immer wieder angesprochen und hoffentlich auch deutlich geworden. In Bezug auf Experimente kommen zu den allgemeinen Überlegungen zur Validität von Untersuchungen zwei spezifische Aspekte hinzu: Die **interne** und die **externe Validität.** Der Gesichtspunkt der internen Validität ist im Zusammenhang mit den verschiedenen im vorigen Abschnitt betrachteten experimentellen Designs schon implizit angesprochen worden. Dort ging es

immer wieder um die Weiterentwicklung eines Designs im Hinblick auf den Ausschluss alternativer Erklärungsmöglichkeiten für ein Untersuchungsergebnis. Dem entsprechend bezieht sich **interne Validität** darauf, alternative – auf den Messvorgang zurückzuführende – Erklärungen für die beobachteten Zusammenhänge auszuschließen. Interne Validität ist also in diesem Sinne die „Validität von Schlüssen bezüglich der Kausalität einer Beziehung zwischen zwei Variablen" (Shadish et al., 2002, S. 508).

In Abb. 6.11 wird der Unterschied zwischen Konstruktvalidität (siehe Abschn. 4.3.2.6) und interner Validität verdeutlicht. **Konstruktvalidität** bezieht sich auf die Übereinstimmung der erhobenen Messwerte (x, y) mit den „wahren" Werten der Konzepte/Konstrukte (X, Y), die gemessen werden sollten. **Interne Validität** bezieht sich dagegen auf den (Kausal-) Zusammenhang zwischen den untersuchten Variablen (x → y), bzw. den Ausschluss alternativer Erklärungsmöglichkeiten für einen aufgetretenen Zusammenhang.

Hinsichtlich der internen Validität steht also die – für Kausalaussagen zentrale – Frage im Mittelpunkt, ob die Veränderung einer abhängigen Variablen tatsächlich auf die vermutete Ursache, also die Veränderung der jeweils interessierenden unabhängigen Variablen, zurückzuführen ist, oder ob Unzulänglichkeiten der Untersuchungsanlage oder der Durchführung der Messungen dafür ausschlaggebend sein können. Daneben stellt sich dann die Frage, inwieweit man die Ergebnisse einer Untersuchung *generalisieren* kann. Welche Aussagekraft hat z. B. eine Untersuchung, die bei deutschen Frauen im Alter von 30 bis 50 Jahren durchgeführt wurde, für deutsche Frauen allgemein, für Frauen allgemein oder für Konsumenten schlechthin? Was sagen die Ergebnisse eines Experiments mit 100 amerikanischen Studierenden für Konsumenten oder die Menschheit generell aus? Diese Fragestellungen gelten der externen Validität von Experimenten. Zur

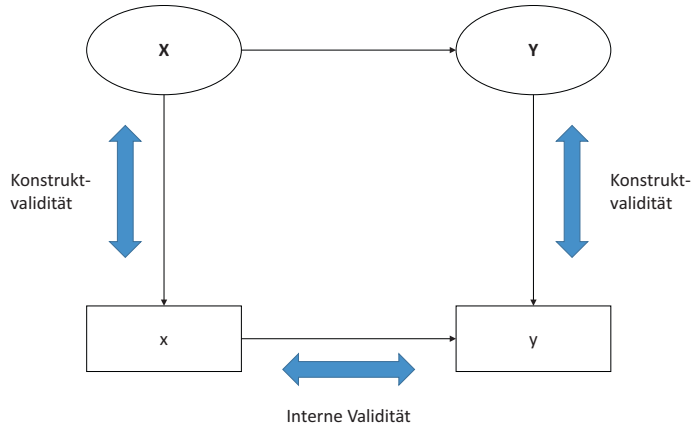


Abb. 6.11 Interne Validität und Konstruktvalidität bei Experimenten. (Nach Viswanathan, 2005, S. 344)

Definition: **Externe Validität** bezieht sich auf die **Generalisierbarkeit** (siehe dazu auch Abschn. 2.3) von Ergebnissen über verschiedene Personen, Situationen, Untersuchungsmethoden etc. Externe Validität ist also die „Validität von Schlüssen hinsichtlich des Bestands der Kausalbeziehung bei verschiedenen Personen, Situationen, und verschiedenen Messungen der Variablen“ (Shadish et al., 2002, S. 507).

Hintergrundinformation

Campbell und Stanley (1963, S. 5) formulieren zentrale Gesichtspunkte zur internen und externen Validität:

> „Grundlegend … ist die Unterscheidung zwischen interner Validität und externer Validität. Interne Validität ist die minimale Grundlage, ohne die jedes Experiment nicht interpretierbar ist: Haben tatsächlich die unabhängigen Faktoren bei diesem Experiment zu einem unterschiedlichen Ergebnis geführt? Externe Validität gilt der Frage nach der Generalisierbarkeit: Auf welche Personengruppen, Situationen, unabhängige Variablen und Messungen kann der Effekt generalisiert werden? Beide Arten von Kriterien sind offenkundig wichtig, obwohl sie häufig im Widerspruch stehen, weil Merkmale, die dem Einen dienen, das Andere gefährden können.“

Im Hinblick auf praktische (Marketing-)Fragestellungen wird teilweise betont, dass die externe Validität unverzichtbar sei, weil es eben darum geht, von den Ergebnissen einer Untersuchung auf die realen Verhältnisse in den Märkten, für die die Entscheidungen getroffen werden, zu schließen (vgl. Calder et al., 1981, 1982). Wesentliche Bedeutung für die externe Validität haben offenkundig die *repräsentative Auswahl* von Versuchspersonen (analog zur Vorgehensweise bei repräsentativen Befragungen, siehe Kap. 4) und die *realitätsnahe* („natürliche“ oder auch „biotische“) *Untersuchungssituation* (Winer, 1999), auf die anschließend im Zusammenhang mit Feldexperimenten noch kurz eingegangen wird.

In Anlehnung an Campbell und Stanley (1963, S. 5 f.) und Shadish et al., (2002, S. 54 ff.) sollen einige typische Fehler bzw. Effekte, die die *interne Validität* eines Experiments beeinträchtigen können, kurz gekennzeichnet werden:

- **„Auswahl- bzw. Zuordnungs-Fehler“**
  Hier geht es um das im vorigen Abschnitt schon angesprochene Problem, dass die Unterschiede zwischen Ergebnissen bei Versuchs- und Kontrollgruppen möglicherweise nicht auf die unabhängige Variable, sondern auf die Unterschiedlichkeit der Zusammensetzung beider Gruppen zurückzuführen sind (siehe dazu als Beispiel Design 2 in Abschn. 6.1). Gängige Vorgehensweisen zum Ausschluss dieses Problems (bzw. dieser alternativen Erklärungsmöglichkeit) sind – wie schon erläutert – Randomisierung oder Matching bei der Zuordnung von Versuchspersonen zu Versuchs- und Kontrollgruppen.
- **„Treatment-Effekt“**
  Dieser Effekt ist dadurch gekennzeichnet, dass die Untersuchungssituation die Wirkung der unabhängigen Variablen beeinflusst (oftmals verstärkt), z. B. durch höhere

Aufmerksamkeit der Versuchsperson oder deren Anpassung an vermeintliche Erwartungen des Untersuchungsleiters. Als Gegenmaßnahmen werden die Tarnung des Untersuchungsgegenstandes und/oder die Ablenkung der Versuchsperson empfohlen.

- **„Test-Effekt“**
  Damit ist der Einfluss einer Vormessung auf die Wirkung einer unabhängigen Variablen gemeint (siehe dazu Designs 4 und 5 in Abschn. 6.1). Zur Kontrolle dieses Effekts werden häufig Versuchs- und Kontrollgruppen vorgesehen, bei denen keine Vormessung stattfindet.
- **„Entwicklungs-Effekt“**
  Dieser bezieht sich auf die Veränderung der Umwelt und ihrer Wirkungen auf die abhängige Variable während der Dauer des Experiments (siehe dazu Design 2 im Abschn. 6.1). In dieser Hinsicht ist es wesentlich, die Vergleichbarkeit der Bedingungen bei Versuchs- und Kontrollgruppe zu sichern.

Hintergrundinformation

Zumindest bei Laborexperimenten (siehe unten) ist es für Versuchspersonen klar erkennbar und bewusst, dass sie und ihr Verhalten Gegenstand der Untersuchung sind. Das kann dazu führen, dass sie Vermutungen über Ziele der Untersuchung anstellen und ihr Verhalten auf vermutlich „erwünschte“ Ergebnisse ausrichten oder auch versuchen, die Untersuchung durch bewusst verfälschtes Verhalten bzw. verfälschte Angaben zu stören. Daneben kann es auch sein, dass das Verhalten der Wissenschaftler, die die Untersuchung durchführen und die Ziele und Hypothesen der Untersuchung kennen, einen Einfluss auf die Versuchspersonen hat und somit zur Verfälschung der Ergebnisse beiträgt. Als wichtigste Gegenmaßnahme wird dazu empfohlen, Versuchspersonen über Ziele und Hypothesen einer Untersuchung im Unklaren zu lassen oder diese durch Täuschung (z. B. mit „cover stories“) zu verdecken. Dies widerspricht zwar dem im ESOMAR Code aufgestellten Grundsatz der Transparenz gegenüber dem Befragten gerade auch zum Zweck der Befragung (vgl. ESOMAR, 2016). Ist eine Täuschung aus methodischen Gründen jedoch erforderlich, so ist diese erlaubt, sofern die befragte Person vor dem Ende der Untersuchung über die Täuschung aufgeklärt wird und dann fordern kann, dass ihre Daten gelöscht werden.

Hinsichtlich des Verhaltens des Untersuchungspersonals versucht man, die Durchführung so anzulegen, dass für den Versuchsleiter nicht erkennbar ist, zu welcher Versuchs- oder Kontrollgruppe eine Versuchsperson gehört. Damit soll ein bewusster oder unbewusster Einfluss auf die Ergebnisse (s. o.) verhindert werden. Wenn man auf solche Weise wichtige Aspekte der Anlage von Experimenten *verdeckt,* spricht man auch vom **„Blinding“**.

Analog zur Einteilung der Beobachtungsverfahren in Feld- und Laborbeobachtungen (siehe Abschn. 4.4.2) spielt auch für den praktischen Einsatz experimenteller Anordnungen die Unterscheidung von Feld- und Laborexperimenten eine Rolle. Vorgehensweisen, bei denen die Wirkung der unabhängigen Variablen und die Messungen der abhängigen Variablen sich in einer natürlichen, realistischen Umgebung vollziehen, werden als **Feldexperimente** bezeichnet, während man bei Experimenten in einer künstlichen, stark vom Forscher beeinflussten (bzw. kontrollierten) Situation von **Laborexperimenten** spricht. Diese beiden Typen sind aber keine eindeutig abgrenzbaren Klassen von Experimenten, sondern lediglich die extremen Ausprägungen einer Vielfalt von Gestaltungsmöglichkeiten. Bei zahlreichen experimentellen Untersuchungen mischen

sich Elemente von Labor- und Feldexperimenten, sodass die Zuordnung zu den beiden Kategorien häufig nicht eindeutig ist, was aber für die Aussagekraft der Ergebnisse kein Nachteil sein muss.

▶ **Definition** Seymour Sudman und Edward Blair (1998, S. 215) definieren Labor- und Feldexperimente auf folgende Weise:

> „Ein *Laborexperiment* ist ein Experiment, mit dem die interessierenden Phänomene außerhalb der natürlichen Situation untersucht werden. Der Begriff ‚Labor' bezieht sich dabei nicht auf einen bestimmten Ort für die Untersuchung, sondern einfach auf einen Kontext, der nicht der natürliche Kontext ist. (…) Im Gegensatz dazu werden mit *Feldexperimenten* die interessierenden Phänomene in einem natürlichen Kontext untersucht."

„In Feldexperimenten sind sich die Teilnehmer nicht bewusst, dass sie an einer Untersuchung teilnehmen oder – wenn sie sich dessen bewusst sind – sie verhalten sich so, wie sie das normalerweise tun, unabhängig vom Experiment (…). Typischerweise bemerken die Teilnehmer am Experiment nicht, dass der Forscher Rahmenbedingungen manipuliert sowie Verhaltensweisen und Ergebnisse misst." (Gneezy, 2017, S. 140) Bei *Feldexperimenten* steht dem Vorteil der Realitätsnähe der Untersuchungssituation das Problem der häufig (nicht immer!) eingeschränkten Kontrollierbarkeit der Randbedingungen des Experiments gegenüber. Die Aussagekraft von Feldexperimenten wird in manchen Fällen dadurch eingeschränkt, dass sie auf beobachtbares Verhalten (z. B. Kauf*verhalten*) ausgerichtet sind und eine Befragung zu Motiven, Einstellung etc. (z. B. *Gründe* für ein bestimmtes Kaufverhalten) nicht möglich ist (Gneezy, 2017). Ein weiteres Problem von Feldexperimenten in der praktischen Marktforschung besteht darin, dass für Konkurrenten erkennbar wird, welche Fragestellungen untersucht werden. Eine klassische Einsatzmöglichkeit des Feldexperiments in der Marktforschung ist in der entsprechenden Anlage von Testmärkten etc. (siehe Abschn. 6.4) zu sehen.

Beispiel

Gneezy (2017, S. 141) gibt ein Beispiel für ein Feldexperiment, indem sie eine entsprechende Untersuchung zusammenfassend darstellt:

> „Gneezy et al. (2014) untersuchten die Preis-Qualitäts-Heuristik mit einem Experiment, bei dem sie die Qualität von Wein (hoch vs. niedrig) und den Preis ($10, $20, $40) variierten. Teilnehmer waren Besucher eines Weinguts, die zu einer Weinprobe kamen. Diese bekamen eine Liste mit Namen und Preisen von neun Weinen, von denen sie sechs Weine für die Probe auswählten. Dieses Design ermöglichte es den Autoren, die Wirkung von Preis und Qualität auf Akzeptanz, Nachfrage, Umsatz und Gewinn zu untersuchen. Das Hauptergebnis bezog sich auf den Wein hoher Qualität: Die Nachfrage wuchs, wenn der Preis von $10 auf $20 stieg und ging leicht zurück, wenn der Preis bei $40 lag. Im Gegensatz dazu ging die Nachfrage nach Wein niedriger Qualität durchgehend mit sinkendem Preis zurück. Die profitabelste Kombination lag bei $20 für den Wein hoher Qualität. Basierend auf diesen Ergebnissen bot das Weingut seinen besseren Cabernet Sauvignon für $20 an, was zu einer Steigerung des Gewinns um 11 % führte." ◀

Die Probleme bei der Anlage von *Laborexperimenten* stellen sich umgekehrt zu denen des Feldexperiments dar: Einerseits lassen sich in einer stark vom Forscher beeinflussten Untersuchungssituation die Randbedingungen des Experiments gut kontrollieren, andererseits wird die Generalisierbarkeit von Ergebnissen, die unter künstlichen Bedingungen gewonnen wurden, natürlich häufig (nicht immer !) fraglich (siehe Abb. 6.12). Daneben bieten Laborexperimente Vorteile hinsichtlich der Anzahl und Art (z. B. Manipulation von Konkurrenzpreisen) von Manipulationsmöglichkeiten, die im realen Umfeld nicht möglich wären. Weit verbreitet ist der Einsatz von Laborexperimenten für Produkt-, Packungs- und Werbemittel-Tests in der Markforschung (siehe Abschn. 6.5).

Hintergrundinformation

Seymour Sudman und Edward Blair (1998, S. 226) geben folgende Empfehlung für die Abwägung zwischen Feld- und Laborexperimenten:

> „Ein Feldexperiment sollte durchgeführt werden, wenn (1.) im Experiment nur ein oder zwei Faktoren mit einer geringen Anzahl von Gruppen untersucht werden, wenn (2.) diese Faktoren im „Feld" manipuliert werden können, wenn (3.) die Störung eines laufenden Marketing-Programms durch ein Feldexperiment nicht bedeutsam ist, wenn (4.) die Kosten eines Feldexperiments gerechtfertigt sind und wenn (5.) externe Validität wichtiger ist als interne Validität."

Auch hinsichtlich der **Auswahl von Versuchspersonen** ergeben sich oft deutliche Unterschiede zwischen Feld- und Laborexperimenten. Bei Laborexperimenten ist oft wegen aufwendiger Messungen und besonderer Untersuchungssituationen eine ausgeprägte Bereitschaft zur Mitarbeit seitens der Versuchspersonen erforderlich, die nicht selten durch finanzielle Anreize gefördert wird. Vor diesem Hintergrund ist eine Zufallsauswahl kaum zu realisieren, weil nur bei einem begrenzten Teil der Zielgruppe die Bereitschaft zur freiwilligen Mitwirkung besteht. Die Anzahl von Versuchspersonen, die für Experimente benötigt werden, hängt wesentlich von der Anzahl von Untersuchungsgruppen ab, die sich wiederum aus der Zahl unabhängiger Variabler und der Zahl unterschiedlicher Ausprägungen dieser Variablen ergibt. Als gängige Faustregel für die Stich-

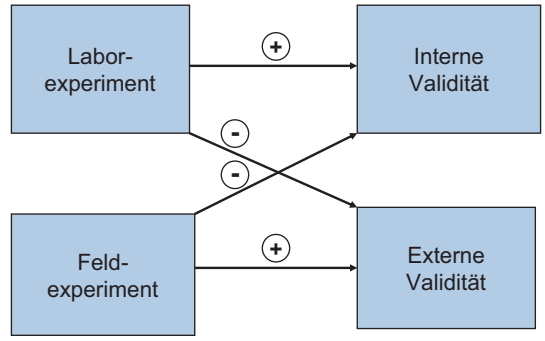


Abb. 6.12 *Tendenzielle* Vor- und Nachteile von Labor- und Feldexperimenten im Hinblick auf interne und externe Validität

probengröße werden oft etwa 30 Personen pro Versuchs- und Kontrollgruppe genannt (Koschate-Fischer & Schandelmeier, 2014).

Bei Feldexperimenten in der praktischen Marktforschung (siehe Abschn. 6.4 bis 6.8) findet man oft deutlich höhere Fallzahlen und auch das Bemühen um eine repräsentative Auswahl, weil damit die externe Validität einer Untersuchung wächst. Untersuchungen in der Praxis sind ja typischerweise darauf ausgerichtet, Entscheidungen vorzubereiten. Deswegen spielt in diesem Zusammenhang die Frage, inwieweit die Ergebnisse eines Experiments auf eine reale Marktsituation übertragbar sind, eine besondere Rolle. Allerdings ist die zufällige (→ repräsentative) Auswahl von Versuchspersonen nur ein (wesentlicher) Aspekt der Sicherung externer Validität. Es spielt eben auch der möglichst realitätsnahe Untersuchungskontext eine Rolle.

▶ Wenn man an die *zufällige Zuordnung* von Versuchspersonen zu Versuchs- und Kontrollgruppen (Randomisierung, s. o.) denkt, dann kann man sagen, dass diese zufällige Zuordnung eher zur *internen Validität* eines Experiments beiträgt. Dagegen leistet die *zufällige Auswahl* von Versuchspersonen aus einer Grundgesamtheit eher einen Beitrag zur *externen Validität* von Experimenten.

Abschließend sei noch auf eine Form von Experimenten hingewiesen, die für manche praktischen Fragestellungen und für bestimmte Aspekte der Methodenforschung besonders geeignet ist, so genannte **Befragungsexperimente.** Deren Grundidee besteht darin, dass man eine (repräsentative) Stichprobe per Zufall in zwei (oder mehr) Teilgruppen aufteilt, und diese jeweils mit unterschiedlichen Stimuli konfrontiert. Für praktische Fragestellungen könnten das z. B. unterschiedliche Entwürfe einer Verpackung oder einer Anzeige sein. Dann werden jeweils dazu die Einschätzungen der Auskunftspersonen (z. B. „Gefällt mir“, „Finde ich schön“, „Wirkt modern“) gemessen. Aus den Unterschieden der Angaben in den verschiedenen Gruppen (z. B. 68 % vs. 42 % „Gefällt mir“) wird auf die unterschiedliche Wirkung der entsprechenden Alternativen geschlossen. Man erkennt deutlich das Grundmuster von Experimenten: Eine unabhängige Variable (hier: ein Stimulus) wird variiert und aus dem Vergleich der Ergebnisse wird auf die Wirkung der unabhängigen Variablen geschlossen. Da die Vorgehensweise auf der (zufälligen) Aufteilung von Stichproben basiert, spricht man hier auch von **Split-Ballot-Experimenten** (Groves et al., 2009, S. 267 f.). Dieser Ansatz ist auch in zahlreichen Untersuchungen zur Methodik bei Umfragen verwendet worden. Wenn man die Wirkung unterschiedlicher Frageformulierungen auf das Antwortverhalten feststellen will, dann kann man dies tun, indem man bei verschiedenen Teilstichproben die verschiedenen Formulierungen verwendet und die Ergebnisunterschiede entsprechend interpretiert. Zahlreiche der im 4. Kapitel vorgestellten Ergebnisse sind auf diese Weise entstanden.

In den Abschnitten 6.1 und 6.2 wurde nur ein Überblick über wesentliche Gestaltungsmöglichkeiten. Für die Durchführung von Experimenten stellen sich weit mehr und weit komplexere Fragen der Untersuchungsanlage, der Datenerhebung und der Datenanalyse. Dazu existiert insbesondere in der Psychologie eine kaum überschaubare

Fülle an Spezialliteratur. Für die Marketingforschung seien insbesondere die Artikel von Koschate-Fischer und Schandelmeier (2014) und von Spilski et al. (2018) empfohlen, die nicht nur eine Fülle speziellerer Probleme kurz umreißen, sondern auch auf die jeweils relevante Literatur verweisen.

6.3 Quasi-Experimente

Typisch für die vorstehend gekennzeichneten experimentellen Designs sind der vom Untersuchungsleiter kontrollierte Einsatz der unabhängigen Variablen und die *zufällige Zuordnung* von Versuchspersonen zu Versuchs- und Kontrollgruppen mit dem Ziel, systematische Unterschiede zwischen diesen Gruppen, die die Wirkung der unabhängigen Variablen überlagern könnten, auszuschließen. Nun gibt es Untersuchungssituationen, in denen diese Bedingungen nicht realisiert werden können. Zwei Beispiele mögen dieses Problem illustrieren:

- Es soll untersucht werden, ob bei Menschen, deren Eltern Raucher sind/waren, die Neigung, selbst Raucher zu werden, stärker entwickelt ist als bei anderen Menschen. Hier ist offenkundig, dass eine zufällige Zuordnung zu den beiden zu vergleichenden Gruppen („Eltern Raucher“ und „Eltern Nichtraucher“) nicht nur praktisch unmöglich ist, sondern auch ethisch höchst bedenklich wäre.
- Es soll untersucht werden, in welchem Maße ein hoher Preisnachlass (>10 %) beim Kauf eines Autos die Markenpräferenz auch langfristig (5 Jahre und mehr) bestimmt. Hier wird man kaum einen Marketing-Manager finden, der dem Marktforscher 5 Jahre Zeit lässt, um die Entwicklung der Markenpräferenzen bei den verschiedenen Käufergruppen (hoher oder geringer Preisnachlass) sorgfältig zu beobachten. Man müsste wohl eher bei jetzigen Autokäufern rückschauend feststellen, welchen Preisnachlass sie früher bekommen haben, entsprechende Vergleichsgruppen bilden und bei diesen Markenpräferenzen messen. Das wäre sicher keine zufällige Zuordnung, würde aber das Problem der Untersuchungsdauer lösen.

Campbell und Stanley (1963, S. 34) sprechen in Situationen, in denen man wesentliche Prinzipien experimenteller Untersuchungen anwendet, ohne allen entsprechenden Anforderungen gerecht werden zu können, von **Quasi-Experimenten.** Da bei Quasi-Experimenten durch den notwendigen Verzicht auf die zufällige Zuordnung von Untersuchungsobjekten zu Versuchs- und Kontrollgruppen entsprechende Fehler nicht ausgeschlossen werden können, sind andere Wege zum Ausschluss alternativer Erklärungsmöglichkeiten notwendig. Shadish et al., (2002, S. 105) heben dazu u. a. die „Identifizierung und Analyse möglicher Bedrohungen der internen Validität“ durch kritische Überprüfung in Frage kommender Einflussfaktoren hervor. Andererseits haben Quasi-Experimente oftmals Vorteile im Hinblick auf die externe Validität, weil die verwendeten Daten in „natürlichen“ Situationen erhoben wurden.

Hintergrundinformation
Campbell und Stanley (1963, S. 34) zu Quasi-Experimenten:

> „Es gibt viele reale Situationen, in denen der Forscher so etwas wie ein experimentelles Design bei seiner Untersuchung anwenden kann (z. B. beim „wann" und „bei wem" der Messungen), obwohl er nicht die volle Kontrolle über den Einsatz der experimentellen Stimuli hat (das „wann" und „bei wem" des Einsatzes der Stimuli und dessen Randomisierung), was ein wirkliches Experiment ermöglicht".

Kerlinger und Lee (2000, S. 536) kennzeichnen die Gründe für die Durchführung von Quasi-Experimenten:

> „Das wirkliche Experiment bedarf der Manipulation mindestens einer unabhängigen Variablen, der zufälligen Zuordnung der Ausprägungen der unabhängigen Variablen zu den Gruppen. Wenn eine oder mehrere dieser Voraussetzungen aus dem einen oder anderen Grund nicht gegeben ist, haben wir es mit einem „Kompromiss-Design" zu tun. Kompromiss-Designs sind bekannt als quasi-experimentelle Designs".

Zwei gängige Arten von Quasi-Experimenten seien kurz charakterisiert:

- **Zeitreihen-Design**
  Hier beobachtet man den Verlauf einer Zeitreihe der interessierenden abhängigen Variablen (z. B. Marktanteil). Wenn zu einem bestimmten Zeitpunkt eine Marketing-Maßnahme (z. B. Preissenkung) erfolgt ist, dann führt man eine signifikante Änderung des Verlaufs der Zeitreihe (z. B. deutliche und nachhaltige Steigerung des Absatzes) auf diese Marketing-Maßnahme zurück. Auch dabei sollen soweit wie möglich andere Einflussfaktoren (z. B. Konkurrenzaktivitäten) ausgeschlossen werden können (Abb. 6.13).
- **Designs mit Kontrollgruppe ohne Einflussfaktor**
  Grundidee dieses Designs ist es, diejenigen Personen, die einem Einflussfaktor (z. B. Kontakt zur Werbung) ausgesetzt waren, mit denen zu vergleichen, bei denen das nicht der Fall war. Bei den im folgenden Abschnitt skizzierten Markttests findet man häufig entsprechende Anwendungen. Aus einem Unterschied bei einer (vermuteten) abhängigen Variablen (z. B. Kaufintensität) schließt man auf die Wirkung des Einflussfaktors (z. B. Sonderpreise in verschiedenen Gebieten). Je besser man andere Erklärungsmöglichkeiten für den Unterschied ausschließen kann, desto aussagekräftiger sind die Ergebnisse.

In der empirischen Analyse werden hier häufig Verfahren angewendet, die den bereits oben (Abschn. 6.1) angesprochenen Matching-Ansatz aufgreifen. Bei solchen (post-hoc) Matchingverfahren werden Kontrollgruppe und Behandlungsgruppe anhand von Werten einer oder mehrerer Matchingvariablen gepaart zusammengestellt (Stuart & Rubin, 2008). Die statistischen Paare sollen sehr ähnliche – möglichst identische – Werte in relevanten, charakteristischen Variablen (z. B. Alter, Bildungsstand, Kaufhäufigkeit) haben, die mit der abhängigen Variablen in Beziehung stehen. Nur dann kann davon aus-

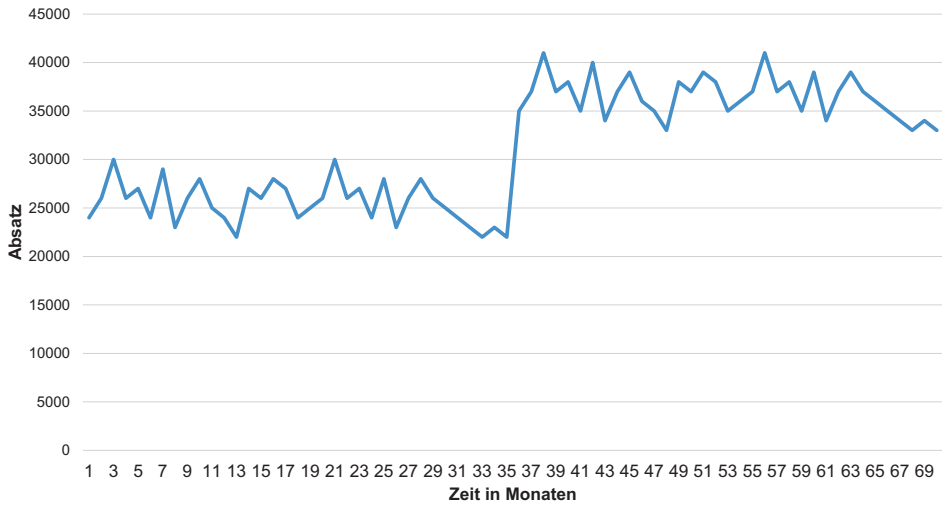


Abb. 6.13 Beispielhafter Absatzverlauf über etwa 5 Jahre: Nach einer Preissenkung zu Beginn des 3. Jahres kommt es zu einer nachhaltigen Absatzsteigerung (bei Fortführung des saisonalen Verlaufs)

gegangen werden, dass Unterschiede im Ergebnis auf den interessierenden Einflussfaktor zurückzuführen sind. Sofern also alle relevanten Variablen beim Matching zum Einsatz kommen, wird der Selektionsbias verringert und die Vergleichbarkeit der Gruppen erhöht.

Da häufig ein exaktes und vollständiges Matching nicht möglich ist, wurde eine Vielzahl statistischer Methoden zum Matching mit möglichst geringen Abweichungen entwickelt (zum Überblick: Steiner & Cook, 2014).

Wenn die vorhandene Kontrollgruppe im Ausgangsdatensatz um ein Vielfaches größer ist als die Experimentalgruppe, begünstigt dies den Matchingprozess, da dann die Wahrscheinlichkeit groß ist, passende „Zwillingspaare“ zu finden. Bei ähnlicher Gruppengröße ist ein vollständiges Matching häufig nicht möglich, sodass die Stichprobe kleiner wird, da nicht zu jedem Teilnehmer der Experimentalgruppe ein passender „Zwilling“ gefunden werden kann. Mit der Anzahl der eingeschlossenen Variablen steigt die Schwierigkeit, einen „statistischen Zwilling“ zu identifizieren. Um Verzerrungen zu vermeiden, dürfen allerdings keine relevanten Variablen ausgelassen werden. Eine kritische Diskussion verschiedener Matching Methoden findet sich bei Stuart und Rubin (2008).

6.4 Tests in der Marktforschung – Grundlagen und Überblick

Während die Panels und Wellenbefragungen, die im vorhergehenden Kapitel dargestellt sind, die Frage beantworten, wie der Zustand und die Veränderungen in einem Markt sich *tatsächlich darstellen,* beantworten Tests die Frage, *was wäre,* wenn eine

bestimmte Maßnahme ergriffen wird. So soll ein *Neuprodukttest* die Frage nach dem möglichen Erfolg der Einführung eines Neuprodukts in den Markt beantworten oder ein *Kommunikationstest* soll darüber Auskunft geben, welche Folgerungen für den Markterfolg sich aus der Änderung im Kommunikationsverhalten eines Unternehmens ergeben, um nur zwei Beispiele zu nennen. Hier wird schon erkennbar, dass es sich um Anwendungen der eingangs dieses Kapitels skizzierten Grundidee handelt, Hypothesen durch direkte Intervention bzw. durch Manipulation unabhängiger Variabler zu überprüfen. Bei den in den folgenden Abschnitten behandelten „Tests" handelt es sich um häufig angewandte Untersuchungsdesigns für wichtige Marketing-Probleme, die nicht mit den bei der Datenanalyse angewandten „statistischen Tests" (siehe Abschn. 8.2) verwechselt werden dürfen. Der übliche Sprachgebrauch in Praxis und Wissenschaft ist hier ein wenig unübersichtlich.

Dabei sind Tests nur dann sinnvoll, wenn abhängig vom Ergebnis des Tests unterschiedliche Maßnahmen ergriffen werden. Diese Abhängigkeit sollte vorher in einer Hypothese formuliert sein. Eine Hypothese ist eine Behauptung über die Folgen einer Maßnahme, die wahr oder falsch sein kann (siehe auch Abschn. 2.2.2). Sie sollte so formuliert sein, dass beim Zutreffen der Hypothese eine andere Konsequenz gezogen wird als beim Nichtzutreffen.

Ein Beispiel soll das verdeutlichen

Eine Getränkefirma überlegt, ob sie zusätzlich zur bisherigen Einwegflasche eine Mehrwegflasche in den Markt einführen soll. Das Controlling der Firma hat errechnet, dass sich die Einführung der Mehrwegflasche dann lohnt, wenn dadurch der Getränkeabsatz in Litern bei gleicher Distribution um 30 % gesteigert werden kann. Die Hypothese würde dann lauten: *„Wenn die neue Mehrwegflasche eingeführt und wie die bisherige Einwegflasche distribuiert wird, dann wird der Getränkeabsatz in Litern um 30 % oder mehr gesteigert."*

Hypothesen zerfallen formal in einen Voraussetzungsteil („Wenn … wird,") und einen Folgenteil („dann … gesteigert."). Beide Teile sind konkret. So werden im Voraussetzungsteil die Bedingungen konkret genannt, unter denen das Produkt eingeführt wird, und im Ergebnisteil wird der mindestens zu erreichende Erfolg zahlenmäßig benannt. Es heißt z. B. nicht, dass im Falle einer Einführung ein hoher Absatz oder ein guter Verkaufserfolg erzielt wird, was eben nicht konkret wäre. Wird die Maßnahme wie geplant ergriffen, dann lässt sich anschließend auch eindeutig und ohne weitere Diskussion feststellen, ob die Hypothese zutreffend ist oder nicht. ◀

Das Beispiel zeigt auch, dass bei einem Test verschiedene *Arten von Variablen* beteiligt sind:

- Die *Testvariable* oder unabhängige Variable wird im Voraussetzungsteil genannt. Im Beispiel ist das die vorgenommene oder unterlassene Einführung der neuen Mehrwegflasche mit gleicher Distribution.

- Die *Ergebnisvariable* oder abhängige Variable wird im Ergebnisteil der Hypothese genannt. Im Beispiel ist das der Mehrverkauf in Litern.
- Daneben gibt es noch *intervenierende Variable* oder Störvariable, die nicht explizit genannt sind, die aber auch die Ergebnisvariable beeinflussen können. Im Beispiel ist denkbar, dass der Verkauf des Getränks auch vom Wetter beeinflusst wird. Hypothesen sind stets so zu verstehen, dass die intervenierenden Variablen konstant gehalten werden.

Tests in der Marktforschung bestehen aus mehreren *Elementen:*

- Die *Testfragestellung,* die sich aus der Testhypothese ergibt. Diese würde im Beispiel lauten: Kann durch die zusätzliche Einführung einer gleich distribuierten Mehrwegflasche der Getränkeabsatz in Litern um mindestens 30 % gesteigert werden?
- Das *Testdesign,* mit dem festgelegt wird, welche Daten wann und unter welchen Bedingungen zu erheben sind. Dies muss so geschehen, dass die Testfragestellung beantwortet werden kann. Dazu muss das Testdesign den Einfluss der Testvariablen auf die Ergebnisvariable vom Einfluss der intervenierenden Variablen isolieren.
- Die *Testdurchführung,* bei der die im Testdesign definierten Bedingungen hergestellt und die definierten Daten erhoben werden.
- Die *Testauswertung,* bei der die erhobenen Daten so miteinander verrechnet werden, dass die Testfragestellung beantwortet werden kann.

Für das **Testdesign** gibt es mehrere Strategien, mit denen der Einfluss der Testvariablen auf die Ergebnisvariable isoliert werden kann. Eine Strategie besteht darin, die intervenierenden Variablen konstant zu halten. So ist es im Beispiel des Tests eines Getränks möglich, dass das Getränk in einem Studio bei einer im Testdesign festgelegten Raumtemperatur getestet wird. Eine weitere Strategie besteht darin, das Testdesign so festzulegen, dass der Einfluss der Störvariablen bei der Testauswertung herausgerechnet werden kann. Unter Umständen können hier gerade auch die in Abschn. 4.4.4 thematisierten impliziten Verfahren Anwendung finden.

Dies soll an Beispiel des Getränks und eines Storetests erläutert werden. Der Storetest wird weiter unten (vgl. Abschn. 6.8) erläutert. Hier reicht es zu wissen, dass es sich dabei um einen probeweisen Abverkauf in ausgewählten Testgeschäften handelt. Ein mögliches Testdesign zeigt die Abb. 6.14 (sie entspricht dem Design 2 der Abb. 6.3). Dabei wird in zwei Gruppen von Geschäften erhoben, der Testgruppe und der Kontrollgruppe. Weiter wird die gesamte Erhebungsdauer in zwei Perioden, die Vor- und die Testperiode unterteilt. Das Testdesign legt fest, dass nur in der Testperiode und der Testgruppe sowohl die Einweg- als auch die Mehrwegflasche verkauft wird, ansonsten wird nur die Einwegflasche verkauft. Die Zahlen in Klammern zeigen den Abverkauf für das Getränk gesamt pro Woche in Litern.

Der **Testeffekt** lässt sich dann wie folgt bestimmen:

$$\text{Testeffekt} = (120/100)/(90/80) = 1{,}067$$

Abb. 6.14 Testdesign zum Test einer zusätzlichen Mehrwegflasche

	Vorperiode	Testperiode
Testgruppe	Einweg (100)	Einweg + Mehrweg (120)
Kontrollgruppe	Einweg (80)	Einweg (90)

→ Steigerung Abverkauf durch die neue Flasche um 6,7 %.

Insbesondere die Einführung eines neuen Produkts ist für die Hersteller mit einem erheblichen Risiko verbunden. Von daher ist es nicht erstaunlich, dass für den Test neuer Produkte eine ganze Reihe von Testverfahren existiert. Das beginnt bereits beim *Konzepttest,* bei dem überprüft wird, ob eine Produktidee erfolgversprechend ist. Ziel ist es, die weitere Produktentwicklung in die richtige Richtung zu lenken. Beim *Produkttest* wird dagegen ein fertiger Prototyp des Produkts getestet. Dabei können das Produkt als Ganzes oder aber einzelne Aspekte des Produkts (z. B. Geschmack oder Verpackung) getestet werden. Die Kommunikation und der Preis können in eigenen *Kommunikations- und Preistests* überprüft werden.

Schließlich wird bei den *Markttests* überprüft, ob die einzelnen Elemente des Marketingmix auch zusammenpassen. So kann jeder Teil für sich optimiert sein, die einzelnen Teile passen aber nicht zusammen. Zum Beispiel müssen bei einem Premiumprodukt die Verpackung, die Kommunikation und der Preis die Hochwertigkeit signalisieren, die dann das Produkt auch halten muss. Abb. 6.15 gibt einen Überblick über die verschiedenen Testverfahren.

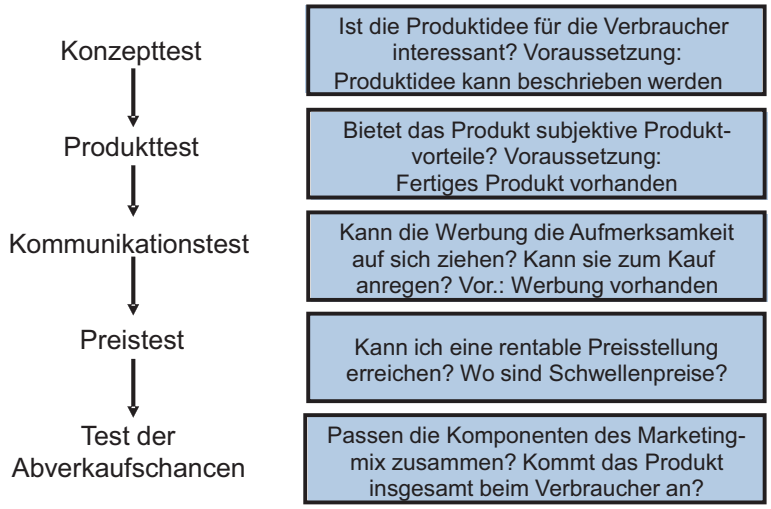


Abb. 6.15 Überblick über die Testverfahren in der Marktforschung

6.5 Konzept- und Produkttests

Konzepttests dienen der Überprüfung eines Produktkonzepts. Dieses muss zutreffend beschrieben werden können. Beim **Produkttest** ist dagegen das Produkt bereits vorhanden und wird getestet. Dabei geht es beim Produkttest stets um das subjektive Produkterlebnis, nicht um objektive Produkteigenschaften. Das unterscheidet ihn vom *Warentest,* der kein Marktforschungsinstrument darstellt (siehe z. B. www.test.de). So kann es sein, dass die Marke das subjektive Produkterlebnis beeinflusst.

Konzept- und Produkttests können aufeinanderfolgen. Abhängig vom Produkt wird man sich jedoch oft auch nur auf eine der Testarten beschränken. Manche Produkte lassen sich nur als Konzept prüfen, dies trifft z. B. für eine Lebensversicherung zu. Andere Produkte können nur schwer zutreffend beschrieben werden. Dies gilt vor allem für Produkte, bei denen sensorische Eigenschaften wie Geruch, Geschmack etc. wichtig sind. Diese eignen sich besser für einen Produkttest.

Beide Testarten werden mit Verbrauchern der Zielgruppe durchgeführt, wobei die Stichprobengröße i. d. R. zwischen 200 und 1000 Personen liegt. Die Größe hängt vor allem davon ab, ob und welche Untergruppen analysiert werden sollen. So wird man bei einem Relaunch die Beurteilung der Verwender des bisherigen Produkts sehen wollen. Auch die nachfolgend aufgeführten wesentlichen Fragestellungen sind bei beiden Testarten gleich.

- Die *Akzeptanz* ermittelt, ob das Produkt bzw. das Produktkonzept insgesamt beim Verbraucher auf positive Resonanz stößt. Diese wird häufig als *Kaufbereitschaft* abgefragt. Bei der Auswertung ist zu beachten, dass diese in der Regel deutlich nach oben verzerrt ist. Sie muss daher mit der auf gleiche Weise ermittelten Kaufbereitschaft anderer Produkte verglichen werden, deren Erfolg bekannt ist, um sachgerecht beurteilt werden zu können.
- Die *Stärken- und Schwächenanalyse* liefert Hinweise darauf, was auf keinen Fall geändert werden sollte bzw. wo Verbesserungspotenzial besteht.
- Die *Glaubwürdigkeit* gibt Auskunft, ob das ausgelobte Produktversprechen glaubwürdig ist. Dies ist insbesondere dann wichtig, wenn das Produktversprechen nicht oder nur sehr schwer überprüft werden kann, z. B. das Versprechen, dass ein Waschmittel die Wäsche ganz besonders schont.
- Die *Uniqueness* eines Produkts zeigt, ob das Produkt als einzigartig empfunden wird oder ob der Verbraucher der Ansicht ist, dass es viele vergleichbare Produkte gibt.
- Die *Relevanz* zeigt, wie wichtig dem Verbraucher das Produkt bzw. das Konzept ist.
- Schließlich gibt die *Stimmigkeit* Auskunft darüber, ob das Produkt bzw. Produktkonzept zu der Marke passt. So wird es zu einer als umweltfreundlich positionierten Marke nicht passen, wenn unter diesem Markennamen ein Waschmittel mit besonders großer Reinigungskraft angeboten wird.

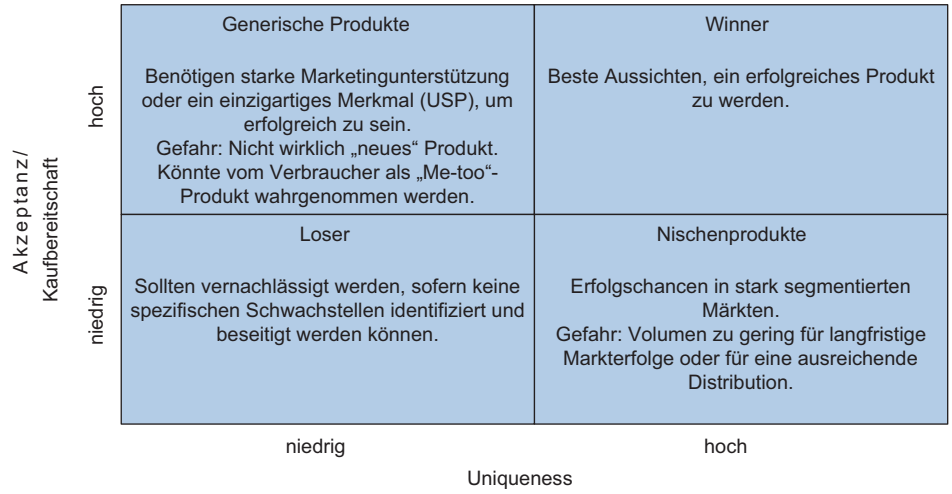


Abb. 6.16 Die Aussagekraft von Uniqueness und Akzeptanz bei Konzept- und Produkttest

Insbesondere die Kombination aus Akzeptanz und Uniqueness gibt deutliche Hinweise darauf, ob und wie ein Produkt in den Markt erfolgsversprechend eingeführt werden kann (vgl. Abb. 6.16).

Beim Produkttest gibt es je nach Fragestellung verschiedene Unterscheidungen:

- *Studiotest vs. Inhometest:* Wird das Produkt in einem Studio zum Probieren verabreicht, so hat das den Vorteil, dass die Situation besser kontrollierbar ist. Auf der anderen Seite eignen sich nicht alle Produktarten für einen Studiotest. Waschmittel oder Duschbäder müssen daher inhome getestet werden.
- *Monadischer Test vs. Vergleichstest:* Diese Unterscheidung zielt darauf ab, ob ein Produkt für sich oder im Vergleich zu anderen Produkten getestet wird. Grundsätzlich ist ein monadischer Test vorzuziehen, da das Ergebnis beim Vergleichstest in hohem Maße von den anderen mit getesteten Produkten abhängt. Dennoch kann es Fälle geben, wo ein Vergleichstest angezeigt ist. Dies ist z. B. dann der Fall, wenn von mehreren entwickelten Produkten nur das am meisten erfolgversprechende weiterverfolgt werden soll.
- *Blindtest vs. identifizierender Test:* Beim Blindtest wird der Markenname des getesteten Produkts dem Probanden nicht genannt. Sein Ziel ist die Beurteilung der Produktqualität unabhängig vom Markennamen. Beim identifizierenden Test ist der Markenname bekannt. Dadurch kann einerseits das Produkterlebnis an sich beeinflusst werden, da eine starke Marke eine ganze Reihe von Vorstellungen beim Verbraucher aktivieren kann, andererseits wird dadurch die Beurteilung der Qualität verzerrt.

6.6 Kommunikationstests

Kommunikationstests finden meist in der Form statt, dass eine Werbung hinsichtlich ihrer Gestaltung auf verschiedene Aspekte geprüft wird. Die wichtigsten Aspekte sind meist die *Aufmerksamkeit,* die eine Werbung erzielen kann und/oder inwieweit eine Werbung geeignet ist, den Empfänger zum Kauf zu *motivieren.* Weitere Aspekte können sein, ob die Werbung gefallen hat, welche Aspekte erinnert werden und anderes mehr. Da diese Prüfung i. d. R. erfolgt, bevor die Werbung geschaltet wird, werden diese Tests auch als **Werbepretests** bezeichnet. Daneben gibt es auch die Möglichkeit, außer der Gestaltung die komplette Mediastrategie in einem Testmarkt zu testen (siehe Abschn. 6.8).

Eine erste Möglichkeit der Überprüfung einer Werbung ist die Blickregistrierung. Ausgangspunkt ist die Tatsache, dass der Mensch nur einen sehr kleinen Bereich scharf und farbig sieht. Damit er einen größeren Bereich visuell erfassen kann, scannt das menschliche Auge den Bereich in einer Serie von Fixationen ab, wobei pro Sekunde etwa fünf Fixationen möglich sind. Die Blickregistrierung erfasst diese Fixationen und kann dadurch sagen, welche Bereiche einer Werbung in welcher Reihenfolge und wie lange betrachtet wurden. Dies geschieht entweder durch eine Spezialbrille oder durch ein Zusatzgerät an einem Monitor, mit denen jeweils die Augenbewegungen erfasst und den entsprechenden Bereichen der Werbebotschaft zugeordnet werden, wenn die Werbung auf einem Computerbildschirm gezeigt wird.

Abb. 6.17 zeigt das Ergebnis eines solchen Tests. Dabei wurden Probanden in ein Studio eingeladen. Nachdem ihnen die Spezialbrille aufgesetzt wurde, wurde ihnen gesagt, dass der eigentliche Test gleich los geht, sie müssten nur noch ein wenig warten und können sich in dieser Zeit eine Zeitschrift durchblättern, die sie erhalten. In dieser Zeitschrift ist dann auch die zu testende Werbung enthalten. Diese verdeckte Vorgehensweise ist notwendig, um ein valides Ergebnis zu bekommen. Würde nämlich gesagt, es gehe darum, wie Werbung in Zeitschriften betrachtet wird, würde die Werbung auf sehr untypische Weise betrachtet.

Die linke Anzeige ist die vor der Optimierung. Für die Anzeige insgesamt und verschiedene Bereiche der Anzeige ist vermerkt, welcher Anteil der Probanden die Anzeige bzw. den betreffenden Bereich überhaupt betrachtet hat und wenn ja wie lange im Mittel. Die Werbung hat ja auch die Aufgabe, die Aufmerksamkeit auf die Anzeige zu lenken.

So ist erkennbar, dass 98 % die linke Anzeige gesehen haben und diese im Mittel 1,21 s betrachtet haben. Die Frau betrachten 69 % im Schnitt 0,49 s lang. Dies ist zunächst nicht negativ, da das Bild der Frau Aufmerksamkeit auf die Anzeige lenken soll. Negativ ist jedoch zu werten, dass schon das Haus nur noch von 15 % der Betrachter gesehen wird und der Schriftzug „Ich geh den Leonberger Weg“ nur von 39 % der Betrachter. Damit ist zu fürchten, dass ein Großteil der Betrachter gar nicht erkennt, dass es sich um eine Anzeige der Leonberger Bausparkasse handelt.

Abb. 6.17 Beispiel für Ergebnisse der Blickregistrierung mit einer Anzeige vor und nach der Optimierung. (Erläuterung siehe Text. Mit freundlicher Genehmigung der GfK SE)

Bei der Neugestaltung der Anzeige (Abb. 6.17, rechtes Bild) wurde der Balken nach rechts genommen, dadurch die Frau etwas nach links verschoben und das Haus in das Zentrum gerückt, mit der Folge, dass nun 55 % das Haus betrachten. Weiter wurde der Schriftzug des Slogans „Ich geh den Leonberger Weg“ deutlich vergrößert, sodass er nun von 67 % der Betrachter betrachtet wird. Die optimierte Anzeige kann also wesentlich mehr Aufmerksamkeit für den Absender der Werbung generieren, auch wenn sich die Werte für die Anzeige insgesamt kaum verändert haben.

Eine zweite Testart versucht zusätzlich, die Fähigkeit einer Werbung zu erfassen, Präferenz für die beworbene Marke zu generieren. Werbepretests werden von verschiedenen Instituten unter verschiedenen Namen angeboten. Typischerweise laufen sie wie folgt ab:

Der Test findet online am PC oder in einem Studio statt. Bei der online-Variante wird den Probanden gesagt wird, dass es sich um eine Untersuchung zu Multimedia-PCs handelt. Auch hier wird also vom eigentlichen Zweck des Tests abgelenkt, um ein untypisches Betrachten der Werbung zu vermeiden. Zu Beginn wird der Proband nach seiner Präferenz in einem Produktfeld gefragt. Anschließend wird er gebeten, verschiedene Aufgaben zu erfüllen. Dabei wird er auch mit dem zu testenden sowie weiteren Werbefilmen konfrontiert. Nach dem ersten Sehen wird er nach den in der Werbung gezeigten Marken gefragt, nach dem zweiten Sehen erneut nach seiner Präferenz. Aus dem Prozentsatz der Erinnerer ergibt sich als erste Kennzahl die „Awareness“ bzw. das Durch-

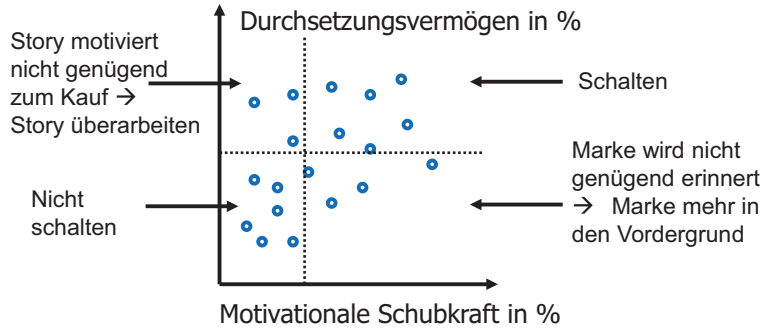


Abb. 6.18 Mögliche Ergebnisse und Konsequenzen aus einem Werbepretest

setzungsvermögen, aus der Präferenzänderung die „Motivationale Schubkraft" der Werbung. Erst dann wird für die Probanden offensichtlich, dass es sich um einen Werbetest handelt und sie werden detailliert zu weiteren Informationen befragt, z. B. ob die Werbung gefallen hat, ob sie überzeugend ist, ob die Geschichte verstanden wurde und ähnliches mehr. Auf die Ähnlichkeit dieser praktischen Anwendung mit den im Abschn. 6.1 erörterten Beispielen von Untersuchungsdesigns sei hingewiesen.

Die beiden Parameterwerte „Awareness" und „Motivationale Schubkraft" werden mit den entsprechenden Werten früherer Tests im gleichen Produktfeld verglichen. Sind beide Parameterwerte niedrig, so ist von einer Schaltung der Werbung anzuraten. Sind beide Werte hoch, dann lautet die Empfehlung, die Werbung zu schalten. Ist die „Awareness" gering, die „Motivationale Schubkraft" aber überdurchschnittlich, dann wird die Marke nicht genügend erinnert. Sie sollte dann z. B. durch eine zusätzliche oder längere Produktabbildung mehr in den Vordergrund gerückt werden. Ist die „Motivationale Schubkraft" gering, die „Awareness" aber hoch, dann wird die Marke zwar gut erinnert, die Geschichte motiviert aber nicht genügend zum Kauf. In diesem Fall muss in der Regel der Werbefilm neu gedreht werden. Abb. 6.18 verdeutlicht diese Zusammenhänge. Der Test kann verbunden werden mit Eye-Tracking oder mit Facial Coding, bei dem die Mimik auf Emotionen hin analysiert wird.

6.7 Preistests

Der Preis war ein lange Zeit in der Testmarktforschung vernachlässigtes Instrument des Marketings. Dabei besteht kein Zweifel, dass eine falsche Preissetzung die Profitabilität eines Produkts stark beeinträchtigen bzw. zerstören kann (vgl. Diller, 2008, S. 21).

Eine relativ einfache Form der Preisanalyse ist der Price Sensitivity Meter oder die Van-Westendorp-Analyse (van Westendorp, 1976) die die Ober- und Untergrenzen für eine Preissetzung ermittelt. Hier soll eine gegenüber dem Original leicht veränderte Version dargestellt werden, die zu engeren Preisbändern führt. Zuerst wird einer Stichprobe

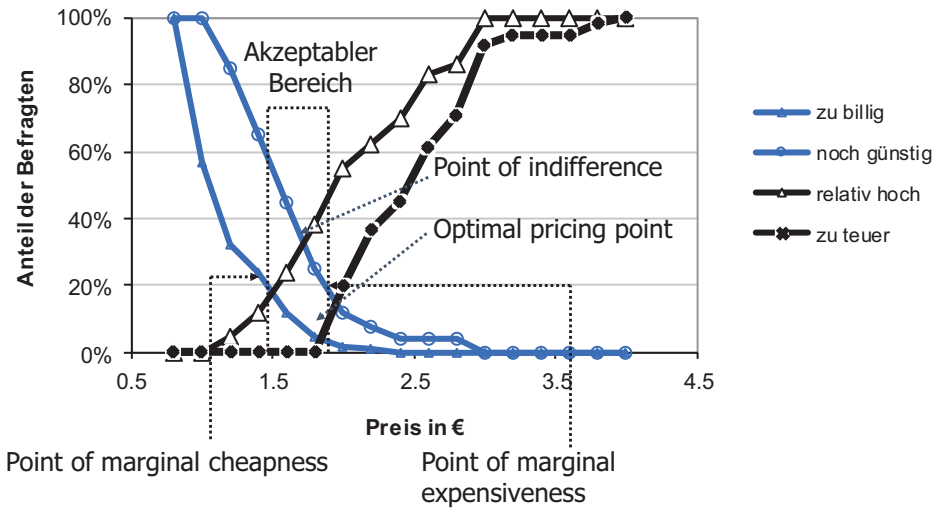


Abb. 6.19 Auswertung des Price-Sensitivity-Meters. (Quelle: Wildner, 2003)

von etwa 300 oder mehr potenziellen Kunden das Produkt vorgestellt. Dann werden die Befragten gebeten, vier Preise zu nennen (offene Abfrage):

1. Einen Preis, der angemessen, aber noch günstig ist.
2. Einen Preis, der relativ hoch, aber noch vertretbar.
3. Den Betrag, ab dem der Preis zu hoch wird
4. Den Betrag, ab dem der Preis so niedrig ist, dass Zweifel an der Qualität geweckt werden.

Die Werte der Fragen 1 und 4 werden nun abwärts, die Kurven 2 und 3 aufwärts kumuliert dargestellt. So finden in Abb. 6.19 etwa 57 % einen Preis von 1 € als zu niedrig und 55 % finden einen Preis von 2 € relativ hoch.

Wichtig sind nun die Punkte, bei denen sich die Kurven „zu billig“ und „relativ hoch“ schneiden. Es ist nicht ratsam, einen noch günstigeren Preis zu wählen, da sonst mehr Personen sagen, dass das Produkt zu billig ist, als Personen, die den Preis als relativ hoch einstufen. Auf der anderen Seite ist es auch nicht ratsam, den Preis über den Schnittpunkt der Kurven „zu teuer“ und „noch günstig“ zu erhöhen. Diese beiden als „Point of marginal cheapness“ und „Point of marginal expensiveness” bezeichneten Punkte grenzen also den akzeptablen Bereich für den Preis ein.

Das Verfahren kann schnell und preiswert erhoben und ausgewertet werden. Die Tatsache, dass jeweils nur ein Produkt unabhängig von anderen Produkten untersucht wird, erlaubt es auf der einen Seite, die Methode auch bei völlig neuartigen Produkten einzusetzen. Auf der anderen Seite ist weder der Einfluss der Preise anderer Produkte noch der

Einfluss auf die verkauften Mengen der anderen Produkte bekannt. Es liefert also lediglich den Preisbereich, für den es am wenigsten Probleme gibt.

Sollen Preiselastizitäten und Kreuzpreiselastizitäten bestimmt werden, so hat sich ein 1998 vorgestelltes Instrument bewährt, das von der GfK als „Price Challenger" angeboten wird (vgl. Wildner, 1998, 2003). Vorbereitend wird jedem Produkt eine Reihe von Preisen zugeordnet. Die Preise sollen einen relativ großen Bereich abdecken, wobei dieser von Produkt zu Produkt verschieden ist.

Im Interview werden zunächst das Produkt und die wichtigsten Konkurrenzprodukte vorgestellt, die zusammen etwa 80 % des Marktes abdecken sollten. Zu jedem Produkt wird gefragt, ob es bekannt ist. Dann werden von den bekannten Produkten, diejenigen eliminiert, die für den Kauf nicht infrage kommen. Ergebnis ist das individuelle „Relevant Set". Anschließend werden die Produkte des **Relevant Sets** mit Preisen gezeigt, die zufällig aus den dem jeweiligen Produkt zugeordneten Preisen ausgewählt wurden. Der Proband wird nun gebeten zu sagen, welches Produkt er in dieser Situation wählen würde oder ob er kein Produkt kaufen würde. Anschließend wird dem Probanden gesagt, dass er nun in ein Geschäft kommt, in dem andere Preise gelten. Wieder werden ihm die Produkte mit zufällig ausgewählten Preisen gezeigt und er wird nach seiner Kaufabsicht gefragt. Dies wird mehrfach wiederholt.

Aus den Daten lassen sich mit einer logistischen Regression (vgl. Abschn. 9.4) die Preis-Absatzfunktionen berechnen, welche für jedes Produkt des Relevant Sets die Preisabhängigkeit des Produkts vom eigenen und den Konkurrenzpreisen beschreibt. Da die wiederholte Simulation zu überhöhten Preiselastizitäten führt, muss das Ergebnis noch auf die Werte der ersten Preissimulation angepasst werden (vgl. Wildner, 2003). Werden die Daten aller Probanden zusammengefasst, dann liefert das Modell Preis- und Kreuzpreiselastizitäten und darüber hinaus auch Preisschwellen. Dies gestattet auch eine gemeinsame Preisoptimierung von mehreren Produkten, was für Firmen wichtig ist, die in einem Produktfeld mehrere Produkte anbieten, wie z. B. Henkel im Waschmittelbereich mit den Marken „Persil", „Spee" und „Weißer Riese". Auf der anderen Seite führt das Erfordernis der Abfrage in einem „Relevant Set" auch dazu, dass das Verfahren nur für Produkte geeignet ist, für die es auch eine definierte Warengruppe gibt. Völlig neuartige Produkte, die noch ohne Konkurrenz sind, können folglich damit nicht untersucht werden.

6.8 Markttests

Markttests überprüfen, ob die einzelnen Teile des Marketingmix sich zu einem stimmigen Ganzen verbinden. Sie unterscheiden sich danach, welches Marketingmix eingesetzt und getestet werden kann und wie tief das Ergebnis analysiert werden kann. Die im Folgenden beschriebenen Markttests wurden für die täglichen Verbrauchsgüter entwickelt.

Der einfachste Test ist der sogenannte **Storetest,** der schon im Abschn. 6.4 erwähnt wurde. Dabei werden je nach Testdesign ein oder mehrere Gruppen von Geschäften angeworben. Dort kann dann ein neues Produkt zum Verkauf angeboten werden. Weiter ist es möglich, unterschiedliche Preise, Platzierungen oder Handelsaktionen einzusetzen. Es kann aber keine klassische Werbung eingesetzt werden. Dabei übernimmt das Marktforschungsinstitut die Anwerbung der Geschäfte, die Platzierung der Testware, die sonstige Einhaltung der Testbedingungen und die Erhebung der Verkaufsdaten. Storetests liefern Ergebnisse, die besonders geeignet sind, Handelspartner zu überzeugen, nämlich Abverkäufe, Abverkaufsanteile und Regalumschlag. Sie können aber nur bedingt Aussagen zum langfristigen Produkterfolg treffen, da die dafür wichtige Trennung zwischen Erst- und Wiederkauf nicht vorhanden ist. Die Dauer eines Storetests beträgt in der Regel 20–26 Wochen, die Kosten liegen bei ca. 70 Tsd. €.

Ein heute nur noch selten eingesetzter Test ist der **regionale „Testmarkt“**. Dabei führt der Hersteller das neue Produkt unter möglichst realistischen Bedingungen in einem Testgebiet ein. Bis 1989 war West-Berlin ein wegen seiner Isolation beliebtes Testgebiet, heute ist es wegen der zentralen Lage vor allem Hessen. Die Distribution erfolgt durch den Außendienst. Das Marktforschungsinstitut erhebt die Daten in einer Stichprobe von Geschäften. In der Praxis wird oft ein schon bestehendes Handelspanel genutzt, das noch durch zusätzliche Geschäfte aufgestockt wird. Die Kosten für die Marktforschung sind dabei eher gering (ca. 50 Tsd. €). Teuer (ab ca. 1 Mio. €) sind dagegen die Produktion der Testprodukte und die Listung der Produkte beim Handel. Dabei können Handelsaktionen und auch klassische Werbung eingesetzt werden. Ergebnisse sind auch hier Verkäufe und Verkaufsanteile. Zusätzlich erhält der Hersteller Hinweise zur Schnelligkeit des Distributionsaufbaus. Doch ebenso wie beim Storetest gibt es keine Trennung zwischen Erst- und Wiederkauf, weswegen es nicht möglich ist, die langfristigen Erfolgschancen eines Produkts abzuschätzen.

Elektronische Mikrotestmärkte wurden seit den 1970er Jahren von ERIM in Frankreich und seit 1980 von IRI in den USA unter dem Namen „BehaviorScan“ entwickelt. 1977 wurden das ERIM-Panel und 1985 BehaviorScan von der GfK in Deutschland eingeführt (vgl. Feldenkirchen & Fuchs, 2009, S. 159 f.). Für GfK „BehaviorScan“ wurden in Haßloch (Rheinland-Pfalz) mit Ausnahme eines Aldi-Marktes alle wichtigen Geschäfte zur Kooperation gewonnen. In diesen Geschäften können Testprodukte distribuiert werden. Außerdem können hier alle Formen von Handelsaktionen durchgeführt werden. Weiter wurden in Haßloch 3400 Testhaushalte angeworben, von denen 2400 über Kabel mit spezieller TV-Werbung versorgt werden können. Die Testhaushalte erhalten wöchentlich auch die Fernsehzeitschrift „HörZu“. In diese kann Werbung für Testprodukte eingebunden werden. Schließlich sind auch Plakatwerbung oder Zeitungswerbung möglich.

Die teilnehmenden Haushalte sind mit Identifikationskarten ausgestattet, die sie in den kooperierenden Geschäften an der Scannerkasse vorzeigen. Auf diese Weise wer-

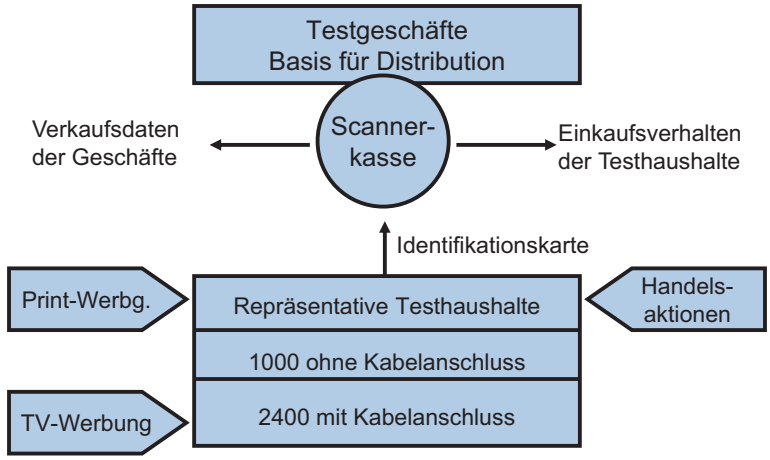


Abb. 6.20 Testsystem des elektronischen Mikrotestmarkts in Haßloch. (Mit freundlicher Genehmigung der GfK SE.)

den ihre Einkäufe erfasst (siehe auch Abb. 6.20). Neben den Verkaufsdaten der Testgeschäfte, welche die Grundlage der Analyse der Wirksamkeit von Handelsaktionen sind, werden demnach auch Einkaufsdaten erfasst. Somit wird erfasst, wie hoch Erst- und Wiederkauf sind. Damit ist auch eine Messung des langfristigen Erfolgs möglich. Weiter ermöglicht ein Vergleich des Einkaufsverhaltens beworbener mit dem nicht beworbener Haushalte eine direkte Analyse der Wirksamkeit der Werbung.

Positiv ist somit zu werten, dass der elektronische Mikrotestmarkt einen nahezu vollständigen Test des Marketingmix erlaubt und ein umfassendes Datenspektrum zur Analyse liefert. „BehaviorScan“ gestattet damit nicht nur, Neuprodukte unter realistischen Bedingungen einzuführen. Es ermöglicht auch den Test verschiedener Kommunikationsstrategien. Nachteilig ist jedoch, dass die Neuprodukte selbst nicht geheim bleiben und dass für einen Test mindestens drei Kaufzyklen der betreffenden Warengruppe erfasst werden müssen. Ein Test dauert somit je nach Warengruppe oftmals ein halbes bis ein Jahr, was vielen Marketingentscheidern zu lange ist. Dies führte dazu, dass der Testmarkt inzwischen eingestellt wurde.

Beide Nachteile versucht die **Testmarktsimulation** zu vermeiden, die von verschiedenen Instituten angeboten werden.

Die Testmarktsimulation ist ein Studiotest, der auf der Prognose nach Parfitt und Collins (1968) beruht. Parfitt und Collins führen eine mathematische Zerlegung des Marktanteils durch. Es gilt nämlich:

$$MA = P \times B \times KI$$

mit:

MA	=	*Marktanteil* (Menge)
P	=	*Käuferpenetration* = Anteil der Käufer des Produkts an den Käufern der Warengruppe. Die Penetration drückt aus, wie viele der Warengruppenkäufer das Produkt zumindest probieren
B	=	*Bedarfsdeckung* = Anteil (Menge), den das neue Produkt an den Warengruppenkäufen der Käufer des Produkts auf sich vereint. Die Bedarfsdeckung ist damit – ähnlich wie die Wiederkaufsrate und die Wiederkäuferpenetration – ein Maß für Treue zu dem neuen Produkt
KI	=	*Kaufintensitätsindex* = durchschnittliche Warengruppenmenge eines Käufers des Produkts dividiert durch die durchschnittliche Warengruppenmenge eines Warengruppenkäufers. Ist der Kaufintensitätsindex größer als 1, dann handelt es sich bei den Käufern des Produkts um Intensivkäufer, ist er kleiner als 1, dann sind es Extensivkäufer. Der Index sagt damit etwas über die Wertigkeit der erreichten Zielgruppe

Das folgende Beispiel macht die Zusammenhänge deutlich

Gegeben sei ein Haushaltspanel mit 5 Haushalten 1 bis 5, in dem eine Warengruppe erhoben wird, die aus den drei Produkten A, B und C besteht. Weiter wird angenommen, dass bei jedem Kaufakt nur ein Stück gekauft wird und dass alle Produkte gleichen Inhalt haben. Die folgende Tabelle zeigt die Einkäufe in diesem Panel in dieser Warengruppe, wobei ein x für einen Einkaufsakt steht. So hat Haushalt 1 zweimal Produkt A und einmal Produkt B eingekauft.

	Haushalt				
Produkt	1	2	3	4	5
A	xx	x			
B	x	xx	x		
C		x		xx	

Damit ergibt sich für Produkt A:

- Penetration = (2 Käufer von A, nämlich Haushalte 1 und 2)/(4 Warengruppenkäufer, nämlich Haushalte 1, 2, 3 und 4) = 0,5
- Bedarfsdeckung = (3 Kaufakte der Käufer von A für A)/(7Kaufakte der Käufer von A in der Warengruppe) = 0,43
- Kaufintensität = (7/2)/(10/4) = 1,4 (die 2 Käufer von A kaufen in der Warengruppe 7 Mal, die 4 Warengruppenkäufer insgesamt 10 Mal).
- Der Marktanteil von A ergibt sich damit wie folgt: $0.5 \times 0{,}43 \times 1{,}4 = 0{,}30 = 30\ \%$. ◄

Es werden entweder etwa 300 Käufer der Warengruppe in ein Studio eingeladen oder - und sann sind auch größere Stichproben möglich - zu einem entsprechenden Online-Interview. Diese erhalten zunächst etwas mehr Geld, als das teuerste Produkt in der Warengruppe kostet. Dann werden sie zu ihrem Kaufverhalten in der Warengruppe befragt. Dabei wird erfasst, wie oft und wie viel sie in der Warengruppe einkaufen. Daraus lässt sich der Kaufintensitätsindex ermitteln. Dann wird das *Relevant Set* (siehe Abschn. 6.7) bestimmt und es werden Chips auf die Produkte des Relevant Set so verteilt, wie es den gekauften Anteilen entspricht. Damit ergibt sich die Bedarfsdeckung im bestehenden Markt. Anschließend erhalten sie Werbung für Produkte in der Warengruppe, u. a. auch für das Testprodukt. Damit ist gewährleistet, dass alle das Testprodukt kennen.

Nun werden die Versuchspersonen zu einem im Studio aufgebauten bzw. online erzeugten Regal geführt und gebeten, ein Produkt einzukaufen. Dieses Produkt wird dann an einer Kasse bezahlt, wobei das eingangs erhaltene Geld verwendet wird. Anschließend werden die Personen gefragt, welches Produkt sie gekauft hätten, wenn das gewählte Produkt nicht verfügbar wäre. Dies wird so oft wiederholt, wie das jeweilige Relevant Set groß ist. Aus dem Anteil der Käufer, die das Testprodukt gekauft hätten, ergibt sich die Käuferpenetration.

Anschließend erhalten die Versuchspersonen (bzw. deren Haushalte) das Testprodukt und das am meisten bevorzugte Konkurrenzprodukt bzw. es wird ihnen zugeschickt. Sie werden gebeten, in der folgenden Zeit beides zu probieren. Nach etwa vier Wochen werden sie angerufen bzw. zu einem weiteren online- Interview gebeten und zunächst nach ihren zukünftigen Kaufabsichten befragt, indem sie wieder Chips auf die Produkte des Relevant Set verteilen. Dadurch erhält man die Bedarfsdeckung des Testprodukts. Anschließend wird zu den Vor- und Nachteilen des Testprodukts befragt. Mit diesen Ergebnissen lässt sich nach Parfitt-Collins der Marktanteil bei 100 % Distribution und 100 % Awareness bestimmen. Aus dem Mediaplan des Herstellers kann die tatsächlich im Markt erreichbare Awareness geschätzt werden. Weiter wird die tatsächlich erreichbare Distribution vom Hersteller geschätzt. Daraus lässt sich ein Korrekturfaktor bestimmen, der den Marktanteil auf ein realistisches Maß korrigiert.

Die Testmarktsimulation ist damit ein sehr schnelles Verfahren, das den wichtigen Vorzug hat, dass das Testprodukt weitgehend geheim gehalten wird. Es liefert darüber hinaus Hinweise auf Verbesserungsmöglichkeiten des Produkts. Aus dem Vergleich des bestehenden Marktes mit dem Markt einschließlich Testprodukt lässt sich auch ermitteln, welche Produkte durch das neue Produkt besonders verlieren. Nachteilig ist jedoch, dass das Parfitt-Collins-Modell, das die Basis der Testmarktsimulation bildet, eine vorhandene Warengruppe voraussetzt. Damit können ganz neuartige Produkte, die eine neue Warengruppe schaffen, nicht getestet werden. Weiter geschieht die Produkteinführung unter künstlichen Bedingungen. So können keine Handelsaktionen eingesetzt werden. Auch die Werbung wird unter wenig realistischen Bedingungen gezeigt. Dennoch haben die Vorteile der Testmarktsimulation in den vergangenen Jahren dazu geführt, dass diese einen höheren Anteil an den Tests für Neuprodukteinführungen auf sich vereinigen konnten.

Literatur

Arabatzis, T. (2008). Experiment. In S. Psillos & M. Curd (Hrsg.), *The Routledge companion to philosophy of science* (S. 159–170). Routledge.

Aronson, E., Wilson, T. , & Brewer, M. (1998). Experimentation in social psychology. In D. Gilbert, S. Fiske, & G. Lindzay (Hrsg.), *The handbook of social psychology* (4. Aufl., S. 99–142). McGraw-Hill.

Baron, R., & Kenny, D. (1986). The moderator–mediator variable distinction in social psychology research: Conceptual, strategic, and statistical considerations. *Journal of Personality and Social Psychology, 51*, 1173–1182.

Calder, B., Phillips, L., & Tybout, A. (1981). Designing research for applications. *Journal of Consumer Research, 8*, 197–207.

Calder, B., Phillips, L., & Tybout, A. (1982). The concept of external validity. *Journal of Consumer Research, 9*, 240–244.

Campbell, D., & Stanley, J. (1963). *Experimental and quasi-experimental designs for research*. Rand-McNally.

Chalmers, A. (1999). *What is this thing called science?* (3. Aufl.). Open University Press.

Crasnow, S. (2019). Bias in Social Science Experiments. In L. McIntyre & A. Rosenberg (Hrsg.), *The routledge companion to philosophy of social science* (S. 191–201). Routledge.

Diller, H. (2008). *Preispolitik* (4. Aufl.). Kohlhammer.

ESOMAR. (Hrsg.). (2016). ICC/ESOMAR International Code on Market, Opinion and Social Research and Data Analytics. https://www.esomar.org/uploads/public/knowledge-and-standards/codes-and-guidelines/ICCESOMAR_Code_English_.pdf. Zugegriffen: 29. Feb. 2020.

Feest, U., & Steinle, F. (2016). Experiment. In P. Humphreys (Hrsg.), *The Oxford handbook of philosophy of science* (S. 274–295). Oxford University Press.

Feldenkirchen, W., & Fuchs, D. (2009). *Die Stimme des Verbrauchers zum Klingen bringen*. Piper.

Gneezy, A. (2017). Field experimentation in marketing research. *Journal of Marketing Research, 54*, 140–143.

Gneezy, A., Gneezy, U., & Lauga, U. (2014). Reference-dependent model of the price-quality heuristic. *Journal of Marketing Research, 51*, 153–164.

Groves, R., Fowler, F., Couper, M., Lepkowski, J., Singer, E., & Tourangeau, R. (2009). *Survey Methodology* (2. Aufl.). Wiley.

Jaccard, J., & Becker, M. (2002). *Statistics for the behavioral sciences* (4. Aufl.). Wadsworth.

Jaccard, J., & Jacoby, J. (2020). *Theory construction and model-building skills – A practical guide for social scientists* (2. Aufl.). Guilford Press.

Jacoby, J. (2013). *Trademark surveys – Designing, implementing, and evaluating surveys* (Bd. I). American Bar Association.

Kerlinger, F., & Lee, H. (2000). *Foundations of behavioral research* (4. Aufl.). Thomson Learning.

Koschate, N. (2008). Experimentelle Marktforschung. In A. Herrmann, C. Homburg, & M. Klarmann (Hrsg.), *Handbuch Marktforschung* (3. Aufl., S. 107–121). Gabler.

Koschate-Fischer, N., & Schandelmeier, S. (2014). A guideline for designing experimental studies in marketing research and a critical discussion of selected problem areas. *Journal of Business Economics, 84*, 793–826.

Parfitt, J., & Collins, B. (1968). Use of consumer panels for brand share prediction. *Journal of Marketing Research, 5*, 131–146.

Shadish, W., Cook, T., & Campbell, D. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Cengage Learning.

Simon, H., & Fassnacht, M. (2016). *Preismanagement* (4. Aufl.). SpringerGabler.

Spilski, A., Gröppel-Klein, A., & Gierl, H. (2018). Avoiding pitfalls in experimental research in marketing. *Marketing ZFP, 40*(2), 58–90.

Steiner, P., & Cook, D. (2014). Matching and propensity scores. In T. Little (Hrsg.), *The Oxford handbook of quantitative methods* (Bd. 1, S. 237–259). Oxford University Press.

Stuart, E., & Rubin, D. (2008). Best practices in quasi-experimental designs: Matching methods for causal inference. In J. Osborne (Hrsg.), *Best practices in quantitative social science* (S. 155–176). Sage.

Sudman, S., & Blai, E. (1998). *Marketing research – A problem-solving approach*. McGraw-Hill.

Viswanathan, M. (2005). *Measurement error and research design*. Sage.

Voeth, M., & Niederauer, C. (2008). Ermittlung von Preisbereitschaften und Preisabsatzfunktionen. In A. Herrmann, C. Homburg, & M. Klarmann (Hrsg.), *Handbuch Marktforschung* (3. Aufl., S. 1073–1095). Gabler.

Westendorp, P. v. (1976). NSS price sensitivity meter – A new approach to the study of consumer perception of price. In *Proceedings of the 29th ESOMAR Congress, Amsterdam*.

Wildner, R. (1998). The introduction of the Euro – The importance of understanding consumer reactions. *Marketing and Research Today, 11*, 141–147.

Wildner, R. (2003). Marktforschung für den Preis. In *Jahrbuch der Absatz- und Verbrauchsforschung 1/2003* (S. 4–26). Duncker & Humblot.

Winer, R. (1999). Experimentation in the 21st century: The importance of external validity. *Journal of the Academy of Marketing Science, 27*, 349–358.

Worrall, J. (2002). What evidence in evidence-based medicine? *Philosophy of Science, 69*, 316–330.

Deskriptive Datenanalyse

7

Zusammenfassung

In den bisherigen Kapiteln stand die Erhebung von Daten im Vordergrund. Im Kap. 7 geht es erstmals um die Datenanalyse. Grundvoraussetzung dafür ist die Feststellung, welches „Messniveau" die vorliegenden Daten haben. In der Regel ist die Anwendung insbesondere leistungsfähiger Methoden der Statistik nur zulässig, wenn bestimmte Messniveaus gegeben sind. Messniveaus von Daten werden deshalb hier am Anfang behandelt. Eine der Aufgaben statistischer Methoden besteht darin, Daten über eine Vielzahl von Einzelfällen (z. B. Konsumenten, Unternehmen) zusammenzufassen, also zu „verdichten". Dazu bedient man sich statistischer Maßzahlen und Darstellungen in Form von Tabellen und Graphiken, auf die in diesem Kapitel eingegangen wird.

7.1 Überblick

In diesem und den folgenden Kapiteln wird die vorletzte Phase im Ablauf einer Marktforschungsuntersuchung behandelt, die Datenanalyse. Hier geht es einerseits darum, die gesammelten und aufbereiteten Daten zu aussagekräftigen Tabellen, Maßzahlen, Graphiken etc. zusammenzufassen, und andererseits darum zu prüfen, inwieweit die Ergebnisse, die auf der Grundlage einer Stichprobe zustande gekommen sind, auf die eigentlich interessierende Grundgesamtheit übertragen werden können. Auf den sich an die Analyse anschließenden letzten Schritt im Untersuchungsablauf, die Berichterstattung, wird aus den im Abschn. 2.2.1 dargelegten Gründen nicht vertiefend eingegangen.

Bei der Datenanalyse in der Marktforschung handelt es sich fast ausschließlich um Anwendungen des Instrumentariums der **Statistik**. Es wird hier der Versuch unternommen, die Grundideen und Anwendungsbedingungen einiger Verfahren der

H. Kreis et al., *Marktforschung,* https://doi.org/10.1007/978-3-658-44456-3_7

Datenanalyse sehr knapp und möglichst verständlich zu skizzieren. Dabei muss auf eine ausführliche theoretische Ableitung verzichtet werden; es sollen aber wenigstens die wichtigsten Schlussweisen dargestellt werden.

Vor der Erläuterung von Analyseverfahren werden im folgenden Abschn. 7.2 zunächst die unterschiedlichen Messniveaus von Daten charakterisiert, die für die Anwendbarkeit statistischer Methoden grundlegende Bedeutung haben. Danach werden im Abschn. 7.3 einige einfache Methoden vorgestellt, die zur **Verdichtung von Datensätzen** geeignet sind. Unter „Verdichtung" wird hier die Kennzeichnung des Inhalts von möglicherweise sehr umfangreichen Datensätzen durch überschaubare Tabellen, geeignete graphische Darstellungen und statistische Maßzahlen verstanden. Dabei wird auf das Instrumentarium der „beschreibenden" Statistik zurückgegriffen.

Im folgenden Kap. 8 werden **Schlüsse von Stichprobenergebnissen auf Grundgesamtheiten** erläutert. Diese haben in der Marktforschung zentrale Bedeutung, weil eben auf der Basis relativ kleiner Zahlen von Auskunfts- oder Versuchspersonen Aussagen über ganze Märkte oder Marktsegmente gemacht werden sollen. **Multivariate Analyseverfahren,** die im Kap. 9 behandelt werden, gehören inzwischen zum methodischen Standard der Marketingforschung in Wissenschaft und Praxis. Der wichtigste Grund dafür liegt darin, dass Phänomene des Marketing-Bereichs typischerweise so komplex sind, dass die gleichzeitige Analyse einer größeren Zahl von Variablen zu deren Erklärung notwendig ist.

7.2 Messniveau von Daten

Im Abschn. 2.2.2 ist skizziert worden, in welcher Weise Messungen dazu verwendet werden, Aus-prägungen theoretisch interessierender Phänomene in der Realität widerzuspiegeln. Im Abschn. 4.5.3 wurde im Zusammenhang der Codierung erläutert, dass erhobene Daten für die Zwecke der Datenaufbereitung und -analyse in der Regel in ein numerisches System übersetzt werden. Hier geht es darum, dass Zahlen sehr unterschiedliche Aussagekraft haben können und daher auch nur bestimmte Verarbeitungen der Zahlen zulässig sind.

Dies soll anhand eines praktischen Beispiels dargestellt werden, dem folgendes Szenario zugrunde liegt: Ein Hersteller von Luxusartikeln möchte in dem bayerischen Regierungsbezirk Mittelfranken einen neuen Flagship-Store eröffnen und ist auf der Suche nach einem geeigneten Standort. Dazu hat er sich von der Regionalforschung Daten zu allen Stadt- und Landkreisen Mittelfrankens besorgt (s. Tab. 7.1).

Es ist unmittelbar zu sehen, dass die Zahlen der obigen Tabelle unterschiedliche Aussagekraft haben und in der Folge auch unterschiedliche Verarbeitungen möglich sind. So haben die Zahlen in der Spalte „Einwohner" eine andere Aussagekraft als die Zahlen in der Spalte „Kennziffer Kreis". In der Folge ergibt die Addition der Einwohner aller mittelfränkischen Kreise mit 1.681.620 eine sinnvolle Information, nämlich die Einwohnerzahl von Mittelfranken gesamt, während die Addition der mittelfränkischen

Tab. 7.1 Regionale Kennziffern der Kreise und kreisfreien Städte Mittelfrankens. (Quelle: GfK Basisdaten 2023 Teil 1 von GfK GeoMarketing GmbH)

Kennziffer Kreis	Kreis	Einwohner	Bevölkerungsdichte	Arbeitslosenquote	Kaufkraft in Mio. EUR	Kaufkraft pro Einwohner Index	Rang Kaufkraft-Index
9561	SK Ansbach	41.662	417	4,2	1067,8	97,6	11
9562	SK Erlangen	113.292	1.472	3,8	3547,5	119,2	2
9563	SK Fürth	129.122	2.038	5,1	3559,5	104,9	7
9564	SK Nürnberg	510.632	2.739	5,6	13.781,1	102,7	8
9565	SK Schwabach	41.146	1.008	3,4	1168,5	108,1	5
9571	LK Ansbach	186.279	94	2,6	4800,5	98,1	10
9572	LK Erlangen-Höchstadt	139.323	247	2,6	4383,0	119,8	1
9573	LK Fürth	119.432	388	2,8	3528,8	112,5	3
9574	LK Nürnberger Land	171.424	214	2,6	4940,7	109,7	4
9575	LK Neustadt a.d.Aisch-Bad Windsheim	101.788	80	2,3	2649,7	99,1	9
9576	LK Roth	127.520	142	2,3	3532,1	105,4	6
9561	SK Ansbach	41.662	417	4,2	1067,8	97,6	11

Kreiskennziffern zwar auch eine Zahl ergibt (105.256), diese jedoch nicht sinnvoll interpretiert werden kann.

Diese Problematik wird in der Statistik unter dem Begriff „Messniveau“ oder auch „Skalenniveau“ von Daten erfasst. Dabei werden vier Messniveaus unterschieden:

Nominalskalen Hier drückt die Zahl lediglich eine Identität aus, ähnlich wie der Name. Ein Beispiel aus der Tabelle ist die Kreiskennziffer. So sagt die Kreiskennziffer 9564

lediglich aus, dass die zugehörigen Informationen zur Stadt Nürnberg gehören, sie sagt z. B. nichts aus über die Größe von Nürnberg. So haben sowohl der Vorläufer Stadtkreis Fürth als auch der Nachfolger Stadtkreis Schwabach weniger Einwohner als Nürnberg.

Weitere Beispiele für nominalskalierte Daten sind Kundennummern, Autokennzeichen oder Kontonummern.

Rechenoperationen sind mit normalskalierten Daten nicht zulässig, da sie – wie oben gezeigt – zu keinen interpretierbaren Ergebnissen führen. Dennoch sind für die Marktforschung nominalskalierte Daten sehr wichtig. Sie dienen oft zur Bildung von Gruppen von Merkmalsträgern, die dann bezüglich der Häufigkeit ihres Auftretens oder bezüglich der Ausprägungen ihrer anderen Merkmale verglichen werden. So werden Abverkäufe oft nach Geschäftstypen oder Regionen und Einkäufe von Personen nach Geschlecht oder Berufsgruppe dargestellt.

Ordinalskalen Hier drückt die Zahl einen Platz in einer Rangfolge aus. In der obigen Tabelle ist dies beim „Rang Kaufkraftkennziffer“ der Fall. So bedeutet die Zahl „1“ beim Kreis 9572, dass dieser die höchste Kaufkraft pro Einwohner verfügt. Die Rangordnung sagt jedoch nichts über Unterschiede aus. Z.B. ist die Differenz der Kaufkraft pro Einwohner zwischen den Nummern 1 und 2 deutlich kleiner als die entsprechende Differenz zwischen den Nummern 2 und 3. Deswegen sind auch bei ordinalskalierten Daten keine Rechenoperationen zulässig.

Weitere Beispiele für ordinalskalierte Daten sind alle auf Rangordnungen basierende Daten wie z. B. Hitlisten von Produkten, Einkommens-, Umsatz- oder Gemeindegrößenklassen mit unterschiedlichen Klassenbreiten. Ähnlich wie die nominalskalierten Daten werden diese oft zur Gruppenbildung verwendet und die Gruppen bezüglich der Häufigkeit ihres Auftretens und ihrer anderen Merkmale miteinander verglichen.

Intervallskalen: Diese Skalen liegen dann vor, wenn der Nullpunkt einer Skala mehr oder weniger willkürlich festgelegt ist. Beispiele sind Temperaturen in Grad Celsius oder Kalenderjahre.

Bei intervallskalierten Daten lassen sich die Unterschiede interpretieren, nicht jedoch die Verhältnisse. Die Temperaturdifferenz zwischen 5 und 10 Grad Celsius ist mit 5 Grad die gleiche wie die zwischen 15 und 20 Grad Celsius. Es macht aber keinen Sinn zu sagen, dass es bei 20 Grad Celsius viermal so warm ist wie bei 5 Grad Celsius.

Berechnungen sind also mit Einschränkungen möglich. Die Werte dürfen subtrahiert werden, die Werte dürfen jedoch nicht durch einander dividiert werden. Diese Einschränkung hat jedoch nur geringe Bedeutung. In der Praxis lassen sich alle wichtigen uni- und multivariate Verfahren wie z. B. Mittelwert, Streuung, Korrelation, Regression, Faktorenanalyse oder Clusteranalyse anwenden.

	Gleichheit / Ungleichheit	Ordnung	Abstand, Differenz	Verhältnis
	a = b	a > b	a - b	a / b
Nominalskala	+	-	-	-
Ordinalskala	+	+	-	-
Intervallskala	+	+	+	-
Verhältnisskala	+	+	+	+

+': Aussage möglich, '-': Aussage nicht möglich

Abb. 7.1 Analysemöglichkeiten der Skalenniveaus

Ratioskalen oder Verhältnisskalen: Dies ist glücklicherweise das in der Praxis am häufigsten anzutreffende Messniveau. In der obigen Tabelle sind alle Merkmale von „Einwohner" bis „Kaufkraft je Einwohner Index" verhältnisskaliert. Weitere in der Marktforschung häufig vorkommende verhältnisskalierte Daten sind Umsätze, Preise, Abverkäufe, Zahl der Käufer eines Produkts oder der Nutzer eines Mediums wie die Leser einer Zeitschrift. Hier sind alle Rechenoperationen zulässig. So lassen sich bei den Einwohnerzahlen sowohl Differenzen (z. B. „SK Fürth hat 15.830 Einwohner mehr als SK Erlangen") als auch Verhältnisse (z. B. „Der LK Ansbach hat fast viereinhalb mal so viel Einwohner wie der SK Ansbach") sinnvoll interpretieren. Damit sind auch alle Berechnungen ohne Einschränkungen zulässig.

Die Abbildung (Abb. 7.1) fasst die Skalen und die auf sie basierenden möglichen Aussagen zusammen.

Es zeigt sich, dass die Aussagemöglichkeiten einer höheren Skala auch die Aussagemöglichkeiten der niedrigeren Skalen beinhalten. So ist z. B. bei den Temperaturen in Grad Celsius auch die Relation der Ordnung (z. B. „gestern war es wärmer als heute") als auch der Gleichheit/Ungleichheit enthalten (z. B. „vorgestern war es ebenso warm wie heute). Und bei den verhältnisskalierten Merkmalen sind alle Aussagen möglich, wie das folgende Beispiel bezüglich der Einwohnerzahlen zeigt: „Die Einwohnerzahlen von Fürth und Nürnberg unterscheiden sich" (Gleichheit/Ungleichheit), „Nürnberg hat mehr Einwohner als Fürth" (Ordnung), „Nürnberg hat 381.510 mehr Einwohner als Fürth" (Abstand/Differenz) und eben auch „Nürnberg hat fast viermal so viel Einwohner wie Fürth" (Verhältnis).

In den meisten Fällen ist eine Zuordnung eines Merkmals zu einem Messniveau einfach möglich. Es gibt jedoch eine sehr wichtige Ausnahme davon, nämlich die in Abschn. 4.3.2 behandelten und in der Marktforschung sehr wichtigen **Ratingskalen**. Dabei werden die Befragten gebeten, zu einem Statement den Grad ihrer Zustimmung anzugeben, wobei z. B. die folgenden Antwortmöglichkeiten bestehen:

- Stimme voll und ganz zu
- Stimme eher zu
- Stimme teil zu, teils nicht zu

- Stimme eher nicht zu
- Stimme voll und ganz nicht zu

Streng genommen kann man nicht beweisen, dass der Abstand zwischen „Stimme voll und ganz zu“ und „Stimme eher zu“ gleich groß ist wie der Abstand zwischen „Stimme eher zu“ und „Stimme teils zu/teils nicht zu“. Demzufolge ist von einer Ordinalskala auszugehen. Weil Ratingskalen jedoch sehr häufig eingesetzt werden und weil z. B. alle Multi-Item-Skalen (vgl. Abschn. 4.3.2) eine Form der Verrechnung von verschiedenen Skalen zu einem Wert erfordern, wäre dies für die Marktforschung sehr misslich.

Was in Abschn. 4.3.2 für die Likert-Skala – die wichtigste Form der Ratingskalen – gesagt wurde, nämlich dass allgemein davon ausgegangen wird, dass diese hinreichend gut den Anforderungen einer Intervallskalierung entsprechen, kann für die Ratingskala insgesamt gesagt werden. Mayer (2008, S. 83) schreibt hierzu: „Genau genommen liefern Rating-Skalen lediglich ordinale Daten. Bei einer genügend großen Anzahl von Ausprägungen, kann jedoch angenommen werden, dass die Anstände auf der Skala von den Befragten als gleich Abstände aufgefasst werden.“ Er empfiehlt wie üblich 5 bis 7 Abstufungen einzusetzen. Auch Döring und Bortz (2016) kommen zu dieser Einschätzung: „Ratingskalen („rating scales“) werden meist als Intervallskalen aufgefasst. Somit können mit den Daten dann z. B. sinnvoll interpretierbare Mittelwerte gebildet und die für intervallskalierte Daten vorgesehenen statistischen Verfahren verwendet werden.“

In der Praxis werden daher Ratingskalen wie intervallskalierte Daten behandelt, sofern die Beschriftung der Skalenpunkte und Gestaltung der Skala dieser Annahme nicht widersprechen. Dies rechtfertigt sich auch dadurch, dass die Ergebnisse uni- und multivariater Verfahren in aller Regel robust auf die Verletzung der Annahme der gleichen Abstände reagieren. Nachfolgend eine Gestaltungsmöglichkeiten für Ratingskalen mit entsprechenden Kommentaren (siehe dazu auch Lehmann et al., 1998, S. 243 ff.):

- **Anzahl der Skalenpunkte:** Hier ist zunächst zwischen gerader und ungerader Zahl von Skalenpunkten zu unterscheiden. Bei ungerader Zahl bietet man meist den Auskunftspersonen ein mittlere bzw. neutrale Antwortkategorie an, z. B. bei einer Frage nach der Zufriedenheit „sehr unzufrieden; eher unzufrieden, teils – teils, eher zufrieden, sehr zufrieden“. Eine gerade Anzahl verwendet man oft, um der Auskunftsperson das Ausweichen auf eine (bequeme) neutrale Angabe zu verwehren und diese dazu zu drängen, sich zu entscheiden. Eine immer wieder diskutierte Frage bezieht sich darauf, wie viele Antwortkategorien vorgesehen sind (z. B. 5 oder 7). Das Spektrum der Möglichkeiten erscheint zunächst als ziemlich unbegrenzt. Einer geringen Zahl (z. B. drei Skalenpunkte) werden zwar von einigen Autoren (z. B. Jacoby & Matell, 1971) zufriedenstellende Messeigenschaften attestiert, es gibt aber erhebliche Zweifel, ob eine so „grobe“ Skala einer Intervallskalierung hinreichend gut entspricht (Döring & Bortz, 2016, S. 249). Andererseits dürfte eine zu große Zahl von Skalenpunkten (>10) in den meisten Fällen einen erheblichen Teil der Auskunftspersonen insofern über-

fordern, als ein sehr differenziertes Urteil mit entsprechend vielen Abstufungen gefordert wird. Vor diesem Hintergrund liegen die gängigen (ungeraden) Anzahlen von Skalenpunkten meist bei 5, 7 oder 9. Damit sollte auch eine akzeptable Annäherung an eine Intervallskalierung möglich sein.

- **Sprachliche oder numerische Kennzeichnung von Skalenpunkten:** Häufig werden die einzelnen (z. B. 5 oder 7) Antwortmöglichkeiten bei einer Ratingskala durch entsprechende Begriffe bezeichnet (z. B. „sehr wichtig", „eher wichtig", „weder noch", „eher unwichtig", „sehr unwichtig"). Dabei kommt man allerdings bei einer größeren Zahl von Skalenpunkten (>9) oftmals an die Grenze, dass man keine angemessenen (und von den Auskunftspersonen entsprechend verstandenen) Begriffe findet, um solche relativ feinen Unterschiede zu verdeutlichen. Deutlich günstiger ist die Situation bei der Verwendung von Ziffern zur Kennzeichnung der Skalenpunkte. Dabei werden nur noch die Extrempunkte verbal charakterisiert und alle einzelnen Punkte mit Ziffern bezeichnet. Damit ist klar, dass die Abstände zwischen diesen Punkten genau gleich („äquidistant") sind; ein zentrales Merkmal der Intervallskalierung. Dagegen ist bei rein verbalen Bezeichnungen nicht so klar, ob diese wirklich äquidistant sind bzw. von den Auskunftspersonen so empfunden werden.
- **Graphische Gestaltung von Ratingskalen:** Auch graphische Hilfsmittel können dabei helfen, den Auskunftspersonen zu verdeutlichen, dass die Abstände zwischen den Antwortmöglichkeiten gleich sein sollen. Zum einen sollte das durch gleiche räumliche Abstände zwischen den Antwortmöglichkeiten verdeutlicht werden. Daneben kann man auch die Größe bzw. die Flächen für die Angaben entsprechend variieren (z. B. kleine oder größere „Kästchen" für niedrige bzw. höhere Messwerte).

7.3 Verdichtung von Daten

7.3.1 Tabellierung und graphische Darstellung von Daten

Eine erste und gleichzeitig die wichtigste Datenverdichtung in der Marktforschung ist die Tabelle. Bei der Untersuchung nur eines Merkmals ergibt sich eine Häufigkeitstabelle Tab. 7.2 zeigt alle in der Praxis vorkommende Arten von Häufigkeiten. Dabei werden die Häufigkeiten oder Fallzahlen auch als absolute Häufigkeit, die Prozentwerte

Tab. 7.2 Einkommensverteilung monatliches Haushaltsnettoeinkommen in Euro 2018. (Quelle: Statist. Bundesamt 2021, S. 207)

	Anteile in %	Kumulierte %
bis unter 1300	13,3	13,3
1300 b. u. 2600	29,7	43,0
2600 b. u. 3600	17,8	60,8
3600 b. u. 5000	16,9	77,7
5000 und mehr	22,2	100,0

auch als relative Häufigkeit und die kumulierten Prozente auch als kumulierte relative Häufigkeit bezeichnet. Absolute und relative Häufigkeit benötigen nur nominalskalierte Daten, kumulierte Häufigkeiten und kumulierte relative Häufigkeiten benötigen mindestens ordinalskalierte Daten.

In Tab. 7.2 werden die Häufigkeiten einer stetigen Variablen dargestellt. Die absoluten Zahlen der Haushalte (die sogen. **absoluten Häufigkeiten**) wurden nicht aufgeführt, da es hier um die Struktur geht. Die Anteile in % werden auch als **relative Häufigkeiten** bezeichnet, die kumulierten % auch als **relative kumulierte Häufigkeiten.**

Das Einkommen ist klassifiziert dargestellt. Diese Klassen müssen vollständig sein und dürfen sich nicht überlappen. Die Ausdrucksweise „bis unter" wird dem gerecht. Dagegen ist die Klassifizierung „bis 1300" und „1301 bis 2600" zwar häufig angewandt, aber streng genommen nicht zulässig, da ein Haushaltsnettoeinkommen von 1300,50 € dann nicht eingeordnet werden kann.

Zur grafischen Darstellung nur eines Merkmals eignen sich **Säulendiagramme** oder **Kreisdiagramme** (vgl. Abb. 7.1.) Dabei ist ein Kreisdiagramm zu bevorzugen, wenn die Aufteilung eines Gesamtwerts dargestellt werden soll. Soll dagegen auf die Unterschiede in den Klassen abgestellt werden, so ist ein Säulendiagramm besser.

Wird ein Kreisdiagramm eingesetzt, so sollte auf eine 3-dimensionale Darstellung verzichtet werden, da diese zu einer verzerrten Wahrnehmung führt. In Abb. 7.2c entsteht der Eindruck, dass es mehr Haushalte mit einem Nettoeinkommen von 2600 bis unter 3600 € gibt als solche mit mehr als 5000 €, obwohl das Gegenteil richtig ist. Der Grund ist, dass bei der vorne stehenden Einkommensklasse die untere Fläche, welche die 3. Dimension darstellt, zu der Kreisfläche unbewusst addiert wird.

In Tab. 7.3 werden zwei Variable gleichzeitig dargestellt. Dass die privaten Konsumausgaben in Euro dargestellt sind, die Verwendungszwecke aber in Prozent dargestellt werden, wird nicht explizit genannt, sondern muss aus der Tabelle erschlossen werden. Als grafische Darstellung der Struktur Verwendungszwecke eignet sich ein **gestapeltes Säulendiagramm** (siehe Abb. 7.2).

Die Darstellung (Abb. 7.3) lässt unmittelbar erkennen, dass mit höherem Einkommen die Ausgabenanteile für die Grundbedürfnisse Wohnung und Nahrung zurückgehen, während die Ausgabenanteile für Freizeit, Verkehr, Bekleidung und Sonstige (u. a. Innenausstattung, Haushaltsgeräte, Gaststätten- und Beherbergungsdienstleistungen, Gesundheit, Post und Telekommunikation und Bildungswesen) zunehmen.

Bei der Darstellung von zwei Merkmalen können auch Streudiagramme sehr sinnvolle Darstellungsweisen sein. Als Beispiel wird auf die Tab. 7.1 der Regionaldaten zurückgegriffen und der Zusammenhang zwischen der Bevölkerungsdichte und der Arbeitslosenquote dargestellt (Abb. 7.4). Hier ist ein sehr enger linearer Zusammenhang erkennbar (Korrelationskoeffizient $r = 0{,}92$, vgl. Abschn. 7.3.2), was jedoch keine Aussage über eine Ursachen-Wirkungsbeziehung zulässt.

Zur Analyse des Zusammenhangs von zwei Variablen wird beim Vorliegen von Daten auf niedrigem Messniveau in erster Linie die **Kreuz**- bzw. **Kontingenztabelle** angewandt. Darauf wird im Kap. 8 im Zusammenhang mit Schlüssen von Stichproben-

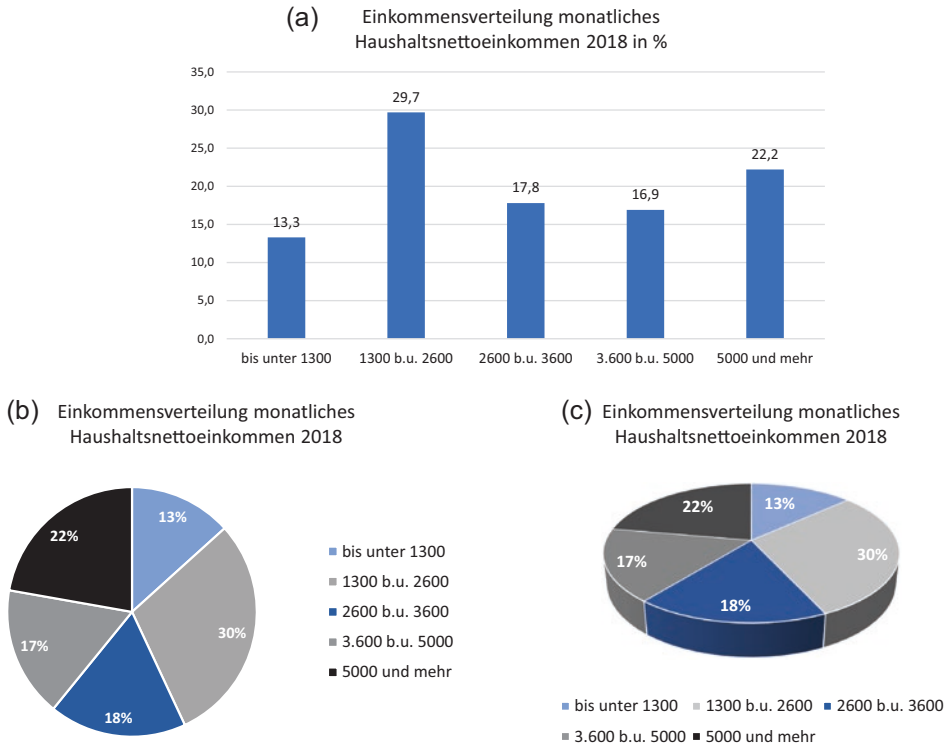


Abb. 7.2 Darstellung der Anteile von Tab. 7.2 als Säulendiagramm (**a**), Kreisdiagramm (**b**), dreidimensionales Kreisdiagramm (**c**). (Quelle der Daten: Statist. Bundesamt 2021, S. 207, eigene Darstellung)

Tab. 7.3 Konsumausgaben privater Haushalte nach dem monatlichen Haushaltsnettoeinkommen 2018. (Quelle: Statist. Bundesamt 2021, S. 212)

	Monatliches Haushaltsnettoeinkommen in Euro				
	Unter 1300	1300–2600	2600–3600	3600–5000	5000 und mehr
Private Konsumausgaben	1059	1761	2551	3253	4657
Wohnen, Energie, Wohnungsinstandhaltung	44,5	38,5	35,4	33,0	29,1
Nahrungsmittel, Getränke, Tabakwaren	17,4	15,1	13,9	13,3	11,6
Verkehr	8,2	11,0	13,2	14,6	16,4
Freizeit, Unterhaltung, Kultur	8,2	10,5	11,2	11,4	12,0
Bekleidung, Schuhe	3,5	4,2	4,3	4,5	4,9
Sonstige	18,2	20,8	22,1	23,0	26,0

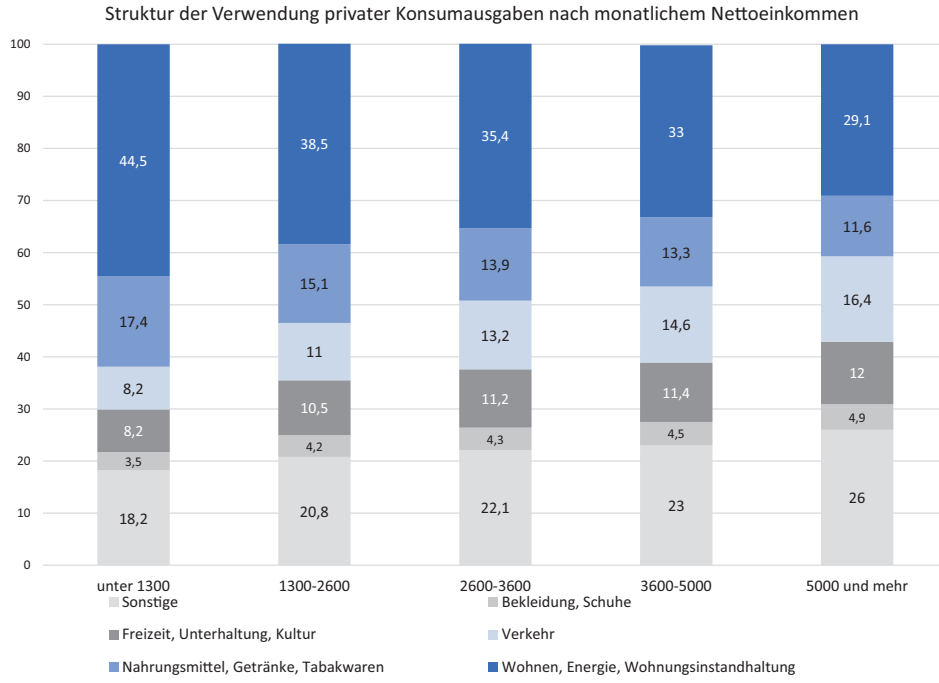


Abb. 7.3 Darstellung der Struktur der Konsumausgaben aus Tab. 7.3. (Quelle der Daten: Statist. Bundesamt 2021, S. 212, eigene Darstellung)

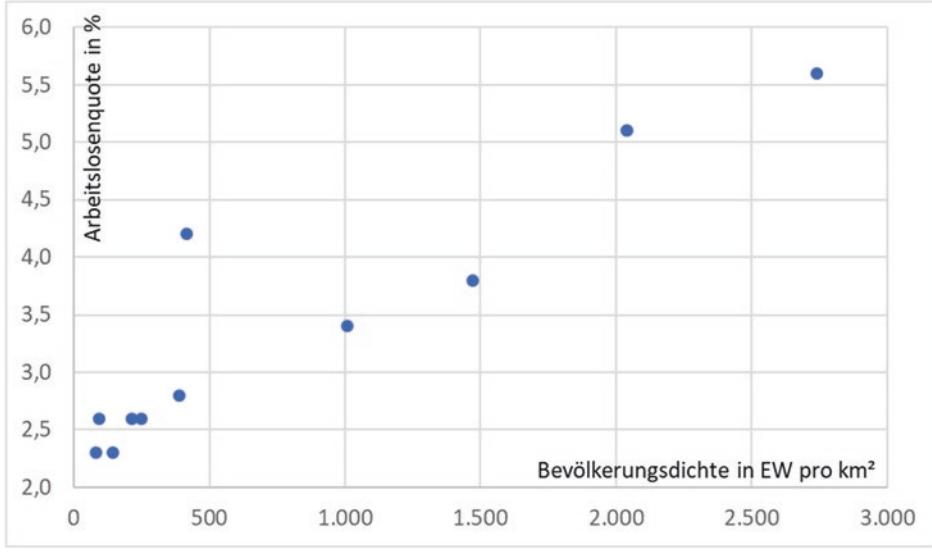


Abb. 7.4 Streudiagramm zu Bevölkerungsdichte und Arbeitslosenquote aus Tab. 7.1

ergebnissen auf die entsprechenden Grundgesamtheiten, im Rahmen des χ^2-Tests (oder auch Chi2-Test), nochmals eingegangen. Die zentrale Idee bei der Interpretation von Kontingenztabellen besteht im *Vergleich von Verteilungen*. Dazu ist zunächst (substanzwissenschaftlich) die Frage zu klären, welche Variable unabhängig und welche abhängig ist.

Wenn man bei den verschiedenen Ausprägungen der unabhängigen Variablen jeweils die Verteilungen der abhängigen Variablen betrachtet und diese sich beim Vergleich nicht (wesentlich) unterscheiden, dann spricht nichts dafür, dass die unabhängige Variable den vermuteten Einfluss auf die abhängige Variable hat. Ergeben sich aber bei den verschiedenen Ausprägungen der unabhängigen Variablen deutliche Unterschiede bei den Verteilungen der abhängigen Variablen, dann spricht das für einen Einfluss der unabhängigen Variable.

Beispiel Kontingenztabelle (Abb. 7.5)

Den (hypothetischen) Daten in der Tabelle liegt die Vermutung zugrunde, dass in unterschiedlichen Altersgruppen der Konsum von Cola verschieden ist, dass also das Alter den Cola-Konsum beeinflusst. Die umgekehrte Vermutung – Cola-Konsum beeinflusst das Alter – wäre völlig unsinnig. Zur Überprüfung der Vermutung muss man also die Verteilungen des Cola-Konsums in den verschiedenen Altersgruppen betrachten, also hier die Spalten der Tabelle miteinander vergleichen. Wäre der vermutete Zusammenhang zwischen Alter und Cola-Konsum nicht vorhanden, so müsste sich jede Alterskohorte so verhalten, wie sich das gesamte Sample verhält. Gemäß den Zeilensummen sollte sich also der Cola-Konsum in jeder Alterskohorte aufteilen in ca. 33,7 % „weniger als 0,5 L“ (138/410), ca. 37,3 % „0,5–1 L“ (153/410) und ca. 29 % „mehr als 1 L“ (119/410).

Alter der Auskunftsperson

Konsum von Cola-Getränken pro Woche	< 25 J.	25 – 40 J.	> 40 J.	
< 0,5 ltr.	24 20%	30 20%	84 60%	138
0,5 – 1 ltr.	36 30%	75 50%	42 30%	153
> 1 ltr.	60 50%	45 30%	14 10%	119
	120 100%	150 100%	140 100%	410

Abb. 7.5 Beispiel einer Kontingenztabelle

An den entsprechenden Prozentzahlen wird jedoch sofort klar, dass hier deutliche Unterschiede vorliegen. So sieht man beispielsweise, dass bei der jüngsten Altersgruppe 50 % der Auskunftspersonen in der Gruppe mit dem höchsten Cola-Konsum liegen, während das in der Altersgruppe über 40 Jahre nur 10 % sind. Es bestätigt sich somit der vermutete Zusammenhang zwischen Alter und Cola-Konsum. ◀

7.3.2 Statistische Maßzahlen

Die am meisten gebrauchten Maßzahlen zur Charakterisierung von Häufigkeitsverteilungen sind die Lageparameter und die Streuungsmaße. **Lageparameter** sollen angeben, wo der „Schwerpunkt" einer Verteilung von Messwerten liegt; **Streuungsmaße** sollen die Homogenität bzw. Heterogenität der Messwerte wiedergeben.

Beim niedrigsten Messniveau (Nominalskalierung) ist die Angabe des **Modus** zur Beschreibung einer Verteilung üblich. Der Modus ist der Wert, der am häufigsten auftritt, wobei es Fälle geben, in denen der Modus nicht eindeutig festgelegt ist, da mehrere Messwerte die gleiche Häufigkeit haben. Wenn die Daten mindestens ordinalskaliert sind, ist die Verwendung des **Medians** zulässig. Der Median ist der Wert, der eine (nach Größe der Messwerte geordnete) Verteilung in zwei gleich große Teilmengen separiert. Zur Berechnung des Medians gibt es unterschiedliche Vorgehensweisen. Bei einer geraden Anzahl von Messwerten ist der Median das arithmetische Mittel der beiden in der Mitte der Verteilung liegenden Werte. Bei ungerader Anzahl von Messwerten ist der Median der in der Mitte liegende Wert.

Bei intervall- oder ratioskalierten Daten kann das **arithmetische Mittel** als besonders gängige Maßzahl berechnet werden. Es ergibt sich durch

$$\bar{x} = \frac{\sum x_i}{n}$$

mit

$\bar{x}$	arithmetisches Mittel
x_i	Messwerte
n	Zahl der Messwerte

Das arithmetische Mittel ist gegenüber Ausreißern (weit außerhalb des sonstigen Wertebereichs liegenden Messwerten) sehr empfindlich. Deswegen wird oft empfohlen, beim Auftreten von **Ausreißern** eher den Median als Lageparameter zu verwenden, der von extremen Werten nicht beeinflusst wird.

Hintergrundinformation

Bei der Datenanalyse können **Ausreißer** erhebliche Probleme verursachen und die Ergebnisse einer Analyse völlig verfälschen. So kann man sich leicht vorstellen, dass eine Analyse zur sozia-

len und ökonomischen Lage einer kleinen und armen Landgemeinde grundlegend verändert wird, wenn dort ein einzelner Multimillionär seinen Wohnsitz hat und dessen Einkommens- und Vermögensdaten in die örtliche Statistik einfließen. Beispielsweise würde dadurch das örtliche Durchschnittseinkommen sprunghaft ansteigen, indem aus armen Kleinbauern im Durchschnitt bzw. auf dem Papier wohlhabende Landbewohner werden. Die verzerrende Wirkung von Ausreißern wird noch größer, wenn für die Berechnung der statistischen Maßzahlen eine Quadrierung der Messwerte vorgenommen wird (z. B. bei der Berechnung der Varianz, s. u.). Was also tun mit Ausreißern? Eine erste Empfehlung besteht darin, den Datensatz dahingehend zu kontrollieren, ob ein Tippfehler vorliegt. Allzu leicht wird bei der Dateneingabe aus einem Monatseinkommen von 3000,- € eines von 30.000,- €. Eine andere – nur mit großer Vorsicht zu handhabende – Option kann es sein, den betreffenden Fall (als besonders atypisch) aus dem Datensatz zu eliminieren. Die Vorsicht ist deshalb geboten, weil diese Verfahrensweise natürlich mit der Gefahr verbunden ist, dass der Untersuchungsleiter zu großzügig Fälle eliminiert, die nicht mit seinen Vermutungen oder den Erwartungen des Auftraggebers verträglich sind.

Weil die alleinige Angabe von Lageparametern für die Charakterisierung einer Häufigkeitsverteilung nicht ausreicht, wird zusätzlich meist eine Angabe über die Streuung der Messwerte gemacht wird. Das einfachste Streuungsmaß ist die **Spannweite**, die als die Differenz zwischen dem größten und dem kleinsten Messwert definiert ist. Daraus ergibt sich schon, dass Intervallskalierung die Anwendungsvoraussetzung dafür ist. Die Spannweite ist ein recht grobes Streuungsmaß, das außerdem auch sehr empfindlich gegenüber Ausreißern ist.

Ein weiteres Streuungsmaß ist schon aus dem Abschn. 7.3.1 bekannt, die **interquartile Distanz.** Sie gibt an, über welchen Wertebereich die „mittleren 50 %“ der Messwerte verteilt sind. Die Berechnung der zur Bestimmung der interquartilen Distanz notwendigen oberen und unteren Quartile (75- bzw. 25-%-Punkt) vollzieht sich analog zu der des Medians (50-%-Punkt).

Die weitaus gebräuchlichsten Streuungsmaße sind die **Varianz** und die Standardabweichung, bei denen mindestens Intervallskalierung der Daten vorausgesetzt wird. Die Varianz einer Stichprobe definiert als

$$S_x^2 = \frac{\sum (x_i - \bar{x})^2}{\mathrm{n} - 1}$$

mit

s_x^2	Varianz von x
x_i	Messwerte
$\bar{x}$	arithmetisches Mittel der Variablen X
n	Zahl der Messwerte/Fälle

Die Varianz ist also als Mittelwert der quadrierten Abweichungen zwischen den einzelnen Messwerten und dem arithmetischen Mittel interpretierbar. Je weiter Messwerte vom arithmetischen Mittel abweichen, je heterogener also die Verteilung ist, desto größer ist

die Varianz. Dabei ist zu beachten, dass die Varianz in einer anderen Größenordnung liegt als die Ausgangswerte, da sie auf der Basis der quadrierten Abweichungen errechnet wird.

Bei der **Standardabweichung** liegen die Werte dagegen in der Größenordnung der Ausgangsvariablen. Sie ist definiert durch

$$S_x = \sqrt{S_x^2}$$

mit

s_x	Standardabweichung von x
s_x^2	Varianz

Varianz und Standardabweichung sind in besonderem Maße *empfindlich gegenüber Ausreißern* (extrem vom üblichen Wertebereich abweichenden Werten), da diese durch die Quadrierung der Abweichungen vom Mittelwert die resultierenden Maßzahlen stark beeinflussen können.

Im **bivariaten Fall**, wenn es also gilt, den Zusammenhang zwischen zwei Variablen zu beschreiben, wird oftmals als Maßzahl der **Korrelationskoeffizient** r (nach Pearson) verwendet. Dessen Anwendung ist an zwei Voraussetzungen geknüpft:

- beide Variablen müssen mindestens intervallskaliert sein;
- der Zusammenhang zwischen den Variablen muss linear sein.

Die Idee, die dem Korrelationskoeffizienten zugrunde liegt, kann in einigen Schritten leicht verdeutlicht werden. Ausgangspunkt ist der in Abb. 7.6 graphisch dargestellte Zusammenhang zwischen den Variablen X und Y.

Eine Maßzahl zur Beschreibung des Zusammenhangs zwischen x und y könnte vielleicht folgende Form haben:

$$M = \sum (x_i - \bar{x})(y_i - \bar{y})$$

mit

M	Maßzahl für den linearen Zusammenhang
x_i, y_i	Messwerte
$\bar{x}$	arithmetisches Mittel von x
$\bar{y}$	arithmetisches Mittel von y

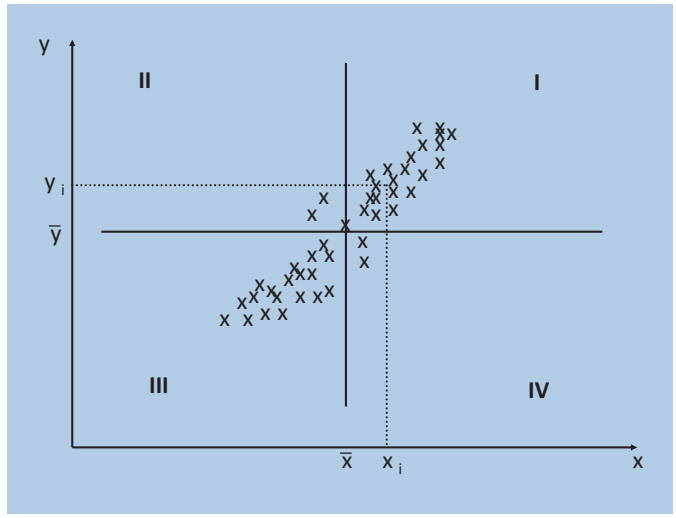


Abb. 7.6 Grafische Darstellung des Zusammenhangs zwischen zwei Variablen X und Y

Wie man leicht sieht, ist das Produkt der Abweichungen vom Mittelwert positiv bei allen Messwerten, die in den Quadranten I und III liegen, und negativ bei den Messwerten, die in den Quadranten II und IV liegen. Für die Maßzahl M ergeben sich daraus folgende Konstellationen:

- Wenn ein positiver linearer Zusammenhang zwischen X und Y vorliegt (wie in Abb. 7.6), dann liegen die meisten Messwerte in den Quadranten I und III und M wird positiv.
- Wenn ein negativer linearer Zusammenhang zwischen X und Y vorliegt, dann befinden sich die meisten Messwerte in den Quadranten II und IV und M wird negativ.
- Wenn kaum ein linearer Zusammenhang zwischen X und Y erkennbar ist, die Messwerte also über alle vier Quadranten relativ gleichmäßig verteilt sind, dann gleichen sich positive und negative Werte von $(x_i - \bar{x})(y_i - \bar{y})$ weitgehend aus und M wird sehr klein.

Die Maßzahl M hat aber noch zwei Nachteile: Durch Hinzufügung von Messwerten (Vergrößerung der Stichprobe) wächst die Maßzahl, obwohl sich an Art und Stärke des Zusammenhangs zwischen den Variablen nichts ändern muss. Außerdem ist die Maßzahl von den gewählten Maßeinheiten abhängig. Wenn man z. B. bei der (mehr oder weniger sinnvollen) Analyse des Zusammenhangs zwischen Körpergröße und Gewicht von Menschen an Stelle der Maßeinheiten „Meter“ und „Kilogramm“ die Maßeinheiten „Zentimeter“ und „Gramm“ verwendete, wären die in die Berechnung von M einfließenden

Zahlen natürlich viel größer und M würde entsprechend wachsen, ohne dass sich am Zusammenhang zwischen Körpergröße und Gewicht etwas verändert hat. Deswegen werden diese beiden störenden Effekte dadurch eliminiert, dass man die gesuchte Maßzahl hinsichtlich der Zahl der Fälle (n) und der Maßeinheiten (hier in Form der Standardabweichungen) normiert. Das Ergebnis ist der **Korrelationskoeffizient r** nach Pearson:

$$r = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

mit

r	Korrelationskoeffizient
x_i, y_i	Messwerte
$\bar{x}$	arithmetisches Mittel von x
$\bar{y}$	arithmetisches Mittel von y
n	Zahl der Fälle
s_x	Standardabweichung von x
s_y	Standardabweichung von y

Abb. 7.7 illustriert, welche Werte r bei unterschiedlichen Konstellationen annehmen kann. Der Korrelationskoeffizient r hat die Werte

- +1 wenn alle Messwerte auf einer Geraden mit positiver Steigung liegen (Abb. 7.7c),
- −1 wenn alle Messwerte auf einer Geraden mit negativer Steigung liegen,
- 0 wenn keinerlei linearer Zusammenhang zwischen den Variablen erkennbar ist (Abb. 7.7d).

Es sei allerdings darauf hingewiesen, dass Teil d der Abbildung zu beachten, bei dem erkennbar ist, dass ein klar erkennbarer *nichtlinearer* Zusammenhang zwischen den Variablen existiert und sich für den Korrelationskoeffizienten r = 0 ergibt, was (wenn man Linearität unterstellt) für keinerlei Zusammenhang stünde. Solche Zusammenhänge gibt es sehr oft. Man denke nur an den Zusammenhang zwischen dem Wohlbefinden von Studenten in einem Hörsaal (auf der y-Achse aufgetragen) und der Temperatur in diesem Hörsaal (auf der x-Achse) oder der Geschmacksqualität einer Suppe auf der y-Achse und der Menge des zugegebenen Salzes auf der x-Achse.

Gängig im Bereich der Markt- und Sozialforschung sind am ehesten Korrelationen, wie sie in Abb. 7.7b ausgewiesen sind. Hier erkennt man, dass ein deutlicher Zusammenhang zwischen beiden Variablen existiert (r ist deutlich von Null verschieden), dass aber nicht die eine Variable die andere weitgehend bestimmt.

Bei der Interpretation eines Korrelationskoeffizienten gibt es zwei verschiedene Sichtweisen. Die eine ist darauf fokussiert, dass man sich auf starke Korrelationen (Koeffizienten nahe +1 oder −1) konzentriert (Abb. 7.7a). In solchen Fällen könnte man jeweils

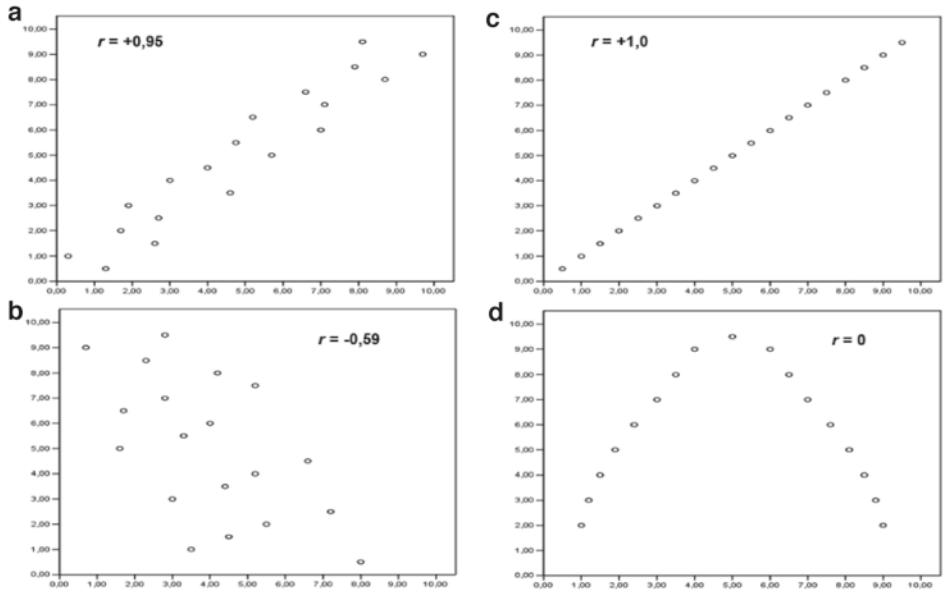


Abb. 7.7 Beispiele für Daten und Korrelationskoeffizienten

eine Variable durch die andere weitgehend erklären. Solche hohen Korrelationen treten im Bereich der Markt- und Sozialforschung nur selten auf. Hier geht es eher um Zusammenhänge zwischen den Variablen, die mehr oder weniger deutlich sind (s. o.), bei denen aber nicht die eine Variable durch die andere Variable weitgehend bestimmt ist.

Die Frage ist also, ob ein Korrelationskoeffizient *„signifikant" von Null verschieden* ist, d. h., dass er so stark (positiv oder negativ) von Null abweicht, dass diese Abweichung nicht mehr durch Zufälligkeiten bei der Stichprobenziehung oder Messfehler zu erklären ist, sondern auf einen systematischen (aber nicht deterministischen!) Zusammenhang zwischen den Variablen hinweist (siehe hierzu auch die Ausführungen in Kap. 8). Dafür gibt es auch geeignete statistische Tests, deren ausführliche Darstellung aber den Rahmen dieses einführenden Lehrbuchs sprengen würde. Meist werden von Programmen zur statistischen Datenanalyse (z. B. SPSS) Ergebnisse solcher Tests bei der Berechnung von Korrelationskoeffizienten automatisch ausgewiesen.

Der hier erläuterte Korrelationskoeffizient r nach Pearson ist wohl die gängigste Maßzahl zur Bestimmung des Zusammenhangs zwischen zwei (intervallskalierten) Variablen. Gleichwohl existiert natürlich eine große Zahl entsprechender Maßgrößen für die unterschiedlichsten Kombinationen von Messniveaus der „beteiligten" Variablen. Als Beispiele seien genannt:

- **Lambda** für zwei nominalskalierte Variable
- **Kendall's tau** für zwei ordinalskalierte Variable
- **Spearman's Korrelationskoeffizient** für zwei ordinalskalierte Variable
- **Eta** für eine Kombination von nominal- und intervallskalierten Variablen

Tab. 7.4 Möglichkeiten der Verdichtung von Daten in Abhängigkeit vom Messniveau

	Messniveau der Daten			
	Nominalskala	Ordinalskala	Intervallskala	Ratioskala
Tabellierung und Grafiken				
Häufigkeitstabelle	X	X	(X)	(X)
Kreisdiagramm	X	X	(X)	(X)
Histogramm	(X)	(X)	X	X
Stem-and-Leaf-Plot			X	X
Boxplot			X	X
Kreuz- bzw. Kontingenztabelle	X	X	(X)	(X)
Lageparameter				
Modus	X	X	X	X
Median		X	X	X
Arithmetisches Mittel			X	X
Streuungsmaße				
Spannweite			X	X
Interquartile Distanz			X	X
Varianz			X	X
Standardabweichung			X	X
Zusammenhangsmaße				
Korrelationskoeffizient r			X	X
Kovarianz			X	X
Lambda	X			
Kendall's tau		X		
Spearman's Korrelationskoeffizient		X		
Eta	(X)		(X)	

Zu Einzelheiten muss auf die Literatur zur deskriptiven Statistik verwiesen werden, z. B. de Vaus (2002, S. 274 ff.) oder Jaccard und Becker (2002).

Eine weitere Maßzahl, die bei manchen komplexen statistischen Analysen (siehe Abschn. 9.5 und 9.6) eine Rolle spielt, die sogenannte **Kovarianz**, verzichtet im Gegensatz zum Korrelationskoeffizienten nach Pearson auf die Normierung durch die Standardabweichung. Für eine Datenreihe von zwei Variablen (x und y) ist korrigierte Stichprobenvarianz folgendermaßen definiert:

$$\widehat{\sigma}_{xy} = \frac{1}{n-1} \sum (x_i - \overline{x})(y_i - \overline{y})$$

Tab. 7.4 gibt eine Übersicht über die in diesem Kapitel dargestellten Möglichkeiten der Verdichtung von Daten in Abhängigkeit vom jeweiligen Messniveau der Daten. Die

Klammern (X) bedeuten dabei, dass eine Darstellung zwar möglich ist, aber beim angegeben Messniveau wenig sinnvoll, da z. B. eine Aggregation von einem hohen (z. B. intervallskalierten) auf ein niedrigeres (z. B. ordinalskaliertes) Messniveau nur durch entsprechende Klassenbildung und somit Informationsverlust erreicht werden kann. Beim Koeffizienten Eta verweisen die Klammern darauf, dass der Koeffizient für eine Kombination aus nominal- und intervallskalierten Variablen berechnet wird.

Literatur

De Vaus, D. (2002). *Analyzing social science data*. Sage.

Döring, N., & Bortz, J. (2016). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (5. Aufl.). Springer-VS.

Jaccard, J., & Becker, M. (2002). *Statistics for the behavioral sciences* (4. Aufl.). Wadsworth.

Jacoby, J., & Matell, M. S. (1971). Three point Likert scales are good enough. *Journal of Marketing Research, 8*(4), 495–500.

Lehmann, D., Gupta, S., & Steckel, J. (1998). *Marketing research*. Addison Wesley Pub Co Inc.

Mayer, H. O. (2008). *Interview und schriftliche Befragung – Entwicklung- Durchführung – Auswertung"*. Oldenbourg Wissenschaftsverlag.

Statistisches Bundesamt. (Hrsg.). (2021). Einkommen und Konsum – Auszug dem Datenreport 2021, Wiesbaden 2021.

Schlüsse auf Grundgesamtheiten

8

Zusammenfassung

Für die Marktforschung ist die Verwendung von Stichproben ganz typisch. Beispielsweise soll auf Basis einer Befragung von wenigen hundert Konsumenten auf Präferenzen, Verhaltensweisen etc. aller Konsumenten geschlossen werden. Dazu bedarf es spezieller statistischer Schlussweisen, auf die im Kap. 8 eingegangen wird. Dabei geht es einerseits um Schätzverfahren, mit deren Hilfe man die Genauigkeit und Sicherheit von Ergebnissen berechnen kann. Mithilfe statistischer Tests trifft man Entscheidungen, insbesondere darüber, ob Zusammenhänge zwischen Variablen oder Unterschiede zwischen Gruppen durch Zufälligkeiten bei der Stichprobenziehung zu erklären oder systematisch begründet sind.

8.1 Schätzungen

Bei der Interpretation von Stichprobenergebnissen gibt es zwei typische Schlussweisen. Bei **Schätzungen** wird versucht, auf der Basis der sich in der Stichprobe ergebenden Werte (z. B. Mittelwerte, Anteilswerte) Aussagen über die entsprechenden Werte in der Grundgesamtheit sowie die Genauigkeit und Sicherheit dieser Schätzungen zu machen. Bei **statistischen Tests** trifft man Entscheidungen. In der Marktforschung sehr gängig sind Entscheidungen hinsichtlich der Annahme bzw. Ablehnung von Hypothesen über Zusammenhänge zwischen Merkmalen (z. B. Einstellung und Kaufabsicht) und über Unterschiede zwischen Gruppen (z. B. Markenpräferenzen bei Männern und Frauen). Nachdem in Abschn. 4.2 die Bedeutung der Stichprobenziehung für die Generalisierbarkeit von Marktforschungsergebnissen bereits diskutiert wurde, geht es hier nun darum zu überprüfen, ob sich ein in einer Stichprobe gefundenes Ergebnis wie z. B. Unterschiede in den Einstellungen von Männern und Frauen auch auf die Grundgesamtheit übertrag-

H. Kreis et al., *Marktforschung,* https://doi.org/10.1007/978-3-658-44456-3_8

bar sind (vgl. Abschn. 2.3). Im vorliegenden Abschnitt soll das Grundprinzip von Schätzungen anhand eines Beispiels der Schätzung des Mittelwerts, das – angeregt durch Jaccard und Becker (2002, S. 182 ff.) – für die Zwecke dieses Buches entwickelt wurde, erläutert werden. Abb. 8.1 bezieht sich auf Schätzungen bei einfachen Zufallsstichproben.

Im Beispiel wird von einer Grundgesamtheit ausgegangen, deren Messwerte in Abb. 8.1 eingetragen sind. Diese Situation – bekannte Grundgesamtheit – ist natürlich für die Forschungspraxis völlig atypisch, da man ja dort Stichproben gerade zieht, um Aufschluss über *unbekannte* Grundgesamtheiten zu bekommen. Die gedanklichen Schritte bei der Entwicklung von Schätzverfahren lassen sich aber bei der Unterstellung einer bekannten Grundgesamtheit leichter nachvollziehen. In dem hypothetischen Beispiel geht es darum, auf der Basis einer Stichprobe der Größe $n = 10$ Aussagen über die gekauften Packungen eines Produkts in der Grundgesamtheit von 100 Haushalten zu machen.

Die Kaufmengen aller 100 Haushalte sind also in Abb. 8.1 eingetragen. Dort findet sich auch der („wahre“) arithmetische Mittelwert der Grundgesamtheit $\mu = 4{,}3$ und die Varianz für diese Grundgesamtheit $\sigma^2 = 8{,}5$. Kennwerte für die Grundgesamtheit werden mit griechischen Buchstaben gekennzeichnet, während Kennwerte für die Stichprobe mit lateinischen Buchstaben bezeichnet werden. Die beiden Werte aus der Grundgesamtheit sind natürlich im Normalfall ebenfalls unbekannt und sollen in realen Untersuchungen mithilfe von Daten aus der Stichprobe geschätzt werden.

8	6	8	6	3	3	7	4	9	5
9	1	1	6	5	8	1	9	1	0
8	5	3	4	3	1	4	8	2	8
9	1	6	8	1	8	6	1	2	4
7	2	0	2	1	4	3	5	1	4
2	5	1	5	3	5	5	3	1	7
0	8	0	2	9	6	0	1	6	5
8	8	6	7	9	0	5	0	3	3
1	2	1	3	3	6	8	2	7	0
4	7	8	4	1	8	5	8	2	4
1	6	2	0	4	5	2	0	1	7
9	5	1	5	2	1	6	8	5	7
5	5	1	0	9	8	3	2	0	8
2	4	0	9	0	0	1	1	7	8
9	9	9	4	0	6	4	7	7	3
8	7	6	5	4	2	9	6	1	4

Mittelwert der Grundgesamtheit μ=4,3

Varianz der Grundgesamtheit σ²=8,5

Abb. 8.1 Anzahl gekaufter Packungen eines Produkts in einer Grundgesamtheit von 100 Haushalten

Tab. 8.1 Ergebnisse verschiedener Stichproben der Größe $n = 10$

Nr	Stichprobenwerte										$\bar{x}$	$\bar{x} - \mu$
1	8	6	8	6	3	3	7	4	9	5	5,9	1,6
2	9	1	1	6	5	8	1	9	1	0	4,1	−0,2
3	8	5	3	4	3	1	4	8	2	8	4,6	0,3
4	9	1	6	8	1	8	6	1	2	4	4,6	0,3
5	7	2	0	2	1	4	3	5	1	4	2,9	−1,4
6	2	5	1	5	3	5	5	3	1	7	3,7	−0,6
7	0	8	0	2	9	6	0	1	6	5	3,7	−0,6
8	8	8	6	7	9	0	5	0	3	3	4,9	0,6
9	1	2	1	3	3	6	8	2	7	0	3,3	−1,0
10	4	7	8	4	1	8	5	8	2	4	5,1	0,8
11	1	6	2	0	4	5	2	0	1	7	2,8	−1,5
12	9	5	1	5	2	1	6	8	5	7	4,9	0,6
13	5	5	1	0	9	8	3	2	0	8	4,1	−0,2
14	2	4	0	9	0	0	1	1	7	8	3,2	−1,1
15	9	9	9	4	0	6	4	7	7	3	5,8	1,5
16	8	7	6	5	4	2	9	6	1	4	5,2	0,9
												$\sum = 0$

Schätzungen des Mittelwerts

Wenn man beispielsweise die folgende Zufallsstichprobe (Nr. 1, Tab. 8.1) der Größe $n = 10$ mit den Messwerten der gekauften Menge 8, 6, 8, 6 3, 3, 7, 4, 9, 5 zieht, würde man dafür den Mittelwert 5,9 erhalten, einen Wert also, der von dem Mittelwert der Grundgesamtheit ($\mu = 4{,}3$) abweicht. Diese Abweichung ist auf den Stichprobenfehler zurückzuführen. Bei einer anderen Stichprobe (Nr. 11, Tab. 8.1) mit den Messwerten 1, 6, 2, 0, 4, 5, 2, 0, 1, 7 erhielte man einen Stichprobenmittelwert $\bar{x} = 2{,}8$. Der erste Schätzwert für den Mittelwert weicht also nach oben, der andere nach unten vom wahren Mittelwert der Grundgesamtheit ab. Das deutet auf eine nützliche Eigenschaft der Stichprobenmittelwerte hin: Diese schwanken um den Mittelwert der Grundgesamtheit. Wenn man aus der Grundgesamtheit alle möglichen Stichproben zieht, so ergibt sich dabei eine durchschnittliche Abweichung der verschiedenen Stichprobenmittelwerte vom Mittelwert der Grundgesamtheit von Null. Da das Ausmaß des Stichprobenfehlers im Durchschnitt gleich Null ist, spricht man davon, dass der Stichprobenmittelwert ein **unverzerrter Schätzwert** für den Mittelwert der Grundgesamtheit ist. Tab. 8.1 illustriert auch diesen Effekt.

Schätzung von Varianz und Standardabweichung

Wenn man versucht, mithilfe der erstgenannten Stichprobe (8, 6, 8, 6 3, 3, 7, 4, 9, 5) die Varianz zu schätzen, so erhält man mit $s^2 = 4{,}09$ einen Wert, der von der Varianz in der Grundgesamtheit (σ^2) deutlich abweicht. Die Varianz in der Stichprobe ist allerdings kein unverzerrter Schätzwert für die Varianz in der Grundgesamtheit. Dafür ist eine Korrektur gegenüber der üblichen Formel für die Varianz notwendig, die im folgenden Ausdruck für den Schätzwert der Varianz ($\hat{S}^2$) berücksichtigt ist:

$$\widehat{s^2} = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

Im obigen Beispiel würde sich also ergeben:

$$\widehat{s^2} = \frac{40{,}90}{9} = 4{,}54$$

Dieser Schätzwert weicht wegen des hier relativ großen Stichprobenfehlers immer noch erheblich von der Varianz in der Grundgesamtheit ab. Der Schätzwert für die Standardabweichung in der Grundgesamtheit ergibt sich dann durch:

$$\widehat{s} = \sqrt{\widehat{s^2}}$$

In dem Zahlenbeispiel erhält man dann

$$\widehat{s} = \sqrt{4{,}54} = 2{,}13$$

Verteilung des Stichprobenmittelwerts

Auf den vorigen Seiten ist (hoffentlich) deutlich geworden, dass sich bei mehreren unterschiedlichen Stichproben aus der gleichen Grundgesamtheit verschiedene Werte für den zu schätzenden Mittelwert der Grundgesamtheit ergeben können. Diese verschiedenen Schätzwerte schwanken um den „wahren" Wert. An dieser Stelle interessiert es besonders, wie diese Werte schwanken, wie also die Verteilung des Stichprobenmittelwertes (bei mehreren Stichproben aus der gleichen Grundgesamtheit) aussieht. Obwohl man in der Praxis kaum Verteilungen von mehreren Stichprobenmittelwerten betrachtet, da man ja in der Regel nur eine Stichprobe zieht, interessiert diese Art von Verteilungen, weil sie Aufschlüsse über die Fehler gibt, die man beim Schluss von Stichprobenergebnissen auf eine Grundgesamtheit macht.

Eine wichtige Eigenschaft der Verteilung des Stichprobenmittelwertes hat sich schon aus den Überlegungen im Zusammenhang mit der Tab. 8.1 angedeutet: Der Mittelwert der Verteilung der Stichprobenmittelwerte ist gleich dem (gesuchten) Mittelwert der Grundgesamtheit. Weiterhin ist die Standardabweichung der Verteilung der Stichprobenmittelwerte von Interesse. Die Standardabweichung ist ja allgemein als durchschnittliche Abweichung der einzelnen Messwerte vom Mittelwert einer Verteilung interpretierbar.

Im vorliegenden Fall (Verteilung des Stichprobenmittelwertes) ist die Standardabweichung ein Maß für die durchschnittliche Abweichung der Mittelwerte der einzelnen Stichproben vom Mittelwert der Grundgesamtheit. Man spricht deshalb auch vom

Standardfehler des Stichprobenmittelwerts. Wenn dieser Standardfehler klein ist, dann wird man erwarten können, dass ein einzelner Stichprobenmittelwert mit recht großer Wahrscheinlichkeit nur wenig vom eigentlich interessierenden Mittelwert der Grundgesamtheit abweicht. Wenn der Standardfehler dagegen groß ist, heißt das, dass man sehr wohl mit einer relativ großen Abweichung des Stichprobenmittelwertes vom Mittelwert der Grundgesamtheit rechnen muss. Deswegen ist es für Schlüsse von einer Stichprobe auf eine Grundgesamtheit bedeutsam, die Standardabweichung der Verteilung des Stichprobenmittelwertes zu kennen.

Wenn die Stichprobe im Verhältnis zur Grundgesamtheit klein ist (<5 %, was bei Marktforschungsuntersuchungen meist der Fall ist,) ist die Standardabweichung für die Verteilung des Stichprobenmittelwertes näherungsweise gegeben durch

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

mit

$\sigma_{\bar{x}}$:	Standardabweichung der Verteilung des Mittelwertes (Standardfehler)
σ:	Standardabweichung der Messwerte in der Grundgesamtheit
n:	Stichprobengröße

Die *Abhängigkeit dieser Standardabweichung* von der *Standardabweichung* der Messwerte und von der *Stichprobengröße* ist recht plausibel. Wenn die Standardabweichung der Messwerte klein ist, wenn also die Messwerte relativ homogen sind, dann dürften sich die Mittelwerte verschiedener Stichproben aus der gleichen Grundgesamtheit recht eng um den „wahren" Mittelwert gruppieren (und umgekehrt). Ein Wachstum der Stichprobengröße (n) führt zu einer Verkleinerung des Standardfehlers des Stichprobenmittelwertes, weil einzelne extreme Messwerte, die in einer Stichprobe auftreten können, den jeweiligen Stichprobenmittelwert weniger beeinflussen.

Normalerweise ist σ, die Standardabweichung der Messwerte in der Grundgesamtheit, unbekannt. Deswegen berechnet man einen Schätzwert für die Standardabweichung der Verteilung des Mittelwertes ($\widehat{s_{\bar{x}}}$), also den Standardfehler, indem man für σ den entsprechenden Schätzwert ŝ verwendet:

$$\widehat{s_{\bar{x}}} = \frac{\widehat{s}}{\sqrt{n}}$$

Konfidenzintervall

Nachdem jetzt einige Informationen über die Verteilung des Stichprobenmittelwertes bekannt sind, fragt man sich, ob man nicht generelle Aussagen über diese Verteilung machen kann. Dann wäre es möglich, Aussagen über die Wahrscheinlichkeiten für das Auftreten von bestimmten Stichprobenmittelwerten zu machen. Daraus ließen sich wiederum Wahrscheinlichkeitsangaben für mögliche Fehler ableiten, die beim Schluss

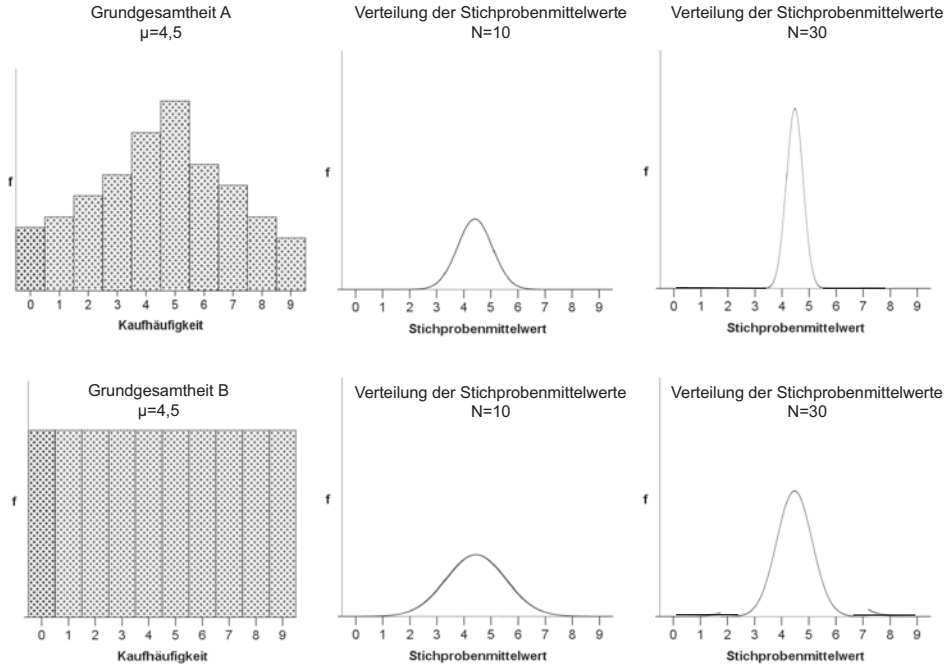


Abb. 8.2 Verteilung des Stichprobenmittelwerts bei verschiedener Homogenität der Grundgesamtheit und verschiedener Stichprobengröße. (Angelehnt an Jaccard & Becker, 2002, S. 196)

vom Stichprobenmittelwert auf den (gesuchten) Mittelwert der Grundgesamtheit auftreten können.

Bei den in der Abb. 8.2 dargestellten Verteilungen deutet sich an, dass diese einer Normalverteilung folgen könnten. Man erkennt, dass (bei gleichem Mittelwert ($\mu = 4{,}5$) sowohl für eine Grundgesamtheit A, in der die Werte der Kaufhäufigkeiten in relativ geringem Maße um den Mittelwert streuen, aber auch für eine Grundgesamtheit B, in der alle Kaufhäufigkeiten in gleicher Anzahl auftreten, sich mit wachsender Stichprobe die entsprechende glockenförmige Kurve ergibt. In der Tat besagt der **zentrale Grenzwertsatz** der Statistik, dass sich die Verteilung der Stichprobenmittelwerte mit zunehmender Stichprobengröße an eine Normalverteilung annähert. Von einer Stichprobengröße $n = 30$ an wird Annäherung als hinreichend angesehen.

Hintergrundinformation

Voß (2000, S. 338) schreibt zu den Eigenschaften der Normalverteilung:

> „Die Normalverteilung ist die wohl wichtigste Wahrscheinlichkeitsverteilung überhaupt. Dies beruht darauf, dass zum einen viele technische (zum Beispiel Fertigungstoleranzen) und biologische (zum Beispiel Körpermaße) Größen normalverteilt sind und zum anderen andere – diskrete wie stetige – Verteilungen in der Praxis oft durch die Normalverteilung

angenähert werden können. Zudem ergibt sich nach den Grenzwertsätzen aus der Überlagerung einer größeren Zahl (in der Regel n > 30) einzelner Zufallsvariablen beliebiger Verteilungen eine näherungsweise Normalverteilung."

Damit sind die Voraussetzungen für die Angabe eines **Konfidenzintervalls** (Vertrauensbereichs) gegeben. Konfidenzintervalle mit einer Konfidenzwahrscheinlichkeit p% haben die folgende Eigenschaft: Wenn man die Stichprobenziehung unendlich oft wiederholt, dann liegt in p% der Stichproben der wahre Wert innerhalb des jeweils berechneten Intervalls. Etwas vereinfacht kann man auch sagen: Der wahre Wert der Grundgesamtheit liegt mit der Wahrscheinlichkeit p% zwischen der Untergrenze und der Obergrenze des Konfidenzintervalls. Wenn man also davon ausgeht, dass die Stichprobenmittelwerte normalverteilt sind, dann kann man das Wissen über die Eigenschaften der Normalverteilung nutzen, um anzugeben, dass in 68 % aller Stichproben der wahre Wert der Grundgesamtheit in einem Intervall liegen wird, das sich von der Untergrenze $\mu - \sigma_x$ bis zur Obergrenze $\mu + \sigma_x$ erstreckt, dass also mit der Spannweite $2\sigma_x$ um den Mittelwert der Grundgesamtheit angeordnet ist.

Die Wahrscheinlichkeit dafür, dass der wahre Wert der Grundgesamtheit in diesen Bereich fällt, ist also $p = 0{,}68$. Die Angabe eines solchen Intervalls ist noch mit einer recht großen Irrtumswahrscheinlichkeit $(1 - 0{,}68 = 0{,}32)$ behaftet. Wenn man das Intervall breiter wählt, dann ist die Wahrscheinlichkeit, dass dieses den Stichprobenmittelwert umschließt, natürlich größer und die Irrtumswahrscheinlichkeit sinkt.

Zu den Eigenschaften der **Normalverteilung** gehört es, dass *95 % aller Werte* in einem Bereich liegen, der in beiden Richtungen um das *1,96-fache der Standardabweichung vom Mittelwert* abweicht. Man kann also über den Stichprobenmittelwert sagen, er liegt bei einem zweiseitigen Test mit einer Sicherheitswahrscheinlichkeit $p = 0{,}95$ in einem Intervall, das von $\mu - 1{,}96\sigma_{\bar{x}}$ bis $\mu + 1{,}96\sigma_{\bar{x}}$ reicht. Dieses Konfidenzintervall wird in folgender Weise formal dargestellt:

$$\mu - 1{,}96\sigma_{\bar{x}} < \bar{x} < \mu + 1{,}96\sigma_{\bar{x}}$$

Nun ist es in der Forschungspraxis, wo man ja nicht von einem bekannten Mittelwert der Grundgesamtheit auf einen zu erwartenden Mittelwert einer Stichprobe schließt, sondern *umgekehrt* vorgeht, notwendig, auf Basis des Stichprobenmittelwertes $(\bar{x})$ ein Konfidenzintervall für den Mittelwert der Grundgesamtheit (μ) zu bestimmen. Durch Umformung erhält man aus der vorstehenden Ungleichung das gewünschte Konfidenzintervall:

$$\bar{x} - 1{,}96\sigma_{\bar{x}} < \mu < \bar{x} + 1{,}96\sigma_{\bar{x}}$$

Dieses Konfidenzintervall ist folgendermaßen zu interpretieren: Der Bereich, der von $\bar{x} - 1{,}96\sigma_{\bar{x}}$ bis $\bar{x} + 1{,}96\sigma_{\bar{x}}$ reicht, wird mit einer Sicherheitswahrscheinlichkeit von $p = 0{,}95$ (in 95 % aller Fälle, in denen eine solche Stichprobe gezogen wird) den gesuchten (unbekannten) Mittelwert der Grundgesamtheit umschließen. Das Konfidenzintervall für den Mittelwert lässt sich natürlich auch allgemein darstellen:

$$\bar{x} - z\sigma_{\bar{x}} < \mu < \bar{x} + z\sigma_{\bar{x}}$$

Für verschiedene gewünschte Sicherheitswahrscheinlichkeiten, bei einem zweiseitigen Test, ergeben sich nach der Normalverteilung verschiedene Werte für z:

Sicherheitswahrscheinlichkeit	z-Wert
$p = 0{,}80$	$z = 1282$
$p = 0{,}90$	$z = 1645$
$p = 0{,}95$	$z = 1960$
$p = 0{,}99$	$z = 2576$

Da ja die Standardabweichung der Stichprobenmittelwerte ($\sigma_{\bar{x}}$) normalerweise nicht bekannt ist, verwendet man an ihrer Stelle den entsprechenden Schätzwert aus der Stichprobe ($\hat{s}_{\bar{x}}$), der in diesem Abschnitt schon erörtert worden ist. Streng genommen ist dann nicht die Normalverteilung, sondern die Studentverteilung zu verwenden. Da ab einem Stichprobenumfang von $n = 30$ die Studentverteilung durch die Normalverteilung approximiert werden kann, ist für die Marktforschung mit ihren meist viel größeren Stichproben in der Regel die Verwendung der Normalverteilung möglich.

Wichtig

David de Vaus (2002**, S. 191 f.) zur Interpretation von Konfidenzintervallen:** ***„Ein enges*** Konfidenzintervall ermöglicht genauere Schätzungen für die Grundgesamtheit. Die Weite des Konfidenzintervalls ist eine Funktion von zwei Faktoren:

- der Sicherheitswahrscheinlichkeit;
- der Größe des Stichprobenfehlers.

Je größer die Sicherheitswahrscheinlichkeit, desto größer das Konfidenzintervall. Wenn wir die Sicherheitswahrscheinlichkeit als festgelegt ansehen, dann besteht der einzige Weg zu einem präziseren Schätzwert für die Grundgesamtheit durch die Verengung des Konfidenzintervalls darin, den Stichprobenfehler zu minimieren.

Der Stichprobenfehler wird gemessen durch die Maßzahl „Standardfehler". Die Größe des Stichprobenfehlers hängt ab von zwei Faktoren:

- der Stichprobengröße;
- dem Ausmaß der Varianz in der Grundgesamtheit.

Wir können nichts tun, um die Varianz in der Grundgesamtheit zu verändern; deshalb besteht die einzige Möglichkeit zur Reduzierung des Stichprobenfehlers und damit zur Verkleinerung des Konfidenzintervalls darin, die Stichprobe zu vergrößern. Als generelle Regel gilt, dass die Stichprobe vervierfacht werden muss, um den Standardfehler zu halbieren."

Das vorstehend dargestellte Grundprinzip der Schätzung von Maßzahlen auf der Basis von Stichprobenergebnissen lässt sich auch auf andere Parameter als den Mittelwert übertragen, nicht zuletzt auf die für die Marktforschung wichtigen Anteilswerte (Prozentwerte oder relative Häufigkeiten).

Bei der Berechnung eines Konfidenzintervalls für Anteilswerte (z. B.: 50 % der Befragten sind der Meinung, dass …) bzw. relative Häufigkeiten liegt der wesentliche Unterschied darin, dass dafür die Standardabweichung auf folgende Weise geschätzt wird:

$$\widehat{s} = \sqrt{\frac{p(1-p)}{n}}$$

Wenn also ein Konfidenzintervall für einen Anteilswert von 50 % (bzw. eine relative Häufigkeit von 0,5) berechnet werden soll, das angeben soll, in welchem Bereich der „wahre“ Anteilswert (in der Grundgesamtheit) mit einer vorgegebenen Sicherheitswahrscheinlichkeit liegt, dann würde sich bei einer Stichprobengröße von $n = 100$ als Standardabweichung des Anteilswerts ergeben:

$$\widehat{s}_p = \sqrt{\frac{0{,}5(1-0{,}5)}{100}} = \sqrt{\frac{0{,}25}{100}} = 0{,}05$$

Wenn man wieder eine Sicherheitswahrscheinlichkeit von $p = 0{,}95$ verlangt, dann ergibt sich das entsprechende Konfidenzintervall durch:

$$0{,}5 - 1{,}96 \times 0{,}05 < \text{Anteilswert in der Grundgesamtheit} < 0{,}5 + 1{,}96 \times 0{,}5$$

$$0{,}5 - 0{,}098 < \text{Anteilswert in der Grundgesamtheit} < 0{,}5 + 0{,}098$$

$$0{,}402 < \text{Anteilswert in der Grundgesamtheit} < 0{,}598$$

Dieses Konfidenzintervall lässt sich folgendermaßen interpretieren: Wenn ein Anteilswert in einer (relativ kleinen) Stichprobe der Größe $n = 100$ bei 50 % (bzw. 0,5) liegt, dann kann man mit einer Sicherheitswahrscheinlichkeit von $p = 0{,}95$ sagen, dass der tatsächliche Anteilswert in der Grundgesamtheit zwischen 40,2 und 59,8 % liegen wird.
Anhand der vorstehend skizzierten Überlegungen lassen sich nun auch Angaben über die **Stichprobengröße** machen (siehe Abschn. 4.2), die erforderlich ist, um bei einer vorgegebenen Sicherheitswahrscheinlichkeit eine bestimmte Genauigkeit der ermittelten Werte zu erreichen, d. h. eine bestimmte Breite des Konfidenzintervalls nicht zu überschreiten. Beispielsweise kann es für eine politische Partei durchaus wichtig sein, dass die Angaben über die Wählerstimmen, die bei einer Befragung ermittelt werden, möglichst genau sind. Wenn die Schwankungsbreite hier bei ±3 % läge, dann würde das Ergebnis, dass eine Partei 5,5 % der Wählerstimmen erreichen würde, wenig darüber aussagen, ob eine Partei die sogenannte 5 %-Hürde geschafft hat oder nicht. Eine geringere Schwankungsbreite (ein kleineres Konfidenzintervall) wäre also wünschenswert.

Hintergrundinformation
Die Forschungsgruppe Wahlen weist deshalb zu Recht auf folgende Problematik zur Ausweisung kleinerer Parteien im „Politbarometer“ hin (www.forschungsgruppe.de):

> „Das Politbarometer ist eine repräsentative Umfrage, das heißt, von den dort erhobenen Ergebnissen kann auf ein Ergebnis in der Gesamtheit, also bei allen Wahlberechtigten, geschlossen werden. Dies allerdings innerhalb eines bestimmten statistischen Fehlerintervalls. Dieser – unvermeidliche – statistische Fehlerbereich ist relativ gesehen bei niedrigeren Anteilswerten größer als bei höheren, also zum Beispiel bei einem Umfrageergebnis für eine Partei von 5 % größer als bei einem Umfrageergebnis von 40 %. Für sehr kleine Parteien sind die relativen Fehlerbereiche inakzeptabel, weshalb für sie keine verlässlichen und seriösen Aussagen möglich sind.“

Wie lässt sich eine *optimale Stichprobengröße* nun ermitteln? Wenn man von dem Problem ausgeht, einen Mittelwert einer Grundgesamtheit auf Basis einer Stichprobe schätzen zu wollen, dann knüpft man zunächst daran an, dass die Stichprobenmittelwerte näherungsweise normalverteilt um den Mittelwert der Grundgesamtheit liegen. Die entsprechende Standardabweichung war als $s_{\bar{x}} = \hat{s}/\sqrt{n}$ geschätzt worden. Darauf aufbauend war das Konfidenzintervall, in dem der „wahre“ Mittelwert (der Grundgesamtheit) liegt, durch $\bar{x} \pm \mathrm{z}s_{\bar{x}}$ bestimmt worden. Unter Verwendung der entsprechenden Schätzwerte ergibt sich für die gewünschte Genauigkeit bzw. Breite des Intervalls der Mittelwertschätzung.

$$\text{Maximale Abweichung vom wahren Wert} = z\frac{\hat{s}^2}{\sqrt{n}}$$

Daraus lässt sich folgende erforderliche Stichprobengröße bei gegebener Schwankungsbreite, Sicherheitswahrscheinlichkeit und Standardabweichung bestimmen:

$$n = \frac{z^2\hat{s}^2}{(\text{Maximale Abweichung vom wahren Wert})^2}$$

Bei einer Sicherheitswahrscheinlichkeit von $p = 0{,}95$ ($\mathrm{z} = 1{,}96$), einer Standardabweichung von 2000 und einer tolerierten Schwankungsbreite (Genauigkeit) von 100 des zu ermittelnden Mittelwerts würde sich ergeben:

$$n = \frac{(1{,}96)^2(2000)^2}{(100)^2} = 1536$$

Eine Stichprobe der Größe $n = 1536$ würde unter diesen Umständen also zu der gewünschten Genauigkeit und Sicherheit der Ergebnisse führen. Bei Anteilswerten („Prozenten“) würde man analog vorgehen und die Stichprobengröße bestimmen durch:

$$n = \frac{z^2 p(1-p)}{(\text{Maximale Abweichung vom wahren Wert})^2}$$

Die obigen Berechnungen setzen voraus, dass man die Standardabweichung des interessierenden Mittel- oder Anteilswerts (z. B. aus früheren Untersuchungen oder Pretests) kennt oder schätzt.

Aus den vorstehenden Formeln lässt sich der schon erwähnte Zusammenhang zwischen Stichprobengröße und Genauigkeit (im Sinne der Enge des Konfidenzintervalls) der Aussagen entnehmen. Da die Stichprobengröße im Nenner des Ausdrucks steht, durch den die Standardabweichung nach Ziehung der entsprechenden Wurzel ermittelt wird, kann man folgende „Faustregeln" ableiten:

- Doppelte Stichprobengröße → 1,41-fache Genauigkeit
- Vierfache Stichprobengröße → doppelte Genauigkeit
- Sechzehnfache Stichprobengröße → vierfache Genauigkeit.

▶ **Wichtig**

Hier eine Zusammenfassung zentraler Aussagen über Konfidenzintervalle:

- Bei einem Konfidenzintervall wird die Angabe eines Wertebereichs mit Wahrscheinlichkeit dafür, dass ein gesuchter Wert von diesem Intervall umschlossen wird, verbunden.
- Mit der Vergrößerung des Konfidenzintervalls (also mit der Verringerung der Genauigkeit der Aussage) steigt die Sicherheitswahrscheinlichkeit (und umgekehrt).
- Bei gegebener Sicherheitswahrscheinlichkeit und gegebener Standardabweichung der Messwerte in der Grundgesamtheit wird das Konfidenzintervall enger (steigende Genauigkeit) bei Vergrößerung der Stichprobe n.
- Bei gegebener Sicherheitswahrscheinlichkeit und gegebener Stichprobengröße wird das Konfidenzintervall bei geringerer Standardabweichung (größerer Homogenität der Daten) der Messwerte enger („genauer").

8.2 Statistische Tests

8.2.1 Grundlagen zu Hypothesentests, Signifikanz und Effektstärken

Statistische Hypothesentests sollen anhand vorliegender Beobachtungen eine begründete Entscheidung über die Korrektheit einer Hypothese ermöglichen. In diesem Sinne geht es also darum festzustellen, ob ein Ergebnis, dass in einer Stichprobe erzielt wurde, sich gemäß den Ausführungen in Abschn. 2.3 auf die Grundgesamtheit generalisieren lässt. Da Stichprobendaten in der Regel Realisationen von Zufallsvariablen sind, lässt sich in den meisten Fällen nicht mit Sicherheit sagen, ob eine Hypothese stimmt oder nicht. Man versucht daher, die Wahrscheinlichkeiten für Fehlentscheidungen zu kontrollieren,

was einem Test zu einem vorgegebenen Signifikanzniveau entspricht. Aus diesem Grund spricht man auch von einem Hypothesentest oder einem Signifikanztest.

Beispiel

Die Grundidee von solchen Tests lässt sich in verkürzter Form in Anlehnung an Benesch (2012, S. 156) anhand einer Gerichtsverhandlung verdeutlichen:

> Ein Angeklagter steht vor Gericht in einem Indizienprozess. Nach der Basisannahme (hier: Unschuldsvermutung) geht man davon aus, dass der Angeklagte unschuldig ist. Im Laufe des Verfahrens bringt der Staatsanwalt Indizien für die Schuld des Angeklagten vor. Nur wenn diese Indizien als hinreichend befunden werden, wird die anfängliche Unschuldsvermutung verworfen und nach der Alternativannahme, der Angeklagte ist schuldig, entschieden. Reichen die Indizien nicht aus, wird der Angeklagte als nicht schuldig angesehen und die Unschuldsannahme wird nicht verworfen. Das Gericht kann bei diesem Fall zwei Fehler machen. Es kann eine Person für schuldig erklären, obwohl sie in Wahrheit unschuldig ist. Oder aber, es kann eine Person auf freien Fuß setzen, obwohl sie in Wahrheit schuldig ist. Letztendlich handelt es sich also um ein Abwägen dieser zwei möglichen Fehler vor dem Hintergrund der vorgebrachten Indizien. ◀

Ganz ähnlich wie im obigen Beispiel verhält es sich bei einem **statischen Hypothesentest.** Es soll z. B. untersucht werden, ob es zwischen zwei Variablen einen Zusammenhang gibt. Die Basisannahme lautet, dass es in der Grundgesamtheit keinen Zusammenhang zwischen den beiden Variablen gibt (Nullhypothese: Es gibt keinen Effekt in der Grundgesamtheit). Statt in einer Gerichtsverhandlung werden nun in der Stichprobe Indizien für einen Zusammenhang zwischen den Variablen gesucht. Nur wenn die Indizien sehr stark gegen die Nullhypothese sprechen, wird diese verworfen. Es spricht dann viel dafür, dass die Alternativhypothese (hier. Es gibt einen Zusammenhang zwischen den Variablen) gilt. Die Stärke der Indizien ist von entscheidender Bedeutung. Ein schwacher Zusammenhang zwischen den Variablen in der Stichprobe muss nicht bedeuten, dass es diesen Zusammenhang auch in der Grundgesamtheit gibt. Es könnte sich um ein lediglich zufälliges Ergebnis in der Stichprobe handeln. Falls die Indizien nicht ausreichen, gilt weiterhin die Nullhypothese, dass kein Zusammenhang vorliegt. Eine 100 %ige Sicherheit, dass die Entscheidung richtig ist, kann der Hypothesentest naturgemäß niemals bieten, da von einer Stichprobe auf die Grundgesamtheit geschlossen wird. Analog zur Gerichtsverhandlung können also auch hier zwei Fehler gemacht werden. Die Nullhypothese kann fälschlicherweise verworfen oder fälschlicherweise nicht verworfen werden.

Wird die Nullhypothese fälschlicherweise abgelehnt, nennt man diesen Fehler den „Fehler 1. Art“ bzw. alpha-Fehler. Die Fragen, die sich in diesem Zusammenhang stellen, lauten also: Wie groß wäre die Wahrscheinlichkeit, eine solches Ergebnis durch Zufall zu erhalten? Wie klein muss diese Wahrscheinlichkeit mindestens sein, damit die Nullhypothese mit hinreichender Sicherheit verworfen werden kann? Oder anders. Wieviel Unsicherheit ist der Forscher bereit, bei der Entscheidung in Kauf zu nehmen? Die Wahrscheinlichkeit einen Fehler 1. Art zu begehen, heißt **Signifikanzniveau** oder

Irrtumswahrscheinlichkeit alpha. Die entsprechende Gegenwahrscheinlichkeit ist die verwendete Sicherheitswahrscheinlichkeit.

Wird die Nullhypothese fälschlicherweise nicht verworfen, d. h. es gibt in Wahrheit (also in der Grundgesamtheit) einen Zusammenhang zwischen den zwei Variablen, wird dieser Fehler wird als „Fehler 2. Art" bzw. beta-Fehler bezeichnet. Diese Wahrscheinlichkeit können wir nicht kontrollieren, sie ist abhängig von der Art des Tests und des Signifikanzniveaus alpha. Man muss sich vor Durchführung des Tests also auf ein Signifikanzniveau alpha festlegen, das die maximale Wahrscheinlichkeit festlegt, mit der uns so ein Fehler 1. Art passieren darf. Je sicherer man bei der Entscheidung sein will, desto niedriger muss diese Fehlerwahrscheinlichkeit gewählt werden. Im Bereich Marketing und Marktforschung wird dieser Wert alpha häufig auf 5 % festgelegt. Wenn also die Irrtumswahrscheinlichkeit alpha geringer als 5 % ist, spricht man von einem signifikanten Ergebnis und in diesem Fall von einem signifikanten Zusammenhang zwischen den Variablen. Die Irrtumswahrscheinlichkeit wird in statistischen Softwareprogrammen direkt als p-Wert ausgegeben. Natürlich kann es in anderen Bereichen (bspw. in der Medizin) sinnvoller sein, andere (kleinere) alpha-Niveaus zu wählen.

Im Gegensatz zum Fehler 1. Art, lässt sich die Wahrscheinlichkeit für den Fehler 2. Art in der Regel nicht einfach berechnen. Man kann diesen ausschließlich berechnen, wenn man für die Alternativhypothese eine andere Wahrscheinlichkeit als für die Nullhypothese annimmt. Im Allgemeinen gilt allerdings: Je kleiner die Wahrscheinlichkeiten für Fehler der 1. und 2. Art, desto besser.

In der folgenden Tabelle sind diese beiden Fehlerarten statistischer Tests zusammenfassend dargestellt bzw. gegenübergestellt:

	Hypothese ist richtig	Hypothese ist falsch
Hypothese nicht abgelehnt	Richtige Entscheidung	Fehler 2. Art
Hypothese abgelehnt	Fehler 1. Art	Richtige Entscheidung

Die Anlage statistischer Tests im Hinblick auf entsprechende Fehlerwahrscheinlichkeiten ist eine anspruchsvolle Aufgabe, für die auf die speziellere Literatur verwiesen werden muss (siehe Sawyer & Peter, 1983). Ein kurzer Überblick findet sich auch bei Eisend und Kuß (2017, S. 162 ff.), ausführlichere Darstellungen findet man in zahlreichen Statistik-Lehrbüchern (z. B. Jaccard & Becker, 2002). Entscheidend ist jedoch, dass man die Nullhypothese nie mit Sicherheit beweisen oder verwerfen und dass man auch „nur" Indizien für die Alternativhypothese findet. Aus diesem Grund ist es wichtig, dass man die Hypothesen richtig herum formuliert: Der Fall, den man bestätigen möchte, ist Gegenstand der Alternativhypothese (entsprechend der Metapher der Gerichtsverhandlung).

Ein statistisch signifikantes Ergebnis besagt lediglich, dass ein Effekt oder ein Zusammenhang in einer Stichprobe nicht zufällig zu Stande gekommen ist, sondern dass sich das Ergebnis mit großer Wahrscheinlichkeit auf die Grundgesamtheit übertragen bzw. generalisieren (siehe Abschn. 2.3) lässt. Signifikanz *allein* sagt nichts darüber aus,

ob das Ergebnis auch inhaltlich bedeutsam bzw. relevant ist (siehe dazu das Beispiel in Abschn. 8.2.3).

Vielfach wird in der praktischen Anwendung statistischer Tests ein kleiner p-Wert, d. h. eine geringe Irrtumswahrscheinlichkeit mit einer vergleichsweise hohen **Effektstärke** assoziiert. Zwar ist es tatsächlich der Fall, dass unter Beibehaltung der anderen Parameter eines Tests (u. a. Stichprobengröße, gewähltes Signifikanzniveau) ein kleinerer p-Wert mit einer größeren Effektstärke assoziiert ist. Dies ist aber einfach ein spezifisches Merkmal des statistischen Tests (bzw. der zugrundeliegenden Verteilungen) und lässt eine Interpretation der Irrtumswahrscheinlichkeit p als Effektstärke (im Hinblick auf die *Relevanz* eines Effekts bzw. Zusammenhangs) nicht zu (vgl. hierzu auch Cohen, 1988). Um die Bedeutsamkeit von Forschungsergebnissen zu beurteilen reichen Signifikanztests also nicht aus. Es sollten zusätzlich Effektstärkemaße herangezogen werden, die die Bedeutung von Auswirkungen quantifizieren. Bei der Integration von Resultaten verschiedener Untersuchungen (Metaanalysen) sind diese unverzichtbar (Bortz & Döring, 1995). Bekannte Effektstärkemaße sind der Korrelationskoeffizient r als Maß des Zusammenhangs und das Differenzmaß d als Maß von Mittelwertunterschieden. Nach Cohen (1988, S. 82) gelten Zusammenhänge unter r = ,10 als unbedeutend, ab r = ,30 als mittel und ab r = ,50 als groß. Im folgenden Abschnitt und auch in Kap. 9 werden weitere ausgewählte und methodenspezifische Effektstärkenmaße kurz erläutert.

8.2.2 Der Chi-Quadrat Test

Am Ende des Abschn. 7.3.1 ist bereits auf die in der Datenanalyse sehr verbreiteten **Kontingenztabellen** (auch Kreuztabellen genannt) hingewiesen und deren weitere Erläuterung angekündigt worden. Diese soll nun hier im Zusammenhang mit der Diskussion der grundlegenden Schlussweise bei statistischen Tests erfolgen und mithilfe eines fiktiven Beispiels aus Lehmann et al. (1998) verdeutlicht werden: Es liegt eine Stichprobe ($n = 100$) zur Markenwahl von Haushalten in einer bestimmten Produktkategorie (bspw. Waschmittel: Marke A, Marke B, Marke C) vor. Die Haushalte lassen sich ihrem Wohnort gemäß drei verschiedenen Regionen zuordnen (Region Nord, Region Ost & West, Region Süd) zur Verfügung.

Es wird also von einer Umfrage ausgegangen, deren hier interessierendes Ergebnis in der Tab. 8.2 dargestellt ist. Einhundert Haushalte aus drei verschiedenen Regionen (Nord, Ost & West, Süd) sind gefragt worden, welche Waschmittelmarke (A, B oder C) sie üblicherweise kaufen. In Tab. 8.2 sind die absoluten und die relativen Häufigkeiten eingetragen, die sich dabei ergeben haben. In jedem Feld der Tabelle stehen an erster Stelle die absoluten Häufigkeiten für die jeweiligen Kombinationen von Merkmalsausprägungen. Es folgen relative Häufigkeiten (in Prozent), die auf die entsprechenden Zeilensummen (an zweiter Stelle) bzw. die entsprechenden Spaltensummen (an dritter Stelle) bezogen sind. In der Randspalte bzw. -zeile findet man die Zeilen- bzw. Spalten-

Tab. 8.2 Beispiel einer Kontingenztabelle

			Marke			
			A	B	C	Gesamt
	Nord	Anzahl	10	3	23	36
		% innerhalb von Region	27,8 %	8,3 %	63,9 %	100,0 %
		% innerhalb von Marke	28,6 %	10,3 %	63,9 %	36,0 %
	Ost & West	Anzahl	17	15	5	37
		% innerhalb von Region	45,9 %	40,5 %	13,5 %	100,0 %
		% innerhalb von Marke	48,6 %	51,7 %	13,9 %	37,0 %
	Süd	Anzahl	8	11	8	27
		% innerhalb von Region	29,6 %	40,7 %	29,6 %	100,0 %
		% innerhalb von Marke	22,9 %	37,9 %	22,2 %	27,0 %
Gesamt		Anzahl	35	29	36	100
		% innerhalb von Region	35,0 %	29,0 %	36,0 %	100,0 %
		% innerhalb von Marke	100,0 %	100,0 %	100,0 %	100,0 %
		Anzahl				

summen, die die eindimensionalen Häufigkeitsverteilungen der beiden Merkmale darstellen. Hinzugefügt sind außerdem die dazugehörigen relativen Häufigkeiten.

Wie interpretiert man nun eine solche Tabelle? Dabei gilt immer das Prinzip, dass *Verteilungen miteinander verglichen werden*. Welche das sind, hängt ab von der Art der Aussagen, die gemacht werden sollen, genauer gesagt von der Art der bezüglich der beiden Merkmale unterstellten Abhängigkeiten. Im vorliegenden Beispiel ist diese Frage eindeutig zu beantworten: Man geht davon aus, dass durch das Merkmal „Region" und die damit verbundenen ökonomischen, sozialen und kulturellen Besonderheiten die Wahl der Marke beeinflusst wird. Die umgekehrte Unterstellung (Abhängigkeit der Wohnregion von der Markenwahl) wäre wohl einigermaßen unsinnig. Es gibt aber durchaus Merkmalskombinationen, bei denen diese Fragestellung nicht so klar zu beantworten ist.

Wenn man vermutet, dass in den drei im Beispiel aufgeführten Regionen die Markenwahl unterschiedlich ist, dann muss man für die verschiedenen Regionen deren Verteilungen miteinander vergleichen. Es müssen also die sich in den Zeilen der Tabelle widerspiegelnden Häufigkeitsverteilungen betrachtet werden. Man sieht sofort, dass diese voneinander abweichen. In Region Nord hat man eine Verteilung von 27,8, 8,3 und 63,9 %, in Region Ost & West von 45,9, 40,5 und 13,5 % und in Region Süd von 29,6, 40,7 und 29,6 % für die drei Kategorien der Kaufintensität. Danach könnte man die Hypothese eines Unterschiedes der Markenwahl in den drei Regionen bestätigen. Diesen Schluss könnte man allerdings nur sicher ziehen, wenn es sich um eine Vollerhebung der Grundgesamtheit handeln würde. Im vorliegenden Fall wird die/der kritische LeserIn sofort einwenden, dass die Tabelle auf der Basis von Stichprobendaten zustande ge-

kommen ist und deswegen die Unterschiede in den Verteilungen auch nur zufällig bei der gezogenen Stichprobe auftreten könnten, während es in der Grundgesamtheit möglicherweise gar keine Unterschiede gibt. Sind also die Unterschiede der Verteilungen in den Regionen durch *systematische Unterschiede* zwischen den Regionen oder durch den *Stichprobenfehler* zu erklären?

Bei der Analyse muss man deshalb eine Entscheidung zwischen zwei **Hypothesen** (systematische Unterschiede oder Unterschiede, die durch Stichprobenfehler zu erklären sind) treffen. Statistische Tests sind Hilfsmittel, um derartige Entscheidungen in einer begründeten und formal festgelegten Weise zu treffen. Einer der am stärksten verbreiteten Tests, der χ^2 **-Test** (Chi^2-Test), soll hier exemplarisch erläutert und zur Lösung des in dem bisher verwendeten Beispiel aufgetretenen Entscheidungsproblems herangezogen werden. Es geht also um die Frage, ob zwischen den Merkmalen „Region" und „Markenwahl" eine *Abhängigkeit oder Unabhängigkeit* besteht.

Bei der vorstehend skizzierten Interpretation einer Kontingenztabelle ist wegen der Unterschiede zwischen den Verteilungen von Merkmalsausprägungen für verschiedene Gruppen (Regionen) vermutet worden, dass dieses Merkmal das andere Merkmal (Markenwahl) beeinflusst, also eine Abhängigkeit vorliegt. Wenn dagegen die relativen Häufigkeiten in den Zeilen und Spalten einer Tabelle exakt den entsprechenden Randverteilungen gleichen, dann besteht offenbar keinerlei Zusammenhang zwischen den Merkmalen. Beim vorliegenden Beispiel würde man ja, wenn die relativen Häufigkeiten der Markenwahl für die drei Regionen gleich sind und damit auch gleich der in der Randverteilung wiedergegebenen Verteilung für die gesamte Stichprobe sind (also jeweils 35, 29, 36 %), sagen, dass das Merkmal „Region" offenbar keinen Einfluss auf das Merkmal „Markenwahl" hat, weil ja in allen Regionen die Verteilungen der Markenwahl gleich sind. Das Merkmal „Markenwahl" wäre also *unabhängig* vom Merkmal „Region".

Der entscheidende **Grundgedanke des χ^2 -Tests** auf Unabhängigkeit besteht in diesem Sinne darin, dass man eine gegebene Häufigkeitsverteilung in einer Tabelle mit einer Häufigkeitsverteilung vergleicht, die zustande gekommen wäre, wenn zwischen den betrachteten Merkmalen Unabhängigkeit vorläge. Die Tabelle der unter der Annahme der Unabhängigkeit erwarteten Häufigkeiten erhält man dadurch, dass man die (erwarteten) Besetzungen der einzelnen Felder aus den relativen Häufigkeiten in den Randverteilungen und der Stichprobengröße errechnet. Für das in Tab. 8.2 dargestellte Beispiel würde sich für das Feld links oben (Region Nord, Marke A) eine erwartete Häufigkeit von 12,6 ergeben (36×35 %). Entsprechende Berechnungen für alle Tabellenfelder führen zu der um die erwarteten Häufigkeiten erweiterte Tab. 8.3.

Jetzt hat man die Grundlage für einen Vergleich der *beobachteten* mit den (unter der Annahme der Unabhängigkeit) *erwarteten* Häufigkeiten in den Feldern der Tabelle. Das Ergebnis dieses Vergleichs überrascht nicht. Die beobachteten Häufigkeiten unterschieden sich von den erwarteten. Wie aber ist dieses Ergebnis zu interpretieren? Soll man die Hypothese, dass ein systematischer Zusammenhang zwischen den Variablen vorliegt, oder die Hypothese, dass die Abweichungen zwischen beobachteten und er-

Tab. 8.3 Vergleich von beobachteten mit den erwarteten Häufigkeiten

			Marke			
			A	B	C	Gesamt
Region	Nord	Anzahl	10	3	23	36
		Erwartete Anzahl	12,6	10,4	13,0	36,0
	Ost & West	Anzahl	17	15	5	37
		Erwartete Anzahl	13,0	10,7	13,3	37,0
	Süd	Anzahl	8	11	8	27
		Erwartete Anzahl	9,5	7,8	9,7	27,0
Gesamt		Anzahl	35	29	36	100
		Erwartete Anzahl	35,0	29,0	36,0	100,0

warteten Häufigkeiten durch den Stichprobenfehler zu erklären sind und kein systematischer Zusammenhang vorliegt, annehmen?

Die Grundidee des χ^2-Tests lautet: Man betrachtet für jedes Feld der Tabelle die **Abweichungen zwischen beobachteten und erwarteten Häufigkeiten.** Wenn diese Abweichungen insgesamt (d. h. die Summe der Abweichungen über alle Felder) groß sind, dann entscheidet man sich eher für die Hypothese der Abhängigkeit (und umgekehrt). Man benötigt also eine Maßzahl, in der die genannten Abweichungen für eine ganze Tabelle zusammengefasst werden, um die Entscheidung über die Annahme einer der Hypothesen zu treffen. Die Maßzahl heißt χ^2 (oder eben auch Chi2) und hat folgende Form:

$$\chi^2 = \sum_i \sum_j \frac{(\text{Beobachtete Häufigkeit in Feld i,j} - \text{erwartete Häufigkeit in Feld i,j})^2}{\text{Erwartete Häufigkeit in Feld i,j}}$$

Gegenüber der bisher skizzierten Idee, die Summe der Abweichungen in den einzelnen Tabellenfeldern als Maßzahl zu verwenden, weicht χ^2 in zwei Aspekten ab:

- Es wird das Quadrat der Abweichungen verwendet. Anderenfalls würden sich positive und negative Abweichungen ausgleichen.
- Die Abweichungen werden hinsichtlich der erwarteten Häufigkeiten normiert. Das liegt daran, dass z. B. eine Abweichung von 20 bei einer erwarteten Häufigkeit von 50 ein anderes Gewicht hat als bei einer erwarteten Häufigkeit von 1000.

Wenn man den χ^2-Wert für die Tab. 8.3 berechnet, so ergibt sich:

$$\begin{aligned}\chi^2 = &\frac{(10-12{,}6)^2}{12{,}6} + \frac{(3-10{,}4)^2}{10{,}4} + \frac{(23-13)^2}{13} + \frac{(17-13)^2}{13} + \frac{(15-10{,}7)^2}{10{,}7} \\ &+ \frac{(5-13{,}3)^2}{13{,}3} + \frac{(8-9{,}5)^2}{9{,}5} + \frac{(11-7{,}8)^2}{7{,}8} + \frac{(8-9{,}7)^2}{9{,}7} \\ &= 23{,}6\end{aligned}$$

Für die Verwendung dieses Ergebnisses bezüglich einer Entscheidung über die Hypothese fehlt jetzt noch ein gedanklicher Schritt, um zu beurteilen, ob ein χ^2-Wert von 23,6 als hoch oder niedrig im Hinblick auf die Aussage eines Zusammenhangs zwischen den betrachteten Merkmalen angesehen wird. Der Maßstab dafür ist eine Verteilung, die angibt, mit welchen Wahrscheinlichkeiten verschiedene χ^2-Werte zu erwarten sind, wenn Unabhängigkeit der Merkmale in der Grundgesamtheit vorliegt. Im Idealfall müsste der χ^2-Wert unter dieser Voraussetzung gleich Null sein. Wenn man Stichproben aus einer solchen Grundgesamtheit zieht, muss man auch beim χ^2-Wert aufgrund des Stichprobenfehlers mit Abweichungen vom „Idealwert" Null rechnen. Kleine Abweichungen treten häufig auf (haben eine relativ hohe Wahrscheinlichkeit), große Abweichungen treten selten auf (haben eine geringe Wahrscheinlichkeit).

Diese **Wahrscheinlichkeitsverteilung** (die **χ^2-Verteilung**) ist der Maßstab für die Beurteilung eines aufgetretenen χ^2-Wertes (z. B. des Wertes 23,6). Wenn man feststellt, dass der Wert in einer Größenordnung liegt, die mit großer Wahrscheinlichkeit wegen des Stichprobenfehlers auch bei vollständiger Unabhängigkeit der Merkmale zu erwarten ist, lehnt man die Hypothese eines systematischen Zusammenhanges ab. Bei einem relativ großen χ^2-Wert ist es entsprechend unwahrscheinlich, dass er bei einer in der Grundgesamtheit vorhandenen Unabhängigkeit durch Zufall zustande gekommen ist. Man entscheidet sich deshalb in diesem Fall also für die Annahme der Hypothese eines systematischen Zusammenhangs zwischen den Merkmalen.

Mithilfe einer **Tabelle der χ^2-Verteilung** (deren Werte natürlich heutzutage von jedem Statistik-Programm oder auch von MS Excel berechnet werden) kann man jetzt entscheiden, welche der Hypothesen angenommen wird. Dazu ist zunächst eine Irrtumswahrscheinlichkeit, die bei der Entscheidung in Kauf genommen werden soll, festzulegen. Ferner gilt es zu beachten, wie viele Felder die Kontingenztabelle hat, die der Ausgangspunkt für die Überlegungen war. Bei „großen" Tabellen (mit vielen Feldern) muss man mit einem größeren χ^2-Wert rechnen als bei „kleinen" Tabellen. Wenn in dem verwendeten Beispiel die Region nicht in drei, sondern in vier Kategorien (z. Bsp.: Nord, Ost, West und Süd) erhoben worden wären, hätte man an Stelle der Tabelle mit 9 Feldern eine Tabelle mit 12 Feldern erhalten, bei denen natürlich mehr „Gelegenheiten" für Abweichungen zwischen erwarteter und beobachteter Häufigkeit existieren, was einen höheren χ^2-Wert zur Folge hat, ohne dass sich am Zusammenhang der Merkmale etwas verändert hat. Der gleiche Effekt ergibt sich natürlich auch für eine erweiterte Auswahl an betrachteten Marken. Die Größe der Tabelle wird in der Fachsprache durch die „**Zahl der Freiheitsgrade**" angegeben. Die Zahl der Freiheitsgrade ergibt sich durch (Spaltenzahl – 1) × (Zeilenzahl – 1). Im angeführten Beispiel hat man also $(3-1)\times(3-1)=4$ Freiheitsgrade.

Wichtig

Das Konzept der Freiheitsgrade kann an folgendem Beispiel erörtert werden: 3 Menschen aus einer Menschenmenge werden gewogen. Das Ergebnis lautet 62 kg, 70 kg und 78 kg. Der arithmetische Mittelwert ist (62 kg + 70 kg + 78 kg)

/3 = 210 kg/3 = 70 kg. Nun könnte man aus der Menschenmenge 2 andere Menschen frei auswählen (z. B. mit 64 kg und 76 kg), wäre dann aber gezwungen, einen Menschen mit 70 kg auszuwählen, damit der Mittelwert von 70 kg konstant bleibt.

Durch die Berechnung und Verwendung des arithmetischen Mittelwerts sind also nur 2 Messwerte frei, der andere ist unfrei bzw. kann nicht frei geändert werden und nimmt einen bestimmten Wert an. Allgemein ist hier die Anzahl der Freiheitsgrade n − 1, mit n als Anzahl der Messwerte, also 3 − 1 = 2.

Das gilt auch, wenn das arithmetische Mittel nur ein Zwischenergebnis ist, das für andere statistische Parameter verwendet wird, z. B. für die Varianz bzw. Standardabweichung.

Sedlmeier und Renkewitz (2008, S. 332) erklären die Idee der Freiheitsgrade im Zusammenhang mit der Berechnung der Stichprobenvarianz, die in Abschn. 7.1 vorgestellt wurde: „Der Begriff Freiheitsgrade steht für die Anzahl der Werte, die in einem statistischen Ausdruck frei variieren können. Ein solcher statistischer Ausdruck ist beispielsweise die Stichprobenvarianz. Bei der Berechnung von s^2 können von den n Messwerten nur n − 1 frei variieren, da die Summe aller Abweichungen der n Einzelwerte von ihrem Mittelwert immer 0 ist. Wenn also z. B. $n = 4$ und die ersten drei Abweichungen vom Mittelwert $6, -9,$ und -1 betragen, muss die vierte Abweichung $0 - 6 + 9 + 1 = 4$ sein.“ Allgemein ist hier die Anzahl der Freiheitsgrade n − 1, mit n als Anzahl der Messwerte, also 4 − 1 = 3.

Analog gilt dies auch für Kreuztabellen zu denen auch die oben erwähnte Tabelle der χ^2 -Verteilung: wenn ein Wert vorgegeben ist (z. B. 9 von 12 Teilnehmern sind z. B. „im Sportverein"), dann ergibt sich daraus der andere Wert bzw. dieser ist unfrei und damit nicht variierbar (dann müssen die anderen 3 Teilnehmer „nicht im Sportverein“ sein). Die Anzahl der Freiheitsgrade in einer Kreuztabelle ist (Anzahl der Zeilen – 1) × (Anzahl der Spalten – 1), bei einer Vierfeldertafel also $(2 - 1) \times (2 - 1) = 1 \times 1 = 1$.

Einer Tabelle der χ^2-Verteilung für vier Freiheitsgrade und einer Irrtumswahrscheinlichkeit von 0,05 kann man entnehmen, dass der entsprechende Grenzwert für X^2 bei 9,49 liegt. Bei χ^2 -Werten, die größer als 9,49 sind (wie im Beispiel der Fall), ist die Wahrscheinlichkeit, dass sie trotz der Unabhängigkeit der Merkmale per Zufall zustande gekommen sind, sehr klein und man kann mit großer Sicherheitswahrscheinlichkeit ($p = 0{,}95$) bzw. kleiner Irrtumswahrscheinlichkeit ($p = 0{,}05$) die Aussage machen, dass ein *systematischer* Zusammenhang vorliegt. Damit ist die Entscheidung zwischen den Hypothesen gefallen und das Problem gelöst, für das der Test als Entscheidungsregel konzipiert wurde.

Die wichtigsten Schritte des Tests seien in drei Schritten noch einmal zusammengefasst und in Abb. 8.3 verdeutlicht:

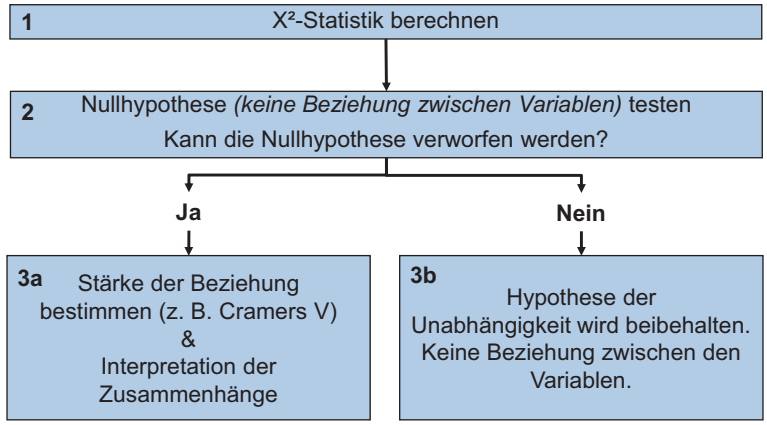


Abb. 8.3 Zusammenfassung zum Ablauf des χ^2 -Tests

1. **Berechnung einer Maßzahl** χ^2 für die Abweichungen der beobachteten Häufigkeiten von den bei Unabhängigkeit der Merkmale zu erwartenden Häufigkeiten.
2. **Vergleich der Maßzahl mit einer Wahrscheinlichkeitsverteilung**, die angibt, wie groß die Wahrscheinlichkeit dafür ist, dass ein bestimmter χ^2-Wert in der Stichprobe zufällig zustande kommt, obwohl in der Grundgesamtheit Unabhängigkeit der Merkmale vorliegt.
3. **Entscheidung:**
   a) Wenn die Wahrscheinlichkeit für ein zufälliges Auftreten des errechneten oder eines größeren χ^2-Wertes gering ist (kleiner als das vorgegebene Signifikanzniveau), wird die Hypothese der Unabhängigkeit verworfen (→ systematischer Zusammenhang zwischen den Merkmalen). In diesem Fall lassen sich in den gängigen Statistiksoftwarepaketen auch Maße wie z. B. Cramers V berechnen, die Aufschluss über die Stärke des Zusammenhangs geben (Effektstärke).
   b) Wenn die Wahrscheinlichkeit für ein zufälliges Auftreten des errechneten oder eines größeren χ^2-Wertes groß ist (größer als das Signifikanzniveau), wird die Hypothese der Unabhängigkeit beibehalten (kein systematischer Zusammenhang zwischen den Merkmalen).

In diesem Abschnitt ist ein in der Praxis der Marktforschung sehr verbreiteter Test recht ausführlich dargestellt worden, um die Prinzipien statistischer Tests zu verdeutlichen. Andere Tests für andere Fragestellungen und/oder mit anderen Voraussetzungen unterscheiden sich vom χ^2-Test natürlich in vieler Hinsicht. Die Grundprinzipien (Berechnung einer Maßzahl → Vergleich der Maßzahl mit einer Wahrscheinlichkeitsverteilung → Entscheidung) sind aber die gleichen.

Wie bereits im einführenden Abschn. 8.2.1, ist auch bei der vorstehenden Darstellung des χ2-Tests immer wieder von Wahrscheinlichkeiten bzw. von Irrtumswahrscheinlich-

keiten die Rede gewesen. Damit wird unterstrichen, dass man mit statistischen Tests nicht zu sicheren Aussagen über die Korrektheit von Hypothesen kommen kann. Es können vielmehr die bereits angesprochenen Fehler der 1. und 2. Art auftreten.

8.2.3 Der t-Test

t-Test Ein weiterer (auch) in der Marktforschung häufig angewandter statistischer Test ist der **t-Test**. Auch dabei lässt sich die grundlegende Idee relativ leicht plausibel machen; Einzelheiten der theoretischen Begründung und der Anwendung seien den (spezialisierten) Autoren aus der Statistik überlassen.

Der t-Test wird hier angewandt auf Entscheidungen im Hinblick auf den **Vergleich zweier Mittelwerte**. Wenn man zwei Mittelwerte – z. B. 1,72 und 1,78 oder 5,9 und 9,8– vergleicht, dann stellt sich wiederum die Frage, ob die erkennbaren Unterschiede durch die Zufälligkeiten der Stichprobenziehung erklärbar sind oder ob man von einem systematischen Unterschied ausgehen kann. Welche Gesichtspunkte müssen bei einer solchen Entscheidung (→ Test) eine Rolle spielen, wenn man sie auf der Basis des „gesunden Menschenverstandes“ trifft? Sicher die beiden folgenden Aspekte:

- Wie groß ist der Unterschied zwischen den beiden Mittelwerten? Je größer der Unterschied, desto weniger würde man erwarten, dass er zufällig zustande gekommen ist.
- Wie stark schwanken die Mittelwerte, wenn man mehrere Stichproben zieht, d. h. wie groß ist der Standardfehler der Stichprobe (vgl. Abschn. 8.1)? Je kleiner der Standardfehler ist, je geringer die Stichprobenmittelwerte bei der Ziehung zahlreicher Stichproben also um den Mittelwert der Grundgesamtheit streuen, desto besser wird im Allgemeinen der wirkliche Mittelwert der Grundgesamtheit durch die Stichprobe geschätzt.

An diese beiden einfachen Überlegungen knüpft der t-Test direkt an. Man bildet die *Differenz der beiden Mittelwerte*, über deren Signifikanz eine Entscheidung getroffen werden soll, und setzt diese in Beziehung zur *Streuung der Mittelwerte*. Wenn man prüfen will, ob sich ein Stichprobenmittelwert von einem bestimmten – vielleicht theoretisch erwarteten – Mittelwert einer Grundgesamtheit μ systematisch unterscheidet, dann ergibt sich die entsprechende Maßgröße durch.

$$t = \frac{\overline{x} - \mu}{S_{\overline{x}}}$$

Hier kann man an Überlegungen zur Schätzung des Standardfehlers aus dem Abschn. 7.1 anknüpfen und kommt zu.

$$t = \frac{\overline{x} - \mu}{\frac{s}{\sqrt{n}}}$$

Wie lässt sich diese Maßgröße t nun interpretieren?

- Bei einer großen Differenz $\bar{x} - \mu$ ist t (natürlich) relativ groß,
- Bei geringer Streuung der Stichprobenmittelwerte bzw. bei geringer Streuung der beobachteten Variablen ist der Nenner relativ klein und die Maßzahl t relativ groß,
- Bei großer Stichprobe n wird der Nenner der Maßzahl t klein und die Maßzahl insgesamt groß.

Man kommt zu dem – plausiblen – Ergebnis, dass die Maßzahl t groß ist (und tendenziell eher zur Entscheidung für einen „signifikanten“ Zusammenhang führt), wenn der Mittelwert-Unterschied groß und/oder die Streuung gering und/oder die Stichprobe groß ist. Auch hier wird die Entscheidung wieder so getroffen, dass man die ermittelte Maßzahl t mit einer entsprechenden Wahrscheinlichkeitsverteilung (näherungsweise Normalverteilung) vergleicht und seine Entscheidung trifft. Daher gelten – ähnlich wie beim χ^2 -Test – folgende Grundsätze:

- Wenn ein t-Wert so klein ist, dass sein Zustandekommen mit großer Wahrscheinlichkeit zufällig erfolgt sein kann, dann lehnt man die Hypothese eines systematischen Unterschieds zwischen den Mittelwerten ab.
- Wenn ein t-Wert so groß ist, dass sein Zustandekommen nur mit geringer Wahrscheinlichkeit durch Zufälligkeiten der Stichprobenziehung zu erklären ist, dann geht man von einem systematischen Unterschied aus.

Analog ist die Vorgehensweise, wenn man zwei Mittelwerte aus verschiedenen Grundgesamtheiten (z. B. Nutzer und Nichtnutzer eines Produktes) vergleichen will. Wieder betrachtet man die Differenz der beiden Stichprobenmittelwerte $\bar{x}_1$ und $\bar{x}_2$ und setzt sie in Beziehung zur entsprechenden Streuung. Es ergibt sich

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_{x_1}^2}{n_1} + \frac{s_{x_2}^2}{n_2}}}$$

Auch hier wird die Entscheidung über die Signifikanz des Mittelwertunterschiedes mithilfe einer entsprechenden Wahrscheinlichkeitsverteilung (Studentverteilung mit n1 + n2 − 2 Freiheitsgraden) getroffen.

Beispiel

Das folgende Beispiel verdeutlicht die Abhängigkeit der Signifikanz vom Stichprobenumfang:

In Rahmen eines Markttests (siehe hierzu auch Abschn. 6.8) soll überprüft werden, ob sich das Ausprobieren eines Produkts positiv auf die Kaufbereitschaft auswirkt. Dazu wird die Kaufbereitschaft vor und nach dem Produkttest bei denselben zufällig ausgewählten Befragten auf einer sieben-stufigen Ratingskala erhoben. Im

Mittel ergibt sich nun z. B. eine um 0,2 Skalenpunkte höhere Kaufbereitschaft nach dem Ausprobieren. Es stellt sich die Frage, ob diese Verbesserung signifikant ist oder nur zufällig in dieser Stichprobe aufgetreten ist. Bei einer gegebenen sehr großen Grundgesamtheit ergeben sich für verschiedene Stichprobengrößen ganz unterschiedliche p-Werte:

p-Werte für verschiedene Stichprobenumfänge

Stichprobenumfang	p-Wert
n = 100	0,11
n = 150	0,06
n = 200	0,04

Im Fall von einer Stichprobengröße von n = 100 und n = 150 ist die Verbesserung der durchschnittlichen Kaufbereitschaft nicht signifikant, da sich p-Werte größer als 0,5 ergeben. Für n = 200 ist sie dagegen signifikant. Bei einem größeren Stichprobenumfang ist es also wahrscheinlicher, dass ein Effekt signifikant ist, als bei einem geringeren Stichprobenumfang (siehe hierzu auch die Ausführungen zur Effektstärke in Abschn. 8.2.1). ◀

Zur Berechnung der Effektstärke bei Mittelwertunterschieden steht eine Reihe von Maßzahlen, wie etwa Cohen's d oder auch Hedges' g zur Verfügung (siehe hierzu auch Cohen, 1988), die in der Regel auch direkt von statistischen Softwarepaketen (wie SPSS) ausgegeben werden. Auf der Website www.psychometrika.de/effektstaerke.html findet sich außerdem eine deutschsprachige Übersicht sowie die Möglichkeit zur direkten Berechnung dieser und weiterer Maßzahlen.

Hintergrundinformation

Die vorstehend skizzierten Überlegungen zu den Schlussweisen bei statistischen Tests setzen voraus, dass diese nur auf eine (sehr) begrenzte Zahl von Hypothesen angewendet werden, die *vor* einer entsprechenden Untersuchung formuliert und begründet wurden. Bei den meisten Untersuchungen der Marktforschung wird gleichzeitig (z. B. in einem Fragebogen) eine größere Zahl von Variablen erhoben, die wiederum – rein formal und ohne substanzielle Begründung – die Bestimmung von hunderten von *statistischen* Zusammenhängen (z. B. Korrelationen) ermöglichen. Vor diesem Hintergrund kann es für unseriöse Forscher verlockend sein, aus einer Vielzahl – heutzutage leicht „per Knopfdruck" zu erstellender – Ergebnisse entsprechender Tests die auszusuchen, die *scheinbar* (!) signifikant sind und sich dazu mehr oder weniger passende Hypothesen nachträglich auszudenken. Man spricht bei einem solchen Vorgehen vom **„HARKing"**: *H*ypothesizing *A*fter the *R*esults are *K*nown (Kerr, 1998). Die Versuchung zu solch einem Verhalten ist nicht gering, weil in der Wissenschaft (wenn auch nur scheinbar) signifikante Ergebnisse leichter zu publizieren sind und auch in der Praxis finden (scheinbar) signifikante Zusammenhänge und Unterschiede größeres Interesse.

Wo liegt das Problem? Bei einer sehr großen Zahl (z. B. mehrere hundert) von – beispielsweise – Korrelationen finden sich fast immer einige relativ hohe (→ scheinbar „signifikante") Werte, die nicht durch einen substanziellen Zusammenhang zwischen zwei Variablen zustande

gekommen sind, sondern durch Zufälligkeiten (z. B. Ausreißer, siehe Abschn. 7.3.2) bei den Daten. Die Interpretation solcher Zufallsergebnisse ist somit ohne wissenschaftliche Substanz und eher irreführend. Kerr (1998, S. 205) spricht davon, dass durch HARKing Fehler 1. Art (siehe Abschn. 8.2.1) „in Theorie übersetzt" werden. Weiterhin ist es leicht nachvollziehbar, dass auf solche Weise zustande gekommene Post-hoc „Hypothesen" keinen Sinn haben, weil es sich bei bereits vorhandenen Ergebnissen nicht mehr um Vermutungen handelt und eine Falsifizierung solcher (Schein-) Hypothesen natürlich nicht mehr möglich ist. Einige Einzelheiten zum HARKing-Problem findet man bei Eisend und Kuß (2017, S. 165 ff.) und vor allem in dem grundlegenden Aufsatz von Kerr (1998).

Die Schlussweise bei statistischen Tests ist also nur sinnvoll, wenn man diese auf eine *sehr begrenzte Zahl von Hypothesen* anwendet, die vor dem Vorliegen der Ergebnisse formuliert und begründet wurden.

Literatur

Benesch, T. (2012). *Schlüsselkonzepte zur Statistik*. Springer.
Bortz, J., & Döring, N. (1995). *Forschungsmethoden und Evaluation*. Springer.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Erlbaum.
De Vaus, D. (2002). *Analyzing social science data*. Sage.
Eisend, M., & Kuß. (2017). *Grundlagen empirischer Forschung*. Springer Gabler.
Jaccard, J., & Becker, M. (2002). *Statistics for the behavioral sciences* (4. Aufl.). Wadsworth.
Kerr, N. L. (1998). HARKing: Hypothesizing after results are known. *Personality and Social Psychology Review, 2*(3), 196–217.
Lehmann, D., Gupta, S., & Steckel, J. (1998). *Marketing research*. Addison Wesley Pub Co Inc.
Sawyer, A. G., & Peter, J. P. (1983). *The significance of statistical significance tests in marketing research*. *Journal of Marketing Research, 20,* 122–133.
Sedlmaier, P., & Renkewitz, F. (2008). *Forschungsmethoden und Statistik in der Psychologie*. Addison-Wesley.
Voß, W. (Hrsg.). (2000). *Taschenbuch der Statistik* (S. 338). Carl Hanser Verlag.

Multivariate Analyseverfahren

9

Zusammenfassung

Im Marketing hat man es in der Regel mit komplexen Zusammenhängen zwischen zahlreichen Variablen zu tun. So sind Aspekte des Konsumentenverhaltens (z. B. Markenwahl, Art von Bedürfnissen) kaum durch nur eine Variable erklärbar und der Erfolg oder Misserfolg eines Produkts hängt niemals von nur einem Einflussfaktor (z. B. Werbebudget oder Preis) ab. Dem entsprechend spielen im Marketing die sogenannten „multivariaten Analyseverfahren", die eben dazu geeignet sind, eine große Zahl von Variablen gleichzeitig zu analysieren, seit langem eine wesentliche Rolle. Hier gibt es Verfahren („Dependenzanalysen"), die darauf ausgerichtet sind, eine „abhängige" Variable durch eine gewisse Zahl „unabhängiger" Variablen zu erklären, beispielsweise den Marktanteil eines Produkts durch Werbebudget, Preis, Kaufkraft der Zielgruppe, relative Produktqualität, usw. Bei anderen multivariaten Verfahren, den sogenannten „Interdependenzanalysen", stehen Zusammenhänge zwischen einer größeren Zahl von Variablen im Vordergrund. Im Einzelnen werden in diesem Kapitel die grundlegenden Ideen, die wesentlichen Anwendungsvoraussetzungen und Anwendungsmöglichkeiten der Verfahren Varianzanalyse, Regressionsanalyse, logistische Regressionsanalyse, Conjoint-Analyse, Faktoranalyse, Strukturgleichungsmodelle und Clusteranalyse charakterisiert. Das Kapitel schließt mit einem Ausblick zu der Nutzung künstlicher Intelligenz in der Marktforschung.

9.1 Überblick

Multivariate Analyseverfahren, die im vorliegenden Kapitel vorgestellt werden, gehören inzwischen zum methodischen Standard der Marketingforschung in Wissenschaft und Praxis. Der wichtigste Grund dafür liegt darin, dass Phänomene des Marketing-

H. Kreis et al., *Marktforschung,* https://doi.org/10.1007/978-3-658-44456-3_9

Bereichs typischerweise so komplex sind, dass die gleichzeitige Analyse einer größeren Zahl von Variablen zu deren Erklärung notwendig ist. Hinzu kommt die breite Verfügbarkeit entsprechender Software in den gängigen kommerziellen Programmpaketen zur statistischen Datenanalyse (z. B. SPSS, SAS/JMP, Stata) sowie die rasant gewachsene Rechnerkapazität, wodurch die bis vor wenigen Jahrzehnten äußerst aufwendigen entsprechenden Berechnungen auf jedem gängigen Computer leicht und schnell durchgeführt werden können. Zum gegenwärtigen Zeitpunkt stehen außerdem eine Reihe von kostenfrei Softwarealternativen zur Verfügung. Besondere Beachtung verdient sicherlich das R-Programmsystem (www.r-project.org), das an SPSS angelehnte in der Anwendung eingeschränkte Jamovi (www.jamovi.org) sowie das in MS-Office enthaltene MS-Excel.

Gleichwohl ist für eine Anwendung multivariater Analyseverfahren das Verständnis der Methoden einschließlich ihrer Anwendungsvoraussetzungen und Interpretationsmöglichkeiten natürlich unabdingbar.

Die wohl gängigste Einteilung der multivariaten Verfahren ist die in Dependenz-Analyse und Interdependenz-Analyse. Bei der **Dependenz-Analyse** wird die Menge der untersuchten Variablen eingeteilt in (so genannte) *abhängige und* (so genannte) *unabhängige Variablen*. Man versucht dann festzustellen, ob ein vermuteter Zusammenhang zwischen unabhängigen und abhängigen Variablen anhand des Datenmaterials bestätigt werden kann. Ist das der Fall, dann spricht man davon, dass die abhängige durch die unabhängigen Variablen (in einem gewissen Ausmaß) erklärt werden kann. Eine solche Aussage bedeutet aber nicht unbedingt, dass ein Kausalzusammenhang vorliegt. Die Annahme von Kausalbeziehungen ist an wesentlich strengere Voraussetzungen gebunden (vgl. Abschn. 2.4.1). Ferner soll die Art des Zusammenhangs durch die Schätzung von Parametern eines entsprechend vorab bestimmten Modells beschrieben werden.

Bei der **Interdependenz-Analyse** geht man von einer ungeteilten Variablenmenge aus. Man verzichtet also auf die Untersuchung von Beziehungen zwischen unabhängigen und abhängigen Variablen und beschränkt sich auf die Feststellung von *Zusammenhängen zwischen Variablen* (explorative Faktorenanalyse) oder *Ähnlichkeiten von Objekten* (Clusteranalyse). Eine Übersicht mit Beispielen für die verschiedenen Arten multivariater Verfahren findet sich in Tab. 9.1.

Im vorliegenden Kapitel werden zunächst (Abschn. 9.2) die Varianzanalyse und die Regressionsanalyse (Abschn. 9.3) dargestellt. Diese beiden Formen des allgemeinen linearen Modells sind in der Markt- und Sozialforschung besonders stark verbreitet und relativ leicht nachvollziehbar. Das größere Gewicht der Regressions- und Varianzanalyse in der vorliegenden Darstellung im Vergleich zu anderen Verfahren ist auch dadurch zu erklären, dass bei Letzteren häufig auf Grundideen des linearen Modells zurückgegriffen wird. In den darauffolgenden Abschnitten werden einige weitere gängige und wichtige multivariate Methoden dargestellt.

Auch im vorliegenden Kapitel liegt der Schwerpunkt wieder bei einer möglichst leicht verständlichen Darstellung grundlegender Ideen der verschiedenen Verfahren in Verbindung mit der Erläuterung der wesentlichsten Anwendungsvoraussetzungen und Anwendungsmöglichkeiten. Für eine verständige eigene Anwendung der verschiedenen

Tab. 9.1 Einteilung multivariater Analyseverfahren

Art des multivariaten Analyseverfahrens	Beispiele entsprechender Analyseverfahren
Dependenz-Analyse	Varianzanalyse (Abschn. 9.2)
	Regressionsanalyse (Abschn. 9.3)
	Logistische Regressionsanalyse (Abschn. 9.4)
	Conjoint-Analyse (Abschn. 9.5)
Interdependenz-Analyse	Faktorenanalyse (Abschn. 9.6)
	Strukturgleichungsmodelle (Abschn. 9.7)
	Clusteranalyse (Abschn. 9.8)

Methoden ist sicher ein tieferes Eindringen in deren theoretische Grundlagen notwendig. Die Behandlung der entsprechenden Einzelheiten würde den Rahmen des vorliegenden einführenden Lehrbuchs sprengen. Es wird deshalb auf die entsprechende Spezialliteratur verwiesen (siehe dazu auch die Literaturempfehlungen am Ende dieses Kapitels).

Als wesentliches Kennzeichen multivariater Verfahren gilt (nicht ganz überraschend) die gleichzeitige Analyse einer größeren Zahl ($\gg 2$) von Variablen. Dennoch wird im Folgenden gelegentlich auf den bivariaten Fall mit einer abhängigen und nur einer unabhängigen Variablen Bezug genommen, weil sich manche Gesichtspunkte dabei leichter (auch graphisch) erläutern lassen. Die Übertragung der entsprechenden Ideen auf den multivariaten Fall stellt dann kein entscheidendes Problem mehr dar.

9.2 Varianzanalyse

9.2.1 Grundidee und Voraussetzungen der Varianzanalyse

Die Varianzanalyse (auch: „Analysis of Variance“, kurz ANOVA) kann als eine Erweiterung des zuvor erörterten t-Tests bei unabhängigen Stichproben verstanden werden, bei der es möglich ist, auch mehr als zwei Gruppenmittelwerte miteinander zu vergleichen. Die Varianzanalyse kann also verwendet werden, um zu überprüfen, ob zwischen zwei oder mehr Mittelwerten ein signifikanter Unterschied besteht. Es handelt sich gleichsam allerdings um ein lineares Modell, in dem (substanzwissenschaftlich) zwischen abhängigen und unabhängigen Variablen unterschieden wird. Folglich gibt die Varianzanalyse Aufschluss über den Effekt einer oder mehrerer nominal skalierter unabhängiger Variablen (Gruppenzugehörigkeit) auf eine mindestens intervallskalierte abhängige Variable. Deswegen eignet sich die Varianzanalyse besonders zum Vergleich zwischen Gruppen (Gruppenzugehörigkeit als nominalskaliertes Merkmal), wodurch sich wiederum deren Anwendung zur Auswertung von **Experimenten** erklärt, wo ja Vergleiche zwischen Messwerten aus Versuchs- und Kontrollgruppen vorgenommen werden müssen (Abschn. 6.1).

Eine Varianzanalyse könnte also beispielsweise im Rahmen von Werbemitteltests (siehe auch Kap. 6) Anwendung finden. Wenn zum Beispiel eine Lebensmitteleinzelhandelskette die Absatzwirkung einer Promotionmaßnahme (z. B. die Einrichtung eines Probierstands) untersuchen möchte, könnten dieser in ausgewählten Filialen eingesetzt werden. Nach Ablauf der Testphase würde man dann den Absatz mit Probierstand und ohne (Kontrollgruppe) vergleichen.

Im Folgenden wird zunächst eine solche Varianzanalyse mit einer unabhängigen Variablen skizziert, die eine relativ einfache Erläuterung der grundlegenden Ideen ermöglicht. Man spricht in einem solchen Fall von einer „einfaktoriellen Varianzanalyse". Auf den Fall mehrerer unabhängiger Variabler („mehrfaktorielle Varianzanalyse"), der eine multivariate Analyse ermöglicht, wird am Ende dieses Abschnitts eingegangen. Die sprachliche Analogie zu („faktoriellen") **experimentellen Designs** (siehe Kap. 6) ist natürlich nicht zufällig, sondern unterstreicht den Zusammenhang dieses Untersuchungsdesigns mit der Varianzanalyse. Man geht bei der (einfaktoriellen) Varianzanalyse von folgendem Grundmodell aus:

$$y_{ij} = \mathrm{GM} + a_i + e_{\mathrm{ij}}$$

mit

y_{ij}:	Messwert der abhängigen Variablen für die Beobachtung j, die zur Gruppe i gehört
GM:	Gesamtmittelwert (über alle Messwerte der Untersuchung)
a_i:	Wirkung der Zugehörigkeit zur Gruppe i (→ unabhängige Variable)
e_{ij}:	Fehler („error"), der angibt, inwieweit der Messwert y_{ij} von dem für i typischen Wert abweicht

Bei dem Grundmodell ist leicht erkennbar, dass man die einzelnen Messwerte der abhängigen Variablen gedanklich in die Komponenten Gesamtmittelwert, Mittelwert der jeweiligen Gruppe und Abweichung des Einzelfalls vom Gruppenmittelwert zerlegt. Daran wird gleich angeknüpft, wenn es um die Analyse erklärter und unerklärter Abweichungen geht.

Bei der Varianzanalyse wird (daher kommt ihr Name!) zwischen der erklärten und unerklärten Varianz der abhängigen (intervallskalierten) Variablen unterschieden. Der Einfluss der unabhängigen Variablen (Gruppenzugehörigkeit) wird anhand der *Relation zwischen erklärter Varianz und unerklärter Varianz* beurteilt. Die Grundidee sei mithilfe der Abb. 9.1 illustriert.

In beiden Teilen der Abb. 9.1 sind Verteilungen der abhängigen Variablen Y und ihre Mittelwerte eingetragen. Diese sind für jeweils zwei getrennte Gruppen (A und B bzw. C und D) sowie für den Gesamt-Datensatz dargestellt. Man erkennt sofort, dass im linken Teil der Abbildung die Messwerte deutlich stärker streuen (größere Varianz haben). Der Gesamtmittelwert, die Gruppenmittelwerte sowie die Abweichungen der Gruppen-Mittelwerte vom Gesamt-Mittelwert entsprechen aber denen im rechten Teil der Graphik.

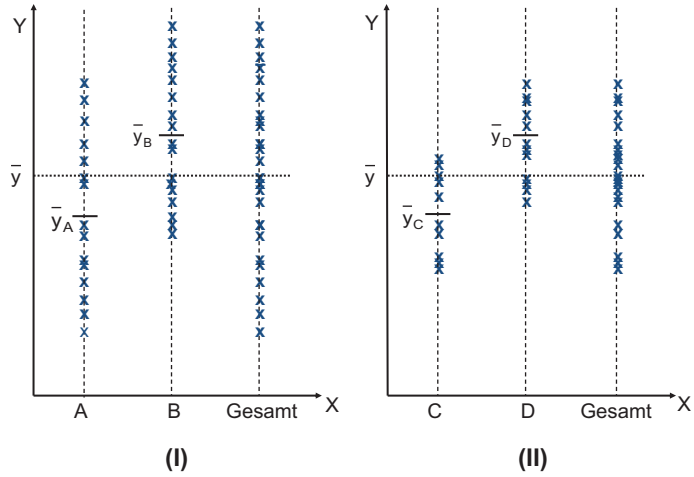


Abb. 9.1 Beispiel für unterschiedliche Varianzen bei verschiedenen Gruppen

Offenbar werden im rechten Teil der Graphik die (relativ geringen) Abweichungen der einzelnen Messwerte vom Gesamt-Mittelwert durch die unabhängige Variable "Gruppenzugehörigkeit" in größerem Ausmaß erklärt, als das bei der relativ großen Varianz der Messwerte im linken Teil der Abbildung der Fall ist. Man spricht auch davon, dass im links dargestellten Datensatz die Varianz *zwischen* den Gruppen klein ist im Vergleich zur Varianz *innerhalb* der Gruppen. Obwohl im Fall II die Abweichungen der Gruppen-Mittelwerte voneinander und vom Gesamt-Mittelwert genauso groß sind wie im Fall I, würde man hier (im Fall II) weit eher davon ausgehen, dass die unabhängige Variable einen systematischen Einfluss auf die Variable Y hat. Im links abgebildeten Datensatz könnte man nicht so eindeutig sagen, ob ein systematischer Unterschied zwischen den Gruppen oder die allgemein recht starken Schwankungen der Messwerte den Unterschied der Gruppen-Mittelwerte verursachen.

Hier einige „Faustregeln" zum Begriff der „erklärten Varianz"

- **Worum geht es?** Die Gründe für die Unterschiedlichkeit von Werten (→ Varianz) der abhängigen Variablen sollen gefunden werden.
- **Was bedeutet „erklärte Varianz"?** Je besser bzw. genauer man die Varianz der abhängigen Variablen – also die Abweichungen der einzelnen Beobachtungen vom Mittelwert für alle Beobachtungen – erklären kann, desto mehr weiß man offenbar über die Einflüsse auf die abhängige Variable.
- **Welche praktische Relevanz hat die „erklärte Varianz"?** Je mehr man über das Zustandekommen der abhängigen Variablen weiß, desto besser kann man diese schätzen bzw. prognostizieren bzw. beeinflussen.

Eine der *zentralen Ideen* der Varianzanalyse besteht also darin, Varianzen der abhängigen Variablen innerhalb der Gruppen mit Varianzen zwischen den Gruppen (→ Abweichungen der Gruppen-Mittelwerte vom Gesamt-Mittelwert) zu vergleichen. Wenn die Varianz zwischen den Gruppen im Vergleich zur Varianz innerhalb der Gruppen groß ist, dann spricht das für einen deutlichen Einfluss der unabhängigen Variablen, die ja die Gruppenzugehörigkeit (z. B. Versuchs- oder Kontrollgruppe) bestimmt. Abb. 9.2 illustriert diesen Ansatz.

Die Grundidee der Varianzanalyse, den Einfluss unabhängiger Variablen auf eine abhängige Variable an Hand der durch die unabhängigen Variablen erklärten Anteile der Gesamtvarianz zu beurteilen, sei mithilfe der Abb. 9.3 zusätzlich verdeutlicht.

In der Abbildung sind die Mittelwerte für den gesamten Datensatz $\bar{y}$ und für die beiden Teilgruppen A und B, $\bar{y}_A$ und $\bar{y}_B$, eingetragen. Ferner findet man dort zwei Beispiele für Messwerte aus den Gruppen A und B,y_{Ai} und y_{Bi}, die beide deutlich von $\bar{y}$ abweichen. Weiterhin ist angegeben, welcher Teil dieser Abweichungen durch die Zugehörigkeit zu den Gruppen A bzw. B erklärt wird.

Diese Betrachtung der Anteile erklärter und nicht erklärter Abweichungen einzelner Messwerte vom Mittelwert wird auf die entsprechende *Zerlegung der Gesamtabweichungen* in erklärte und nicht erklärte Abweichungen übertragen. Hier sei darauf hingewiesen, dass man im Zusammenhang mit der Varianzanalyse meist die Summen quadrierter Abweichungen vom Mittelwert betrachtet, die sich von der Varianz nur dadurch unterscheiden, dass keine Division durch die jeweilige Fallzahl (genauer: N – 1)

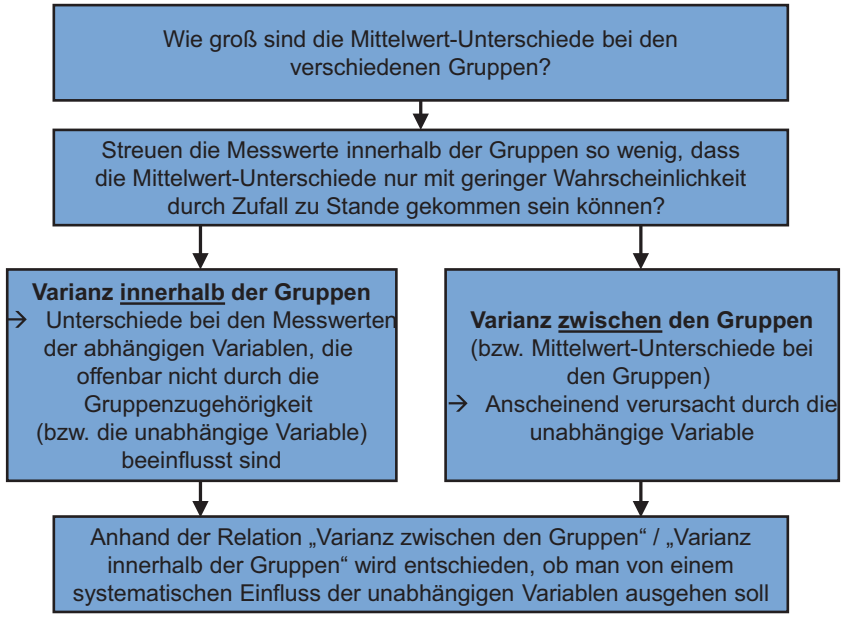


Abb. 9.2 Illustration der Grundidee der Varianzanalyse

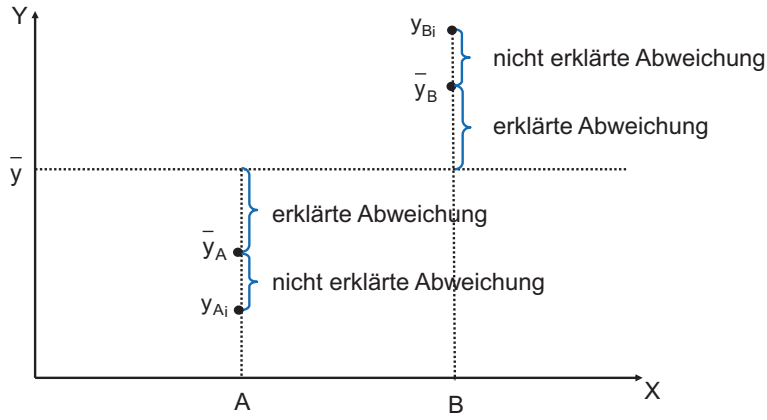


Abb. 9.3 Beispiel zur Zerlegung in erklärte und nicht erklärte Abweichungen von Mittelwerten. (Nach Backhaus et al., 2016, S. 180)

vorgenommen wird. Die Zerlegung der Gesamtabweichungen lässt sich formal recht einfach darstellen (vgl. z. B. Backhaus et al., 2016, S. 173 ff.; Jaccard & Becker, 2002, S. 329 ff.):

Gesamtabweichung = erklärte Abweichung + nicht erklärte Abweichung

Summe der quadrierten Gesamtabweichungen = (Summe der quadrierten Gesamtabweichungen zwischen den G Gruppen) + (Summe der quadrierten Abweichungen von jeweils K Messwerten innerhalb der G Gruppen)

$$\sum_{g=1}^{G}\sum_{k=1}^{K}\left(y_{gk}-\bar{y}\right)^2 = \sum_{g=1}^{G}K\left(\bar{y}_g-\bar{y}\right)^2 + \sum_{g=1}^{G}\sum_{k=1}^{K}\left(y_{gk}-\bar{y}_g\right)^2$$
$$\mathrm{AQ}_{g(\mathrm{esamt})} = \mathrm{AQ}_{z(\mathrm{wischen})} + \mathrm{AQ}_{i(\mathrm{nnerhalb})}$$

Dabei stehen AQ für die Abweichungsquadrate und die Indizes g, z und i für die entsprechenden Werte für den gesamten Datensatz, zwischen und innerhalb der Gruppen.

Vor der Analyse und Interpretation der Schätzergebnisse sind folgende wichtige Anwendungsvoraussetzungen für die Varianzanalyse zu überprüfen:

- Die jeweiligen Fehlervarianzen müssen normalverteilt sein, was gleichzeitig eine Normalverteilung der Messwerte in der jeweiligen Grundgesamtheit voraussetzt.
- Die Fehlervarianzen müssen zwischen den Gruppen gleich bzw. homogen sein (Homoskedastizität).
- Die Messwerte müssen unabhängig voneinander sein.

Wenn ein großer Teil der Gesamtvarianz durch die Gruppenzugehörigkeit erklärt wird, wenn also die verschiedenen Gruppen relativ homogen sind und sich relativ deutlich bei ihren Mittelwerten unterscheiden, dann geht man eher davon aus, dass die unabhängigen

Variablen einen systematischen Einfluss auf die abhängige Variable haben. Für eine entsprechende Entscheidung bedient man sich (ebenfalls analog zur Regressionsanalyse) des *F-Tests*. Nach der Grundidee dieses statistischen Tests wird der Anteil erklärter Varianz zum Anteil nicht erklärter Varianz in Beziehung gesetzt. Je größer dieser F-Wert ist, desto weniger wahrscheinlich ist es, dass die gefundene Beziehung durch Zufall zu erklären ist. Es geht also um die Relation von (durch die unabhängige(n) Variable(n)) erklärten und nicht erklärten Abweichungen. Dazu berechnet man zunächst die mittleren quadratischen Abweichungen, die die Abweichungsquadrate anhand der Anzahl der Gruppen (G) und der Messwerte (K) normieren:

$$MAQ_g = \frac{AQ_g}{G \times K - 1}$$

$$MAQ_z = \frac{AQ_z}{G - 1}$$

$$MAQ_g = \frac{AQ_i}{G \times (K - 1)}$$

der Messwerte (K) normieren:

Die jeweils durch „−1“ ausgedrückten (meist geringen) Abweichungen von den entsprechenden Fallzahlen ergeben sich durch die Festlegung der entsprechenden Freiheitsgrade, zu deren Berechnung auf die weiterführende Literatur verwiesen werden muss. Der empirische F-Wert lässt sich jetzt durch

$$F_{\text{emp}} = \text{MAQ}_z / \text{MAQ}_i$$

bestimmen. Dieses ist also die Relation von erklärter („zwischen den Gruppen“) und unerklärter („innerhalb der Gruppen“) Abweichung. Der Vergleich dieses empirischen **F-Wertes** mit einem theoretischen F-Wert für die entsprechende Sicherheitswahrscheinlichkeit und die jeweilige Zahl von Variablen und von Fällen zeigt, ob man von einem signifikanten Einfluss der unabhängigen Variablen auf die abhängige Variable sprechen kann. In der praktischen Anwendung nimmt die Statistik-Software einen solchen Vergleich automatisch vor.

Beispiel

Sedlmeier und Renkewitz (2008, S. 446) erklären das Funktionsprinzip der Varianzanalyse an einem sehr plastischen Beispiel:

Bei einer unveränderten Anzahl von Zähler- und Nennerfreiheitsgraden gilt stets: Je größer die systematische Variation im Vergleich zur Fehlervariation ausfällt, umso größer wird der F-Wert und umso eher können wir schließen, dass in der Population ein Effekt vorhanden ist. Diese Vorgehensweise der Varianzanalyse ist vergleichbar mit dem Versuch, ein Signal (beispielsweise die Worte, die ein Freund an Sie richtet) vor dem Hintergrund eines Rauschens (die Geräuschkulisse in ihrem Lieblings-Club)

zu entdecken. Je deutlicher das Signal das Rauschen übertrifft, umso leichter wird es erkennbar sein (Ihr Freund wird in besagtem Lieblings-Club sehr viel lauter sprechen müssen als in einer ruhigen Umgebung, um für Sie verständlich zu sein). ◀

Das beliebteste Maß der *Effektstärke* für Varianzanalysen ist das sogenannte Eta-Quadrat das den Anteil an aufgeklärter Varianz an der Gesamtvarianz darstellt. Dieser Anteil lässt sich auch in Cohen's d umrechnen, in der Regel wird aber lediglich der Anteil erklärter Varianz des Modells berichtet (Cohen, 1988, S. 273 ff.).

Das Grundprinzip der Varianzanalyse lässt sich (natürlich) auch auf Modelle mit mehreren unabhängigen Variablen übertragen. Deswegen wird diese Form der Varianzanalyse, die **mehrfaktorielle Varianzanalyse,** den multivariaten Verfahren zugerechnet. Damit hat man auch die Möglichkeit, nicht nur Effekte einzelner unabhängiger Variabler zu untersuchen, sondern auch deren **Interaktionen.** Das eingangs dieses Abschnitts dargestellte Grundmodell würde sich für den einfachsten Fall einer mehrfaktoriellen Varianzanalyse, der zweifaktoriellen Varianzanalyse, folgendermaßen darstellen:

$$y_{ijk} = GM + a_i + b_j + (ab)_{ij} + e_{ijk}$$

mit

y_{ijk}:	Messwert der abhängigen Variablen für die Beobachtung k, die zur Gruppe i (bzgl. der unabhängigen Variablen a) und zur Gruppe j (bzgl. der unabhängigen Variablen b) gehört
GM:	Gesamtmittelwert (über alle Messwerte der Untersuchung)
a_i:	Wirkung der Zugehörigkeit zur Gruppe i (→ unabhängige Variable a)
b_j:	Wirkung der Zugehörigkeit zur Gruppe j (→ unabhängige Variable b)
$(ab)_{ij}$:	Interaktionswirkung der Zugehörigkeit zu den Gruppen i (Variable a) und j (Variable b)
e_{ijk}:	Fehler („error"), der angibt, inwieweit der Messwert y_{ijk} von dem für die gleichzeitige Zugehörigkeit zu den Gruppen i bzw. j typischen Wert abweicht

Bei der entsprechenden Analyse geht es wieder um die Relation der Abweichungen der Gruppenmittelwerte (bezogen auf die einzelnen unabhängigen Variablen und deren Interaktionen) vom Gesamtmittelwert zu den Abweichungen innerhalb der Gruppen vom jeweiligen Mittelwert. Die Schlussweise ist analog zur einfaktoriellen Varianzanalyse.

Varianzanalysen lassen sich auch für mehr als zwei unabhängige Variablen durchführen, wobei die Interpretation von Drei- oder Vierfachinteraktionen nicht immer ganz einfach ist. Besonders berücksichtigt werden müssen auch wiederholte Messungen, etwa wenn die Erinnerung an eine Werbung nach einem und nach drei Tagen bei den gleichen Versuchspersonen erhoben wird. Dies geschieht im Rahmen der im Abschn. 6.1 schon angesprochenen „within subjects" Designs (bzw. **„repeated designs"**). Details zu diesen technischen Varianten der Varianzanalyse finden sich in der entsprechenden Literatur, z. B. bei Cortina und Nouri (2000) oder Jaccard und Becker (2002). Darüber hinaus kön-

nen auch mehrere abhängige Variable gleichzeitig berücksichtigt werden (so genannte **„Multivariate Analysis of Variance“**, kurz MANOVA).

Mit Blick auf die Effektstärke kommt bei komplexeren Varianzanalysen mit mehreren unabhängigen Variablen neben dem bereits erwähnten Eta-Quadrat auch das partielle Eta-Quadrat hinzu, welches den Anteil an Varianzerklärung durch jeweils eine unabhängige Variable umfasst. Gegebenenfalls besteht dabei jedoch das Problem einer Überschätzung der erklärten Varianz durch die partielle Maßzahl. Näheres hierzu findet sich bei Bortz (2005). Für eine einfaktorielle ANOVA sind Eta-Quadrat und partielles Eta-Quadrat identisch.

9.2.2 Varianzanalyse mit Kovariaten

Bei der Varianzanalyse ist (sind), wie oben bereits erläutert, die abhängige(n) Variable(n) mindestens intervallskaliert, während die unabhängigen Variablen in der Regel nominalskaliert sind. Die Berücksichtigung von intervallskalierten Variablen, sogenannten **Kovariaten,** ist allerdings trotzdem möglich und wird in der Literatur auch Kovarianzanalyse genannt. Die Grundidee einer solchen **„Analysis of Covariance“** (ANCOVA) soll mit der Abb. 9.4 (Venndiagramm) veranschaulicht werden.

In Abb. 9.4 werden die Erklärungsbeiträge zur Varianz der abhängigen Variablen illustriert. Hier wird die abhängige Variable (AV) durch zwei unabhängige nominalskalierte Variablen (UV1 und UV2) sowie durch eine metrisch skalierte Kovariate (K) erklärt. UV1 erklärt einen relativ großen Anteil (a) an Varianz im Vergleich zu UV2 (b). Die Schnittmenge (c) von UV1 und UV2 mit AV illustriert einen Interaktionseffekt. In Bezug auf die Kovariate wird deutlich, dass K zusätzliche Varianz der AV erklärt und somit im Vergleich zu einer herkömmlichen ANOVA, in der nur UV1 und UV2 berücksichtigt würden, sich die Fehlervarianz (erheblich) verringert. Um dies zu gewährleisten sind an die Kovariate zwei Anforderungen zu stellen:

- Die Kovariate muss mit der abhängigen Variablen in Beziehung stehen (korrelieren). Ist dies nicht der Fall, entspricht das Ergebnis der ANCOVA dem der ANOVA.

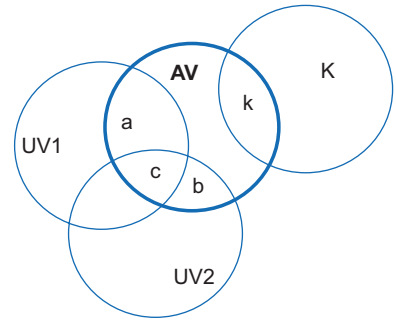


Abb. 9.4 Grundidee der ANCOVA

- Die Kovariate darf nicht in Beziehung zu den unabhängigen Variablen stehen, da sonst die Varianzerklärung der unabhängigen Variablen beeinflusst bzw. eliminiert würde, wodurch sich eine verzerrte Lösung ergeben würde.

In dem zu Beginn des Kapitels kurz charakterisierten Beispiel zum Werbemitteltest eines Probierstandes könnte man sich als sinnvolle Kovariate beispielsweise die Filialgröße (m^2) vorstellen. Sicherlich hat die Größe der Ladenfläche einen Einfluss auf den Absatz in den Filialen und kann somit zusätzliche (Fehler-) Varianz erklären. Damit jedoch keine Verzerrungen auftreten, muss sichergestellt sein, dass die Filialgröße (Kovariate) unabhängig von der unabhängigen Variable „Probierstand" ist. Der Händler sollte also die Testgeschäfte zufällig oder zumindest unabhängig von deren Fläche ausgewählt haben.

Das Ziel der Aufnahme von einer (oder auch mehreren) Kovariaten ist es also die Fehlervarianz zu reduzieren, was genauere Angaben zum Effekt der abhängigen Variablen möglich macht.

9.3 Regressionsanalyse

9.3.1 Grundidee und Ablauf der Regressionsanalyse

Die Regressionsanalyse dürfte wohl das am stärksten etablierte Verfahren der multivariaten Datenanalyse sein. So wie die Varianzanalyse aufgrund des geforderten Messniveaus (nominale unabhängige Variablen und intervallskalierte abhängige Variable) in die methodische Umgebung des t-Tests eingeordnet wurde, lässt sich die Regressionsanalyse mit dem Konzept der Korrelation in Verbindung bringen, da hier die Beziehung von Variablen untersucht werden, die mindestens intervallskaliert sind. Der Begriff „Regression" (Zurückführung) kennzeichnet schon die zentrale Idee des Verfahrens: Die unterschiedlichen Werte einer abhängigen Variablen (z. B. Einkommen) sollen zurückgeführt werden auf andere (unabhängige) Variable (z. B. Ausbildungsdauer, Berufserfahrung). In diesem Sinne wird die abhängige Variable erklärt durch die unabhängigen bzw. erklärenden Variablen.

Mithilfe der Regressionsanalyse lassen sich derartige Zusammenhänge darstellen. Weiterhin kann ein empirisch bewährtes Regressionsmodell dazu dienen, bei Kenntnis von Werten der unabhängigen Variablen den sich voraussichtlich ergebenden Wert der abhängigen Variablen zu prognostizieren. Wenn man beispielsweise ein Regressionsmodell entwickelt hat, mit dem man den Zusammenhang von Ausbildungsdauer und Berufserfahrung auf der einen Seite und Einkommen auf der anderen Seite hinreichend exakt beschreiben kann, dann kann man dieses nutzen, nun auf der Basis von Angaben über Ausbildungsdauer und Berufserfahrung das zu erwartende Einkommen zu schätzen.

Beispiel

Backhaus et al. (2016, S. 67) heben folgende Anwendungsmöglichkeiten der Regressionsanalyse in der Marktforschung hervor:

- **Ursachenanalysen** Wie stark bzw. wie unterschiedlich sind die Wirkungen der unabhängigen Variablen auf die abhängige Variable? (z. B.: Wirken sich Preisänderungen oder Änderungen des Werbebudgets stärker auf den Marktanteil aus?)
- **Wirkungsprognosen** Welchen Wert erhält die abhängige Variable bei einer Änderung von unabhängigen Variablen? (z. B.: Welcher Marktanteil ist zu erwarten, wenn der Außendienst-Einsatz um 20 % gesteigert wird?)
- **Zeitreihenanalysen** Welche Entwicklung einer abhängigen Variablen ist (ceteris paribus) zu beobachten, wenn man als unabhängige Variable die Zeit verwendet? Welche Entwicklung zeigt sich, wenn man den bisher ermittelten Zusammenhang in die Zukunft projiziert? (z. B.: Wie wird sich der Weltmarktpreis für Kohle in den nächsten drei Jahren entwickeln, wenn sich der bisherige Trend fortsetzt?) ◀

Die Grundzüge der **Durchführung einer Regressionsanalyse** sollen im Folgenden in drei Schritten skizziert werden. Am Anfang steht die Formulierung des Regressionsmodells, d. h. die Auswahl der einzubeziehenden Variablen und die Festlegung der abhängigen Variablen. Es folgt die Schätzung der Parameter des Regressionsmodells auf der Grundlage der vorliegenden Daten. Anschließend wird überprüft, ob das geschätzte Regressionsmodell durch bestimmte Fehler und Zufallsschwankungen beeinflusst sein kann (siehe Abb. 9.5).

Formulierung des Regressionsmodells Am Beginn steht also die Formulierung des Regressionsmodells. Grundlage dafür sind zunächst und vor allem *substanzwissenschaft-*

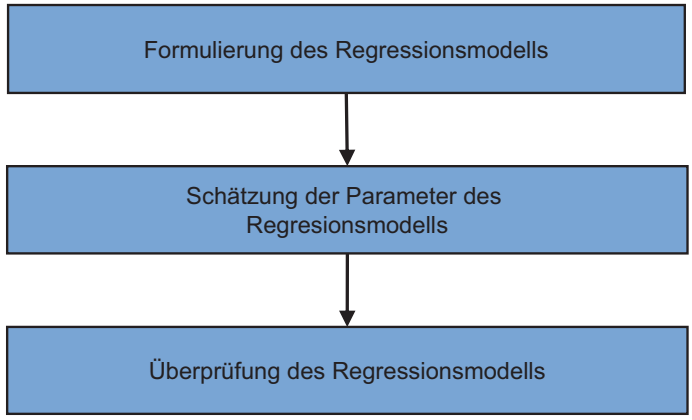


Abb. 9.5 Ablauf der Regressionsanalyse

liche Überlegungen. Auf der Basis theoretischer und empirischer Erkenntnisse sowie bisheriger Erfahrungen muss festgelegt werden, welche unabhängigen Variablen die interessierende (abhängige) Variable erklären könnten. Statistische Methoden sind entscheidend für die Schätzung und Überprüfung des Modells, nutzen aber bei dessen Formulierung wenig. Immerhin gilt es aber schon bei der Formulierung des später zu schätzenden Regressionsmodells zu prüfen, ob eine Regressionsanalyse bei den gegebenen Daten durchgeführt werden darf. Wie bei allen statistischen Verfahren gibt es auch bei der Regressionsanalyse eine Reihe von Anwendungsvoraussetzungen, deren Einhaltung gegeben sein muss, um zu aussagefähigen Ergebnissen zu gelangen. Grundsätzlich ist davon auszugehen, dass bei der Verletzung der Prämissen die Schätzwerte verzerrt sein können, d. h. die geschätzten Parameter streben nicht gegen den wahren Wert, und dass verzerrte Standardfehler auftreten, die zur Berechnung falscher Konfidenzintervalle und damit auch zu ungültigen Signifikanztests führen können. Umfassendere Diskussionen zu diesem Komplex finden sich bei Backhaus et al. (2016, S. 97 ff.), Fox (1997, S. 265 ff.) und Skiera und Albers (2008, S. 478 ff.) sowie in anderen einschlägigen Lehrbüchern. Hier eine überblicksartige Zusammenstellung der wichtigsten Prämissen:

- **Abhängige und unabhängige Variable sind mindestens intervallskaliert:** Diese Voraussetzung ist ganz elementar, weil Regressionskoeffizienten, die sich auf nominal- oder ordinalskalierte Daten beziehen, sinnlos wären. Eine Ausnahme bildet die sogenannte Dummy-Regression, siehe Abschn. 9.3.2. Zusätzlich gibt es Anwendungen der Regressionsanalyse im sozialwissenschaftlichen Bereich, in denen Variable verwendet werden, die an die Anforderungen einer Intervallskalierung (strenggenommen) nur angenähert sind, beispielsweise die im Abschn. 4.3.2 erläuterten Multi-Item-Skalen.
- **Es sollte ein linearer Zusammenhang zwischen den unabhängigen und der abhängigen Variablen bestehen:** Das heißt, dass der Zusammenhang proportional ist. Bei einer Veränderung einer unabhängigen Variablen um eine Einheit verändert sich der Wert der abhängigen Variablen immer um die gleiche (durch den entsprechenden Regressionskoeffizienten geschätzte) Zahl von Einheiten. Es dürfte unmittelbar einsichtig sein, dass ein lineares Modell (Regressions*gerade*) einen nichtlinearen Zusammenhang nur unzureichend wiedergeben kann.
- **Die Zahl der Beobachtungen („Fälle") ist größer als die Zahl der zu schätzenden Parameter (Regressionskoeffizienten und konstantes Glied):** Das absolute Minimum der Zahl erforderlicher Beobachtungen ist die Zahl der zu schätzenden Parameter (= Zahl der unabhängigen Variablen + konstantes Glied). Die üblichen Angaben über ein auf Basis einer Stichprobe geschätztes Modell für die entsprechende Grundgesamtheit (→ F-Wert, t-Wert) sind aber nur aussagekräftig, wenn die Zahl der Beobachtungen noch deutlich höher liegt.
- **Vollständigkeit des Modells:** Damit ist gemeint, dass in dem Modell alle unabhängigen Variablen enthalten sind, die gemäß der zu prüfenden theoretischen Vorstellung die abhängige Variable bestimmen (Berry, 1993, S. 364 ff.). Anderenfalls

wäre damit zu rechnen, dass die Wirkung der nicht berücksichtigten Variablen die Regressionskoeffizienten der im Modell berücksichtigten Variablen beeinflusst und somit verzerrt. Entscheidender Maßstab hinsichtlich der Vollständigkeit des Modells ist die Entsprechung von theoretischer Vorstellung und Spezifikation des Modells.

- **Keine perfekte Multikollinearität bei unabhängigen Variablen:** Erklärende (unabhängige) Variable sollen untereinander *nicht hoch korreliert* sein. Im Extremfall könnte es ansonsten sein, dass sich eine (unabhängige) Variable als Linearkombination aus anderen unabhängigen Variablen darstellen lässt (perfekte Multikollinearität). In derartigen Fällen lässt sich die Regressionsanalyse rechnerisch nicht durchführen. Aber auch ein hoher Grad an Multikollinearität kann zu Problemen führen, da sich die Streuungen der unabhängigen Variablen dann stark überschneiden und sich die Informationen aus dem Regressionsmodell nicht mehr eindeutig den Variablen zuordnen lassen.
- **Erwartungswert der Residuen ist Null:** Hier geht es nur darum, dass der Erwartungswert des Fehlers „e“ (siehe unten) bei 0 liegt, das heißt, dass sich die Abweichungen der beobachteten Werte gegenüber den geschätzten Werten insgesamt ausgleichen, was normalerweise durch die Kleinste-Quadrate-Schätzung sichergestellt ist, denn dabei werden ja positive und negative Abweichungen zwischen beobachteten und geschätzten Werten im Mittel ausgeglichen. Es kann aber vorkommen, dass die Messwerte der abhängigen Variablen durch einen systematischen Messfehler zu hoch oder zu niedrig liegen. Dies ist z. B. der Fall, wenn eine gekrümmte Kurve durch eine Gerade angenähert wird. Dann schlägt sich diese Verzerrung im Schätzwert des konstanten Glieds nieder, der aber bei der Interpretation der Regressionsanalyse eine eher geringe Rolle spielt.
- **Normalverteilung der Residuen:** Dieser Gesichtspunkt spielt für die Anwendbarkeit der gängigen bis unverzichtbaren F-Tests und t-Tests (s. o.) eine wichtige Rolle. Diesen Tests liegt die sogenannte *Normalverteilungsannahme* zugrunde, die in der Regel bei hinreichend großer Stichprobe gegeben ist.
- **Homoskedastizität der Residuen (Gleiche Varianz der Residuen):** Hier geht es um die Annahme der gleichen Varianz bei allen Residuen, d. h. die Differenzen von beobachteten und geschätzten Werten verändern sich nicht systematisch über die Reihenfolge der Beobachtungen hinweg.
- **Keine Autokorrelation:** Unter Autokorrelation versteht man eine *Korrelation unter den Residuen*. Diese tritt vor allem bei Zeitreihen auf, wenn zeitliche (z. B. saisonale) Zyklen nicht ausreichend durch die unabhängigen Variablen abgedeckt werden. Die Abweichungen der geschätzten Werte der Regressionsgeraden sind dann nicht mehr zufällig.
- **Kein wesentlicher Einfluss von Ausreißern:** Ausreißer sind Werte der abhängigen Variablen, die angesichts der Werte der unabhängigen Variablen weit außerhalb des „üblichen“ Wertebereichs liegen. Diese können eine starke bis verzerrende Wirkung auf die Schätzung der Regressionskoeffizienten haben, weil durch die Kleinste-Qua-

drate-Schätzung einzelne sehr weit außerhalb des üblichen Bereichs liegende Werte besonders großes Gewicht erhalten.

Schätzung der Parameter des Regressionsmodells Nun zum zweiten Schritt, der Schätzung der Parameter eines Regressionsmodells. Ein solches Modell hat im bivariaten Fall die Form

$$\hat{Y} = b_0 + b_1 X$$

bzw. im multiplen (multivariaten) Fall die Form

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \cdots + b_m X_m$$

mit

Y:		abhängige Variable
$\hat{Y}$:		Schätzung der abhängigen Variablen)
b_0:		Parameter (konstantes Glied)
$X_1, \ldots, X_m$:	unabhängige Variable	
$b_1, \ldots, b_m$:	Regressionskoeffizienten	

Hier sei nochmals hervorgehoben, dass sich das vorgestellte Modell auf einen Schätzwert für die abhängige Variable bezieht. Für den jeweils beobachteten Wert der abhängigen Variablen Y müsste das Modell folgendermaßen lauten:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + \ldots b_m X_m + e$$

mit

e: Fehler „error“ (Differenz zwischen geschätztem und beobachtetem Wert von Y)

Beispiel

Ein stark vereinfachtes Beispiel eines Regressionsmodells könnte folgendermaßen aussehen:

$$\hat{Y} = 5{,}2 + 0{,}7 X_1 - 1{,}9 X_2$$

mit

Y:	Marktanteil (geschätzt: $\hat{Y}$)
X_1:	Differenz eigenes Werbebudget/Werbebudget der Konkurrenz in Mio. €
X_2:	Relativer Preis

◀

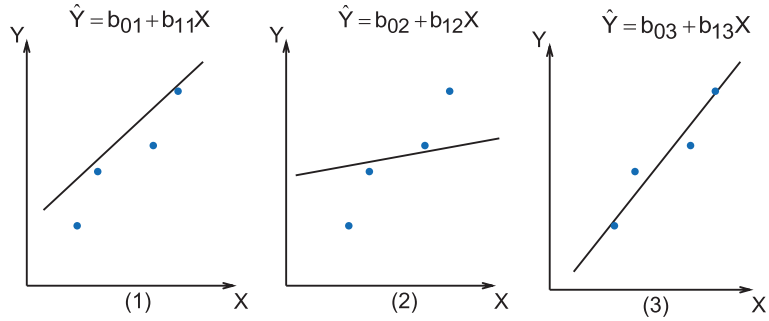


Abb. 9.6 Beispiele für die Schätzung von Regressionsgeraden

In Abb. 9.6 sind drei einfache (bivariate) Beispiele für die Schätzung einer Regressionsgeraden bei einem sehr kleinen Datensatz (4 Fälle, 2 Variable), der natürlich nur zur Demonstration dient und kein realistisches Beispiel sein soll, dargestellt.

Welches der in Abb. 9.6 dargestellten Modelle beschreibt den Zusammenhang zwischen X und Y am besten bzw. am sinnvollsten? Wenn man diese Frage ohne jegliche statistischen Hilfsmittel, sondern nur auf Basis des „gesunden Menschenverstandes" bzw. mit der „Methode des scharfen Hinsehens" beantwortet, würde man sich wohl für Modell 3 entscheiden. Warum? Bei Modell 1 findet man nur Abweichungen in einer Richtung von der Regressionsgeraden, während man intuitiv eher erwartet, dass sich positive und negative Abweichungen ausgleichen. Dieser Anforderung ist bei Modell 2 entsprochen, aber die Abweichungen der beobachteten Werte von den durch das Modell geschätzten Werten sollten deutlich geringer sein. Damit sind schon die zentralen Anforderungen an die **Bestimmung der Parameter** eines Regressionsmodells formuliert:

- Die Abweichungen der beobachteten Werte von den geschätzten Werten sollen *minimiert* werden.
- Die positiven und negativen Abweichungen der beobachteten von den geschätzten Werten sollen sich *ausgleichen*, d. h. die Summe der Abweichungen soll gleich 0 sein.

Diesen Anforderungen genügt man, indem man den Ausdruck $\sum_i (y_i - \hat{y}_i)^2$ minimiert, wobei y_i für den jeweils beobachteten Wert der abhängigen Variablen y steht und $\hat{y}_i$ dem auf Basis eines Regressionsmodells geschätzten Wert der abhängigen Variablen Y entspricht. Man entscheidet sich dann aus der Vielzahl denkbarer (aber unterschiedlich geeigneter) Regressionsmodelle für das, bei dem diese Minimierung gegeben ist bzw. wählt die Parameter b_0 und b_1 (bzw. b_1, …,b_m im multiplen Fall) so, dass die *Abweichungen minimiert* werden. Die Abweichungen $(y_i - \hat{y}_i)$ werden quadriert, weil sich ansonsten bei vielen Regressionsgeraden positive und negative Abweichungen zu Null addieren würden und eine eindeutige Minimierung damit nicht mehr durchführbar ist. Man spricht deshalb bei dem angewandten Verfahren von einer **„Kleinste-Quadrate-Schätzung"**.

Die Einzelheiten der rechnerischen Bestimmung der Parameter müssen hier nicht diskutiert werden, sondern können getrost dem benutzten Statistikprogramm überlassen werden.

Die somit bestimmten Parameter – die sogenannten **Regressionskoeffizienten** – legen also die Beziehung zwischen unabhängigen und der abhängigen Variablen für den untersuchten Datensatz fest. Durch Verwendung dieser Parameter und der jeweiligen Variablenwerte der unabhängigen Variablen lässt sich für jeden Fall der Wert der abhängigen Variablen *schätzen.*

Die Gegenüberstellung von geschätzten ($\hat{y}_i$) und tatsächlichen beobachteten Werten (y_i) ermöglicht es, die *Güte des Regressionsmodells* zu beurteilen. Je geringer die sich ergebenden Abweichungen sind, desto besser beschreibt offenbar das Modell den gegebenen Zusammenhang.

Bevor auf die Überprüfung des Regressionsmodells weiter eingegangen wird, noch ein Hinweis zur **Interpretation der Regressionskoeffizienten.** Bei vielen Anwendungen der multiplen Regressionsanalyse bzw. bei der Interpretation ihrer Ergebnisse ist der Vergleich der verschiedenen Regressionskoeffizienten der unabhängigen Variablen von Interesse. Je größer ein Koeffizient ist, desto stärker wirkt sich offenbar eine Veränderung der entsprechenden unabhängigen Variablen um eine Einheit auf den resultierenden Wert der abhängigen Variablen aus. Diese Art der Interpretation spielt vor allem eine wichtige Rolle, wenn man feststellen will, wovon die interessierende abhängige Variable (z. B. Marktanteil, Marktvolumen) am stärksten abhängt.

Allerdings sind die bisher erläuterten Regressionskoeffizienten dafür nur bedingt geeignet, weil deren Größe von den verwendeten Maßeinheiten der unabhängigen Variablen abhängt. Wenn man einen Marktanteil u. a. durch das eingesetzte Werbebudget erklären will, dann ergeben sich natürlich unterschiedliche Werte für den entsprechenden Regressionskoeffizienten in Abhängigkeit davon, ob das Werbebudget z. B. in „Mio. $“ oder in „Tsd. €“ gemessen wurde. Deswegen werden Regressionskoeffizienten von den jeweiligen Maßeinheiten unabhängig gemacht, indem man eine **Standardisierung** anhand der Standardabweichungen von x und y vornimmt. Die standardisierten Regressionskoeffizienten **Beta** ergeben sich durch

$$\text{Beta} = b\frac{s_x}{s_y}$$

Standardisierte Regressionskoeffizienten erhält man auch, wenn man vor der Berechnung der Regression alle Variablen standardisiert, d. h. den Mittelwert abzieht und durch die Standardabweichung dividiert. Sie erlauben es, die Einflussstärke der verschiedenen unabhängigen Variablen zu interpretieren bzw. zu vergleichen. Wenn man also beispielsweise ein Regressionsmodell mit den standardisierten Variablen Y, X_1 und X_2aufgestellt und

$$Y = \text{Beta}_1 X_1 + \text{Beta}_2 X_2$$

mit den Parametern $Beta_1 = 0{,}7$ und $Beta_2 = 0{,}2$ geschätzt hat, dann kann man daran erkennen, dass die Variable X_1 offenbar eine stärkere Wirkung (genau: eine 3,5-fach stärkere Wirkung) auf die Werte der abhängigen Variablen Y hat als die Variable X_2. Standardisierte Regressionskoeffizienten werden in den aktuellen statistischen Analyseprogrammen automatisch berechnet.

Überprüfung des Regressionsmodells und seiner Parameter Nun zur Überprüfung von Regressionsmodellen. Oben ist schon angesprochen worden, dass die Abweichung von tatsächlichen und geschätzten Werten der abhängigen Variablen sicher ein Indikator für die Güte eines Regressionsmodells ist. Dazu in Abb. 9.7 ein Beispiel. Beide Regressionsgeraden in Abb. 9.7 haben identische Parameter (konstantes Glied, Steigung). Offenkundig wird der Datensatz im linken Teil der Abbildung dadurch besser beschrieben als der andere Datensatz, da die Abweichungen zwischen geschätzten Werten (auf der Regressionsgeraden) und beobachteten Werten hier viel geringer sind. Eine Maßgröße für dieses Beurteilungskriterium ist die „erklärte Varianz" bzw. der „Anteil erklärter Varianz", der schon im Abschnitt zur Varianzanalyse (Abschn. 9.2) verdeutlicht wurde.

Beispielsweise lässt sich für die Erklärung von Einkommensunterschieden (abh. Variable Einkommen) die Variable „Alter" heranziehen. Es ist leicht nachzuvollziehen, dass das Alter zwar einige Einkommensunterschiede zu erklären vermag, aber eben nicht alle. Die Variable „Alter" erklärt also nur einen Teil der Varianz des Einkommens.

▶ **Skiera und Albers (2008, S. 472): Erläuterung der Bedeutung des Anteils der erklärten Varianz:** Zur Beurteilung der Anpassungsgüte der linearen Regressionsanalyse lässt man sich von der Überlegung leiten, dass ohne die Kenntnis der unabhängigen Variablen die beste Schätzung des zu erwartenden Werts der abhängigen Variablen durch die Bestimmung des

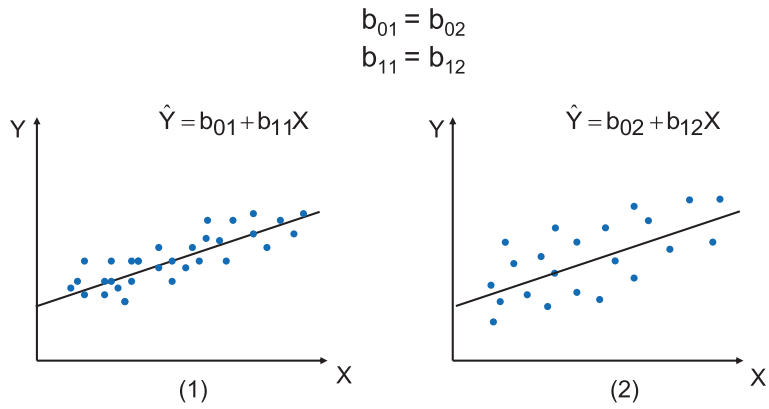


Abb. 9.7 Gleichartige Regressionsgeraden, die verschiedene Datensätze unterschiedlich gut beschreiben

Mittelwerts der abhängigen Variablen erfolgt. Die Güte einer Regressionsanalyse wird nun daran gemessen, um wie viel sich die Aussage durch die Betrachtung von unabhängigen Variablen gegenüber der ausschließlichen Betrachtung der abhängigen Variablen und der damit verbundenen „einfachen" Schätzung in Form des Mittelwerts verbessert. Gemessen wird dies durch das Bestimmtheitsmaß R^2, das den Anteil der durch die Regressionsgleichung erklärten Varianz an der Varianz der „einfachen" Aussage in Form des Mittelwerts erfasst.

Eine wesentliche und besonders gängige Maßzahl zur Beurteilung eines Regressionsmodells ist also der **Anteil** (durch das Modell) **erklärter Varianz,** auch **Bestimmtheitsmaß** genannt und mit $\mathbf{R^2}$ bezeichnet. Die weiter unten dargestellte Berechnung von R^2 und deren Begründung lässt sich durch die Abb. 9.8 veranschaulichen. In Abb. 9.8 findet sich das (fiktive) Beispiel einer Erklärung von Marktanteilen durch die über die Zeit kumulierten Werbebudgets für die entsprechenden Produkte. Ein Wert y_i ist herausgehoben und sein Abstand zum Mittelwert $\bar{y}$ eingetragen und mit $y_i - \bar{y}$ bezeichnet worden. Diese Abweichungen für alle n Messwerte würden nach Quadrierung, Summierung und Teilung durch n für alle Messwerte der Varianz von Y entsprechen. Weiterhin findet sich für den gleichen Wert von x auf der Regressionsgeraden der geschätzte Wert für y, also $\hat{y}$. Auch hier ist die Abweichung von $\bar{y}$ mit $\hat{y}_i - \bar{y}$ dargestellt. Wenn man die Abweichung von Y_i vom Mittelwert $\bar{y}$ betrachtet, dann wird offenbar der Teil $\hat{y}_i - \bar{y}$ durch das Regressionsmodell erklärt, der Rest bleibt unerklärt.

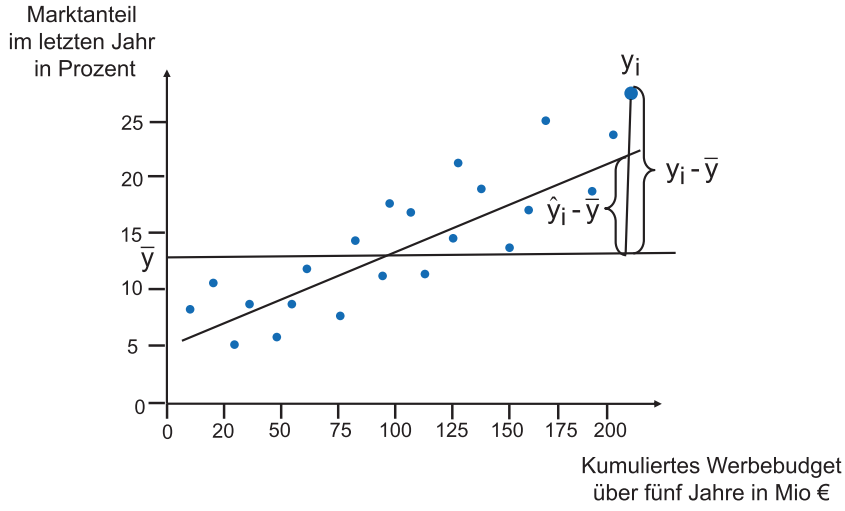


Abb. 9.8 Beispiel eines Regressionsmodells zur Erklärung von Marktanteilen durch das kumulierte Werbebudget

Wenn man diese Grundidee auf alle Werte anwendet und dabei die übliche Art zur Berechnung der Varianz zugrunde legt, kommt man zu folgendem Ausdruck

$$\frac{\text{Erklärte Varianz}}{\text{Gesamtvarianz}} = \frac{1/n \sum \left(\hat{y}_i - \bar{y}\right)^2}{1/n \sum (y_i - \bar{y})^2} = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = R^2$$

R^2(„Bestimmtheitsmaß“) liegt *zwischen 0 und 1*. Bei $R^2 = 0{,}47$ spricht man beispielsweise von „47 % erklärter Varianz“. Die extremen Werte 0 und 1 stehen dafür, dass ein Modell gar keine Varianz erklärt (also ohne jede Erklärungskraft ist) bzw. die abhängige Variable vollständig erklärt (siehe Abb. 9.9). Beides sind extreme Fälle, die äußerst selten vorkommen. Nur wenn eine unabhängige Variable völlig unsinnig ausgewählt wurde, wird R^2 bei 0 liegen. Auch ein R^2 das nahe bei 1 liegt, tritt im sozialwissenschaftlichen Bereich mit typischerweise großer Komplexität von Wirkungszusammenhängen kaum einmal auf.

Typisch sind eher Fälle, in denen R^2 irgendwo zwischen 0 und 1 liegt. Dabei stellt sich – für den nächsten Schritt der Überprüfung des Regressionsmodells – zunächst die Frage, ob der Anteil der durch das Modell erklärten Varianz tatsächlich durch einen Zusammenhang zwischen den Variablen zu Stande gekommen ist oder auch durch Zufälligkeiten bei der Datenerhebung zu erklären wäre. Beispielsweise könnte es bei kleinen R^2-Werten (z. B. 0,02) ja sein, dass diese bei einem anderen Datensatz (bzw. einer anderen Stichprobe) nicht mehr auftreten, weil der „wahre“ R^2-Wert gleich 0 ist. Man überprüft also die **Signifikanz von R^2,** d. h. man prüft die Frage, ob R^2 (mit großer Wahrscheinlichkeit) *systematisch von 0 verschieden* ist. Dazu verwendet man – wie bei den statistischen Tests in Abschn. 8.2 – eine geeignete Maßgröße, hier den sogenannten **F-Wert**:

$$F = \frac{\frac{R^2}{v}}{\frac{(1-R^2)}{(n-v-1)}}$$

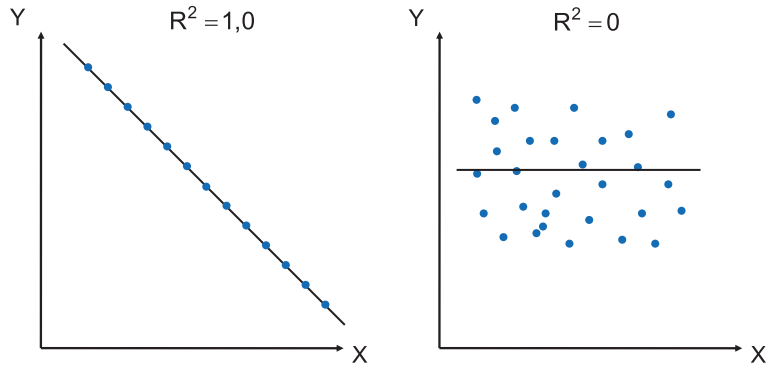


Abb. 9.9 Beispiele für Bestimmtheitsmaße

mit

R^2:	Bestimmtheitsmaß (zwischen 0 und 1)
v:	Anzahl der unabhängigen Variablen im Modell
n:	Stichprobengröße

Die Grundidee des F-Tests wird schnell deutlich, wenn man die verschiedenen Komponenten der Maßzahl F betrachtet. Zunächst findet man die Relation $R^2/(1-R^2)$. In Analogie zur Varianzanalyse wird hier also erneut der Anteil erklärter Varianz zum Anteil nicht erklärter Varianz in Beziehung gesetzt und anschließend auf Signifikanz geprüft. Der R^2-Wert über dem Bruchstrich wird noch durch v (Anzahl der einbezogenen unabhängigen Variablen) normiert, weil bei einem Modell mit hohem R^2-Wert auf Basis nur weniger unabhängiger Variabler eher auszuschließen ist, dass dieser Wert per Zufall entstanden ist, als bei einem Modell mit einer größeren Zahl von unabhängigen Variablen. Im Nenner wird der Ausdruck $(1-R^2)$ ebenfalls normiert, durch einen Ausdruck $(n-v-1)$, der bei relativ großer Stichprobe (z. B. 500 oder 1000) und kleiner Zahl unabhängiger Variablen (z. B. 2 oder 3) fast identisch mit der Stichprobengröße ist. D. h., dass bei großen Stichproben der Nenner klein und der F-Wert groß wird.

Der Test auf **Signifikanz des Regressionsmodells** insgesamt erfolgt durch Vergleich der empirisch ermittelten Maßzahl F mit einer entsprechenden Wahrscheinlichkeitsverteilung **(F-Verteilung)**. Wenn die Wahrscheinlichkeit für ein zufälliges Zustandekommen des gegebenen Anteils erklärter Varianz bei gegebener Anzahl unabhängiger Variabler und gegebener Stichprobengröße relativ klein (bzw. groß) ist, dann ist das Regressionsmodell insgesamt signifikant (bzw. nicht signifikant).

Mit dem F-Test kann also überprüft werden, ob das Regressionsmodell insgesamt Aussagekraft hat. Eine weitere Frage bezieht sich auf die Aussagekraft bzw. **Signifikanz der einzelnen Parameter** b_0 bis b_m des Regressionsmodells. Wenn einer der Steigungsparameter nicht signifikant von 0 verschieden wäre, dann würde das ja bedeuten, dass die entsprechende unabhängige Variable keinen Einfluss auf die abhängige Variable hätte. Für einen solchen Test kann man wieder auf den **t-Test** zurückgreifen. Hier werden natürlich nicht Mittelwerte verglichen, sondern Regressionsparameter. Es ergibt sich:

$$t = \frac{b_j - \beta}{s_{b_j}}$$

mit

b_j:	Regressionskoeffizient (empirisch) für Variable j
β_j:	„wahrer“ (unbekannter) Regressionskoeffizient für Variable j
s_{b_j}:	Standardfehler für Regressionskoeffizient

Nun ist es schwierig, einen Vergleich mit einem unbekannten Regressionskoeffizienten β_j durchzuführen. Hier stellt sich aber die Frage, ob dieser nicht *in Wirklichkeit 0* sein könnte, was bedeutet, dass die entsprechende unabhängige Variable die abhängige Variable nicht beeinflusst. Wenn man also testen will, ob b_j signifikant von 0 verschieden ist, ergibt sich

$$t = \frac{b_j - 0}{s_{b_j}} = \frac{b_j}{s_{b_j}}$$

Unter Berücksichtigung von Freiheitsgraden und Sicherheitswahrscheinlichkeit kann man wieder einen Vergleich mit der entsprechenden Wahrscheinlichkeitsverteilung vornehmen, was in der praktischen Durchführung natürlich von dem angewendeten Datenanalyse-Programm erledigt wird. Als sehr vereinfachte Faustregel gilt, dass ein t-Wert $> +2$ bzw. t-Wert < -2 für eine signifikante ($p < 0{,}05$) Abweichung des jeweiligen Regressionskoeffizienten von 0 spricht. Wie bereits in Abschn. 8.2 thematisiert ist neben der Signifikanz und der erklärten Gesamtvarianz auch immer noch die Effektstärke, also der Erklärungsbeitrag des Modells insgesamt und der Beitrag der einzelnen unabhängigen Variablen zu betrachten und zu bewerten. Das oben thematisierte R-Quadrat, kann für das gesamte Regressionsmodell direkt berichtet werden aber auch in die Effektstärke f nach Cohen (1992) umgerechnet werden. In diesem Fall ist der Wertebereich der Effektstärke zwischen 0 und ∞. Um den Einfluss der einzelnen unabhängigen Variablen zu bewerten, wird jeweils der standardisierte Regressionskoeffizient verwendet. Dieser sogenannte Beta-Koeffizienten ist im Fall der bivariaten Regression identisch mit der Korrelation, wodurch er ein geeignetes Maß für die Effektstärke ist. Der standardisierte Regressionskoeffizient wird von Statistiksoftware wie SPSS automatisch berechnet.

9.3.2 Moderation und Mediation

Sowohl bei Moderation als auch bei Mediation geht es zunächst um die Zusammenhänge zwischen drei Variablen X, Y und M. Untersucht wird der Effekt eines Prädiktors oder Faktors X auf Y. Dieser kann natürlich mit einem Regressionsmodell mit X als unabhängige und Y als abhängige Variable untersucht werden. Zusätzlich gibt es eine dritte Variable M. Sie ist entweder der Moderator oder der Mediator (siehe auch Abschn. 6.1).

Bei der Moderation wirkt die als Moderator auf die Beziehung zwischen X und Y. Der Einfluss von M ändert also den Effekt von X auf Y. Dies äußert sich so, dass die Beziehung zwischen X und Y je nach Ausprägung von M unterschiedlich ausfällt. Statistisch gesprochen liegt eine Interaktion zwischen M und X vor. In der praktischen Anwendung wird also eine Regressionsanalyse mit den drei Faktoren X, M und der Interaktion zwischen X und M (Moderatoreffekt) zur Erklärung von Y gerechnet. Wird in diesem Modell die Interaktion signifikant, so liegt eine signifikante Moderation vor.

Beispiel

Teilnehmern eines Seminars wird ein Filmbeitrag zur Vorzugswürdigkeit von Bioprodukten gezeigt. Der Effekt des Beitrags (X) wird dadurch untersucht, dass beobachtet wird, wie viele Bioprodukte im Anschluss gekauft werden (Y). Moderiert wird der Einfluss des Seminars z. B. durch das Einkommen der Teilnehmer (M). Bioprodukte sind teurer als vergleichbare nicht-biologische Produkte. Je höher das Einkommen, umso eher kann sich der Teilnehmer auch Bioprodukte leisten. ◀

Bei der Mediation steht die Variable M als Mediator sowohl zu X als auch zu Y in Beziehung. Der direkte Effekt zwischen X und Y wird durch den indirekten Effekt über M erklärt, also durch X -> M -> Y. Untersucht wird auf Mediation, indem mehrere Regressionsmodelle gerechnet werden.

Beispiel

In einer Studie wird der Einfluss der Anzahl von Pausen (X) auf den Lernerfolg (Y) untersucht. Im Ergebnis zeigt sich, dass Versuchspersonen mit mehr Pausen einen größeren Lernfortschritt erzielt haben. Vermittelt wird dieser Effekt jedoch darüber, dass die häufigeren Pausen sinnvoll genutzt wurden zum Rekapitulieren der Lehrinhalte (M). ◀

Moderations- und Mediationsanalysen lassen sich mithilfe des PROCESS-Moduls von Hayes (2013), einfach in SPSS integrieren und durchführen (siehe www.processmacro.org). Der Autor bietet eine große Vielzahl an unterschiedlichen und in Teilen sehr komplexen Modelldesigns zur einfachen Analyse an.

9.3.3 Regression mit Dummy-Variablen

Eingangs des Abschn. 9.3.1 ist schon auf den besonderen Fall von Regressionsanalysen mit einzelnen nominalskalierten unabhängigen Variablen, die sogenannte Dummy-Regression (Dummy = „Strohmann“, hier: Hilfsvariable), hingewiesen worden. Bei diesen **Dummy-Variablen** handelt es sich um *binäre Variable*, also um Variable, die nur die beiden Ausprägungen 0 und 1 haben können. Dabei steht der Wert 1 dafür, dass eine bestimmte Ausprägung eines qualitativen Merkmals gegeben, der Wert 0 dafür, dass diese Ausprägung nicht gegeben ist. Wenn man in ein Modell

$$\hat{Y} = b_0 + b_1 X$$

zusätzlich eine Dummy-Variable (D) einbezieht, dann würde sich

$$\hat{Y} = b_0 + b_1 X_1 + b_2 D$$

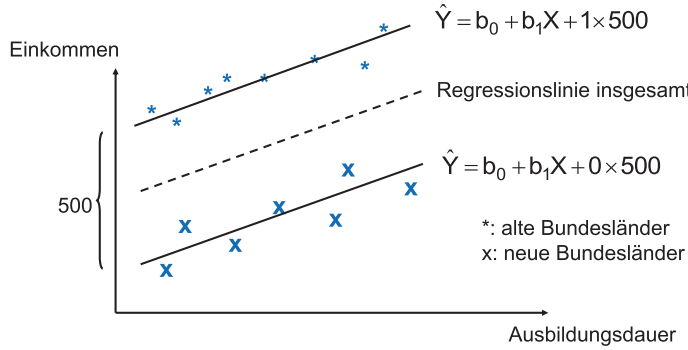


Abb. 9.10 Beispiel einer Regressionsanalyse mit einer Dummy-Variablen

ergeben und in Abhängigkeit davon, welche Ausprägung der qualitativen Variable auftritt,

$$\hat{Y} = b_0 + b_1X_1 + b_20 \text{ bzw. } \hat{Y} = b_0 + b_1X_1 + b_21$$

Diese Vorgehensweise sei an einem Beispiel illustriert. Wenn man erneut den Zusammenhang zwischen Ausbildungsdauer und Einkommen betrachtet und zusätzlich vermutet, dass sich zwischen alten und neuen Bundesländern die Einkommensniveaus unterscheiden, dann könnte man durch eine Dummy-Variable mit dem Wert 0 für Personen aus den neuen Bundesländern und dem Wert 1 für Personen aus den alten Bundesländern diesen Aspekt in die Analyse einbeziehen. In Abb. 9.10 ist das fiktive Beispiel dargestellt.

Man erkennt in Abb. 9.10, dass gewissermaßen separate Regressionsgeraden für beide Teilgruppen des Datensatzes (Personen aus den neuen und den alten Bundesländern) geschätzt werden, die sich in ihrem Achsenabschnitt, nicht jedoch in ihrer Steigung unterscheiden. Als Parameter b_2 hat sich 500 ergeben, was so interpretiert werden kann, dass das Einkommensniveau in den alten Bundesländern im Mittel um 500 höher liegt. Auch qualitative Variable mit mehr als zwei Ausprägungen lassen sich durch eine entsprechende Zahl von binären Dummy-Variablen (Zahl der Dummy-Variablen = Zahl der Ausprägungen − 1) darstellen.

9.4 Logistische Regression

Die bisher dargestellten Verfahren der Varianz- und Regressionsanalyse erlauben die Untersuchung einer metrischen abhängigen Variablen anhand von unabhängigen Variablen unterschiedlichen Messniveaus. Mit der logistischen Regression wird nun ein Verfahren dargestellt, mit dem auch Einflüsse auf eine abhängige nominalskalierte Variable untersucht werden können. Früher wurde bei derartigen Fragestellungen in erster Linie

die **Diskriminanzanalyse** (siehe z. B. Backhaus et al., 2016, S. 215 ff.) eingesetzt. Jetzt spielt hier die logistische Regression die größere Rolle, nicht zuletzt wegen ihrer größeren Robustheit im Hinblick auf ihre Anwendungsvoraussetzungen (Backhaus et al., 2016, S. 287).

Hintergrundinformation

Hair et al. (2010, S. 436) vergleichen die Stärken und Schwächen von Diskriminanzanalyse und logistischer Regression:

> „Die Anwendung der Diskriminanzanalyse ist geeignet, wenn die abhängige Variable nicht metrisch ist. Wenn die abhängige Variable nur zwei Gruppen umfasst, dann ist die logistische Regression aus zwei Gründen zu bevorzugen. Erstens beruht die Diskriminanzanalyse auf strengeren Annahmen bezüglich der multivariaten Normalverteilung und einer gleichen Varianz-Kovarianzmatrix in den verschiedenen Gruppen – Annahmen, die in vielen Situationen nicht erfüllt sind. Die logistische Regression unterliegt nicht diesen Annahmen und ist sehr robust, wenn diese Annahmen nicht erfüllt sind, so dass ihre Anwendung unter verschiedensten Bedingungen geeignet ist. Zweitens bevorzugen viele Forscher die logistische Regression, selbst wenn alle genannten Anforderungen erfüllt sind, weil sie der multiplen Regression sehr ähnlich ist. Sie verwendet einfache statistische Tests, ähnliche Ansätze bei der Einbeziehung metrischer und nicht-metrischer Variablen sowie nicht-linearer Effekte und bietet eine Vielzahl von anderen Diagnostiken."

Bei der **logistischen Regression** ist die abhängige Variable also nominalskaliert. Im üblichen und einfachsten Fall wird davon ausgegangen, dass die abhängige Variable dichotom ist, also nur zwei Werte annehmen kann (z. B. 0 und 1 für Kauf bzw. Nicht-Kauf oder für Erfolg bzw. Misserfolg). Mithilfe mehrerer unabhängiger Variablen werden die Wahrscheinlichkeiten für das Auftreten der Ausprägungen der abhängigen Variable (eines „Ereignisses") geschätzt. In einer multinomialen logistischen Regression können auch kategoriale abhängige Variablen mit mehr als zwei Ausprägungen analysiert werden, dieses Verfahren wird hier nicht weiter behandelt, die Grundidee entspricht jedoch dem hier vorgestellten dichotomen (binären) Fall.

Da in einem solchen Fall, bei einer Verwendung der linearen Regression die Varianzen der Residuen nicht gleich sind (Heteroskedastizität), wird für die Schätzung der Wahrscheinlichkeiten eine geeignete Wahrscheinlichkeitsfunktion, die sogenannte logistische Verteilung (daher der Name logistische Regression) verwendet, die zusätzlich gewährleistet, dass die *Vorhersagewerte der abhängigen Variablen innerhalb des für Wahrscheinlichkeiten zulässigen Bereichs* [*0, 1*] liegen. Die Parameter dieser Verteilung werden durch eine Linearkombination der unabhängigen Variablen bestimmt. Ziel ist es dabei die Parameter so zu schätzen, dass die Wahrscheinlichkeit einer korrekten Zuordnung möglichst groß ist (siehe auch Backhaus et al., 2016, S. 306). Die beobachteten Erhebungsdaten, also die empirischen Werte, nehmen die Ausprägungen 0 oder 1 (z. B. Kauf oder Nicht-Kauf) an. Die Parameterschätzung der logistischen Regression ergibt nun für jeden Fall einen Wahrscheinlichkeitswert (einen z-Wert oder auch Logit genannt) zwischen 0 und 1. Üblicherweise wird bei einem Wahrscheinlichkeitswert größer als

0,5 eine Zuordnung zur Ausprägung 1 und bei einem Wahrscheinlichkeitswert kleiner als 0,5 eine Zuordnung zur Ausprägung 0 vorgenommen. Abb. 9.11 zeigt den Verlauf einer logistischen Funktion sowie die Zuteilung der Werte der Schätzfunktion zu den Beobachtungswerten. Der Schätzalgorithmus, der dazu verwendet wird, versucht nun die Parameter des logistischen Regressionsmodells so zu schätzen, dass die Zuordnung optimiert wird.

Da hier kein linearer Zusammenhang getestet wird, lassen sich die Regressionskoeffizienten auch nicht genau so wie bei einer linearen Regression interpretieren. Es ist *lediglich die Richtung des Einflusses interpretierbar*. Z. B. führen positive Regressionskoeffizienten dazu, dass bei steigenden Werten der unabhängigen Variablen (z. B. Einkommen) die Wahrscheinlichkeit für ein Ereignis (z. B. Kauf) steigt. Da die zugrundeliegende logistische Funktion nicht linear ist, lässt sich die Größe des Einflusses nicht interpretieren, etwa derart, dass eine Veränderung der unabhängigen Variable um eine Einheit die abhängige Variable entsprechend dem Regressionskoeffizienten verändert. Eine Überprüfung der Signifikanz der Koeffizienten erfolgt anhand der **Wald-Statistik.**

Bei der Beurteilung der Modellgüte untersucht man, wie gut die unabhängigen Variablen insgesamt zur Trennung der Ausprägungen der abhängigen Variable (z. B. Kauf versus Nicht-Kauf) beitragen. Abb. 9.12 zeigt zwei Beispiele für die Anpassung der logistischen Funktion an empirische Daten, die durch Sternchen dargestellt sind. Offensichtlich ist der „Fit" im rechten Beispiel deutlich höher als im linken Beispiel, denn hier werden die empirischen Werte vollständig durch die Funktion erfasst, d. h. allen Fällen mit der Ausprägung 1 werden Wahrscheinlichkeitswerte größer als 0,5 zugeordnet und allen Fällen mit der Ausprägung 0 werden Wahrscheinlichkeitswerte kleiner als 0,5 zugeordnet. Im linken Beispiel dagegen ist die Zuordnung nicht ganz eindeutig. Ähnlich wie die Fehlerquadratsumme bei der linearen Regression, wird zur Untersuchung des

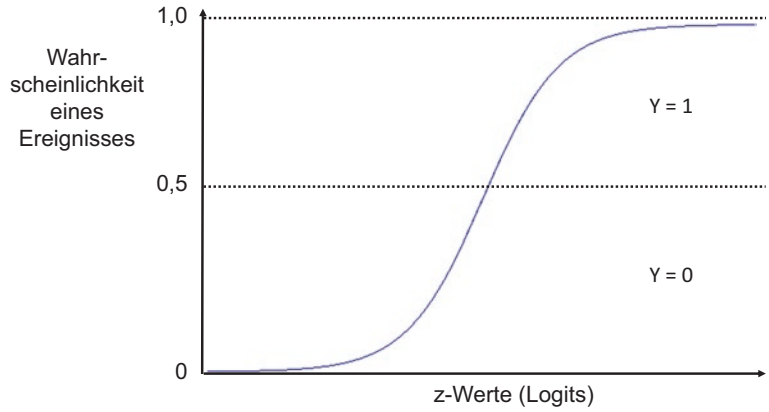


Abb. 9.11 Wahrscheinlichkeitsverteilung und Zuordnung der Schätzwerte der logistischen Regression. (Nach Backhaus et al., 2016, S. 285)

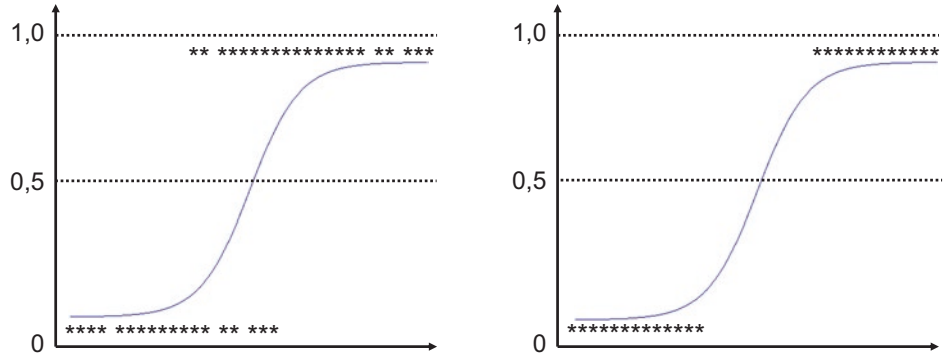


Abb. 9.12 Beispiele für die Anpassung einer logistischen Funktion an empirische Daten. (Nach Hair et al., 2010, S. 417)

Tab. 9.2 Übersicht zu Verfahren der Dependenz-Analyse

	Skalierung	
Verfahren	Abhängige Variable	Unabhängige Variable(n)
Varianzanalyse	Intervall/metrisch	Nominal
Varianzanalyse mit Kovariaten		Nominal, intervall/metrisch
Regressionsanalyse		Intervall/metrisch
Regressionsanalyse mit Dummies		Intervall/metrisch, nominal
Binäre Logistische Regression	Binär (0/1)	Intervall/metrisch, nominal

Fits eine Abweichung vom Idealwert (vollständiger Fit) ermittelt. Der **Likelihood-Ratio-Test** prüft dann, ob die unabhängigen Variablen dazu führen, dass die Abweichung vom Idealwert signifikant verkleinert wird. Außerdem lassen sich auch ähnlich dem Bestimmtheitsmaß R^2 der linearen Regressionsanalyse Bestimmtheitsmaße für die logistische Regression berechnen, sogenannte **Pseudo-R^2-Statistiken**, wobei hier ein vollständiges Modell (alle unabhängigen Variablen werden einbezogen) mit einem Null-Modell (die Regressionskoeffizienten werden alle auf Null gesetzt, d. h. der Einfluss der unabhängigen Variablen wird nicht beachtet) verglichen wird. Schließlich können bei der logistischen Regression auch die Klassifikationsergebnisse beurteilt werden, etwa anhand einer **Klassifikationsmatrix**. Detaillierte Beschreibungen der Einzelheiten dieser Verfahren finden sich in der entsprechenden Literatur (z. B. Backhaus et al., 2016, S. 314 ff.; Menard, 2002).

Zum Abschluss der dependenzanalytischen Verfahren folgt eine Übersicht (siehe Tab. 9.2), aus der hervorgeht, bei welchem Skalenniveau sich welche Methode anbieten. Es wird deutlich, dass man beispielsweise eine Varianzanalyse mit Kovariaten auch als Regressionsanalyse mit Dummy-Variablen rechnen kann. Es wird dasselbe Ergebnis

erzielt. Welches Verfahren man in einem solchen Fall wählt, bleibt dem Forscher überlassen, bzw. richtet sich nach der vorliegenden Forschungsfrage.

9.5 Conjoint-Analyse

Die Conjoint-Analyse hat in den letzten ca. 30 Jahren wegen ihrer besonderen Bedeutung für die Praxis große Beachtung gefunden. Ihr Anwendungsbereich liegt vor allem im Bereich der Produktpolitik (einschließlich der damit verbundenen preispolitischen Entscheidungen). Dabei geht es vor allem um die Einschätzung, welche Bedeutung (welches Gewicht) einzelne Produkteigenschaften im Hinblick auf die Präferenzen bzw. die Kaufentscheidungen von Kunden haben. Für die folgende knappe Darstellung der Grundidee der Conjoint-Analyse soll deshalb auf diesen Anwendungsbereich Bezug genommen werden.

Im Wesentlichen geht es bei der Conjoint-Analyse darum, aus Angaben von Auskunftspersonen zum Nutzen verschiedener Produkte mit verschiedenen Kombinationen von Eigenschaften (bzw. zu ihren Präferenzen hinsichtlich dieser Produkte) zu schließen, welchen *Beitrag die einzelnen Produkteigenschaften zur Gesamtbewertung* des Produkts leisten. Dazu zunächst ein ganz einfaches Beispiel. Eine Person soll Präferenzen äußern im Hinblick auf die Autos A, B, C, D, E, F, G und H, die sich dadurch unterscheiden, ob ABS, Klimaanlage oder Schiebedach vorhanden sind oder nicht (Tab. 9.3).

Wenn die Person nun angibt, dass sie die Autos A, C, E und G, also die Autos mit ABS, deutlich präferiert, dann lässt das erkennen, dass für diese Person die Produkteigenschaft „mit ABS“ offenbar ganz wichtig ist.

Das Verfahren beruht also darauf, dass man Kombinationen von Produkteigenschaften systematisch variiert und aus den dazugehörigen Präferenzen auf die einzelnen Eigenschaften schließt. Insofern handelt es sich um eine Dependenz-Analyse, weil davon ausgegangen wird, dass die Präferenzen von den Produkteigenschaften (als un-

Tab. 9.3 Ein einfaches Beispiel zur Conjoint-Analyse

Eigenschaft Auto	ABS	Klimaanlage	Schiebedach
A	Mit	Mit	Mit
B	Ohne	Mit	Mit
C	Mit	Mit	Ohne
D	Ohne	Mit	Ohne
E	Mit	Ohne	Mit
F	Ohne	Ohne	Mit
G	Mit	Ohne	Ohne
H	Ohne	Ohne	Ohne

Abb. 9.13 Kurz-Beschreibungen von Pauschalreise-Angeboten als Stimuli der Conjoint-Analyse

Tab. 9.4 Präferenz-Rangfolge einer Auskunftsperson bezüglich Pauschalreisen

	Hotel in Ortsmitte		Hotel am Ortsrand	
	Mit Disco	Ohne Disco	Mit Disco	Ohne Disco
Preis € 500	5	3	1	7
Preis € 600	6	4	2	8

abhängigen Variablen) und deren Gewichtungen abhängen. Dabei ist auch eine Analogie zum linearen Modell erkennbar.

Hier ein Beispiel zur Verdeutlichung der *Grundidee* der Conjoint-Analyse (vgl. Sudman & Blair, 1998, S. 229 f.). Man stelle sich vor, dass die Bedeutung der folgenden drei Eigenschaften (mit jeweils zwei Ausprägungen) eines Pauschalreise-Angebots gemessen werden soll:

- Lage des Hotels (Ortsmitte, Ortsrand)
- Disco im Hotel (ja, nein)
- Preis für zwei Wochen Halbpension (HP) (€ 500,–, € 600,–)

Der jeweiligen Auskunftsperson werden Karten („Stimuli“) mit entsprechenden Kurz-Beschreibungen vorgelegt (siehe Abb. 9.13) und sie soll diese ihren Präferenzen gemäß ordnen.

Die Ergebnisse dieser Einschätzungen, also die Präferenz-Rangfolge, werden in Tab. 9.4 wiedergegeben. Niedrige Zahlenwerte stehen für starke Präferenzen und umgekehrt.

Wenn man die (etwas kühne) Annahme macht, dass die Noten zur Präferenz-Rangfolge (annähernd) intervallskaliert sind, dann lässt sich abschätzen, welche Bedeutung die einzelnen Eigenschaften haben (wobei zu beachten ist, dass ein niedriger Zahlenwert für eine hohe Präferenz steht):

Hotel in Ortsmitte:	(5+6+3+4)/4	=4,5
Hotel am Ortsrand:	(1+2+7+8)/4	=4,5
Hotel mit Disco:	(5+6+1+2)/4	=3,5
Hotel ohne Disco:	(3+4+7+8)/4	=5,5
Hotel € 500,–:	(5+3+1+7)/4	=4,0
Hotel € 600,–:	(6+4+2+8)/4	=5,0

Man kann daraus erkennen, dass die Lage des Hotels (Ortsrand, Ortsmitte) offenbar keine Rolle spielt, weil beide Alternativen gleich beurteilt werden. Im Hinblick auf das Vorhandensein einer Disco gibt es eine klare Bevorzugung eines Hotels mit Disco. Selbstverständlich wird der geringere Preis präferiert, wobei aber der Preisunterschied geringeres Gewicht hat als die Frage, ob im Hotel eine Disco vorhanden ist oder nicht. Letzteres wird daraus geschlossen, dass die *Differenzen der jeweiligen Präferenzen* entsprechend unterschiedlich sind.

Derartige Auswertungen beziehen sich nur auf die jeweilige Person, können aber über mehrere Personen aggregiert werden. Im Mittelpunkt steht dabei die Bestimmung der sogenannten **Teilnutzenwerte** (für jede Eigenschaftsausprägung). Dabei wird ein *kompensatorisches Entscheidungsverhalten* der Auskunftspersonen unterstellt, d. h. dass eine Schwäche bei einem Produktmerkmal durch eine Stärke bei einem anderen Merkmal ausgeglichen werden kann. Diese Annahme kann durchaus problematisch sein, beispielsweise bei der Auswahl einer Airline, bei der für viele Menschen mangelnde Sicherheit nicht durch günstige Abflugzeiten, niedrige Preise etc. kompensiert werden kann. Im einfachsten Fall – wenn man annimmt, dass die erhobenen Präferenzwerte intervallskaliert sind – kann man die Teilnutzenwerte der verschiedenen Eigenschaftsausprägungen mit einem linearen Modell (z. B. Regression mit Dummy-Variablen) so bestimmen, dass die aus den Teilnutzenwerten für eine Alternative resultierenden **Gesamtnutzenwerte** möglichst gut den empirisch ermittelten Präferenzwerten entsprechen.

Die **relative Wichtigkeit** der einzelnen Produktmerkmale für die Auskunftsperson (z. B. den Kunden) lässt sich wiederum aus den unterschiedlichen Teilnutzenwerten der verschiedenen Ausprägungen eines Merkmals ermitteln. Die Grundidee dabei ist ein-

fach: Wenn die Differenz zwischen dem höchsten und dem geringsten Teilnutzenwert der verschiedenen Ausprägungen eines Merkmals groß ist, dann hat dieses Merkmal für den Kunden offenbar große Bedeutung, weil ja unterschiedliche Ausprägungen dabei zu deutlich unterschiedlichen Gesamtnutzenwerten führen. Eine detaillierte Beschreibung der Rechenschritte zur Bestimmung der Teilnutzenwerte, der Gesamtnutzenwerte sowie der relativen Wichtigkeiten der einzelnen Eigenschaften findet sich bei Backhaus et al. (2016, S. 529 ff.).

Reale Anwendungen der Conjoint-Analyse sind mit mehr Produktmerkmalen und mehreren Ausprägungen dieser Merkmale meist wesentlich komplexer. Daher kommt es darauf an, die Zahl der zu beurteilenden (hypothetischen) Produkt-Alternativen überschaubar zu halten, um die Auskunftspersonen nicht zu überfordern. Bei der im vorstehenden Beispiel unterstellten sogenannten **Profilmethode** (Kombinationen von je einer Ausprägung aller betrachteten Merkmale, die von der Auskunftsperson in eine Rangfolge gebracht werden müssen) kann die Zahl zu beurteilender Stimuli schnell anwachsen. Im angeführten Beispiel waren es acht Stimuli, die in eine Rangfolge gebracht werden mussten. Hätten alle drei Eigenschaften jeweils drei Ausprägungen gehabt, dann wären bereits 27 Stimuli in eine Rangfolge zu bringen. Eine Alternative zur Profilmethode stellt die **Trade-Off-Analyse** dar, bei der die Auskunftsperson jeweils nur zwei Stimuli miteinander vergleich muss. Ein Ansatz zur Vereinfachung besteht auch darin, nicht alle, sondern nur eine (systematische) Auswahl der unterschiedlichen Kombinationen von Produktmerkmalen bewerten zu lassen („reduziertes Design“).

In jüngerer Zeit hat eine spezielle Form der Conjoint-Analyse, die sogenannte **Choice-Based Conjoint-Analyse,** in Wissenschaft und Praxis große Bedeutung gewonnen. Dabei ist die Vorgehensweise insofern stärker an reales Kaufverhalten angenähert, als die Auskunftspersonen nur eine der Alternativen auswählen müssen oder sich auch für eine Ablehnung aller Alternativen entscheiden können und nicht – wie im obigen Beispiel – alle Alternativen in eine Rangfolge bringen müssen. Bei dieser Form der Conjoint-Analyse erfolgt keine Analyse auf individueller Ebene, deren Ergebnisse dann über alle Auskunftspersonen zusammengefasst werden. Vielmehr werden die notwendigen Parameter auf der Basis der Daten über eine größere Zahl von Auswahlentscheidungen geschätzt. Unabhängig davon muss die Auskunftsperson – wie generell bei der Conjoint-Analyse – auf der Basis der Abwägung von Vor- und Nachteilen der Alternativen zu Einschätzungen bzw. Entscheidungen kommen.

Hintergrundinformation

Hair et al. (2010**, S. 308) zu den Vorteilen der Choice-Based Conjoint-Analyse:**

> „Mit dem vorrangigen Ziel, den Entscheidungsprozess der Auskunftsperson zu verstehen und ihr Verhalten im Markt vorherzusagen, geht die traditionelle Conjoint-Analyse davon aus, dass eine auf Rangordnungen oder Ratings basierende Beurteilung die Auswahlentscheidungen der Auskunftspersonen erfasst. Jedoch weisen Forscher darauf hin, dass dieses nicht der realistischste Weg ist, um den tatsächlichen Entscheidungsprozess der Auskunftsperson darzustellen (…) Es ist eine alternative Conjoint-Methode entstanden, bekannt als

Choice-Based-Conjoint, mit dem Vorteil größerer Inhaltsvalidität, der darauf beruht, dass die Auskunftsperson eine Alternative aus einer Menge von Alternativen auswählen muss."

9.6 Faktorenanalyse

Die exploratorische Faktorenanalyse ist ein Verfahren der Interdependenz-Analyse, das oftmals zur Datenreduktion verwendet wird. Es geht bei der Faktorenanalyse darum, dass sich eine Vielzahl von beobachteten („manifesten") Variablen bzw. die zwischen ihnen existierenden Beziehungen auf wenige dahinterstehende („latente") Faktoren zurückführen lässt. Als Beispiel für die Grundidee soll hier das klassische und allseits bekannte Beispiel der Intelligenzmessung in der Psychologie verwendet werden. Ein Phänomen wie Intelligenz lässt sich natürlich nicht direkt messen. Jeder erkennt, dass eine Frage „Wie intelligent sind Sie?" wohl einigermaßen grotesk wäre. Üblich ist vielmehr die Messung unterschiedlicher Fähigkeiten einer Person (Sprachvermögen, Wissen, Bilder und Zahlenfolgen ergänzen, Problemlösung etc.). Von diesen verschiedenen manifesten Variablen wird dann auf den dahinterstehenden Faktor(!) Intelligenz geschlossen. Entsprechende Anwendungen findet man auch in der Marktforschung, wenn von unterschiedlichsten (manifesten) Angaben zu Einstellungen und Konsumverhalten auf Faktoren wie „Genussorientierung" oder „Innovationsfreude" geschlossen wird.

Beispiel

Philip Zimbardo und Richard Gerrig (2004, s. 412 f.) zum Einsatz der Faktorenanalyse in der Intelligenzforschung

„Die Psychometrie ist das Gebiet der Psychologie, das sich mit dem Testen mentaler Fähigkeiten befasst. Darin eingeschlossen sind die Persönlichkeitsdiagnostik, Intelligenzdiagnostik und Eignungsprüfungen. Insofern sind psychometrische Ansätze eng mit den Testmethoden verwandt. Diese Theorien untersuchen statistische Beziehungen zwischen den verschiedenen Maßen geistiger Fähigkeit, (…). Auf der Basis dieser Beziehungen werden dann Schlussfolgerungen über die Beschaffenheit der menschlichen Intelligenz gezogen. Die am häufigsten verwendete Technik ist die Faktorenanalyse, ein statistisches Verfahren, das eine kleinere Zahl von Faktoren aus einer größeren Menge unabhängiger Variablen extrahiert. Ziel der Faktorenanalyse ist es, die grundlegenden psychologischen Dimensionen der untersuchten Konstrukte zu identifizieren…

Die Anwendung der Faktorenanalyse auf das Forschungsgebiet der Intelligenz durch Charles Spearman war eine der ersten und einflussreichsten. Spearman fand heraus, dass die Leistungen von Personen in je verschiedenen Intelligenztests hoch miteinander korrelierten. Er zog aus diesem Muster den Schluss, dass es einen Faktor allgemeiner Intelligenz gibt, den so genannten g-Faktor oder Generalfaktor der Intelligenz, der jeder Intelligenzleistung zugrunde liegt." ◀

Voraussetzung für die Anwendung der Faktorenanalyse sind intervallskalierte Daten. Wenn Variablen stark korreliert sind, dann kann man versuchen, anstelle einer größe-

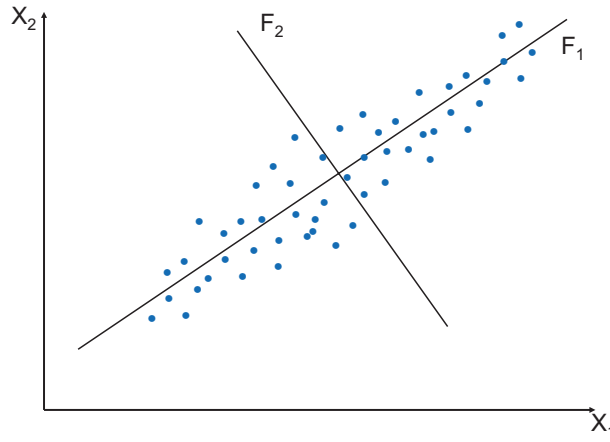


Abb. 9.14 Datenreduktion (im einfachsten Fall) durch Wahl geeigneter Faktoren

Tab. 9.5 Beispiel zur Korrelationsmatrix

	V1	V2	V3	V4	V5	V6
V1	1					
V2	0,66	1				
V3	0,05	0,12	1			
V4	0,52	0,77	0,16	1		
V5	0,10	0,02	0,78	0,08	1	
V6	0,72	0,56	0,09	0,69	0,15	1

ren Zahl von Variablen eine (deutlich) kleinere Zahl von Faktoren zu identifizieren, die bei Inkaufnahme eines gewissen Informationsverlusts den Datensatz weitgehend beschreiben. Die **Grundidee** sei an einem einfachen Beispiel mit nur zwei Variablen gezeigt, das in Abb. 9.14 dargestellt ist.

In der Abb. 9.14 ist erkennbar, dass die Variablen X_1 und X_2 offenbar hoch korreliert sind. Wenn man nun in der dort dargestellten Weise statt der ursprünglichen Variablen X_1 und X_2 neue „künstliche" Variablen ($\rightarrow$ Faktoren) F_1 und F_2 verwendet, also das Koordinatensystem X_1, X_2 durch das Koordinatensystem F_1, F_2 ersetzt, dann kann man möglicherweise bei einer späteren Interpretation der Daten den Faktor F_2 vernachlässigen. Faktor F_1 ist ja so gewählt worden, dass er den größten Teil der Varianz des Datensatzes erklärt. Bei realen Anwendungen geht es natürlich nicht darum, zwei Variable durch eine zu ersetzen, sondern darum, eine große Zahl von Variablen durch wenige Faktoren zu ersetzen.

Die *typische Vorgehensweise der Faktorenanalyse* ist durch folgende Schritte gekennzeichnet:

1. Berechnung einer Korrelationsmatrix

2. Faktorextraktion und Bestimmung der Kommunalitäten
3. Bestimmung der Zahl der Faktoren
4. Faktorrotation und -interpretation
5. Ermittlung von Faktorwerten

Ausgangspunkt ist im Gegensatz zu den bisher betrachteten Verfahren also nicht die Datenmatrix, sondern eine **Korrelationsmatrix**, in der die Korrelationen der für die Analyse ausgewählten Variablen abgebildet werden. In Tab. 9.5 findet sich ein Beispiel für eine Korrelationsmatrix. Offensichtlich reicht eine Dreiecksmatrix aus, da die Matrix selbst symmetrisch ist. In der Diagonale steht der Wert 1, der die Korrelation einer Variablen mit sich selbst kennzeichnet. An dem (sicherlich sehr vereinfachten) Beispiel lässt sich schon eine Zwei-Faktoren-Lösung auf der Basis der Korrelationen vermuten, denn die Variablen V1, V2, V4 und V6 sind stark miteinander korreliert ebenso wie die Variablen V3 und V5. Ob eine Korrelationsmatrix überhaupt geeignet ist für die Anwendung der Faktorenanalyse, lässt sich vorab anhand unterschiedlicher Tests ermitteln. Eine ausführliche Beschreibung dieser Verfahren findet sich z. B. bei Backhaus et al. (2016, S. 395 ff.).

Für die Extraktion der Faktoren nutzt man die inzwischen schon fast vertrauten Konzepte linearer Modelle. *Jede Variable* lässt sich als *Linearkombination von Faktoren* darstellen. Die sogenannten **Faktorladungen** geben an, wie stark der jeweilige Faktor mit den (manifesten) Variablen korreliert ist. Diese Beziehung wird durch das sogenannte **Fundamentaltheorem der Faktorenanalyse** erfasst, das besagt, dass sich die Korrelationsmatrix der Variablen durch die Matrix der Faktorladungen reproduzieren lässt.

Es gibt unendlich viele Möglichkeiten zur Darstellung der verschiedenen Variablen durch Linearkombinationen von Faktoren. Eine zentrale Idee der Faktorenanalyse besteht nun darin, den 1. Faktor aus der Vielzahl von Möglichkeiten so zu wählen, dass er möglichst viel Varianz der (manifesten) Variablen erklärt, den 2. Faktor so, dass ein möglichst großer Teil der noch nicht erklärten Varianz durch diesen 2. Faktor erklärt wird usw. Geht man von standardisierten Variablen aus, d. h. jede Variable hat einen Mittelwert von 0 und eine Varianz von 1, dann lässt sich der Anteil der Varianz einer Variable, der durch die Gesamtheit aller Faktoren erklärt wird, auch numerisch erfassen. Man bezeichnet dies als **Kommunalität**. Ist die Kommunalität gleich 1, dann wird die Gesamtvarianz einer Variablen vollständig durch die Gesamtheit aller Faktoren erklärt. Offensichtlich steigt mit dem Wert der Kommunalität aber auch die Anzahl der für die Varianzerklärung benötigten Faktoren. Nun bringt es keinen großen Erkenntniszuwachs, wenn man n Variable durch n Faktoren darstellt. Im Gegenteil: Die Interpretation wäre viel abstrakter und schwieriger. Man macht sich also das vorstehend skizzierte Vorgehen zu Nutze und beschränkt sich unter Inkaufnahme einer gewissen Ungenauigkeit (d. h. einer *Kommunalität kleiner als 1*) auf die Betrachtung der (relativ wenigen) Faktoren, die den größten Teil der Varianz der manifesten Variablen erklären. Damit wäre die Aufgabe der Datenreduktion (viele Variable → wenige Faktoren) weitgehend gelöst.

Wie wird nun die Zahl der zu extrahierenden Faktoren bestimmt? Dazu bedient man sich des sogenannten Eigenwerts. Der **Eigenwert** gibt den Varianzbeitrag eines Faktors im Hinblick auf die Varianz aller Variablen wieder. Geht man von standardisierten Variablen aus, die eine Varianz von 1 besitzen, so wird erwartet, dass ein Faktor einen *Eigenwert größer als 1* besitzt. Andernfalls wäre die Erklärung der Varianz der Beobachtungswerte durch einen Faktor kleiner als durch die Variable selbst, weshalb auf den Faktor im Prinzip verzichtet werden kann. Die Extraktion von Faktoren mit einem Eigenwert größer eins wird als Kaiser-Kriterium bezeichnet. Ein anderes Kriterium („Scree-Test“) besteht darin, solange Faktoren einzubeziehen bis der Eigenwert des nächsten Faktors deutlich geringer ist als beim vorherigen Faktor.

Für die Interpretation der extrahierten Faktoren haben die oben schon erwähnten **Faktorladungen** zentrale Bedeutung. Das sind Korrelationen zwischen den Faktoren und den gemessenen Variablen. Wenn beispielsweise in einer Untersuchung zum Konsumentenverhalten Messungen zu Interesse an Diätprodukten, Häufigkeit von Vorsorgeuntersuchungen, sportlichen Aktivitäten und Interesse an Publikationen zu Gesundheitsfragen hoch mit einem Faktor korreliert sind („auf diesen Faktor laden“), dann wird man diesen Faktor vielleicht als „Gesundheitsbewusstsein“ interpretieren. Um eine klare Interpretationsmöglichkeit zu erhalten, nimmt man oftmals eine sogenannte **Faktorenrotation** vor. Die Extraktion von Faktoren führt nämlich in der Regel dazu, dass mehrere Variable zunächst auf mehreren Faktoren ähnlich gut laden. Ziel aber ist eine Faktorladungsmatrix mit *möglichst vielen Ladungen nahe Null*, sodass jeder Faktor nur mit einigen Variablen korreliert und vor allem jede Variable nur *mit wenigen, am besten mit genau einem Faktor korreliert*. Weitere Einzelheiten zu verschiedenen Rotationsverfahren finden sich bei Backhaus et al., 2016, Kap. 7.

In einem letzten Schritt werden dann die **Faktorwerte** ermittelt. Faktorwerte sind die Ausprägung eines Faktors bei einem Fall, also bei Beispiel der Intelligenzmessung der Intelligenzwert einer Person. Dadurch lassen sich die Faktoren wie alle anderen Variablen in der Datenmatrix für weitere Analyseverfahren verwenden. Beispielsweise könnte man die so erzeugte Variable zum Intelligenzwert einer Person in einer Regression verwenden, um festzustellen, ob es ein systematischer Zusammenhang, bspw. zwischen Intelligenzwert und Einkommen einer Person besteht.

Vorstehend skizziert wurde die sogenannte **explorative Faktorenanalyse,** bei der die verschiedenen Faktoren durch entsprechende Analyse des Datensatzes identifiziert („entdeckt“) werden. Diese wird auch im Rahmen der Multi-Item-Skalenentwicklung eingesetzt, um mögliche Items zu identifizieren, die miteinander stark korrelieren und somit ein gemeinsames Konstrukt bzw. die Dimension eines Konstrukts bilden können. Dagegen geht es bei der **konfirmatorischen Faktorenanalyse** darum zu prüfen, ob eine vorab theoretisch unterstellte Faktorenstruktur mit einem Datensatz hinreichend gut übereinstimmt (siehe z. B. Weiber & Mühlhaus, 2014, S. 143 ff.). Die konfirmatorische Faktorenanalyse ist die Grundlage für die Anwendung von Strukturgleichungsmodellen, auf die im folgenden Abschnitt eingegangen wird.

9.7 Strukturgleichungsmodelle

In den letzten ca. 20 Jahren haben sogenannte Strukturgleichungsmodelle in der wissenschaftlichen Marketingforschung starke Beachtung und vielfältige Anwendungen gefunden und auch in der kommerziellen Marktforschung kommen diese Modelle häufiger zum Einsatz. In der deutschsprachigen Literatur findet sich für entsprechende Ansätze auch die Bezeichnung **„Kausalmodelle“**. Der Begriff „Kausalmodelle“ ist insofern problematisch, als die Anwendungen in der Regel auf Querschnittsdaten beruhen, die keine Überprüfung von Kausalitäten im strengeren Sinn (siehe Abschn. 2.4.1) erlauben. Für die Berechnung von Strukturgleichungsmodellen sind spezielle Softwareprogramme erforderlich. Zu den gängigsten Programmen zählen AMOS (als Modul in SPSS) und LISREL, für die Berechnung von Partial Least Squares-Verfahren arbeitet man häufig mit SmartPLS oder PLS-Graph. Darüber hinaus gibt es auch entsprechende Module für das Open-Source Programmpaket R.

▶ **Hair et al. (2010, S. 679) gehen auf die Problematik der Kausalität bei Strukturgleichungsmodellen ein:**

> „Strukturgleichungsmodelle können Kovariationen auf der Basis von Tests von Variablenbeziehungen im Model•l nachweisen. Strukturgleichungsmodelle können jedoch nicht nachweisen, dass eine Ursache vor einer Wirkung geschieht, weil meist Daten aus Querschnitts-Untersuchungen verwendet werden. Strukturgleichungsmodelle auf der Basis von Daten aus Längsschnitt-Untersuchungen können zeitliche Sequenzen nachweisen. Der Nachweis einer Beziehung von Ursache und Wirkung kann dadurch teilweise geleistet werden. Wenn zusätzliche alternative Gründe die Beziehung zwischen Ursache und Wirkung nicht eliminieren, dann wird der Nachweis der Kausalität stärker ... Strukturgleichungsmodelle sind nützlich bei der Entwicklung von Kausalitäten, aber die Nutzung von Strukturgleichungsmodellen gewährleistet nicht den Nachweis von Kausalitäten.“

Die Grundidee von Strukturgleichungsmodellen besteht darin, dass auf der Grundlage von ermittelten *Varianzen und Kovarianzen von Indikatoren* (manifesten Variablen) Schlüsse im Hinblick auf *Abhängigkeitsbeziehungen zwischen komplexen Konstrukten* (latenten Variablen oder Faktoren) gezogen werden. Hair et al. (2010, S. 634 f.) sehen die charakteristischen Merkmale von Strukturgleichungsmodellen darin, dass eine größere Zahl miteinander verbundener Abhängigkeitsbeziehungen analysiert wird und gleichzeitig nicht direkt beobachtete Konzepte in diese Beziehungen einbezogen werden können, wobei Messfehler explizit berücksichtigt werden können. Es geht um die Prüfung von Theorien bezogen auf die Existenz latenter Variabler und deren Zusammenhänge.

Zunächst zu einer Illustration des Aspekts der Analyse mehrerer Abhängigkeitsbeziehungen. Langer et al. (2008) untersuchen Einflussfaktoren und Konsequenzen der Konsumentenverwirrtheit beim Kauf von Produkten mit ökologischen Gütesiegeln. Das Modell ist in Abb. 9.15 wiedergegeben.

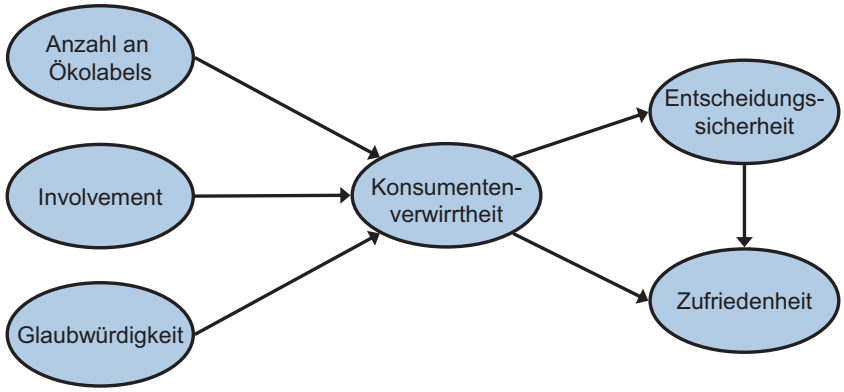


Abb. 9.15 Beispiel eines Strukturmodells. (Quelle: Langer et al., 2008)

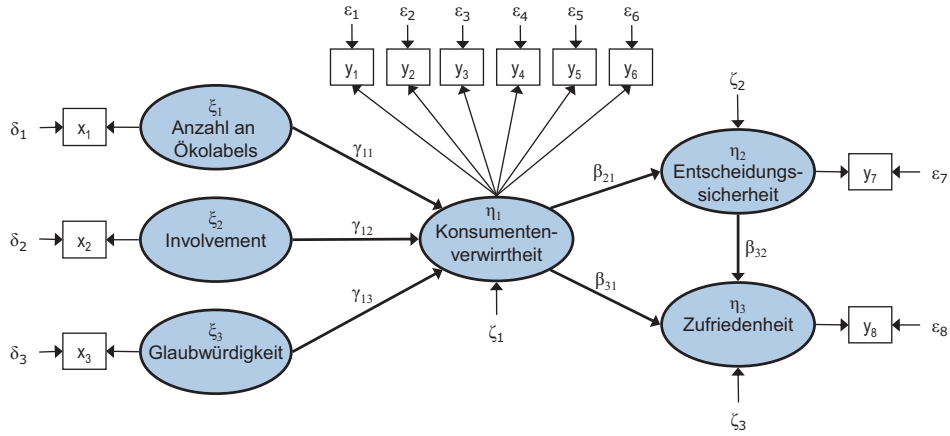


Abb. 9.16 Beispiel eines Struktur- und Messmodells. (Quelle: Langer et al., 2008)

Man erkennt sofort die vielfältigen direkten und indirekten Abhängigkeiten zwischen den betrachteten Konzepten. Ein solches Modell wird als **Strukturmodell** bezeichnet. Es beschreibt *Beziehungen zwischen latenten Variablen*.

Der zweite Aspekt – Berücksichtigung von Messfehlern – wird deutlich, wenn man sich ansieht, wie das Modell bzw. seine Variablen gemessen wurden. In Abb. 9.16 findet sich wieder im Zentrum das Strukturmodell mit den latenten Variablen, die durch griechische Buchstaben gekennzeichnet sind, ebenso wie die Beziehungen, die sogenannten Pfade zwischen den latenten Variablen. Außerdem findet sich für jede der latenten Variablen ein **Messmodell**. Die latenten Variablen, die häufig nicht direkt beobachtbar sind, werden mithilfe von sogenannten Indikatorvariablen geschätzt, die im Modell durch Quadrate und lateinische Buchstaben gekennzeichnet sind. Im vorliegenden Beispiel ist die latente Variable „Konsumentenverwirrtheit“ mit sechs Indikatoren gemessen wor-

den, für alle anderen Variablen wurde jeweils nur ein Indikator verwendet. Schließlich sind noch die Messfehler eingetragen (ebenfalls mit griechischen Buchstaben, aber ohne Kreis oder Quadrat). Sie verweisen darauf, dass bei der Messung einer latenten Variablen anhand eines oder mehrerer Indikatoren in der Regel stets *ein (möglichst kleiner) Fehler* vorliegt.

Wie geht man nun bei der Analyse von Strukturgleichungsmodellen vor? Typischerweise unterscheidet man wieder drei Schritte (Backhaus et al., 2016):

1. Modellformulierung
2. Parameterschätzung
3. Beurteilung der Schätzergebnisse

Modelle werden auf der Basis theoretischer bzw. sachlogischer Überlegungen formuliert. Bei der Formulierung der Messmodelle ist darauf zu achten, in welcher Beziehung latente Variable und Indikator zueinander stehen. Werden die Indikatoren durch eine latente Variable verursacht, dann spricht man von **reflektiven Indikatoren**. Das ist in der Regel bei Einstellungen der Fall, wobei bestimmte Äußerungen („Die Marke X finde ich gut“) auf eine zugrundeliegende Einstellung schließen lassen („Einstellung zur Marke X“). Bei der Verwendung von mehreren Indikatoren müssen diese dann auch hoch korrelieren. Letztendlich hat man es also auch wieder mit einer *faktoranalytischen Vorgehensweise* zu tun. Allerdings werden hier die Indikatoren nicht explorativ ermittelt, sondern sind aufgrund von Vorüberlegungen definiert. Die Stärke der Beziehung zwischen Indikator und latenter Variable (Faktor) lässt sich dann auch testen, weshalb diese Messmodelle dem Ansatz der **konfirmatorischen Faktorenanalyse** folgen.

Verursachen hingegen die Indikatoren die latente Variable, spricht man von **formativen Indikatoren** (vgl. Abschn. 4.3.2.3). Versucht man beispielsweise die soziale Schicht zu ermitteln und bedient man sich der Indikatoren Einkommen, Bildung und Ansehen des Berufs, kann man eher davon ausgehen, dass die Indikatoren die latente Variable beeinflussen und nicht umgekehrt.

Die Schätzung der Parameter kann auf unterschiedliche Weise geschehen. Einmal können sukzessive für die Messmodelle zunächst die Faktorwerte bestimmt werden, die in einem zweiten Schritt als Messwerte für die latenten Variablen in eine Regressionsanalyse eingehen. Dieser Vorgehensweise folgt das (varianzbasierte) **Partial Least Squares** (PLS)- Verfahren. In der Marketingforschung wird auch häufig die *simultane Schätzung aller Parameter* angewendet (kovarianzbasierte Verfahren). Dabei geht man von einer Kovarianz- oder Korrelationsmatrix „S“ der manifesten Variablen aus. Zu schätzen sind nun die unbekannten Parameter (z. B. die Pfadkoeffizienten im Strukturmodell). Wichtige Voraussetzung ist ein *identifiziertes Modell,* d. h. die Zahl der zu schätzenden unbekannten Parameter darf höchstens so hoch sein wie die Zahl der empirisch vorliegenden Korrelationen und Varianzen. Die Korrelationsmatrix lässt sich nun algebraisch auch so darstellen, dass jede Korrelation eine Funktion der unbekannten Parameter ist. Bei der simultanen Parameterschätzung wird nun versucht, mit der

modelltheoretischen Korrelationsmatrix „Σ“ die vorliegende empirische Korrelationsmatrix S möglichst gut zu reproduzieren, d. h. die Differenz zwischen der modelltheoretischen Korrelationsmatrix Σ und der empirischen Korrelationsmatrix S zu minimieren. Dies geschieht natürlich durch die Zuhilfenahme entsprechender Software.

Die *Beurteilung der Schätzergebnisse* erfolgt durch diverse Gütemaße und inferenzstatistische Tests. Man unterscheidet globale von lokalen Gütemaßen. Während *globale Gütemaße* eine Beurteilung der Konsistenz des Gesamtmodells mit den erhobenen Daten ermöglichen, erlauben *lokale Gütemaße* die Überprüfung der Messgüte einzelner Indikatoren und latenter Variablen. Bei den **globalen Gütemaßen** wird zwischen inferenzstatistischen Anpassungsmaßen (Chi^2-Teststatistik, RMSEA) und deskriptiven Anpassungsmaßen (z. B. AGFI, NFI, NNFI, CFI) unterschieden. Deskriptive Anpassungsmaße beruhen im Gegensatz zu den inferenzstatistischen Anpassungsmaßen nicht auf statistischen Tests, sondern im Wesentlichen auf Faustregeln. **Lokale Gütemaße** erlauben die Beurteilung der Messgüte einzelner Indikatoren (Indikatorreliabilität) und Faktoren bzw. latenter Variabler (Faktor- bzw. Konstruktvalidität, durchschnittlich erfasste Varianz). In Tab. 9.6 findet sich eine Übersicht über verschiedene Anpassungsmaße sowie die geforderten Anspruchsniveaus für einen guten „Fit“.

Weitere wichtige Schritte zur Überprüfung der Schätzung eines Strukturgleichungsmodells sind die Prognosegüte und die Kreuzvalidierung. Bei der **Prognosegüte** wird das Modell auf Basis eines Teils der Daten geschätzt und die Güte des Modells durch dessen Anwendung auf den Rest der Daten geprüft. Bei der **Kreuzvalidierung** wird das Modell auf Basis einer Stichprobe geschätzt und dann geprüft, ob das Modell die entsprechenden Strukturen bei einer anderen Stichprobe hinreichend gut beschreibt.

Eine im Zusammenhang der empirischen Marketingforschung stark beachtete Anwendung von Strukturgleichungsmodellen bzw. konfirmatorischen Faktoranalysen besteht in der *Überprüfung von Messinstrumenten* hinsichtlich der durch den Ansatz der Multitrait-Multimethod-Matrix MTMM (siehe Abschn. 4.3.2.6) gegebenen Anforderungen zur Konstruktvalidität, nämlich der Konvergenz- und Diskriminanzvalidität. Dazu formuliert man ein entsprechendes Modell, das den Anforderungen der

Tab. 9.6 Anspruchskriterien ausgewählter Gütemaße in Anlehnung an Weiber und Mühlhaus (2014)

Globale Anpassungsmaße	Anspruchsniveaus	Lokale Anpassungsmaße	Anspruchsniveaus
RMSEA	≤0,05–0,08	Indikatorreliabilität	≥0,4 und ≤0,9
Chi^2/d.f	≤3	Faktorreliabilität	≥0,6
SRMR	≤0,10	Durchschnittlich erfasste Varianz	≥0,5
NFI	≥0,90		
TLI (NNFI)	≥0,90		
CFI	≥0,90		

MTMM hinsichtlich der Korrelationen zwischen den unterschiedlichen Variablen entspricht, und prüft die Übereinstimmung dieses Modells mit den vorliegenden Daten (siehe dazu Marsh & Grayson, 1995; Bagozzi & Yi, 2012). Verschiedene konfirmatorische Faktormodelle werden eingesetzt, um eine Trennung von Trait-, Methoden- und Messfehleranteilen zu ermöglichen und die Gültigkeit zugrundeliegender Annahmen (Eindimensionalität, Korreliertheit, Unkorreliertheit) zu überprüfen (Schermelleh-Engel und Schweizer (2012). Für **Konvergenzvalidität** spricht es, wenn die Indikatoren eines Faktors stark miteinander korreliert sind. Die **Diskriminanzvalidität** kann auf zweierlei Art und Weise überprüft werden. Beim Fornell/Larcker-Kriterium wird gefordert, dass alle quadrierten Korrelationen zwischen den latenten Variablen jeweils unter den durchschnittlich erfassten Varianzen dieser latenten Variablen liegen. Beim Chi^2-Differenztest wird die Differenz der Güte zwischen einem speziellen Modell, bei dem die Korrelation zwischen den latenten Variablen auf eins fixiert wird, und einem allgemeinen Modell, bei dem diese Restriktion nicht gegeben ist, berechnet. Zur Erfüllung des Kriteriums ist eine signifikante Chi^2-Differenz gefordert, d. h. der Modellfit zwischen beiden Modellen soll sich signifikant unterscheiden und zwar so, dass das spezielle Modell einen schlechteren Fit aufweist als das allgemeine Modell.

9.8 Clusteranalyse

Bei der Clusteranalyse handelt es sich um ein Verfahren der **Interdependenz-Analyse,** bei dem unterschiedliche Berechnungsalgorithmen für die verschiedenen Messniveaus von Daten zur Verfügung stehen. Die Zielsetzung dieser Methode besteht darin, ähnliche Objekte (z. B. Personen) zu Gruppen („Clustern“) zusammenzufassen. Im Wesentlichen geht es darum, die Objekte anhand einer Vielzahl von Merkmalen bzw. Merkmalsausprägungen so zu Gruppen zusammenzufassen, dass diese Gruppen in sich möglichst homogen sind und sich untereinander deutlich voneinander unterscheiden. Daraus wird schon erkennbar, dass die *Marktsegmentierung* ein typischer Anwendungsbereich der Clusteranalyse im Marketing ist.

Die *Grundidee* der Clusteranalyse lässt sich am einfachsten im – für die Anwendung dieser Methode natürlich unrealistischen – Fall mit nur zwei Variablen darstellen, die intervallskaliert sind. Abb. 9.17 zeigt ein entsprechendes Beispiel.

Dort sind die Messwerte von einigen Personen in ein Koordinatensystem mit den Variablen „Alter“ und „Kinobesuche pro Jahr“ eingetragen. Man erkennt, dass sich die Ähnlichkeit der Objekte hinsichtlich der betrachteten Variablen durch die Abstände der Objekte im Merkmalsraum ausdrücken lässt. Objekte (Personen), deren Abstände im Merkmalsraum gering sind, sind sich ähnlich und werden zu einer Gruppe zusammengefasst. Bei den typischen Anwendungen der Clusteranalyse mit einer großen Zahl von Variablen lassen sich die verschiedenen identifizierten Gruppen natürlich nicht mehr so einfach wie in dem Beispiel darstellen.

Für den Ablauf einer Clusteranalyse sind ebenfalls drei Schritte typisch:

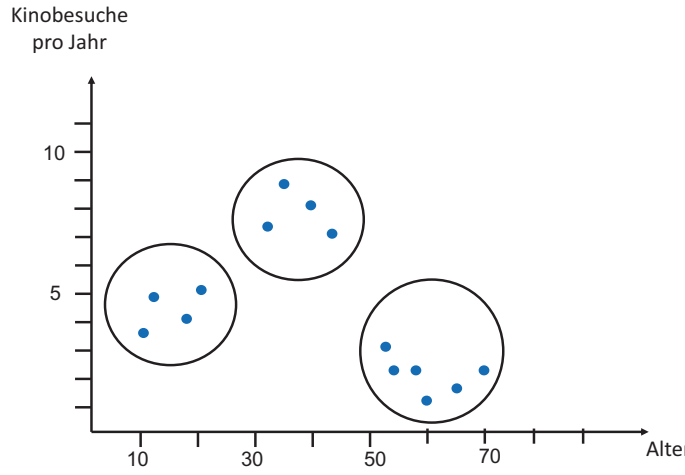


Abb. 9.17 Beispiel zur Clusteranalyse mit (nur) zwei Variablen

1. Auswahl der für die Gruppenbildung heranzuziehenden Variablen (z. B. soziodemographische Merkmale, Einstellungsvariablen, Merkmale des Lebensstils).
2. Quantifizierung von Ähnlichkeiten bzw. Unähnlichkeiten von Objekten anhand eines sogenannten Proximitätsmaßes und Ermittlung einer Distanz- bzw. Ähnlichkeitsmatrix.
3. Zusammenfassung der Objekte zu in sich homogenen Gruppen auf Basis der Werte des Proximitätsmaßes mithilfe der Anwendung eines Fusionsalgorithmus.

Bei der Festlegung des **Proximitätsmaßes** kommt es stark auf das Messniveau der verwendeten Daten (siehe Abschn. 7.2) an. Bei metrischen Daten (Intervall- oder Ratioskalierung) verwendet man *Distanzmaße,* die den Abstand der Objekte anhand der entsprechenden Ausprägungen bei bestimmten Variablen (z. B. Alter, Einkommen, Nutzungsintensität oder Kaufhäufigkeit bei einem Produkt) messen. Bei nicht-metrischen Daten (Nominal- oder Ordinalskalierung) verwendet man *Ähnlichkeitskoeffizienten,* die z. B. den Anteil gleicher Ausprägungen nominalskalierter Variabler (Geschlecht, Familienstand, PKW-Besitz etc.) als Maß für die Ähnlichkeit von zwei Objekten verwenden.

Der letzte wesentliche Schritt bei der Clusteranalyse ist die Anwendung von **Fusionierungsalgorithmen** (auch als Clusteralgorithmen oder –methoden bezeichnet). Darunter versteht man unterschiedliche *Methoden der Zuordnung von Objekten zu Gruppen* im Hinblick auf die schon angesprochenen Kriterien (Homogenität innerhalb der Gruppen, Unterschiede zwischen den Gruppen). Dabei werden die folgenden Verfahren unterschieden:

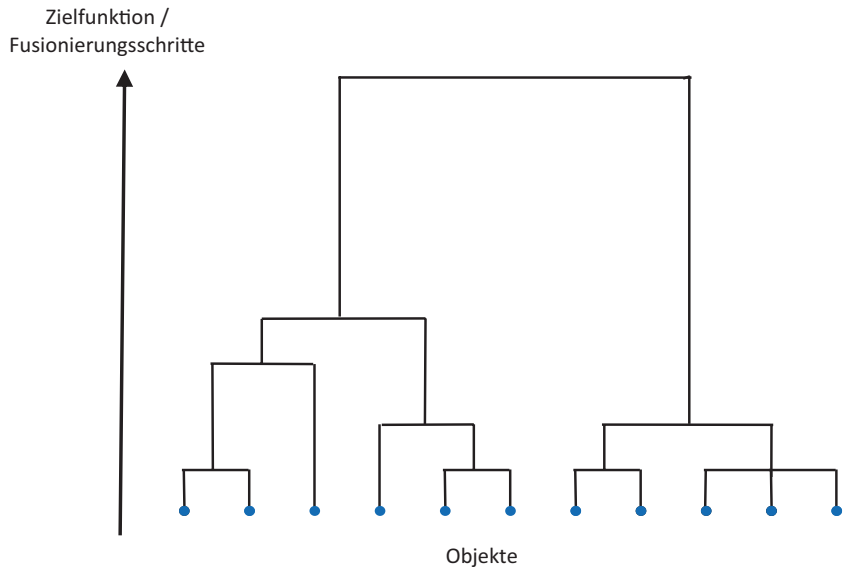


Abb. 9.18 Dendrogramm zur Bestimmung der Anzahl der Cluster

Bei **hierarchischen Verfahren** ist die Zuteilung eines Objekts zu einem Cluster *endgültig*. Man unterscheidet zwischen agglomerativ-hierarchischen und divisiv-hierarchischen Verfahren. Bei agglomerativen Verfahren geht man von der feinsten Partition aus, d. h. jedes Objekt bildet ein eigenes Cluster und fusioniert sukzessive Objekte bzw. Cluster anhand ihrer Ähnlichkeit. Bei divisiven Verfahren geht man von der gröbsten Partition aus, d. h. alle Objekte sind in einem Cluster, und man bildet sukzessive mehrere Cluster, indem die unähnlichsten Objekte bzw. Cluster herausgenommen werden und neue Cluster bilden. Den hierarchischen Verfahren ist gemein, dass die Zuordnung eines Objekts zu einem Cluster endgültig ist und keine nachträglichen „Umsortierungen" vorgenommen werden können.

Die **nicht-hierarchischen** oder **partitionierenden Verfahren** gehen von einer gegebenen Gruppierung der Objekte aus und ordnen die Objekte zwischen den Gruppen so lange um, bis ein *Optimalwert einer Zielfunktion* erreicht ist. Bei den sogenannten „K-means" Algorithmen geschieht das so, dass eine bestimmte – vom Anwender festgelegte – Zahl von Clustern gebildet wird.

Eine typische Vorgehensweise bei der Clusteranalyse ist das Identifizieren der Anzahl der Cluster durch hierarchische Verfahren in einem ersten Schritt. Diese Anzahl dient als Vorgabe für die Anwendung nicht-hierarchischer Verfahren, mit denen man in einem zweiten Schritt die Zuordnung der Objekte zu den Clustern zu optimieren versucht. Bei der Identifizierung der Anzahl der Cluster helfen unter anderem graphische Verfahren,

wie das **Dendrogramm**, das in Abb. 9.18 für ein agglomerativ-hierarchisches Verfahren dargestellt ist (siehe z. B. Backhaus et al., 2016).

Das Dendrogramm zeigt an, wie die einzelnen Objekte *sukzessive zu Clustern fusioniert* werden (in der Abbildung von unten nach oben). Für die Entscheidung, wie viele Cluster zu wählen sind, d. h. welche (möglichst geringe) Anzahl von Clustern die optimale Homogenität in den Gruppen und zugleich ausreichend Heterogenität zwischen den Gruppen gewährleistet, ist die jeweilige Zielfunktion maßgeblich. Dies könnte z. B. die Distanz einzelner Objekte vom Clustermittelpunkt sein, die bei homogenen Clustern natürlich möglichst gering sein sollte. Im dargestellten Beispiel (Abb. 9.18) würde man sich wohl für zwei Cluster entscheiden, denn die Fusionierung von zwei Clustern zu einem einzigen Cluster verschlechtert die Zielfunktion erheblich, während die Fusionierung von vier zu zwei Clustern nur eine geringe Verschlechterung bedeutet.

▶ **Auf den großen Interpretationsspielraum bei der Durchführung einer Clusteranalyse verweisen Hair et al. (2010, S. 515):** „Die Auswahl der finalen Clusterlösung erfordert eine grundlegende Beurteilung durch den Forscher und wird von vielen als zu subjektiv eingeschätzt. Obwohl sehr gründliche Methoden entwickelt wurden, um Clusterlösungen beurteilen zu können, bleibt es immer noch dem Forscher überlassen, zu entscheiden, wie viele Cluster letztendlich angenommen werden. Auch die Entscheidung über die einzubeziehenden Merkmale, die Clustermethode, und auch die Interpretation der Cluster beruhen genauso auf dem Urteil des Forschers… Es ist daher unabdingbar, dass Forscher sich um die bestmögliche Objektivierung bemühen und sich von substantiellen Überlegungen leiten lassen, vor allem bei der Festlegung des Designs und der Interpretation.“

9.9 Ausblick: Nutzung von Künstlicher Intelligenz in der Marktforschung

Will man sich mit künstlicher Intelligenz (KI) in der Marktforschung beschäftigen, so muss man diesen Begriff zunächst definieren. Die Schwierigkeit dabei ist: „Aktuell gibt es noch keine allgemein anerkannte Definition von künstlicher Intelligenz, da hierfür zunächst ‚Intelligenz‘ als Begriff geklärt sein müsste, was ebenfalls noch nicht der Fall ist.“ (Bünte, 2020, S. 54).

Für das Thema ist eine breite Definition ohne Bezug auf den Menschen erforderlich. Eine solche Definition liefert Max Tegmark. Er definiert Intelligenz als „Fähigkeit, kom-

plexe Ziele zu erreichen“ und künstliche Intelligenz als „nicht-biologische Intelligenz“ (Tegmark 2017, S. 45). Dies beinhaltet auch die Fähigkeit, Wissen und Fähigkeiten zu erwerben, also zu lernen. Im Folgenden wird diese sehr breite Definition zugrunde gelegt.

Die derzeit wichtigste Anwendung in der Marktforschung ist die Klassifikation von Daten. Dies war auch die älteste bekannte Anwendung von KI in der Marktforschung, die damals jedoch noch nicht unter diesem Namen bekannt war.

Diese Anwendung entwickelte in den 1980er Jahren durch die US-Firma IRI. Hintergrund war, dass IRI ab 1987 ein auf wöchentlichen Scannerdaten basierendes Handelspanel aufbaute. Dies führte gegenüber den traditionellen Panels, deren Daten im 2-Monatsrhythmus manuell erhoben wurden, zu einer Vervielfachung der Datenmengen. Um diese angemessen prüfen zu können, setzte IRI sehr erfolgreich Künstliche Neuronale Netze (KNN) ein. Dazu wurden Beispieldatensätze zunächst manuell als fehlerhaft einschließlich Fehlertyp oder als fehlerfrei klassifiziert. Liefert z. B. ein Geschäft Scannerdaten, bei denen die Verkaufszahlen von Woche zu Woche stiegen, so ist dies ein Indiz dafür, dass die Verkäufe kumuliert erfasst wurden. Die entsprechenden Datensätze wurden als fehlerhaft mit dem Fehlertyp „Kumulation“ klassifiziert. Die so klassifizierten Datensätze wurden als Trainingsdatensätze verwendet, um ein KNN zu generieren, mit dessen Hilfe ein Großteil der Datensätze automatisch klassifiziert und teilweise auch gleich korrigiert werden konnte. So konnten beim Fehlertyp „Kumulation“ die richtigen Verkäufe durch Subtraktion der Werte der Vorwoche korrigiert werden. Das KNN liefert auch die Information, wie sicher eine Klassifikation ist. Unsichere Klassifikationen wurden an einen Bearbeiter übergeben, der eine Klassifikation vornahm und den Datensatz als weiteren Trainingsdatensatz zur Verbesserung des KNN verwendete. Damit zeigt dieses Beispiel alle Merkmale von KI: Lösung komplexer Aufgaben, die durch Lernen sukzessive zu immer besseren Ergebnissen kommt.

Schon diese sehr frühe Anwendung zeigt die wesentlichen Elemente der KI: Es sind Daten erforderlich, die (oft noch manuell) klassifiziert sind. Diese Daten werden durch ein KNN geschickt, welches dann die wichtigen Zusammenhänge erfasst. Diese Zusammenhänge können dann auf neue Datensätze angewendet werden. Wichtig ist auch, dass bereits bestehende Zusammenhänge durch weitere klassifizierte Daten verbessert werden können.

Seitdem wurden die KNN unter dem Stichwort „Deep Learning“ weiter deutlich verbessert. Zusammen mit der deutlich gesteigerten Computerleistung wurden automatische Bild- und Textanalysen möglich. Entsprechend ist die Zahl der Anwendungen deutlich gewachsen. Die folgenden Beispiele machen dabei die Bandbreite deutlich.

- Der GfK Verein (heute: NIM e. V.) hat 2012 eine Methode vorgestellt, mit der Emotionen beim Betrachten von Werbefilmen aufgrund des Gesichtsausdrucks mit KI erfasst werden können. (Vgl. Echtzeitmessung von Emotionen bei Konsumentscheidungen|NIM e. V.) Die Firma Ipsos hat ein ähnliches System im Einsatz.

- Ebenfalls der GfK Verein hat 2016 eine Methode vorgestellt, bei der Social Media Bilder mit KI auf Markennamen durchsucht werden und die Bilder als positiv, negativ oder neutral für die Marke klassifiziert werden (Kaiser, 2017).
- Das Startup „Deepsight“ analysiert mit KI Antworten auf offene Fragen (Müssigmann, 2021).
- GfK analysiert mit KI Paneldaten zur Unterstützung bei strategischen Entscheidungen, Planungen und Prognosen (ohne Verfasser, 2021)
- Die Firma Produkt + Markt erstellt mit KI Transkripte von qualitativen Interviews und analysiert diese, ermittelt Likes und Dislikes und zeichnet erste Wortwolken (Schomberg & Junker, 2022)
- Auch Ipsos setzt KI zur Verbesserung der qualitativen Forschung ein (vgl. Gespräche mit KI Teil II: Enthüllung der KI-Qualität in qualitativen workstreamslIpsos).
- Die Firma Behaviourally untersucht mit KI die Qualität von Verpackungen (Dössel, 2022).

Trotz dieser Beispiele konstatiert Tiedemann (2022) aufgrund einer Umfrage unter betrieblichen Marktforschern, dass KI in der betrieblichen Marktforschung zwar grundsätzlich positiv beurteilt wird, der Einsatz aktuell aber vor allem in Randtätigkeiten wie z. B. in der Übersetzung von Texten erfolgt. Für die Zukunft wird zwar eine Änderung der Arbeitswelt in der Marktforschung erwartet, eine Reduktion der Stellen für Marktforschung oder einen Wegfall des eigenen Arbeitsplatzes durch KI befürchtet nur eine Minderheit.

Als einen wichtigen Grund für die noch geringe Anwendungsbreite der KI sind laut Tiedemann Datenschutzbedenken. Die Anwendung der KI erfordert in der Regel, dass die Daten in eine Cloud hochgeladen werden. Viele Systeme sehen dann vor, dass diese Daten auch als Trainingsdatensätze verwendet werden können. Damit werden die Daten weitergegeben und weiterverarbeitet, was bei personenbezogenen Daten oder sonst vertraulichen Daten nur unter engen Voraussetzungen möglich ist.

Das Thema KI hat in der Öffentlichkeit insbesondere durch die Veröffentlichung von ChatGPT im November 2022 neue Aufmerksamkeit erfahren. Microsoft hat ChatGPT im Mai 2023 in seine Suchmaschine „Bing“ integriert und kann dort auch kostenlos ausprobiert werden (www.bing.de). Diese sehr generelle Anwendung der KI ist derzeit (Herbst, 2023) aber noch nicht so zuverlässig, dass eine professionelle Anwendung in der Marktforschung angezeigt ist. Ob und wie ein künftiger Einsatz möglich ist, ist derzeit noch nicht abzusehen.

Literatur

Backhaus, K., Erichson, B., Plinke, W., & Weiber, R. (2016). *Multivariate Analysemethoden – Eine anwendungsorientierte Einführung* (14. Aufl.). Springer.

Bagozzi, R., & Yi, Y. (2012). Specification, evaluation, and interpretation of structural equation models. *Journal of the Academy of Marketing Science, 40*, 8–34.
Berry, W. (1993). Understanding regression assumptions. In M. Lewis-Beck (Hrsg.), *Regression analysis* (S. 335–424). Sage.
Bortz, J. (2005). *Statistik für Sozial- und Humanwissenschaftler* (S. 280/281). Springer, Heidelberg.
Bünte, C. (2020). *Die chinesische KI-Revolution*. Springer.
Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). N.J.: L. Erlbaum Associates, Hillsdale.
Cohen, J. (1992). A power primer. *Psychological Bulletin, 112*, 155–159.
Cortina, J., & Nouri, H. (2000). *Effect size for ANOVA designs*. Sage.
Dössel, C. (2022). Packaging design optimiert. In *Planung & analyse 3/2022* (S. 23–25).
Echtzeitmessung von Emotionen bei Konsumentscheidungen|NIM e. V. https://www.nim.org/forschung/uebersicht-forschungsprojekte/forschungsprojekt/echtzeitmessung-von-emotionen-bei-konsumentscheidungen. Zugegriffen: 20. Nov. 2023.
Fox, J. (1997). *Applied regression analysis, linear models, and related methods*. Sage.
Gespräche mit KI. Teil II: Enthüllung der KI-Qualität in qualitativen workstreams|Ipsos. https://www.ipsos.com/de-de/gesprache-mit-ki-teil-ii-enthullung-der-ki-qualitat-qualitativen-arbeits-ablaufen. Zugegriffen: 20. Nov. 2023
Hair, J., Anderson, R., Tatham, R., & Black, W. (2010). *Multivariate data analysis* (7. Aufl.). Prentice Hall.
Hayes, A. F. (2013). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. The Guilford Press.
Jaccard, J., & Becker, M. (2002). *Statistics for the behavioral sciences* (4. Aufl.). Wadsworth.
Kaiser, C. (2017). Ein Bild sagt mehr als 100 Worte. In *Planung & analyse 1/2017* (S. 51–53).
Langer, A., Eisend, M., & Kuß, A. (2008). Zu viel des Guten? Zum Einfluss der Anzahl von Ökolabels auf die Konsumentenverwirrtheit. *Marketing – ZFP, 30,* 19–28.
Marsh, H., & Grayson, D. (1995). Latent variable models of multitrait-multimethod data. In R. Hoyle (Hrsg.), *Structural equation modeling* (S. 177–198). Sage.
Menard, S. (2002). *Applied logistic regression analysis* (2. Aufl.). Sage.
Müssigmann, L. (2021). Datenanalyse in Rekordzeit. In *Planung & analyse 1/2021* (S. 17).
Ohne Verfasser. (2021). *GfK stellt neue Plattform gfknewron vor*. In *Planung&analyse 4/2021* (S. 7).
Schermelleh-Engel, K., & Schweizer, K. (2012). Multitrait-Multimethod-Analysen. In A. Kelava & H. Moosbrugger (Hrsg.), *Testtheorie und Fragebogenkonstruktion* (2. Aufl., S. 345–362). Springer.
Schomberg, J., & Junker, H. (2022). Insights – so schnell wie nie. In *Planung & analyse 3/2022* (S. 56–58).
Sedlmaier, P., & Renkewitz, F. (2008). *Forschungsmethoden und Statistik in der Psychologie*. Pearson.
Skiera, B., & Albers, S. (2008). Regressionsanalyse. In A. Herrmann, C. Homburg, & M. Klarmann (Hrsg.), *Handbuch Marktforschung* (3. Aufl., S. 467–497). Springer Gabler.
Sudman, S., & Blair, E. (1998). *Marketing research – A problem solving approach*. Mcgraw-Hill.
Tegmark, M. (2017). *Life 3.0 – Being human in the age of artificial intelligence*. Allen Lane.
Tiedemann, F. (2022). Keine Sorge, durch KI ersetzt zu werden. In *Planung & analyse, 3/2022* (S. 41–43).
Weiber, R., & Mühlhaus, D. (2014). *Strukturgleichungsmodellierung* (2. Aufl.). Springer.
Zimbardo, P., & Gerrig, R. (2004). *Psychologie* (16. Aufl.). Pearson.

Forschungsethik und Datenschutz

10

Zusammenfassung

Im Zusammenhang der Marktforschung stellen sich ethische Fragen vor allem im Hinblick auf drei Aspekte: Wie stark darf man bei Auskunfts- oder Versuchspersonen Stress verursachen und in deren Intimsphäre eindringen? Wie soll sich das Verhältnis zwischen Auftraggebern für eine Untersuchung und den Marktforschern gestalten? Was geschieht mit den erhobenen Daten und wie wird die Vertraulichkeit gewahrt? Im deutschsprachigen Raum existieren hierzu spezielle Regelungen zum Datenschutz. Auf diese Aspekte wird im 10. Kapitel kurz eingegangen.

Ein großer Teil der in diesem Buch vorgestellten Methoden ist geeignet, um vielfältige Informationen über Personen und Haushalte zu sammeln, die auch die Intimsphäre der Betroffenen berühren können. Darüber hinaus kann der Prozess der Datenerhebung für Auskunfts- und Versuchspersonen auch psychischen Stress und Zeitaufwand bedeuten. Es stellen sich also Fragen, *was der Forscher den Auskunftspersonen zumuten darf* und wie er sich beispielsweise hinsichtlich der Nutzung der gewonnenen Daten verhalten soll. Andererseits ist auch an die *Position der Auftraggeber aus Wirtschaft und Gesellschaft* zu denken, die nicht nur mit teilweise beachtlichen Budgets die Untersuchungen bezahlen, sondern auch weitreichende Entscheidungen auf der Basis entsprechender Untersuchungsergebnisse treffen. Wie muss sich der Marktforscher verhalten, um diesem Vertrauen gerecht zu werden? Welche Mindeststandards müssen vor diesem Hintergrund bei Untersuchungen eingehalten werden? Derartige Fragen berühren das Gebiet der Forschungsethik, das in den letzten Jahrzehnten zunehmende Beachtung gefunden hat.

Nun gibt es Forschungsgebiete in den Natur- und Sozialwissenschaften, in denen sich wesentlich gravierendere ethische Probleme stellen als in der Marketingforschung. Hier sei nur an die breite öffentliche Diskussion zur Genforschung erinnert. Auch in der medizinischen Forschung, z. B. beim klinischen Test von Medikamenten, wird die Ein-

H. Kreis et al., *Marktforschung,* https://doi.org/10.1007/978-3-658-44456-3_10

haltung ethischer Standards sorgfältig beachtet. Im sozialwissenschaftlichen Bereich ist die Sensibilität des Umgangs mit Informationen über psychische Merkmale oder politische Meinungen von Personen leicht nachvollziehbar. Im Vergleich dazu sind viele Fragestellungen der *praktischen* Marketingforschung (z. B. nach der präferierten Kaffee-Marke oder der Häufigkeit des Konsums von Mineralwasser) eher unproblematisch. Gleichwohl gibt es auch in der Marketingforschung Untersuchungsgegenstände und Forschungspraktiken, die zumindest **ethische Fragen** aufwerfen. Dazu einige Beispiele:

- Durch eine Marktuntersuchung soll geklärt werden, wie bei bestimmten Süßwaren, deren Konsum bei Kindern zu Übergewicht und Ausbreitung von Karies beiträgt, der Geschmack verändert werden soll, damit der Konsum durch Kinder steigt. Ist eine Untersuchung mit diesem Ziel ethisch vertretbar?
- Durch **„Mystery Shopping“**, also durch den Einsatz als Kunden getarnter Mitarbeiter und deren Beobachtungen, soll die Beratungsqualität und die Kompetenz von Verkäufern ermittelt werden. Ist eine solche Täuschung zulässig, auch wenn sie für die betroffenen Verkäufer vielleicht zu Nachteilen an ihrem Arbeitsplatz führt?
- Kreditkarten-Anbieter können mit den bei ihnen vorhandenen Daten individuelle Konsumprofile ihrer Kunden erstellen, die zur gezielten Ansprache durch bestimmte Anbieter genutzt werden könnten. Wäre ein solcher Gebrauch der Daten mit einem Eindringen in die Privatsphäre zulässig?

Weitere Beispiele für ethische Probleme und deren Einschätzung durch Marketing- und Marktforschungs-Praktiker stellen z. B. Akaah und Riordan (1989) dar.

Beispiel

Ein in der breiteren Öffentlichkeit recht bekanntes Beispiel für eine ethische Problematik auch bei sozialwissenschaftlichen Untersuchungen ist das so genannte „Milgram-Experiment“, benannt nach Stanley Milgram (1933–1984), der diese Untersuchung konzipiert und durchgeführt hat (Milgram, 1963).

In der Untersuchung ging es um Gehorsam, genauer gesagt um die Frage, inwieweit sich Menschen Autoritäten unterordnen, selbst wenn sie unmenschliche Handlungen ausführen sollen. Hintergrund der Untersuchung von Milgram war die Erfahrung, dass während der Zeit des Nationalsozialismus einzelne ansonsten ganz „normale“ oder „durchschnittliche“ Menschen in der Hierarchie bestimmter (militärischer) Organisationen schlimmste Verbrechen begingen und dies später mit einem so genannten „Befehlsnotstand“ begründeten.

An der Untersuchung von Milgram waren Personen beteiligt, die drei verschiedene Rollen hatten:

- Eine (angebliche) *„Testperson“* (TP). Dabei handelte es sich um einen Schauspieler, der bei den einzelnen Beobachtungen ein bestimmtes Verhalten (s. u.) simulierte.
- Einen (angeblichen) *Versuchsleiter* (VL), der die – damals noch wirksame – Autorität eines seriösen Wissenschaftlers ausstrahlte und der (eigentlichen) Versuchsperson (s. u.) Anweisungen für ihr Verhalten gab.
- Die jeweilige (eigentliche) *Versuchsperson* (VP), der vorgetäuscht wurde, dass sie an einem „Lern-Experiment“ teilnimmt.

Ablauf der einzelnen Beobachtungen:
Die VP kam in ein Labor und erhielt vom VL Anweisungen, der TP Lernaufgaben zu geben. Als Untersuchungsziel wurde der VP vermittelt, dass es darum gehe, ob Bestrafungen zu einem besseren Lernerfolg führen. Nach jeder falschen Antwort erhielt die TP per Knopfdruck durch die VP (scheinbar) einen Stromstoß, der sich von Mal zu Mal angeblich steigerte, bis auf 450 V. Nach jedem (simulierten) Stromstoß zeigte die TP ihre schauspielerischen Fähigkeiten mit entsprechenden Reaktionen – vom etwas schmerzverzerrten Gesicht bei leichten „Stromstößen“ bis zu schweren Reaktionen und Schreien nach starken „Stromstößen“. Der VL insistierte mit seiner Autorität, dass die VP immer weiter fortfährt bis hin zu lebensbedrohlichen Strafen.

Das (erschreckende) Untersuchungsergebnis war, dass 26 von 40 VP den Anweisungen des VL bis zur Verabreichung von lebensgefährlichen Stromstößen in Höhe von 450 V folgten. Die VP zeigten dabei Anzeichen extremen psychischen Stresses. Hier das Beispiel einer Äußerung einer Versuchsperson nach dieser Untersuchung (aufgezeichnet von Heinz Schuler, 1991, S. 332 f.):

> „Ich finde es grausam, so Menschen zu quälen, so ein Experiment zu machen…. Ich finde das schändlich, ich konnte nicht mehr, ich bin in ärztlicher Behandlung, meine Nerven sind nicht die besten … ich bekomme Valium … ich hatte richtige Depressionen. (…) Es ist furchtbar gewesen … Das Geschrei ging mir durch Mark und Bein. Ich könnte keinen Menschen so quälen … Ich habe mir nicht vorgestellt, dass ich so etwas machen muss. Ich glaube, das verfolgt mich noch tagelang.“

Fragen zur Forschungsethik: Darf man – auch bei einer wichtigen Forschungsfrage wie in diesem Fall – Versuchspersonen unter solch extremen Stress setzen? Kann man es verantworten, dass den Versuchspersonen für den Rest ihres Lebens bewusst ist, dass sie bereit gewesen wären, jemand in akute Lebensgefahr zu bringen, um irgendwelchen Anweisungen zu entsprechen? ◄

Nach diesen Beispielen sollte die Relevanz der Forschungsethik deutlich geworden sein. Allgemein geht es hier bei der Forschungsethik um die *Einhaltung moralischer Prinzipien, Werte und Verhaltensweisen in Situationen, in denen durch Marktforschung Schaden entstehen kann*. „Ethik kann definiert werden als ein Untersuchungsfeld, in dem bestimmt wird, welche Verhaltensweisen als angemessen angesehen werden“ (Burns & Bush, 2006, S. 63). In der Ethik wird meist zwischen einem **deontologischen** und einem

teleologischen Ansatz unterschieden (siehe auch Hansen, 1995). Beim erstgenannten erfolgt die Beurteilung einer Handlung nicht nach deren Folgen, sondern danach, ob sie in sich richtig oder falsch ist. Maßstab dafür ist der Schutz der Rechte des Individuums. Eine entsprechende Regel für die Forschungsethik könnte dabei beispielsweise darin bestehen, dass eine Versuchsperson *auf keinen Fall* in psychischen Stress geraten darf. Der teleologische Ansatz bezieht sich dagegen auf eine *Abwägung* von positiven (für eine Gruppe oder eine Gesellschaft insgesamt) und negativen (für Versuchspersonen) Folgen einer Untersuchung. Beispielsweise könnte man damit eine begrenzte Belastung von Versuchspersonen (z. B. durch kurzfristiges Unwohlsein) bei der Entwicklung eines wichtigen Medikaments zur Heilung schwerster Krankheiten rechtfertigen. Es geht gewissermaßen um eine Art von „Kosten-Nutzen-Analyse", wobei die „Kosten" in Form von Belastungen der Versuchspersonen in (engen) Grenzen bleiben müssen.

Nicht zuletzt liegt ethisches Verhalten auch im Interesse der Markt- und Sozialforscher selbst. Die Branche ist auf die Auskunftsbereitschaft der Bevölkerung angewiesen und diese wäre bei unethischem Verhalten gefährdet.

Ein grober Rahmen für Entscheidungen, bei denen Aspekte der Ethik berührt werden, ist durch *Konventionen und entsprechende rechtliche Regelungen in einer Gesellschaft* vorgegeben. Gleichwohl bleiben viele Entscheidungen in unterschiedlichen Situationen dem/der einzelnen ForscherIn und seinem/ihrem persönlichen Urteil überlassen. Trotz der Schwierigkeit bei der Formulierung allgemeiner Regeln kommen Sudman und Blair (1998, S. 644) zu drei einfachen Empfehlungen, die hier (frei übersetzt) wiedergegeben seien:

- „Was Du nicht willst, das man Dir tu, das füg auch keinem anderen zu."
- „Nichts tun, bei dem man Sorge haben müsste, wenn es öffentlich bekannt wird."
- „Wenn Dein Gefühl Dir sagt, dass irgendetwas nicht in Ordnung ist, dann folge diesem Gefühl."

Auch die anwendungsorientierte Marktforschung erhebt ja den Anspruch der Wissenschaftlichkeit und begründet damit die Aussagekraft und Verlässlichkeit ihrer Untersuchungsergebnisse. Insofern gelten einige wesentliche Grundsätze der Forschungsethik (siehe dazu Resnik, 2008; Eisend & Kuß, 2023) hier natürlich ebenfalls. Zunächst ist der Grundsatz der *wissenschaftlichen Ehrlichkeit* zu nennen. Das bedeutet, dass die relevanten Einzelheiten (auch die Schwachstellen) der methodischen Vorgehensweise und der Ergebnisse korrekt berichtet werden. Ebenso gehört *Sorgfalt* zu den elementaren Prinzipen wissenschaftlicher Arbeit, d. h. dass Fehler bei einer Untersuchung (im Rahmen der vorhandenen Ressourcen) minimiert werden sollen. Weiterhin sollen MarktforscherInnen trotz einer gewissen Abhängigkeit von Auftraggebern und deren Interessen um *Objektivität* bemüht sein. Auch den Auftraggebern wäre ja dauerhaft nur durch möglichst korrekte (also möglichst objektive) Ergebnisse gedient. Im Hinblick auf den Gesichtspunkt der *Offenlegung von Ergebnissen* muss man sicher Unterschiede zwischen angewandter Forschung und Grundlagenforschung akzeptieren. Während im wissen-

schaftlich-theoretischen Bereich Untersuchungsergebnisse und methodische Einzelheiten umfassend publiziert werden, um den wissenschaftlichen Austausch zu fördern, ist im Bereich der kommerziellen Anwendungen und des kommerziellen Wettbewerbs die Publikation von Ergebnissen eher die Ausnahme und auch methodische Einzelheiten werden außenstehenden InteressentInnen oft vorenthalten.

Ein wesentlicher Aspekt ist der **Datenschutz in der Marktforschung.** Beim Datenschutz geht es immer um den Schutz personenbezogener Daten. Darunter versteht man alle Informationen, die einer natürlichen Person zugeordnet werden können. Beispielsweise ist eine Telefonnummer ein personenbezogenes Datum, weil sie einer Person, nämlich dem Inhaber des Anschlusses, zugeordnet werden kann. Im Mai 2018 trat die von der EU erlassene Datenschutzgrundverordnung (DS-VGO) nach einer Vorlaufphase von zwei Jahren in Kraft. Substanziell waren die Regelungen des älteren Bundesdatenschutzgesetzes sehr ähnlich. Allerdings wurden die Strafen gegen Verstöße drastisch verschärft.

Hintergrundinformation

Mit der im Mai 2018 in Kraft getretenen Datenschutzgrundverordnung (DS-GVO, z. B. https://dsgvo-gesetz.de/art-89-dsgvo/)) verfolgte die EU auch das Ziel, gleiches Datenschutzrecht innerhalb der Gemeinschaft zu installieren. Dies gelang nicht vollständig. Deshalb muss die DS-GVO immer mit dem neuen Bundesdatenschutzgesetz (BDSG) zusammen gelesen werden. Daneben spielen für die Interpretation die sogen. Erwägungsgründe (https://dsgvo-gesetz.de/erwaegungsgruende/) eine erhebliche Rolle. Zu dem komplexen Gebiet gibt es umfangreiche Literatur, weshalb hier nur ein grober Überblick gegeben werden kann.

Grundsätzlich gilt, dass die Verarbeitung von personenbezogenen Daten verboten ist, es sei denn, es gibt eine spezielle Erlaubnisnorm (Art. 6 DS-GVO). Die für die Marktforschung wichtigsten gesetzlichen Gründe sind:

1. Die Einwilligung der Betroffenen zur Verarbeitung ihrer personenbezogenen Daten. Diese Einwilligung muss freiwillig und informiert erfolgen (Art. 7 und Art. 12). Sie kann widerrufen werden und die Betroffenen können die Löschung ihrer Daten verlangen (Art. 17 und Art. 21).
2. Die Verarbeitung ist zur Erfüllung eines Vertrags notwendig mit den Betroffenen (z. B. müssen Arbeitgeber personenbezogene Daten ihrer Arbeitnehmer speichern).
3. Die Verarbeitung ist aufgrund eines Gesetzes notwendig. So müssen die Daten von Empfängern von Incentives aufgrund von Buchführungsvorschriften 10 Jahre gespeichert werden.

Ist die Verarbeitung grundsätzlich erlaubt, so gelten dennoch wichtige Beschränkungen (Art. 5). Ausgangspunkt ist, dass ein legitimer Zweck der Datenverarbeitung definiert werden muss. Jegliche Verarbeitung personenbezogener Daten muss diesen Zwecken dienen. Das heißt: Es dürfen nur solche Daten verarbeitet werden, die zu Erfüllung des Zwecks notwendig sind. Die Daten dürfen auch nur so lange mit ihrem Personenbezug gespeichert werden, als es zur Erfüllung des Zwecks notwendig ist. Für die Marktforschung bedeutet das z. B., dass Namen und Adressen von Interviewten zu löschen sind, sobald die Datenkontrollen abgeschlossen sind, es sei denn es sind Wiederholungsbefragungen geplant und mit den Betroffenen auch vereinbart. Weiter sind die Daten technisch vor unberechtigten Zugriffen zu schützen. Eine Hilfestellung bietet die „Checkliste für den Datenschutz“ von ESOMAR (2016).

Wichtig für die Marktforschung ist, dass es Erleichterungen für die wissenschaftliche Forschung gibt (Art. 89 DS-GVO) und gemäß Erwägungsgrund 159 wissenschaftliche Forschung breit definiert ist und auch die privat finanzierte Forschung einschließt.

Gegenüber dem früheren Datenschutzrecht drastisch verschärft wurden die möglichen Strafen für Verstöße, die nun bis zu 20 Mio. Euro oder 4 % des weltweiten Jahresumsatzes gehen können, je nachdem, was höher ist (Art. 83 DS-GVO).

Ohne Vertrauen in den Datenschutz wird ein hoher Anteil der kontaktierten Personen es ablehnen, ein Interview zu geben. Ist der Anteil der antwortenden Personen jedoch gering, dann leidet besonders bei Zufallsstichproben die Repräsentativität (vgl. Abschn. 4.2.2). Schon aus diesem Grund ist ein hohes Niveau des Datenschutzes in der Marktforschung wichtig. Darüber hinaus hat auch der Gesetzgeber eine Reihe von Gesetzen zum Datenschutz erlassen. Zentral ist – wie erwähnt – die DS-GVO zusammen mit dem Bundesdatenschutzgesetz (BDSG). Werden Fernmeldeeinrichtungen genutzt (z. B. Telefon- oder Onlinestichproben) ist darüber hinaus das Telekommunikationsgesetz wichtig.

Die in Abschn. 4.4.4. thematisierten impliziten Methoden der Marktforschung sind mit Blick auf ihre ethische Vertretbarkeit differenziert zu betrachten. Auf der einen Seite ermöglichen implizite Methoden, wie beispielsweise die Analyse von biometrischen Messungen oder das Facial Coding, tiefere Einblicke in die Präferenzen und Reaktionen der Verbraucher, was für Unternehmen wertvolle Informationen bedeutet. Auf der anderen Seite stellen diese Methoden jedoch mindestens potenzielle Datenschutz- und Privatsphäreherausforderungen dar. Es ist entscheidend, dass Unternehmen transparent über ihre Datensammlungspraktiken im Sinne der GS-GVO informieren und sicherstellen, dass die Einwilligung der Verbraucher eingeholt wird.

Im Hinblick auf das **Verhältnis Marktforscher/Auftraggeber** stellen sich natürlich andere ethische Fragen. Nicht zuletzt geht es auch um das Vertrauen der Nutzer von Marktforschung und die daraus resultierende Akzeptanz und Relevanz von Ergebnissen (Moorman et al., 1993). Wichtig für die Marktforschungspraxis ist die *Vertraulichkeit von Untersuchungsergebnissen und von Informationen,* die der Auftraggeber für die Untersuchung gegeben hat. Das kann zu Problemen führen, wenn ein Marktforschungsinstitut für mehrere Auftraggeber in einer Branche tätig wird. Dann ist es oftmals schwer zu vermeiden, dass früher erworbenes branchenbezogenes „Hintergrundwissen" bei späteren Untersuchungen für andere Auftraggeber genutzt wird. Jeder Auftraggeber hat natürlich Anspruch auf sorgfältige *Einhaltung der methodischen Standards*, einschließlich der korrekten Durchführung von Untersuchungsteilen, die schlecht einsehbar sind (z. B. Durchführung von Interviews). Eine *verzerrte Darstellung von Untersuchungsergebnissen* ist jedenfalls zu *vermeiden (→ Objektivität)*, auch wenn diese im (vordergründigen) Interesse des Auftraggebers liegen könnte. Beispielsweise könnte ja eine Werbeagentur daran interessiert sein, dass sich nach der Durchführung einer Werbekampagne günstige Werte bei einer Umfrage zu Bekanntheitsgrad oder Image des betreffenden Produkts ergeben. Ferner müssen alle aus Untersuchungsanlage, Stichprobengröße etc. resultierenden *Begrenzungen* (→ Genauigkeit und Sicherheit der Ergebnisse)

Tab. 10.1 Wichtige Organisationen der Markt- und Sozialforschung mit „Codes of Ethics", „Codes of Conduct", „Best Practices" etc.

Organisation	Abkürzung	Internet-Adresse
American Marketing Association	AMA	www.ama.org
American Psychological Association	APA	www.apa.org
ESOMAR	ESOMAR	www.esomar.org
American Association for Public Opinion Research	AAPOR	www.aapor.org
Rat der Deutschen Markt- und Sozialforschung e. V.	–	www.rat-marktforschung.de
Arbeitskreis Deutscher Markt- und Sozialwissenschaftlicher Institute e. V.	ADM	www.adm-ev.de
Arbeitsgemeinschaft Sozialwissenschaftlicher Institute e. V.	ASI	www.asi-ev.org
Berufsverband Deutscher Markt- und Sozialforscher e. V.	BVM	www.bvm.org
Deutsche Gesellschaft für Online-Forschung e. V.	DGOF	www.dgof.de
Verband der Hochschullehrer für Betriebswirtschaft e. V.	VHB	www.vhbonline.org

Abb. 10.1 Arten ethischer Probleme in der Marktforschung

offengelegt werden. Letztlich gilt die Anforderung an die Marktforscher, dass sie keine Untersuchungsergebnisse für *unethische Zwecke* (z. B. betrügerische Verkaufspraktiken) zur Verfügung stellen.

Um die ethischen Grundsätze und methodischen Mindeststandards der empirischen Markt- und Sozialforschung weiter zu konkretisieren, haben wichtige einschlägige Organisationen diese in sogenannten „Codes of Ethics" (o. ä.) zusammengefasst. In der Übersicht (Tab. 10.1) werden einige wichtige Quellen dafür genannt. Der Leserin und dem Leser sei die Beschäftigung damit empfohlen. Bei ethischen Fragestellungen im Zusammenhang der Marktforschung stehen die Beziehungen zu den Auskunfts-/Versuchspersonen auf der einen Seite und zum Auftraggeber auf der anderen Seite im Mittelpunkt (siehe Abb. 10.1). Dazu kommen als dritter Aspekt noch die Pflichten gegenüber dem Berufsstand der Markt- und Sozialforscher, dessen Ansehen zu schützen ist, um weiter die Auskunftswilligkeit der Bevölkerung zu erhalten. Dies beinhaltet beispielsweise das Verbot, andere Markt- und Sozialforscher unberechtigt zu kritisieren.

Bei der Kodifizierung der „Codes of Ethics“ kommt der internationalen Marktforschungsorganisation ESOMAR eine Schlüsselrolle zu. Diese hat zusammen mit der International Chamber of Commerce in 2016 den „ICC/ESOMAR International Code on Market, Opinion and Social Research and Data Analytics“ herausgegeben, der auf der ESOMAR-Website eingesehen werden kann. Dieser wurde von den deutschen Verbänden ADM, ASI, BVM und DGOF angenommen und durch eine „Erklärung für das Gebiet der Bundesrepublik Deutschland zum ICC/ESOMAR Internationaler Kodex zur Markt-, Meinungs- und Sozialforschung sowie zur Datenanalytik“ oder kurz „Deutsche Erklärung“ ergänzt, die auf der Website des ADM (www.adm-ev.de) oder des BVM (www.bvm.org) eingesehen werden kann.

Der ICC/ESOMAR-Code hat drei grundlegende Prinzipien:

> „1. Bei der Erhebung personenbezogener Daten für Forschungszwecke müssen die Forscher klar darüber informieren, welche Arten von Daten sie erheben wollen, den Zweck der Erhebung und an wen die Daten in welcher Form übermittelt werden sollen.
>
> 2. Forscher müssen sicherstellen, dass die in der Forschung verwendeten personenbezogenen Daten sorgfältig vor nicht autorisiertem Zugriff geschützt sind und nicht ohne die Zustimmung der Forschungsteilnehmer offengelegt werden.
>
> 3. Forscher müssen sich stets ethisch korrekt verhalten und alles vermeiden, was einem Forschungsteilnehmer Schaden zufügen oder den Ruf der Markt-, Meinungs- und Sozialforschung beeinträchtigen kann.“ (vgl. ICC/ESOMAR, 2016, S. 7, Übersetzung vom Autor)

Diese werden ergänzt durch weitere Pflichten, u. a.:

1. Sicherstellung, dass die Antworten der Befragten stets freiwillig gegeben werden (Artikel 4).
2. Einhaltung wissenschaftlicher Standards bei der Ermittlung und Publikation der Ergebnisse (Artikel 8).
3. Schutz des guten Rufes der Markt- und Sozialforschung durch ein ethisch einwandfreies geschäftliches Verhalten. Dies beinhaltet auch die Einhaltung der anerkannten Regeln des fairen Wettbewerbs und auch das Verbot, andere Marktforscher ungerechtfertigt zu kritisieren (Artikel 9).

In der „Deutschen Erklärung“ werden die ICC/ESOMAR-Richtlinien ergänzt und verschärft durch das absolute Verbot, die Identität der antwortenden Personen Dritten gegenüber offen zu legen. Dieses Anonymisierungsgebot ist in den meisten Fällen kein Problem, weil Informationen über eine relativ kleine Zahl von Einzelpersonen für die meisten Auftraggeber ohnehin kommerziell kaum nutzbar wären. Da auch der ICC/ESOMAR-Code festlegt, dass strengere lokale Standards einzuhalten sind, sind diese Regeln auch durch ESOMAR gedeckt.

ESOMAR und auch die deutschen Verbände ADM, ASI, BVM und DGOF haben zusätzliche Richtlinien publiziert, die Einzelprobleme der Forschung behandeln. So werden z. B. in der von den deutschen Verbänden herausgegebenen „Richtlinie für telefonische Befragungen“ (vgl. https://www.bvm.org/fileadmin/user_upload/Verbandsdokumente/

Standesregeln_RL_neu_2021/Richtlinie_Telefonische_Befragungen_2021.pdf) bestimmte Zeiten für Anrufe ausgeschlossen, weil ein Anruf z. B. nachts die Angerufenen zu sehr stören würde. Die Einhaltung dieser Regeln wird auch durchgesetzt. Bei ESOMAR ist es das „Professional Standards Committee", das Verstöße gegen den Code untersucht, soweit sie von ESOMAR-Mitgliedern begangen wurden. Schärfste Sanktion von ESOMAR ist ein Ausschluss von ESOMAR verbunden mit einer entsprechenden Veröffentlichung. In Deutschland ist es der „Rat der Deutschen Markt- und Sozialforschung" (www.rat-marktforschung.de), der Verstöße auch von Nicht-Mitgliedern der Verbände untersucht und gegebenenfalls auch öffentlich rügt. Eine solche veröffentlichte Rüge ist sehr geschäftsschädigend.

Literatur

Akaah, I., & Riordan, E. (1989). Judgements of marketing professionals about ethical issues in marketing research: A replication and extension. *Journal of Marketing Research, 26*, 112–120.

Burns, A., & Bush, R. (2006). *Marketing research* (5. Aufl.). Pearson.

Eisend, M., & Kuß, A. (2023). *Grundlagen empirischer Forschung (3. Aufl.)*. SpringerGabler.

ESOMAR. (Hrsg.). (2023). *Data Protection Checklist*, Amsterdam. https://esomar.org/uploads/attachments/clgdr4frf0for7g3v3v8lnw4e-data-protection-checklist-cg4-4-23-1.pdf.

ICC/ESOMAR. (Hrsg.). (2016). *ICC/ESOMAR code on market, opinion and social research and data analytic.*

Hansen, U. (1995). Ethik und Marketing. In B. Tietz, R. Köhler, & J. Zentes (Hrsg.), *Handwörterbuch des Marketing* (2. Aufl., S. 615–628). Schäffer-Poeschel.

Milgram, S. (1963). Behavioral study of obedience. *Journal of Abnormal and Social Psychology, 67*, 371–378.

Moorman, C., Deshpande, R., & Zaltman, G. (1993). Factors affecting trust in market research relationships. *Journal of Marketing, 57*, 81–101.

Resnik, D. (2008). Ethics of science. In S. Psillos & M. Curd (Hrsg.), *The Routledge companion to philosophy of science* (S. 149–158). Routledge.

Schuler, H. (1991). Ethische Probleme der (sozial)psychologischen Forschung. In H. Lenk (Hrsg.), *Wissenschaft und Ethik* (S. 331–355). Reclams.

Sudman, S., & Blair, E. (1998). *Marketing research – A problem solving approach*. McGraw-Hill.

Stichwortverzeichnis

H. Kreis et al., *Marktforschung,* https://doi.org/10.1007/978-3-658-44456-3