Wolfgang Ortmanns / Ralph Sonntag

Umfragen erstellen und auswerten

kompakt und leicht verständlich für Studierende und junge Forschende

mit **Wissens-boxen** und **Beispielen**

UVK

Umfragen erstellen und auswerten

Prof. Dr. Wolfgang Ortmanns lehrt Volkswirtschaftslehre und Finanzmärkte an der Hochschule für Technik und Wirtschaft (HTW) in Dresden.

Prof. Dr. Ralph Sonntag lehrt im Bereich Marketing und Existenzgründung und ist an der Hochschule Stralsund.

Wolfgang Ortmanns / Ralph Sonntag

Umfragen erstellen und auswerten

kompakt und leicht verständlich für Studierende und junge Forschende

UVK Verlag · München

Umschlagabbildung: © AndreyPopov · istock
Autorenfoto Wolfgang Ortmanns: © privat | Ralph Sonntag: © privat

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

Diese Open Access Publikation wurde unterstützt durch das Landesdigitalisierungsprogramm für Wissenschaft und Kultur des Freistaates Sachsen.

DOI: https://doi.org/10.24053/9783739882413

Funder: Konsortium der sächsischen Hochschulbibliotheken

Narr Francke Attempto Verlag GmbH + Co. KG
Dischingerweg 5 · D72070 Tübingen

Internet: www.narr.de
eMail: info@narr.de

CPI books GmbH, Leck

ISBN 978-3-7398-3241-8 (Print)
ISBN 978-3-7398-8241-3 (ePDF)
ISBN 978-3-7398-0633-4 (ePub)

Inhalt

Was Sie vorher wissen sollten

Die **Umfrage** ist zu einer beliebten Forschungsmethode geworden, auch schon für Studierende bei ihrer ersten Haus- oder der Bachelorarbeit. Früher musste man für Umfragen Briefe mit vorbezahlten Rückporto verschicken und Adressen anmieten. Viel zu teuer und allzu viel Aufwand für eine erste selbständige Forschung im Studium. Heute stehen im **Internet** kostenlos nutzbare Umfrageportale zur Verfügung. Die Fragebögen lassen sich über **soziale Netzwerke** breit gestreut verteilen und es gibt zahlreiche Anwendungsprogramme zur Auswertung. Für alle Forschungsfragen, für die nicht unbedingt strenge Repräsentativität gefordert wird, ist dieses eine komfortable Methode.

Leider sind die notwendigen Kenntnisse von **wissenschaftlichen Standards** dazu nicht immer so leicht zu erschließen. Dabei gilt es, einige Rahmenbedingungen und Methoden zu beachten, um belastbare **Schlussfolgerungen aus Umfragen** ziehen zu können. Schon die späteren statistischen Auswertungen mit ihrer Fülle an Tests und Voraussetzungen, stellen für Einsteiger:innen oft eine hohe Hürde dar. Jedoch gilt wohl auch hier *Paretos* 80-20-Regel:

Mit nur 20 % der Tests aus den Statistikbücher lassen sich 80 % aller Umfragen **auswerten**. Dieses Buch beschränkt sich auf das Wesentliche und ist ein guter pragmatischer Ratgeber für Umfragen. Die **relevanten Tests** werden nach heutigen Gepflogenheiten geforderten Größen der Effektstärke und Teststärke dargestellt, natürlich anschaulich erklärt und alles nachrechenbar.

Auswerten kann man aber auch mit der besten Statistik nur gut konstruierte **Fragebögen**. Das Buch gibt einen guten pragmatischen Über- und Einblick, was konkret bei der Erstellung und Umsetzung von Umfragen beachten ist. Des Weiteren werden besondere Befragungs- und Analyse-Methoden vorgestellt. Ein guter Leitfaden und kompaktes Wissen für Befragungen.

◆ **Downloads zum Buch | *Excel*-Tools**

Bevor Sie das Buch lesen, können Sie digitale Zusatzmaterialien herunterladen. Darin finden Sie zahlreiche hilfreiche *Excel*-Tools, auf die die Autoren an verschiedenen Stellen im Buch eingehen. Sie finden die Tools unter:
https://files.narr.digital/9783739832418/Zusatzmaterial.zip

1 Einführung

1.1 Der induktive Schluss

In den Wissenschaften können wir **logisch-deduktive** oder **empirisch-induktive** Schlussfolgerungen ziehen. Ein Beispiel einer deduktiven Schlussfolgerung wäre:

Wenn es regnet, wird die Straße nass
Es regnet

Die Straße wird nass

Es ist ein Schluss von der Allgemeinheit auf einen Einzelfall. Er ist logisch gültig, da es unmöglich ist, dass er bei Wahrheit der beiden gesetzten Prämissen falsch ist.

Aus einer empirischen Forschung ergeben sich üblicherweise aber nur induktive Schlüsse der Art:

Schwan 1 ist weiß
Schwan 2 ist weiß
Schwan 3 ist weiß

Alle Schwäne sind weiß

Hier wird von einigen Einzelfällen auf die Gesamtheit geschlossen. Genau das passiert beispielsweise bei einer Umfrage: Wie schließen von einer Stichprobe der Teilnehmer:innen auf eine Gesamtheit und können damit falsch liegen. Wie auch hier, denn bekanntlich gibt es ja auch schwarze Schwäne. Wann und unter welchen Angaben ist ein **Induktionschluss** in der empirischen Forschung trotzdem zulässig? Sicher nicht nach nur drei Beobachtungen wie im Schwanenbeispiel. Aber wie groß muss dann eine Stichprobe sein und wie muss man diese auswählen? Welche Fehlerwahrscheinlichkeit kann man akzeptieren und wie kann man diese berechnen? Wie sind die wissenschaftlichen Standards dazu? Diese Fragen beantwortet dieses Buch.

> **Erkenntnis** | Induktionsschlüsse aus einer Stichprobe sind nie logisch gültig und können nicht bewiesen (verifiziert) werden, jedoch können diese widerlegt (falsifiziert) werden, hier dadurch, dass man einen einzigen schwarzen Schwan erblickt.

Das Vorgehen der quantitativ-empirischen Forschung liegt meist darin, zu versuchen, das Gegenteil einer vermuteten Hypothese als so unwahrscheinlich zu widerlegen, dass man die Hypothese im Umkehrschluss als plausibel annehmen kann. So könnte man die Hypothese: „Es gib keine blauen Schwäne“ durch Beobachtung einer hinreichend großen Menge an Schwänen (Stichprobengröße), und das in der ganzen Welt verteilt (repräsentativ), testen. Finden wir trotzdem keinen blauen Schwan, ist die Hypothese zwar nicht bewiesen, aber Sie hat sich (vorläufig) bewährt. Eine Hypothese gilt als gut abgesichert, wenn sie sich in möglichst vielen Replikationsstudien immer wieder aufs Neue bewährt hat.

1.2 Marktforschung und Vorgehensweise

Unternehmen agieren auf dynamischen Märkten, mit neuen Innovationen, neuen Geschäftsmodellen, regional und global. Vor diesem Hintergrund treffen Unternehmen jeden Tag Entscheidungen, um die Nachhaltigkeit und den Erfolg des Unternehmens zu sichern. Eine Aufgabe von Analysen und Methoden besteht darin, ein gutes Fundament für diese betrieblichen Entscheidungen zu legen. Hieraus lassen sich wiederum Fragestellungen ableiten, deren Antworten für das Unternehmen wichtig sind.

Neben diesen marktbezogenen Analysen können auch Analysen in anderen Bereichen des Unternehmens durchgeführt werden. Da jedoch die meisten Analysen und Befragungen einen Bezug zu den Marktgegebenheiten und -aktivitäten des Unternehmens haben, liegt der Fokus im Folgenden auf dem Bereich der Marktorientierung und -relevanz.

Wissen | Marktforschung
Unter **Marktanalyse** und **Marktforschung** sind die systematische und methodische Beobachtung und Erfassung von Zuständen und Vorgängen auf wirtschaftlichen Märkten zu verstehen. Die Marktforschung untersucht die Absatz- und Beschaffungsmöglichkeiten eines Unterneh-

mens, um marktbezogene Informationen als Entscheidungsgrundlage zu erhalten.[1]

Die Ziele von **Marktforschung** sind vielfältig, wie beispielsweise:
Ermittlung des **Potenzials**:

- Marktkapazität
- Marktpotenzial
- Marktvolumen
- Marktanteil

Ermittlung von **Trends**:

- Einstellungsänderungen der Zielgruppen
- Konsumentenverhalten
- Bedürfnisse von Kund:innen

Risikominimierung:

- Optimierung von Marktaktivitäten
- Kontrolle von Maßnahmen
- Entscheidungsunterstützung

Dabei werden folgende Aktivitäten durchgeführt, um diese Ziele zu erreichen:

- systematische, zeitpunkt- und zeitraumbezogene Untersuchung des Marktes mit Hilfe (Datenerhebung) von wissenschaftlichen Verfahren
- Beschaffung und Auswertung von existierenden internen und/oder externen Daten

Die Datenerhebung kann auf verschiedenen Wegen erfolgen. Dabei wird zwischen Sekundärforschung und Primärforschung unterschieden.

> Unter **Sekundärforschung** wird die Nutzung und Auswertung von Daten verstanden, die bereits vorliegen und somit zu einem früheren Zeitpunkt und/oder zu einem anderen Erhebungszweck erhoben wurden.

1 vgl.: Böhler, Germelmann, Baier, & Woratschek, 2021, S. 15–16; Meffert, 2013, S. 11–13; Naderer & Balzer, 2011

Der Vorteil liegt in der sofortigen Verfügbarkeit der Daten; die Nachteile liegen darin, dass diese Daten meist nicht passgenau zu den Fragestellungen des Unternehmens passen und entsprechend nicht aktuell sind.

Häufig sind Sekundärdaten bzw. -studien verfügbar, z. B. unternehmensinterne oder externe Quellen. Diese bereits bestehenden Daten können genutzt werden, wenn diese die unternehmerische Fragestellung beantworten bzw. einen Teil zur Beantwortung beitragen können. Eine **Primärdatenerhebung** kann einerseits die bereits bestehenden Daten aus der Sekundärforschung ergänzen bzw. aktualisieren. Zum anderen kann die Datenerhebung gezielt auf die Fragestellung angepasst werden.[2]

Im Folgenden liegt der Fokus auf Marktanalyse mittels Primärdaten.

Wissen | Vorgehen einer Marktanalyse
Das Vorgehen von der Konzeption bis zur Beantwortung der Fragestellungen kann wie folgt grob skizziert werden[3]:

1. Festlegung bzw. Identifikation der betrieblichen Aufgabenstellung bzw. des Entscheidungsproblems
2. Ableitung von Fragestellungen, Studienart und Hypothesen
3. Erstellen eines Erhebungsplans (Methoden, Instrumente, Stichprobenauswahl)
4. Auswahl von Analysemethoden anhand von Kriterien wie Qualität der Antworten sowie Operationalisierbarkeit
5. Planung und Durchführung von Datenerhebung, Konsistenzprüfung, Fehlerbehandlung
6. Auswertung der Daten – qualitativ bzw. quantitativ
7. Zusammenfassung Ergebnisse, Aufzeigen von Limitationen und Besonderheiten
8. Ableitung von Handlungsempfehlungen
9. Reflexion des Vorgehens und Ergebnisse, Grundlage für nachfolgende Analysen

2 vgl.: Hoffmann, Franck, Schwarz, Soyez & Wünschmann, 2018, S. 8–9; Raab, Unger & Unger, 2009, S. 31–32

3 vgl.:Hoffmann, Franck, Schwarz, Soyez, & Wünschmann, 2018, S. 2; Döring & Bortz, 2016, S. 182–183

Dabei muss beachtet werden, dass nicht alle Fragestellungen in Unternehmen durch Analyse-Methoden beantwortet werden können. Jedoch können diese eine Basis für weitere Analyse-Aktivitäten darstellen.

1.3 Primärdatenerhebung und Gütekriterien

Unabhängig von vorliegenden Sekundärdaten ist eine spezifische Erhebung von Daten ideal zur Beantwortung einer Fragestellung. Die Passfähigkeit und damit auch die Aussagekraft von Daten einer **Primärdatenerhebung** sind höher als die Daten einer Sekundärdatenforschung. Die Vorteile liegen auf der Hand:

- genaue Auswahl bzw. selbst definierte Stichprobe in der Zielgruppe
- konkrete Fragestellungen
- Kenntnis und Gestaltung Rahmenbedingungen, z. B. Befragungszeitpunkt
- Aktualität der Daten

Jedoch ist die Primärdatenerhebung in der Regel mit mehr Aufwand und Kosten verbunden. Somit ist hier eine Aufwand-Nutzen-Abwägung zu treffen, die auch die Relevanz der betrieblichen Fragestellung sowie die Aussagequalität der erwarteten Ergebnisse berücksichtigt[4].

> Bei der Primärdatenforschung werden folgende Methoden unterschieden[5]:
>
> - Befragung (mündlich oder schriftlich)
> - Beobachtung
> - Experiment
> - Panel

Bei der Datenerhebung wird zwischen **Gesamterhebung** und **Teilerhebung** unterschieden. Bei einer Gesamterhebung werden sämtliche Personen der Zielgruppe befragt. Dies ist bei kleinen Gruppen oder bei einer leichten Erreichbarkeit dieser Gruppe möglich bzw. denkbar. Meist sind Gesamterhebungen aus zeitlichen und finanziellen Gründen nicht praktikabel. Somit werden in der Regel Teilerhebungen durchgeführt. Entscheidend ist, dass

4 vgl.: Raab, Unger & Unger, 2009, S. 31–32
5 vgl.: Döring & Bortz, 2016, S. 210

die Stichprobe repräsentativ ist. Dies ist gegeben, wenn die Verteilung aller interessierenden Merkmale der Untersuchungselemente der Verteilung in der Grundgesamtheit entspricht.[6]

Bei der Planung und Durchführung einer Untersuchung müssen zudem folgende Gütekriterien berücksichtigt werden, um die Qualität gewährleisten zu können:

Wissen | Objektivität
Der Messvorgang und damit die Messwerte sind objektiv, wenn diese von Einflüssen durch die forschende Person oder das Vorgehen selbst unabhängig sind. Dies bedeutet, dass die gleiche Messung bzw. Befragung durch eine andere Person, die identischen Ergebnisse ergibt und diese somit auch intersubjektiv nachprüfbar sind. Objektivität ist Voraussetzung für die Reliabilität.

Zu differenzieren sind drei Formen der Objektivität:

- **Durchführungsobjektivität** wird durch die Beeinträchtigung des bzw. der Proband:in durch die durchführende Person beeinflusst. Beispielsweise kann Kleidung, Mimik und Gestik eine Beeinflussung und damit einen Interviewer:innen-Bias (Fehler) hervorrufen.
- **Auswertungsobjektivität** basiert auf der Wahl des Analyseverfahrens. Daraus können unterschiedliche Ergebnisse entstehen, die dann wiederum zu unterschiedlichen Interpretationen führen. So können verschiedene Forscher:innen mit den gleichen Basisdaten jedoch verschiedenen Analyseverfahren zu unterschiedlichen Ergebnissen kommen.
- **Interpretationsobjektivität** ist hoch, wenn die Ergebnisse der Interpretation unabhängig von der auswertenden Person gleich sind. Je geringer der Auswertungsspielraum ist, desto objektiver ist diese.[7]

Ein Beispiel ist die Auswertung des Mittelwerts und des Medians, beide Auswertungsmethoden analysieren das Ergebnis der „Mitte“. Der **Mittelwert** ist das arithmetische Mittel, der **Median** ist ein numerischer Wert, der die obere Hälfte von der unteren Hälfte teilt. In dem Kontext ist noch die Interpretationsobjektivität zu betrachten. Hierbei können die Ergebnisse

6 vgl.: Raab, Unger & Unger, 2009, S. 49; Döring & Bortz, 2016, S. 292–300
7 vgl.: Döring & Bortz, 2016, S. 443; Hoffmann, Franck, Schwarz, Soyez & Wünschmann, 2018, S. 12)

unterschiedlich interpretiert werden, d. h. die Beurteilung, ob bestimmte Ergebnisse eher positiv oder negativ zu bewerten sind, hängt zum einen von den Vergleichsmöglichkeiten und zum anderen von einem Freiraum der auswertenden Person ab.

Wissen | Reliabilität
Dieses Gütekriterium bezeichnet die Zuverlässigkeit einer Messmethode. Es kann auch als formale Genauigkeit beschrieben werden. D. h. die Ergebnisse sind in sich konsistent, die Ergebnisse können durch eine wiederholte Durchführung mit den gleichen Kriterien und der gleichen Personengruppe mit derselben bzw. einer anderen Messmethodik zu unterschiedlichen Zeitpunkten erneut erhoben werden.

Wenn in einem Fragebogen Eigenschaften wie Egoismus oder Risikofreude indirekt abgefragt werden, sollte die **interne Konsistenz** der Fragen, also ob alle Fragen wirklich das Gewünschte abfragen, über **Cronbachs**[8] **Alpha**[9] ermittelt werden. Das sei hier an einem Beispiel mit 4 Fragen (Items) und 10 Proband:innen gezeigt. Die Proband:innen haben dabei bei jeder Fragen 1–5 Punkte erhalten, die folgende Tabelle zeigt:

8 nach *Lee J. Cronbach* (1916–2001)
9 vgl.: Heimsch et al, 2018, S. 259ff.

Punkte					
Items	1	2	3	4	Summe
	5	4	4	4	17
	5	3	4	3	15
	4	3	4	3	14
	5	3	4	3	15
	4	3	4	4	15
	1	3	2	4	10
	4	3	5	2	14
	3	2	4	5	14
	3	3	2	1	9
	3	1	1	2	7
Varianzen (σ^2)	1,57	0,62	1,60	1,43	10,22
Korrelation	0,72	0,66	0,91	0,52	

Tabelle 1: Beispiel für Cronbachs Alpha

Die Varianzen der Spalten (σ^2) sind jeweils die mittlere quadratische Abweichung der n = 10 Einzelwerte (x_i) von ihrem Mittelwert ($\bar{x}$):

$$Varianz = \sigma^2 = \frac{1}{n-1} \times \sum \left(x_i - \bar{x}\right)^2$$

Varianzen aus Stichproben können bequem über die Excelfunktion VAR.S ermittelt werden.

Formel | Cronbachs Alpha

$$\text{Cronbachs Alpha} = \frac{m}{m-1} \times \left(1 - \frac{\sum \sigma_i^2}{\sigma^2}\right)$$

Dabei ist m die Anzahl der Items (m = 4), $\sum \sigma_i^2$ sind die aufaddierten Varianzen der 4 Items (1,57 + 0,62 + 1,60 + 1,43 = 5,22) und σ^2 ist die Varianz der Summenspalte = 10,22). Damit ergibt sich Cronbachs Alpha zu = 4/3 (1 – 5,22/10,22) = 0,65.

Cronbachs Alpha ist die durchschnittliche Korrelation der Items zueinander und könnte alternativ auch über eine Korrelationsmatrix ermittelt werden. Alles zu Korrelationen finden Sie später im → Kapitel 8.

Werte über 0,7 gelten als akzeptabel, das ist hier nicht der Fall! Wenn sich ein zu kleiner Wert ergibt, kann man einzelne, nicht passende Fragen (jene mit geringer Korrelation zur Summe) herausnehmen und den Wert erhöhen. Hier ist es die Frage 4, deren geringe Korrelation zeigt, dass sie nicht konsistent zu den anderen Fragen ist.

Wie nehmen die Frage 4 heraus und erhalten:

Punkte				
Items	1	2	3	Summe
	5	4	4	13
	5	3	4	12
	4	3	4	11
	5	3	4	12
	4	3	4	11
	1	3	2	6
	4	3	5	12
	3	2	4	9
	3	3	2	8
	3	1	1	5
Varianzen (σ^2)	1,57	0,62	1,60	7,66
Korrelation	0,86	0,70	0,90	

Tabelle 2: Beispiel für Cronbachs Alpha

Cronbachs Alpha ist nun akzeptabel: $\frac{3 \times \left(1 - \frac{3,79}{7,66}\right)}{2} = 0,76$

Mit der **Retestmethode** wird die Korrelation derselben Stichprobe zu unterschiedlichen Zeitpunkten gemessen. Bei einer hohen Korrelation, auch **Retestkorrelation** genannt, kann von einer Stabilität gesprochen werden.

Mit der **Paralleltestmethode** werden bei derselben Stichprobe innerhalb eines kurzen Zeitraums zwei unterschiedliche Tests bzw. Messverfahrens durchgeführt. Die Korrelation dieser beiden Tests ist die Paralleltestkorrelation und wird auch als Äquivalenz bezeichnet.

Eine Sonderform dieses Paralleltests, der in der Praxis häufig anzutreffen ist, ist die Aufteilung der Stichprobe in zwei Teile. Die Korrelation der Ergebnisse dieser Testhälften werden dann als Split-Half- Reliabilität bezeichnet.[10]

Wissen | Validität
Mit Validität wird ausgedrückt, ob die Messmethodik und die Durchführung geeignet sind, die Werte bzw. Ergebnisse der Merkmalsausprägungen bzw. Fragen im Sinne des Befragungsziel zu messen, was auch als **interne Validität** bezeichnet.

Von **Inhaltsvalidität** wird gesprochen, wenn die zu erhebenden Merkmale bzw. Antworten zu der Hypothese bzw. Ziel der Befragung passen, d. h. es muss beispielsweise für jeden bzw. jede Proband:in klar sein, was genau mit der Frage gemeint ist.

Unter **Kriteriumsvalidität** wird verstanden, dass die Antwortmöglichkeiten bzw. Skalen untereinander vergleichbar sind. So wird ausgeschlossen, dass aufgrund der verschiedenen Antwortmöglichkeiten die Korrelation zwischen verschiedenen Ergebnissen beeinflusst wird. Bei ähnlichen Fragen sollten demnach auch die gleiche Skala bzw. die gleichen Antwortmöglichkeiten gegeben werden.[11]

Die Konstruktvalidität beschreibt den Zusammenhang der messbaren Kriterien zu der Gesamtheit der Untersuchung. Die Untersuchung muss geeignet sein und die Messung muss unabhängig von Störeinflüssen o. ä. sein. Differenziert wird hier zum einen zwischen **Konvergenzvalidität**, d. h. die Indikatoren bzw. Fragen müssen eine Beziehung zu dem Ziel des Tests

10 vgl.: Hoffmann, Franck, Schwarz, Soyez, & Wünschmann, 2018, S. 12–13; Döring & Bortz, 2016, S. 444

11 vgl.: Döring & Bortz, 2016,S. 446

aufweisen. Zum anderen gibt es die **Diskriminanzvalidität**, sie bedeutet, dass die Indikatoren nicht bzw. nur in einem geringen Ausmaß dazu dienen können, andere Fragestellungen bzw. Ziele eines Tests zu beantworten.

Durch diese Berücksichtigung der Validität kann dann auch auf eine Generalisierbarkeit der Ergebnisse geschlossen werden.[12] Diese Generalisierungsfähigkeit der Daten bezeichnet man auch als **externe Validität.** Sie kann durch eine zu kleine Stichprobe (Fehler der unzureichenden Statistik) oder eine verzerrte Stichprobenauswahl, also mangelnden Repräsentativität (Fehler der voreingenommenen Statistik) eingeschränkt sein.

Im Folgenden wird explizit auf die Erhebungsmethode Befragung eingegangen. Einige Anforderungen an diese Methode lassen sich auf die Konzeption und Durchführung von Beobachtungen, Experimenten und Panel übertragen.

12 vgl.: Döring, Bortz, 2016, S. 446

2 Befragung

Unter einer **Befragung** wird eine Erhebungsmethode verstanden, bei der durch Antworten (schriftlich, online, persönlich, telefonisch) von Personen Daten zur Beantwortung einer Aufgabenstellung der Marktforschung gewonnen werden.

Solche Befragungen werden häufig angewendet, um z. B. Daten der Zielgruppe, Bedürfnisse, Wahrnehmungen und Meinungen der Befragten zu erheben. Durch das Internet und die schnelle Kommunikation zu einer Befragungsgruppe werden Befragungen häufig auch online durchgeführt.[13]

2.1 Rahmenbedingungen

Die Rahmenbedingungen, unter denen eine Befragung durchgeführt wird, sind maßgeblich für die Ergebnisqualität der Befragung. Der Aufbau und die Durchführung der Befragung sollten die folgenden Aspekte berücksichtigen und mögliche Fehler oder Verzerrungen ausschließen. Entscheidend dabei ist, ob das Ziel und die Fragen der Untersuchung eine Verbindung zu den Rahmenfaktoren besitzen oder auch besitzen könnten.

Erhebungszeitraum

Der Zeitraum muss unabhängig von dem Betrachtungsobjekt bzw. der Fragestellung sowie von spezifischen Zeiten der Zielgruppe sein. Wenn nach saisonalen Produkten gefragt wird, ist die Bekanntheit und damit auch in der Regel die positive Assoziation in der Saison höher. Um möglichst zeitraum-spezifische Einschränkungen zu vermeiden, sollte der Befragungszeitraum Feiertage, Urlaubszeiten o. ä. berücksichtigen.

Ort

Eine Befragung kann u. a. persönlich an einem Ort durchgeführt werden, auch hier ist die Unabhängigkeit zu gewährleisten. An einem Ort in der Innenstadt mit Einzelhandelsgeschäften und Werbung sind die Befragten

13 vgl.: Meffert, 2013, S. 150; Homburg, 2017, S. 265f.; Kreis, Wildner, & Kuß, 2021, S. 130

durch diese Angebote und Kommunikationsbotschaften bezüglich bestimmter Produkte von vorhinein beeinflusst.

Medium

Je nach Art bzw. Medium der Befragung kann es zu einer Verzerrung der Ergebnisse kommen. **Onlinebefragungen** werden häufig durchgeführt, da diese mit geringen Durchführungskosten verbunden sind, schnell durchgeführt werden können und es keinen Einfluss des bzw. der Interviewer:in auf den bzw. die Proband:in gibt. Nachteilig ist die geringe Rücklaufquote. Wenn Onlinebefragungen durchgeführt werden, kann zudem ein Bias, ein Fehler, leicht entstehen, da nur Personen teilnehmen, die zum einen onlineaffin sind und/oder sich zum anderen die Zeit für die Beantwortung der Onlinebefragung nehmen. Häufig wird dieser Bias bewusst in Kauf genommen, um kostengünstig und schnell zu Ergebnissen zu kommen. Bei anderen Medien, wie der **schriftlichen**, **telefonischen** oder **persönlichen Befragung**, können die Rahmenbedingungen anders gestaltet werden, um solche Fehler bzw. Verfälschungen auszuschließen. Jedoch liegen auch dann Fehler wie der Interviewer:inneneffekt vor, sobald eine Person die Befragung durchführt. Wenn dieser Bias zugrunde liegt, ist dieser relevant für die Ergebnisse und die anschließende Ergebnisinterpretation.[14]

Diese Faktoren und Beispiele zeigen, dass das **Setting der Befragung** so ausgestaltet sein muss, dass es nach Möglichkeit keine Abhängigkeiten und Beeinflussungen gibt. Wenn eine Neutralität nicht gewährleistet kann, sollte dies in der Beschreibung der Befragung und späteren Ergebnisinterpretation berücksichtigt werden.

2.2 Struktur

Im Folgenden liegt der Fokus auf Befragungen sowie deren Durchführung und Auswertungsmöglichkeiten. Eine Befragung kann auf unterschiedlichen Kommunikationswegen durchgeführt werden, wie persönlich (face-to-face), telefonisch oder schriftlich. Inzwischen werden viele Befragung online durchgeführt [15].

14 vgl.: Raab, Unger & Unger, 2018, S. 102; Hoffmann, Franck, Schwarz, Soyez & Wünschmann, 2018, S. 31–33

15 vgl.: Hoffmann, Franck, Schwarz, Soyez & Wünschmann, 2018, S. 31f.; Meffert, Burmann, Kirchgeorg & Eisenbeiß, 2019, S. 193

Zu Beginn der eigentlichen Befragung ist es wichtig, den bzw. die Proband:in aufzuklären. Diese Hinweise sind notwendig, um den Zweck der Befragung darzulegen und zugleich auch Hinweise zum Ausfüllen des Fragebogens zu geben.

Wissen | Das sollten Proband:innen vor der Befragung wissen
Folgende Inhalte sollten demnach vor der Befragung dem bzw. der Proband:in mitgeteilt werden:[16]

- Zweck der Befragung
- Bedeutung der Proband:innen und dem bzw. der einzelnen Proband:in
- Verwendung der Daten, u. a. datenschutzkonforme und ggfs. anonymisierte Erhebung und Verarbeitung der Daten
- Angabe über die Zeit zum Ausfüllen der Befragung
- Antwortbeispiele, um zu zeigen, wie Antwortoptionen aussehen, d. h. Darstellung von Mehrfachantworten oder Freitext-Antworten
- Möglichkeit, die Ergebnisse als Teilnehmer:in der Befragung zu erhalten
- Logo des Unternehmens, das die Befragung durchführt
- Kontaktdaten und Impressum
- Dankesformel

Das Layout einer **Onlinebefragung** sollte sehr strukturiert, klar und kontrastreich sein, um den Proband:innen stets Orientierung zu bieten. Somit bieten sich folgende Funktionalitäten an:

- Fortschrittsanzeige innerhalb des Fragebogens, z. B. Bearbeitung in Prozent oder verbleibende Restzeit
- Layout, welches auch für Personen mit Beeinträchtigungen geeignet ist (Barrierefreiheit)
- Layout im *Responsive Design*, um unabhängig vom Endgerät eine optimale Darstellung zu ermöglichen
- Kontaktdaten und Impressum

Der Fragebogen sollte so strukturiert sein, dass primär die Fragen gestellt werden, die den Zweck erfüllen.

16 vgl.: Theobald, 2017, S.44

> Ein Fragebogen sollte so kurz wie möglich bzw. effizient gestaltet sein. Mit jeder Frage (und damit Zeitaufwand) mehr steigt die Gefahr der Nichtteilnahme bzw. des vorzeitigen Abbruchs der Befragung.

Hier ist zu entscheiden, wie wichtig es ist, dass möglichst viele Fragebögen ausgefüllt werden oder dass gerade die ersten Fragen beantwortet werden.[17] Um bei der Auswertung Verbindungen bzw. Korrelationen zwischen den Fragen aufzuzeigen, ist es sinnvoll, dass möglichst viele Proband:innen den Fragenbogen vollständig ausfüllen.

Wissen | Struktur
Die Struktur bzw. die Abfolge der Fragen ist relevant für den Erfolg der Befragung. Beides beeinflusst die Teilnahmequote und somit die Qualität der Antworten. So lassen sich grundsätzlich zwei Abfolgen unterschieden:

- **Leicht zu beantwortende und leichte Fragen zu Beginn:** Hier wird davon ausgegangen, dass Teilnehmer:innen der Befragung durch leicht zu beantwortende Fragen von Anfang an die Befragung gebunden werden und entsprechend die nachfolgenden Fragen beantworten. Dadurch, so die Annahme, nehmen mehr Proband:innen teil. Aber durch eine mögliche ansteigende Komplexität der Fragen, kann auch die Abbruchquote während des Ausfüllens steigen.
- **Demografische Fragen und komplexe Fragen zu Beginn:** Bei komplexen Fragen zu Beginn ist die Annahme, dass die Proband:innen eher am Anfang bereit sind, sich mit komplexen Fragen zu beschäftigen. Abbrüche wären demnach eher zu Beginn und nicht im Laufe der Beantwortung zu erwarten. Somit wäre die Rücklaufquote komplett ausgefüllter Fragebögen höher als bei der anderen oben erwähnten Abfolge.

Die Entscheidung für eine **Abfolge** sollte auf Basis der Relevanz der einzelnen Fragen für den Zweck des Fragebogens getroffen werden. Dies gilt ebenso in Bezug auf die Aussagekraft der Korrelationen der Antworten zu Beginn und am Ende Fragebogens.

Somit kann eher empfohlen werden, die für die Befragung wesentlichen Fragen zu Beginn zu stellen, um die Aussagegüte dieser Antworten zu stei-

17 vgl.: Theobald, 2017, S.38

gern. Da häufig die Proband:innen eine ausgesuchte bzw. vorherbestimmte Gruppe sind, sind demografische bzw. soziodemografische Fragen eher am Ende der Befragung zu stellen, da diese für viele Befragungen einen nachgelagerten Erkenntnisgewinn besitzen.

Demografische, sensitive und komplexe Fragen sollten in der Regel demnach am Ende des Fragebogens stehen, da Proband:innen durch die Art der Fragen eher abbrechen oder die Fragen unbeantwortet lassen. Wenn diese Fragen später im Fragebogen angeordnet und nicht beantwortet werden, können die Antworten der vorherigen Fragen gut genutzt werden[18].

2.3 Fragetypen

Einstiegsfrage

Der ersten Frage, also der **Einstiegsfrage**, kommt eine entscheidende Bedeutung zu, da hier der bzw. die Proband:in entscheidet, ob die Befragung für sie bzw. ihn relevant ist oder nicht. Die erste Frage sollte leicht zu beantworten sein. Auch können einführende Fragen bzw. sogenannte **Eisbrecherfragen**, die dazu dienen, das Interesse an der Thematik zu wecken, gerade zu Beginn der Befragung sinnvoll sein. Dieses zeigt dem bzw. der Proband:in, dass die Befragung gut zu beantworten ist und steigert die Motivation bei den Befragten.

> Die erste Frage sollte die Motivation steigern, z. B. die Neugier der Proband:innen fördern, die Wichtigkeit der Antworten und damit auch eine Wertschätzung betonen.[19] Einstiegsfragen sollten demnach folgende Eigenschaften aufweisen:
>
> - auf das Thema bezogen sein und damit die Bedeutung des Themas unterstreichen
> - einen Bezug zum bzw. zur Proband:in herstellen
> - einfach zu beantworten sein
> - für alle Proband:innen gleichermaßen relevant sein

18 vgl.: Meyer, 2013, S. 66ff.
19 vgl.: Jacob, et al, 2012, S. 138ff.

Bei einer Befragung zum Ernährungsverhalten von sporttreibenden Personen wäre eine solche Frage: „Wie oft haben Sie sich im Durchschnitt in den letzten zwei Wochen sportlich betätigt?“. Als Antwortmöglichkeiten werden in dem Fall Antworten, wie „0–2-mal“, „3–7-mal“, „8–14-mal“, vorgegeben, um die Beantwortung zu erleichtern.

Filterfrage

Zu Beginn des Fragebogens kann auch eine sogenannte **Filterfrage** eingesetzt werden, um für Teilsegmente der Befragten unterschiedliche Fragen im Laufe der Befragung darzustellen. Somit kann ein Fragebogen mit unterschiedlichen Frageabläufen konzipiert werden und zu Beginn wird durch eine solche Filterfrage entschieden, welche Fragen der bzw. die jeweilige Befragte zu beantworten hat.

> Der Vorteil liegt darin, dass der Fragebogen, gerade der onlinebasierte Fragebogen, nur einmal konzipiert werden muss und dann z. B. unter einer Web-Adresse publiziert bzw. kommuniziert werden kann.[20]

Ein Beispiel für eine solche Filterfrage wäre analog dem obigen Beispiel: „Treiben Sie öfters als 3-mal pro Woche Sport?“ oder „Befindet sich ein Fitnessstudio in erreichbarer Nähe Ihrer Wohnung?“. Dadurch könnten im weiteren Verlauf Fragen gestellt werden, die speziell jeweils die beiden Gruppen betreffen.

Offene, geschlossene und halboffene Fragen

Die Fragestellung und damit der Fragentyp ist für das Erhebungsziel des Fragebogens und der Fragen entscheidend. Die Fragestellung beeinflusst auch die Antworten.

Neben den Erwartungen an die Ergebnisse ist auch relevant, welcher Aufwand für die Ergebnisinterpretation und die Auswertung geleistet werden kann. So ist die Auswertung von offenen Fragen mit deutlich mehr Aufwand verbunden als geschlossene Fragen.

- **offene Fragen:** Bei einer **offenen Frage** werden keine Antwortoptionen zur Verfügung gestellt. Ein Beispiel ist die Abfrage nach einer ungestützten Bekanntheit von Produkten oder nach Vorschlägen. Dadurch können die Befragten ihre Ansichten und Ideen ohne vorgegebene

20 vgl.: Jacob et al, 2012, S. 138ff.

Antwortmöglichkeiten angeben. Die Antworten sind in der Regel ein Text und können nicht direkt statistisch ausgewertet werden. Die freie Texteingabe ist ein Vorteil bei offenen Fragen. Der Nachteil ist, dass meist die Befragten schnell den Fragebogen beantworten möchten und nicht die Zeit aufwenden, selbst Antworten zu formulieren. Ein Beispiel für eine offene Frage ist z. B. „Welches sportliche Ereignis hat Sie beeinflusst?“. Durch den direkten persönlichen Bezug ist davon auszugehen, dass die Befragten diese Fragen beantworten.[21]

- **geschlossene Fragen:** Dies sind Fragen, bei denen konkrete Antwortmöglichkeiten vorgegeben sind. Der bzw. die Befragte kann hierbei aus der Auswahl entsprechende Optionen wählen. Der Vorteil bei geschlossenen Fragen besteht darin, dass es für die Befragten leicht ist, die Fragen durch die vorgebenden Optionen zu beantworten. Der Nachteil liegt darin, dass individuelle Antworten nicht möglich und dass Antwortmöglichkeiten nicht immer passend sind.[22]
- **halboffene Fragen:** Diese Fragen beinhalten spezifische Antwortmöglichkeiten sowie eine freie Textantwort. In der Regel wird im Anschluss an die Antwortmöglichkeiten ein freies Antwortfeld angefügt, um „Sonstiges“ zu erfragen. Halboffene Fragen werden genutzt, um konkrete und Textantworten zu gewinnen. Vorteil ist, dass die Befragten ihre Sichtweise und Ideen angeben können. In der Praxis werden diese offenen Textfelder nicht häufig seitens der Teilnehmer:innen ausgefüllt.[23]

Im Bereich der geschlossenen Fragen existieren eine Vielzahl von Fragetypen, um verschiedene Antwortoptionen je nach Frageziel zu bieten. Im Nachfolgenden werden einige Fragetypen vorgestellt:

Soziodemografische Fragen

Soziodemografische Daten sind bestimmte Merkmale einer Zielgruppe wie Alter, Familienstand, berufliche Situation und Kaufkraft. Bei der Mediaplanung und Umsetzung von Kommunikationsmaßnahmen können diese Daten für eine zielgruppengerechte Auswahl der Medien genutzt werden, da Werbeträger entsprechende soziodemografische Daten ihrer Rezipient:innen erheben[24].

---

21 vgl.: Hoffmann, Franck, Schwarz, Soyez & Wünschmann, 2018, S. 10f.
22 vgl.: Porst, 2014, S. 53ff.; Steiner & Benesch, 2021, S. 49
23 vgl.: Porst, 2014, S. 57ff.; Steiner & Benesch, 2021, S. 49f.
24 vgl.: Jacob, Heinz, Décieux & Eirmbter, 2012, S. 156

> Demografische Daten dienen dazu, grundlegende Daten über die Befragten zu aggregieren, um die Befragten bzw. Zielgruppe in die Gesamtbevölkerung einordnen zu können.

Im Nachfolgenden finden Sie einige Beispiele für soziodemografische Fragen:

- Zu welcher der nachfolgenden Alterskategorien gehören Sie?
  17 oder jünger | 18–20 | 21–29 | 30–39 | 40–49 | 50–59 | 60 oder älter
- Bitte geben Sie Ihr Geschlecht an:
  weiblich | männlich | divers (bitte angeben)
- Wie ist Ihr Familienstand?
  verheiratet | verwitwet | geschieden | getrennt | ledig
- Was ist Ihr höchster Schul- oder Hochschulabschluss?
  Hauptschulabschluss | Realschul-/Oberschulabschluss | Abitur oder gleichwertiger Abschluss | Studium ohne Abschluss | Bachelor | Master | Promotion
- Welche der folgenden Kategorien beschreibt Ihren Beschäftigungsstatus am besten?
  angestellt Teilzeit | angestellt Vollzeit | arbeitssuchend | in Rente/Pension | berufsunfähig
- Wie hoch war das gesamte Einkommen aller Mitglieder Ihres Haushalts im vergangenen Jahr?
  0–10.000 € | bis 20.000 € | bis 30.000 € | bis 50.000 € | bis 70.000 € | bis 90.000 € | über 90.000 €

Zur Vereinheitlichung von Befragungen mit soziodemografischen Daten wurden die Fragestellungen und Antwortoptionen auf nationaler und internationaler Ebene vereinheitlicht.[25] Für den spezifischen Einsatz bei Befragungen in der Praxis ist zu prüfen, welche Daten genau zur Unterstützung von betriebswirtschaftlichen Entscheidungen notwendig. Hierbei sind die situativen und passenden Fragen relevanter als die Standardisierung von soziodemografischen Daten. Die Standards sind eine gute Orientierung für die Formulierung der Fragen.

25 vgl.: Heckel & al., 2016

Dichotomische Fragen

Bei diesem Typ sind nur zwei Antworten möglich. Diese dienen zur eindeutigen Segmentierung der Befragten, z. B. „Halten Sie sich für sportlich?“.[26]

Likert-Skala

Diese Skala wird in der Regel für die Messung von Einstellungen der Befragten genutzt. Ein Beispiel ist dabei die Messung der Zufriedenheit, wie „sehr zufrieden“, „eher zufrieden“, „weder zufrieden noch unzufrieden“, „eher unzufrieden“ und „sehr unzufrieden“.[27]
Die Antwortausprägungen auf einer **Likert-Skala** lassen sich auf zwei Arten darstellen:

- **bipolar:** Die beiden Endpunkte einer bipolaren Likert-Skala stellen Minimum und Maximum bzw. zwei gegensätzliche Meinungen dar, z. B. „schlecht“ und „gut“. Dabei kann es seitens der Befragten zu unterschiedlicher subjektiver Interpretation der Endpunkte kommen.
- **unipolar:** Hier werden die Antwortoptionen in einer Reihenfolge angegeben. Der Vorteil ist, dass die Befragten die Abstände als gleich interpretieren und die beiden Endpunkte als jeweilige Gegenteile verstehen.

Häufig wird dabei eine 5-stufige Skala genutzt. Eine Likert-Skala mit einer ungeraden Anzahl hat einen Mittelpunkt. Somit kann der bzw. die Nutzer:in auch ihre bzw. seine Indifferenz ausdrücken. Ungerade Skalen haben die Eigenschaft, dass sich der bzw. die Befragte konkret für eine positive oder negative Tendenz entscheiden muss. Letztendlich hängt die Wahl von der Fragestellung und der Erwartung an das Entscheidungsverhalten der Befragten ab. Wenn der bzw. die Befragte Unsicherheit bzw. Gleichwertigkeit zwischen den beiden Endpunkten ausdrücken soll, sollte eine ungerade Anzahl von Stufen der Skala genutzt werden, um eine Tendenz erkennen zu können.[28]

Generell haben Skalen mit weniger Stufen den Vorteil, dass diese übersichtlicher sind und somit leichter durch die Befragten zu nutzen sind. Bei Likert-Skalen mit vielen Abstufungen haben die Befragten die Schwierigkeit, sich zu entschieden und wählen möglicherweise die Antworten dann zum

26 vgl.: Steiner & Benesch, 2021, S. 54ff.
27 vgl.: Steiner & Benesch, 2021, S. 54ff.
28 vgl.: Steiner, Bensch, 2021, S. 56

Teil zufällig. Ferner ist zu überlegen, welchen konkreten Mehrwert mehr Stufen bei der Fragestellung haben und ob die Einfachheit der Beantwortung und damit die Ergebnisqualität der Antworten bei der Frage und dem gesamten Fragebogen nicht überwiegt.[29]

Single und Multiple Choice

Diese Fragetypen bieten konkrete Antworten für die Befragten. So kann der bzw. die Befragte eine konkrete Antwort wählen (**Single Choice**) oder Mehrfachantworten geben (**Multiple Choice**). Die Antwortmöglichkeiten können dabei nach einer bestimmten Struktur angeordnet werden, z. B. immer „Ja" oder „Nein" bei Single-Choice-Fragen, um die Orientierung bei der Befragung zu erleichtern. Mehrfachantworten können auch in einer bestimmten Struktur angeordnet werden, z. B. bei Sportarten werden Einzel-, Gemeinschaftsportarten, Ausdauer- und Kraftsport aufgeführt. In vielen Befragungen können die Antwortoptionen zufällig angeordnet werden. Dieses ist zu empfehlen, um einer möglichen Präferenz bei der Auswahl entgegenzuwirken. Die zufällige Reihenfolge von Antwortoptionen erhöht die Beschäftigung und Aufmerksamkeit bei der Beantwortung der Fragen und damit auch die Ergebnisqualität der Antworten.[30]

Reihung von Antworten

Dieser Fragentyp wird verwendet, um Antworten zu priorisieren. Bei der Befragung werden die Antwortoptionen nach Wichtigkeit bzw. einem anderen Ordnungskriterium geordnet. Bei Onlinebefragungen kann der bzw. die Befragte meist durch Ziehen der Antwortoptionen die Reihenfolge festlegen. Durch einen solchen Fragentyp und die damit verbundene notwendige Interaktion wird die Aufmerksamkeit des bzw. der Befragten gesteigert und damit auch das bewusste Antworten dieser Frage. Solche Fragen sorgen für eine Vielfalt von Fragen innerhalb einer Befragung und tragen dazu bei, dass die Fragen nicht als Routine beantwortet werden.[31]

Matrixfragen

Dieser Fragentyp beinhaltet de facto mehrere Fragen inklusive Antwortmöglichkeiten und stellt somit einen Fragenblock mit gleichen Antwort-

29 vgl.: Porst, 2014, S. 94
30 vgl.: Döring, Bortz, 2016 S.:454–455
31 vgl.: Föhl, Friedrich, 2022, S 40ff.

möglichkeiten dar. Die Verwendung dieser **Matrix** für Mehrfachfragen eignet sich in den Situationen, in denen für verschiedenen Fragen die gleichen Parameter als Antworten erfragt werden sollen.[32]

Bewerten Sie die Attraktivität der folgenden Sportarten für Sie?
Die Bewertung erfolgt nach Schulnoten.

	1	2	3	4	5
Laufen	☐	☐	☐	☐	☐
Schwimmen	☐	☐	☐	☐	☐
Radfahren	☐	☐	☐	☐	☐
Skifahren	☐	☐	☐	☐	☐
Badminton	☐	☐	☐	☐	☐
Boxen	☐	☐	☐	☐	☐

Abbildung 1: Beispiel für eine Matrixfrage

Net Promotor Score

Der Net Promotor Score (NPS) ist ein Wert für die Weiterempfehlungsrate und spiegelt sich in einem speziellen Fragentyp wider. Diese Fragestellung und der NPS wurde 2003 von *Bain & Company* entwickelt.

Wissen | Net Promotor Score (NPS)
Die Fragestellung lautet in der Regel: „Wie wahrscheinlich ist es, dass Sie Freund:innen und Bekannten Produkt X weiterempfehlen?“. Dabei wird eine Skala von 0 bis 10 als Antwortmöglichkeit angegeben, wobei 0 für „unwahrscheinlich“ steht und 10 für „sehr wahrscheinlich“ steht.

Der NPS wird genutzt, um durch die Zufriedenheit mit dem Unternehmen, Produkten oder Dienstleistungen auch den Grad der Weiterempfehlung zu messen. Dieses wird gerade im Bereich der Wirkungsmessung von *Word of Mouth* als Indikator genutzt.[33]

Die Antworten beim NPS werden in drei Gruppen eingeteilt[34]:

32 vgl.: Föhl, Friedrich, 2022, S. 47f.
33 vgl: Theobald, 2017, S. 270 ff.
34 vgl.: Reichheld, F., 2011, S. 5ff.

- **0–6: Detraktoren**
  Diese Gruppe der Befragten gibt keine Weiterempfehlungen und rät ggfs. auch Freund:innen sowie Bekannten von dem Produkt ab. Außerdem kann erwartet werden, dass die Neigung zu schlechten Rezensionen entsprechend hoch ist.
- **7–8: Indifferente**
  Diese Befragten sind nicht unzufrieden, aber sprechen keine deutliche Weiterempfehlung aus. Diese Gruppe wird eher als neutral gesehen und spielt bei der Berechnung des NPS keine Rolle.
- **9–10: Promoter**
  Die sogenannten Promoter sind sehr zufrieden und empfehlen das Produkt bzw. das Unternehmen weiter.

> Der **Net Promotor Score (NPS)** ergibt sich durch den prozentualen Anteil der Promotoren abzüglich des prozentualen Anteils der Detraktoren. Der NPS ist ein einfaches Instrument zur Messung der Weiterempfehlungsrate und damit der Zufriedenheit.

So ist auch möglich, in einer Branche den NPS für verschiedene Unternehmen zu ermitteln, um den NPS als Benchmark nutzen zu können. Nutzbringend ist auch die Erfassung des NPS für verschiedene Zielgruppen und im Zeitverlauf. Dadurch kann der NPS einen Beitrag für betriebswirtschaftliche Entscheidungen leisten.

Der **Vorteil** der Einfachheit ist zugleich der **Nachteil** des NPS. Die Weiterempfehlung als Indikator für die Zufriedenheit und Kund:innenbeziehung zu nutzen, ist nicht ausreichend. In dem Kontext sind Fragen zu konkreten Beweggründen der Bewertung notwendig. Des Weiteren gibt es Zielgruppen und Kulturen, in denen aus Höflichkeit oder anderen Gründen immer ein Mindestmaß in der Bewertung angegeben wird bzw. eine bestmögliche Bewertung vermieden wird.

2.4 Formulierung und Ableitungen von Fragen

Die Formulierung von Fragen ist entscheidend, damit zum einen keine Beeinflussung des bzw. der Befragten entsteht und zum anderen, dass das Ziel der Befragung bzw. der Hintergrund der Fragestellung verfolgt wird. Die Fragestellung muss von jeder bzw. jedem Befragten eindeutig interpretiert werden können. Die Frage ist direkt zu stellen, damit sich

der bzw. die Befragte angesprochen fühlt und ihre bzw. seine persönliche Einschätzung gibt.

Wissen | Sprache
Die Sprache muss einfach, präzise und sachlich formuliert sein. Jegliche nicht präzisen Ausdrücke führen zu einer subjektiven Interpretation des bzw. der Befragten. Eine Frage wie „Treiben Sie oft Sport?“ mit einer 5-stufigen Skala ist durch das Wort „oft“ unpräzise und führt zu unterschiedlichen Einschätzungen der Befragten. Auch Wörter wie „nur“, „kaum“, „alle“, „immer“ etc. sind keine exakte Ausdrucksweise.

Generell sind neutrale und für die Zielgruppe verständliche Begriffe zu nutzen. So sind Begriffe aus dem Sport wie „Burpee“ oder „Lunges“ für Sportler:innen gängige Begriffe, aber nicht für weniger sportaffine Personen.

Auch nicht doppelte Negierungen sind zu vermeiden, wie z. B. „Finden Sie nicht, dass Sie nicht zu viel Sport treiben?“. Solche Fragestellungen sind schwer verständlich und erfordern eine Analyse des bzw. der Befragten, die Frage richtig zu verstehen.

> Befragte neigen dazu, bestimmte Sachverhalte eher positiver einzuschätzen oder zu bejahen. Eine Lösung ist hier, in der Frageformulierung sowohl positive als auch negative Formulierungen aufzunehmen. Ein Beispiel wäre „Wie leicht bzw. wie schwer fällt es Ihnen, sportlich aktiv zu sein?“.[35]

Die Ableitung der Fragen aus dem Ziel der Studie ist entscheidend, um mit den Antworten die übergeordnete Frage zu klären. Die sogenannte **Forschungsfrage** ist die grundsätzliche Frage, die Anlass für die Studie ist, z. B. „Welchen Einfluss hat die sportliche Aktivität auf das Ernährungsverhalten?“.

Im nächsten Schritt werden aus der Forschungsfrage Themen bzw. sogenannte **Programmfragen** abgeleitet. Die Frageformulierung generell hilft zur Präzisierung der Frage. Daneben ist es auch hilfreich, die Themen, die zur Beantwortung der Forschungsfrage beitragen, abzuleiten. Aus diesen Themen können dann Fragen für die Befragung entwickelt werden. Somit

35 vgl.: Steiner & Benesch, 2021, S. 51f.; Porst, 2014, S. 99f.; Meyer, 2013, S. 80

ist zu empfehlen primär die affinen Themen abzuleiten, als konkrete Fragen in diesem Schritt zu präzisieren. In diesem Beispiel wären die abgeleiteten Themen z. B. „Sportarten und -aktivität", „Ernährung", „Kaufkraft" und „Bildung".[36]

Programmfragen werden auch Hypothesen genannt. In dem obigen Beispiel können dann direkt aus der Forschungsfrage bzw. aus den Themen folgende Fragen entwickelt werden, wie „Ernähren sich Personen, die Sport treiben, bewusster?" oder „Wird durch die Steigerung der sportlichen Aktivität auch das Ernährungsverhalten verändert?".

Der nächste Schritt ist der Transfer der Hypothesen bzw. der Programmfragen in sogenannte **Konstrukte**. Hier muss überlegt werden, wie die Frage mit einer Kennzahl quantifiziert werden kann. Dabei ist zu beachten, dass das Konstrukt alle relevanten Kennzahlen messen muss, um die Frage zu beantworten, aber auch nicht mehr Kennzahlen erfasst, die zur Klärung der Hypothese notwendig sind. In dem vorliegenden Beispiel wären die Konstrukte Sportaktivität und Ernährung. Die Konstrukte sind eine Untergliederung der Hypothese in Einzelbereiche. Aus diesen Konstrukten werden dann die eigentlichen Fragen formuliert. Den Übergang von Konstrukten zu Testfragen gibt es oft nicht, da bei der Zerlegung in Konstrukte automatisch Fragen gebildet werden.[37]

In dem obigen Beispiel wären die Konstrukte zu der Frage „Ernähren sich Personen, die Sport treiben, bewusster?" zwei konkrete Fragestellungen wie „Wie oft treiben Sie Sport?" mit einer Skala z. B. von 1–5 sowie „Wie bewusst ernähren Sie sich?" ebenfalls mit einer Skala von 1–5. Dies würde die Programmfrage bzw. die Hypothese in zwei Konstrukte überführen, die genau diese Hypothese abbilden. Wichtig in dem Kontext ist, dass exakt diese Hypothese damit beantwortet werden kann, nicht darüberhinausgehende Fragestellungen. Die Antwortmöglichkeiten, in dem Beispiel eine Skala von 1–5, spiegeln dann quantitativ die Kennzahl wider. Um die Hypothese in dem Fall zu beantworten, kann die Korrelation der Antworten dieser beiden konkreten Fragen gebildet werden. Anschließend wäre noch eine ergänzende Auswertung hinsichtlich soziodemografischer Daten o. ä. denkbar.

Validitätseinschränkungen in Umfragen

Bei Befragungen kommt es zu speziellen Validitätseinschränkungen, z. B.:

36 vgl.: Steiner & Benesch, 2018, S. 49–51
37 vgl.: Döring & Bortz, 2016, S. 405f.

- **Hang zur Mitte:** Gerade bei Ratingskalen wird gerne von Proband:innen die Mitte gewählt, wenn einem gerade nichts einfällt. Man kann dem entgegenwirken, indem man eine gerade Zahl von Antwortmöglichkeiten vorgibt. Jedoch ist dies nicht in jeden Fall sachgerecht.
- **willkürliches Ankreuzen:** Um diejenigen zu identifizieren die völlig willkürlich ihre Antworten geben, kann man Unsinnsfragen in die Befragung integrieren: „Wie beurteilen Sie das Sprichwort: Wer hohe Häuser baut, wird festen Willen ernten?“. Wer hier eine positive Beurteilung abgibt, dessen Fragenbogen kann aussortiert werden. Beliebt ist es, auch eine Frage zweimal (ähnlich) zu stellen. Wer unterschiedlich antwortet, wird ebenfalls nicht berücksichtigt.
- **Zustimmungstendenz:** Es empfiehlt sich Fragen einzubauen, die man nur mit nein beatworten kann: „Kennen Sie den Politiker Paul Herdegen?“. Bei einer Antwort mit „Ja“ ist die Antwort nicht zu berücksichtigen.
- **sozial erwünschte Antworten:** Proband:innen antworten gern so, wie sie glauben, dass es von ihnen erwartet wird. Man kann diesem Phänomen nachspüren, wenn man Fragen einbaut, die man ehrlicherweise nicht mit „Ja“ beantworten kann. Ein Beispiel: „Ich war noch nie bei einer Verabredung zu spät“.
- **suggestive Fragestellungen:** Vermeiden Sie Wörter wie „nur“, „noch“, „obwohl“ in der Fragestellung, also nicht: “Würden Sie ein Bio-Produkt kaufen, wenn es nur 0,50 € mehr kostet?“
- **Framing-Effekte:** Sie bekommen unterschiedliche Antworten, wenn Sie fragen: „Würden Sie einer Operation zustimmen, die in 99 % aller Fälle komplikationslos verläuft?“ oder „Würden Sie einer Operation zustimmen, die in 1 % der Fälle zu Komplikation führt?“ Bieten Sie besser beide Varianten an.

Es würde zu weit führen, alle Validitätsbedrohungen hier vorzustellen.[38] Sie sollten allerdings ein Bewusstsein für die skizzierten Risiken entwickeln. Die Formulierung eines Fragebogens ist eine diffizile Angelegenheit.

38 Zu Fehlerquellen bei Umfragen siehe z. B.: Blasius, Thiessen, 2021, S. 106ff.

2.5 Aufbau und Pretest

Bei dem Aufbau einer Befragung ist u. a. das Folgende zu beachten:

- Für das Ziel relevante, wichtige Fragen sind zu Beginn zu stellen. Wenn soziodemografische Daten eines bzw. einer Nutzer:in nicht so wichtig sind, sollten diese zum Ende erfragt werden.
- Der Fragebogen sollte kurz sein und nur die für das Ziel relevanten Fragen beinhalten.
- Die Beantwortung von offenen Fragen ist mit Aufwand verbunden und werden meist nicht beantwortet. Sinnvoll ist es hier Antwortmöglichkeiten vorzugeben.
- Fragen, die eine Beantwortung erfordern, beinhalten das Risiko, dass die Befragung abgebrochen wird. Somit sollte in der Regel dem bzw. der Befragten freigestellt sein, die Fragen zu beantworten. Unerlässliche Fragen wie Filterfragen sind entsprechende Pflichtfragen.

Bevor die Befragung durchgeführt wird, ist ein **Pretest** anzuraten. In einem Pretest werden Personen gebeten, die inhaltliche Validität des Fragebogens zu testen. Im Rahmen dessen geben die Personen Hinweise, bei welchen Fragen die Formulierung und die Antwortoptionen nicht verständlich, falsch oder unvollständig sind. So können Fehlinterpretationen vor der Befragung eliminiert werden. Ein weiterer Vorteil: Die Antworten der Prester:innen liefern erste Daten, mit denen Sie die Auswertung durch ein entsprechendes Anwendungssystem testen können.

3 Stichprobe

3.1 Arten von Stichproben und ihre Repräsentativität

Ein vollständiges Bild ergibt sich in einer **Vollerhebung**, also wenn der gesamte Personenkreis adressiert und entsprechend die Fragen beantwortet werden können.[39]

Eine Befragung kann über verschiedene Medien an die vorgesehene Gruppe von Personen kommuniziert werden. Im Rahmen einer Vollerhebung muss gewährleistet werden, dass jede Person erreicht wird, d. h. nicht onlineaffine Personen sollten über alternative Kommunikationswege angesprochen sowie Personen mit mehreren Wohnsitzen sollten am richtigen Ort erreicht werden.

Solche Vollerhebungen stellen aber die Ausnahme dar. Sie werden nur bei einem sehr begrenzten und gut erreichbaren Personenkreis genutzt, z. B. Befragung von Kinobesucher:innen nach einem Film. In der Regel sind Vollerhebungen aus zeitlichen Gründen und aufgrund des hohen Aufwands nicht möglich. Deswegen beschränkt man sich oft auf eine **Teilerhebung** (**Stichprobe**), bei der nur ein deutlich kleinerer Teil der Grundgesamtheit nach bestimmten Kriterien ausgewählt wird.

Um aus einer Stichprobe eine valide Aussage für die Grundgesamtheit treffen zu können, ist die **Repräsentativität** eine notwenige Bedingung. Eine sehr häufig von Studierenden gestellte Fragen bei empirischen Arbeiten lautet: „Wieviel Personen muss ich befragen, damit die Umfrage repräsentativ wird?“ Diese Frage offenbart jedoch ein Missverständnis, welches den Begriff der Repräsentativität betrifft. Dieser Begriff ist nämlich kein streng definierter Statistikbegriff, sondern im engeren Sinn nur eine qualitative Beschreibung einer Stichprobe.

39 vgl.: Hoffmann, Franck, Schwarz, Soyez & Wünschmann, 2018, S. 23; Bruhn, 2022, S. 88

Wissen | Repräsentativität
Eine Stichprobe ist repräsentativ, wenn die für den Untersuchungsgegenstand relevanten Merkmale in der Stichprobe im selben Verhältnis vorkommen, wie in der Grundgesamtheit.

> Wenn es für eine Fragestellung lediglich relevant ist, dass Männer sowie Frauen befragt werden, wäre eine Stichprobe mit einem Mann und einer Frau zwar *repräsentativ*, aber natürlich trotzdem nicht *valide*, weil der Stichprobenumfang nicht ausreichend ist. Eine Stichprobe mit 1.000 Männern wäre trotz ihres beachtlichen Umfanges auch nicht valide, weil sie nicht repräsentativ ist.

Repräsentativität kann durch die Art der Stichprobe erreicht werden, oder eben nicht. Es gibt nun verschiedene Arten wie wir zu einer Stichprobe kommen können:

Zufallsauswahl

Hierbei ist zwischen einer einfachen, einer geschichteten Zufallsauswahl sowie einer Klumpenauswahl zu unterscheiden. [40]

> Bei einer **geschichteten Zufallsauswahl** werden Untergruppen der Grundgesamtheit gebildet. Anschließend wird aus diesen Gruppen eine Zufallsauswahl vorgenommen.

Somit kann hier von einer Quotenauswahl mit anschließender Zufallsauswahl gesprochen werden. Das Ziel besteht darin, dass bestimmte Gruppen der Grundgesamtheit adäquat, d. h. in der relativen Häufigkeit, berücksichtigt werden. Beispielsweise ist bekannt, wie groß die verschiedenen Altersgruppen einer Bevölkerung sind und diese prozentuale Häufigkeit wird in einer Quote und anschließenden Zufallsauswahl pro Altersgruppe berücksichtigt.[41]

> Bei der **Klumpenauswahl** werden nach bestimmten Kriterien möglichst homogene und repräsentative Klumpen hinsichtlich des Unter-

40 vgl.: Hoffmann; Franck, Schwarz; Soyez & Wünschmann, 2018; S 24f.
41 vgl.: Raab, Unger, & Unger, 2018, S. 55ff.; Meyer, 2013, S. 62

> suchungsziels gebildet. Anschließend werden einige Klumpen zufällig ausgewählt.

Diese Gruppen wiederum fließen vollständig in die Stichprobe ein. Ein Beispiel wäre die Klumpenbildung von Sportler:innen, speziell Läufer:innen, nach geografischen Eigenschaften, um nach der Ernährung zu fragen. In dem Fall kann angenommen werden, dass der geografische Ort unabhängig von den Ernährungsgewohnheiten beim Sport ist. Nach dieser Klumpen- bzw. Gruppenbildung werden einige Klumpen zufällig ausgesucht und diese wiederum vollständig befragt.[42]

> Die **einfache Zufallsauswahl** bedeutet, dass die Proband:innen aus der Grundgesamtheit gezogen werden. Entscheidend ist hier, dass die Grundgesamtheit vollständig bekannt ist.

So sind Kund:innen eines Unternehmens bekannt und können bei Befragungen hinsichtlich der Zufriedenheit der Kund:innen entsprechend ausgewählt werden. Schwieriger bzw. kaum möglich ist es, Proband:innen zu identifizieren, die bei dem Unternehmen schon Produkte ausgesucht hatten, aber dann nicht gekauft haben.[43]

Die Zufallsauswahl, auch **probabilistische Auswahl** genannt, ist streng genommen das einzige Verfahren, um eine generalisierbare Aussage über eine Grundgesamtheit zu erhalten. Bei entsprechender Fallzahl stellt sich die Repräsentativität ganz von allein her. Aber Vorsicht mit dem Begriff Zufall. Wer bei einer Onlinebefragung darauf wartet, wer „zufällig" antwortet, hat keine Zufallsauswahl getroffen, sondern nur eine willkürliche Auswahl. Wann also spricht man von einer Zufallsauswahl?

Wissen | Zufallsauswahl
Eine Auswahl ist dann zufällig, wenn jedes Element der Grundgesamtheit die gleiche Chance hat, Teil der Stichproben zu werden.

Wenn Sie eine generalisierbare Aussage über eine bestimmte Einstellung aller erwachsenen Deutschen finden wollen, so müssten Sie sich ein Verzeichnis aller ca. 60 Millionen erwachsener Menschen Deutschlands besorgen

42 vgl.: Kreis, Wildner, & Kuß, 2021, S. 79f.; Meyer, 2013, S. 63
43 vgl.: Raab, Unger & Unger, 2018, S. 52–54

und daraus über einen Zufallsmechanismus Menschen auswählen und quer durch das Land reisen, um genau diese zu befragen und dabei hoffen, dass diese dann auch antwortwillig sind. Eine Marktforschungsagentur kann das mit ihren Ressourcen leisten, im Rahmen einer Bachelorarbeit ist das jedoch kaum möglich.

Stattdessen werden sie wohl eher mit der **willkürlichen Stichprobe** arbeiten. Dies ist eine Auswahl mit dem Ziel: „Ich nehme, was ich kriegen kann." Man spricht auch von einer **Gelegenheitsstichprobe** oder **Ad-hoc-Stichprobe**. Bei Straßenbefragungen ist das der Fall, bei TED-Befragungen der Fernsehsender oder eben auch bei den beliebten Onlinebefragungen, wenn Sie dort wie üblich ihren Fragbogen in sozialen Netzwerken stellen und warten, wer darauf antwortet. Dieses ist zwar praktisch, aber es muss Ihnen bewusst sein, dass es sich dabei um eine **nicht-repräsentative Stichprobe** handelt.

> Bei Onlinebefragungen ist die **Responsequote** oft gering. Das größere Problem ist aber die **Self-Selection**. Nicht der bzw. die Forscher:in wählt den bzw. die Teilnehmer:in aus, sondern diese entscheiden selbst, ob sie mitmachen. Dadurch entsteht eine starke **Verzerrung** dahingehend, dass diejenigen, die sich über ein Thema besonders aufregen oder stark interessiert sind, bevorzugt teilnehmen. Die meist größere Masse der Geringinteressierten eher nicht. Damit ist deren Meinung, oft die einer (schweigenden) Mehrheit, nicht angemessen repräsentiert.
>
> Die gute Nachricht: Wenn Sie eine **Unterschiedshypothese** testen wollen, ist die Repräsentativität verzichtbar![44] Sie ziehen keine Schlüsse von Ihren Werten auf eine Grundgesamtheit, sondern beziehen sich nur auf einen Unterschied und der sollte annäherungsweise gleicht sein, egal ob die Auswahl repräsentativ ist oder nicht.

Wenn Sie z. B. eine Zustimmungsquote zu einer Aussage von 50 % bei Frauen und 60 % bei Männern ermitteln und die „echte" Werte einer repräsentativen Stichprobe wären aber 58 % und 68 %, so bleibt der Unterschied ja so oder so 10 Prozentpunkte. Die Auswahl sollte lediglich hinsichtlich der für Ihre Forschungsfrage relevanten Merkmale nicht stark verzerrt sein. Die Validierung der Hypothese ist dann nur noch eine Frage der Stichprobengröße.

Für einige Fragestellungen können auch bewusste Stichprobenauswahlmethoden sinnvoll sein. Eine bewusste Auswahl wird auch systematische

44 vgl.: Döring, Bortz, 2016, S 300ff. oder Westermann, 2000, S. 336

Stichprobe genannt und ist eine nicht zufällige Auswahl von Proband:innen. Die Auswahl wird nach bestimmten relevanten Kriterien bestimmt. Dabei können folgende Arten von bewussten Auswahlmöglichkeiten vorgenommen werden:[45]

Quotenauswahl

> Hier werden je Merkmalsausprägung **Quoten**, d. h. prozentuale Häufigkeiten pro relevantes Merkmal festgelegt. Anschließend erfolgt die Auswahl auf Basis dieser vorgegebenen **Quotenvorgaben**.

Ein Beispiel verdeutlicht den Nutzen dieser Quotenauswahl. Bei der Befragung der Zufriedenheit der Kund:innen wird eine bewusste Auswahl nach Kaufkraft getroffen, um das Verhältnis der Kund:innen nach Kaufkraft in der Stichprobe abzubilden. Bei einem Test nach Gründen zur Nutzung des ÖPNV wird bewusst eine Quote von regelmäßigen Nutzer:innen, Wenignutzer:innen und Nichtnutzer:innen vor der Durchführung der Befragung gebildet.[46]

Konzentration und Typische Auswahl

> Hierbei werden nur die Proband:innen gefragt, die eine besondere Bedeutung haben bzw. einen hohen Beitrag für den Erkenntnisgewinn für die Befragung leisten können.

Ein Beispiel wäre, Vielnutzer:innen des ÖPNV über die Stärken und Schwächen des ÖPNV zu befragen.[47]

Bei der **typischen Auswahl** werden typische Personen aus der Grundgesamtheit ausgewählt, welche besonders charakteristisch für die Ziele der Befragung und der Grundgesamtheit sind. So könnten bewusst Personen ausgewählt werden, die sich öffentlich als Vielnutzer:in, Wenignutzer:in und Nichtnutzer:in des ÖPNV in Sozialen Medien darstellen und ihre Einstellung kommunizieren.[48]

---

45 vgl.: Kreis et al, 2021, S. 74
46 vgl.: Raab, Unger, Unger, 2018, S. 62ff.; Kreis, Wildner, & Kuß, 2021, S. 75
47 vgl.: Raab, Unger, Unger, 2018, S. 65f.
48 vgl.: Bruhn, 2022, S. 89

Anhand der verschiedenen Arten der typischen Auswahl ist zu erkennen, dass nicht alle Auswahlmöglichkeiten trennscharf sind. Der Vorteil einer typischen Auswahl ist, dass mit einem Vorwissen über die mögliche Segmentierung, Einstellungen und Antwortmöglichkeiten der Proband:innen ein Set einer Auswahl von Personen getroffen werden kann, die dann der Grundgesamtheit entspricht. Gerade bei Befragungen mit wenig Vorwissen über die Grundgesamtheit ist eine typische Auswahl schwer zu treffen und damit ist die Ergebnisgüte dieser Auswahl auch entsprechend schlechter.

3.2 Konfidenzintervalle bei Stichproben

Um von einem quantitativen Stichprobenergebnis, wie einem Mittelwert oder einem prozentualen Anteil, auf den Wert in der Grundgesamtheit schließen zu können, ist es wissenschaftlicher Standard, ein sogenanntes **Konfidenzintervall** (Vertrauensbereich) anzugeben. Dies ist auch entscheidend um den notwendigen Stichprobenumfang zu bestimmen.[49]

Konfidenzintervall bei Anteilswerten

Nehmen wir an, Sie befragen 20 Biertrinker:innen, ob ihnen eine neue besser schmeckt als die alte Biersorte; 12 Personen, also 60 %, antworten mit „Ja“. Dürfen Sie dann sagen, dass eine Mehrheit von ca. 60 % aller Biertrinker:innen die neue Sorte bevorzugt?

Eher nicht, denn richtig ist ja erstmal nur, dass **60 % der Befragten** das neue Bier besser findet. Nur ist dies uninteressant. Interessant wäre, was alle Biertrinker:innen darüber denken. Eine generalisierte Aussage aus einer Stichprobe ist nur in einer **Intervallschätzung**[50] sinnvoll möglich und **nicht als Punktschätzung!**

Formel | Intervalle (IN) bei Anteilen (π)

$$IN = \pi \pm t_{krit} \times \sqrt{\frac{\pi \times (1 - \pi)}{n}}$$

Das π steht für den Prozentsatz, den Sie ermittelt haben, n ist die Stichprobengröße.

49 Vgl: Döring & Bortz, 2016, S. 631; Schmidt-Atzert & Ameland, 2012, S. 51
50 vgl.: Sedlmeier, Renkewitz, 2018, S. 343ff.

Das $t_{krit.}$ (kritischer t-Wert) steht für einen Tabellenwert aus der Statistik, der für eine bestimmte Wahrscheinlichkeit steht, mit der die Aussage zutreffen soll (Vertrauenswahrscheinlichkeit).

Wissen | Konvention für Konfidenzintervalle
Man bilde ein Vertrauensintervall von 95 % mit $t_{krit.} = 1{,}96$, dies entspricht einer Irrtumswahrscheinlichkeit (Alpha-Fehler) von 5 %

Im Falle unserer Bierumfrage ist das Intervall dann:

$$IN = 0,6 \pm 1,96 \times \sqrt{\frac{0,6 \times 0,4}{20}}$$

$$IN = 0,6 \pm 0,21 = [0,38; 0,81]$$

95 % der durch eine Zufallsstichprobe (mit n = 20) gefundene Intervalle umfassen den wahren Wert der Grundgesamtheit. Näherungsweise heißt dies, dass die wahre Zustimmungsrate für die neue Biersorte bei allen Biertrinker:innen wahrscheinlich in einer Spanne zwischen 39 % und 81 % liegt. Die Irrtumswahrscheinlichkeit dieser Aussage (Alpha-Fehler) liegt bei $\alpha = 5$ %.

Bei diesem breiten Intervall dürfen Sie also nicht argumentieren, dass einer Mehrheit die neue Sorte besser schmeckt, es könnte ja auch weniger als 50 % sein. Aber, je mehr Menschen befragt werden, desto enger wird das Vertrauensintervall. Hier könnte man nun folgende Überlegung zur Stichprobengröße anstellen: Soll die Abweichung maximal 10 Prozentpunkte (Genauigkeit g = 0,1) betragen, damit bei 60 % Zustimmung auf eine Mehrheit auch in der Gesamtpopulation geschlossen werden darf (IN 50–70 %), so lässt sich der nötige Stichprobenumfang n wie folgt berechnen:

$$n = \frac{t^2 \times \pi(1-\pi)}{g^2} = \frac{1,96^2 \times 0,6 \times 0,4}{0,1^2} = 92,2$$

Es müssen dann also mindestens 93 Biertrinker:innen befragt werden. Eigentlich müssen Sie also schon vorher wissen welcher Prozentsatz herauskommt. Eventuell kann dann ein kleiner Test vorab (Pretest) dazu Anhaltspunkte liefern. Sonst müssen Sie vom ungünstigsten Fall ausgehen, das wäre $\pi = 0,5$.

Es handelt sich hier aber um eine Näherungsformel für große Stichproben, die Sie bei Prozentwerten kleiner 10 % oder größer 90 % eher nicht verwenden sollten. In diesen Fällen kann es zu inkonsistenten Intervallen kommen mit Werten über 100 % oder unter 0 %. Sollte das der Fall sein, empfiehlt es sich, mit folgender komplexeren Schätzformel für 95 %-Anteilsintervalle zu rechnen:

$$IN = \frac{n \times \pi + 2}{n + 4} \pm \frac{2\sqrt{n}}{n + 4} \times \sqrt{\frac{\pi \times (1 - \pi)}{n} + \frac{1}{n}}$$

Konfidenzintervalle bei Mittelwerten

Bei metrischen Merkmalen können Mittelwerte (μ) gebildet werden. Um das Konfidenzintervall zu ermitteln, ist noch die Varianz der Werte um den Mittelwert als zusätzliche Größe (s^2) notwendig für eine Intervallschätzung.

Formel | Intervalle (IN) bei Mittelwerten (μ)

$$IN = \mu \pm t_{krit} \times \sqrt{\frac{\sigma^2}{n}}$$

Auch dazu ein Beispiel: Eine Befragung von 50 Absolvent:innen eines Studienganges ergab ein Durchschnitteinkommen von 3.500 € und eine Standardabweichung von σ = 700 €. Dann ist das Konfidenzintervall mit 95 % Vertrauenswahrscheinlichkeit:

$IN = 3500 \pm 1,96 \times \sqrt{490000/50} = 3500 \pm 194$, also zwischen 3306 und 3694.

Die notwendige Stichprobengröße für eine gegebene Genauigkeit g ist wie folgt vorab bestimmbar:

$$n = \frac{(t \times \sigma)^2}{g^2}$$

Es dürfte allerding in den meisten Fällen schwierig sein, die Standardabweichung vorab richtig zu schätzen.

Konfidenzintervalle bei kleinen Grundgesamtheiten

Wird die Stichprobe n aus einer kleinen Grundgesamtheit N von wenigen tausend Einheiten gebildet, sind die Formeln um einen Korrekturfaktor zu erweitern, da sonst die Intervalle zu breit angegeben werden. Die Formeln dazu sind:

Formel | Konfidenzintervalle bei kleinen Grundgesamtheiten
Konfidenzintervalle bei Anteilswerten

$$IN = \pi \pm t_{\text{krit}} \times \sqrt{\frac{\pi(1-\pi)}{n}} \times \sqrt{\frac{N-n}{N-1}}$$

Konfidenzintervalle bei Mittelwerten

$$IN = \mu \pm t_{\text{krit}} \times \sqrt{\frac{\sigma^2}{n}} \times \sqrt{\frac{N-n}{N-1}}$$

Ein Beispiel: Von 400 BWL-Studierenden einer Hochschule wurden 50 nach ihrer Zufriedenheit gefragt; davon gaben 70 % an, zufrieden zu sein. Das Konfidenzintervall für 95 % ist dann:

$$IN = 70\% \pm 1,96 \times \sqrt{\frac{0,7 \times 0,3}{50}} \times \sqrt{\frac{400-50}{400-1}} = 70\% \pm 11,9\%$$

Der notwendige Stichprobenumfang für eine gegeben Genauigkeit g ist:

$n = \frac{t^2 \times N \times \pi(1-\pi)}{t^2 \times \pi(1-\pi) + g^2(N-1)}$ für Anteile und $n = \frac{t^2 \times N \times \sigma^2}{t^2 \times \sigma^2 + g^2(N-1)}$ für Mittelwerte.

4 Grundgedanken zum Testen von Hypothesen

4.1 Art der Daten bestimmt den Test

Daten können grob gesprochen in zwei Formen auftreten: In Worten oder Zahlen. Bei Zahlen spricht man von **metrischen Daten** (auch intervallskalierte Daten) wie beispielsweise „Einkommen“, „Alter“, „Preise“ oder „Aktienkurse“. Das Wesen metrischer Daten ist, dass sie „echte“ Zahlen sind und damit der Abstand zwischen ihnen immer gleich ist, also die Differenz von 1 zu 2 ist die gleiche wie von 2 zu 3 usw. Man kann hier also Auswertungen über Mittelwerte und Standardabweichungen vornehmen.

Daten, die aus Wortkategorien bestehen wie „Geschlecht“, „Familienstand“ oder „belegter Studiengang“ sind **kategoriale Daten** (auch **nominale Daten**). Man kann Wortkategorien bilden und ihnen Häufigkeiten zuordnen. Manchmal können die Kategorien in eine sinnvolle Reihenfolge gebracht werden wie Schulnoten, dann haben wir es mit einer Sonderform kategorialer Daten zu tun, den **ordinalen Daten**. Ordinale Daten können mit Zahlen kodiert und in einen Rang gebracht werden, was bei Schulnoten ja üblich ist mit der „1“ für „sehr gut“ bis zur „5“ für „mangelhaft“. Sie sehen dann so aus wie metrische Daten, obwohl sie das nicht sind, da die Zahlenkodierung willkürlich ist und damit auch nicht gewährleistet sein kann, dass die Abstände zwischen einer „1“ und einer „2“ so groß ist wie zwischen einer „4“ und einer „5“.

> Je differenzierter die ordinale Skala aber ist, desto mehr wird sie der metrischen Skalen ähnlich, so dass auch bei ordinalen Daten die Bildung von **Mittelwerten** akzeptiert wird. Bei Schulnoten erfolgt das ungeachtet aller kritischen Einwände regelmäßig, obwohl hier der Unterschied zwischen einer „1“ und einer „2“ sicher nicht der gleiche ist wie zwischen einer „4“ und einer „5“.

Beim Testen von Hypothesen über Zusammenhänge oder Unterschiede von Daten ist die Wahl des geeigneten Testes nun abhängig von den Datenarten, die man zueinander in Bezug setzt. Hat man zwei Datenspalten metrischer Art, beispielsweise die Punktzahl einer Statistikklausur und einer Matheklausur von Studierenden, so bietet sich dafür eine **Korrelationsanalyse** an,

um zu testen, wie stark der Zusammenhang dieser Datenreihen ist. Untersucht man hingegen Unterschiede der „Geschlechter“ oder „Studiengänge“ mit der Punktzahl einer Klausur, trifft eine kategoriale Größe auf eine metrische. Hier lassen sich pro Kategorie „Männer/Frauen“ bzw. „Studiengang“ nun die durchschnittlich erreichten „Punktzahlen“ (Mittelwerte) mit einem **t-Test** (oder über eine **ANOVA**) auf signifikante Unterschiede testen.

Wollen wir nun wissen, ob die Wahl eines Studienganges abhängig vom Geschlecht ist, sind beide Datenmenge kategorial. Hier kann man nun die Häufigkeiten der Studiengangswahl der Geschlechter in einer Kreuztabelle (auch Kontingenztabelle) darstellen und mit einem **Chi-Quadrat-Test** auf eine unterschiedliche Verteilung hin untersuchen.

	kategoriales Merkmal *Studiengang*	metrisches Merkmal *Punkte Matheklausur*
kategoriales Merkmal *Geschlecht*	Chi-Quadrat-Test (bei je 2 Merkmals ausprägungen auch Anteilstest möglich)	t-Test
metrisches Merkmal *Punkte Statistikklausur*	t-Test (bei mehr als zwei Studiengänge auch ANOVA)	Korrelationsanalyse

Tabelle 3

Wie behandeln wir nun Einstellungsfragen, die wir auf einer Skala zwischen „stimme voll zu“ und „stimme gar nicht zu“ erfragt haben? Hier ist es wie bei den Schulnoten: Ratingskalen sind streng genommen auch nur kategorialen Daten. Sie können jedoch mit Zahlen kodiert werden.

> Es gilt als Konsens, dass bei hinreichender Differenzierung kategorialer Daten (mindesten 5 Skalenstufen, gleiche Anzahl positiver wie negativer Stufen) auch eine Auswertung als metrische Merkmale über einen t-Test erlaubt ist.[51]

Methodisch strenger ist es aber, hier nur die Häufigkeitsverteilung auf die Ratingkategorien mit einem Chi-Quadrat-Test zu testen.

51 vgl.: z. B. Mayer; 2013, S. 83 oder Döring, Bortz, 2016, S. 251

Man sollte sich besser schon bei der Fragebogenentwicklung nicht nur darüber im Klaren sein was, sondern auch wie, also mit welchem präferierten Test, man auswerten will. Die Frage sollten Sie so stellen, dass man die Antworten dazu mit einen der hier vorgestellten, allgemein akzeptierten Tests, leicht auswerten und interpretieren kann.

4.2 Methodische Vorgehensweise bei Signifikanztest

Das Testen von Unterschieds- oder Zusammenhangshypothesen eignet sich sehr gut für Haus- und Bachelorarbeiten, da man hierfür auch nicht-repräsentative Stichproben akzeptieren kann.

> Die grundsätzliche Vorgehensweise bei der Hypothesenprüfung besteht zunächst darin, die Forschungsfrage in zwei überprüfbare, sich gegenseitig ausschließenden Hypothesen, von denen genau eine wahr sein muss (kontradiktorischer Gegensatz), zu überführen:
>
> - Die Alternativhypothese H_A: Es gibt einen Unterschied.
> - Die Nullhypothese H_N: Es gibt keinen Unterschied.

Zunächst wird getestet, ob ein Unterschied in einer Stichprobe **signifikant** ist.

Ein Unterschied ist signifikant, wenn er mit hoher Wahrscheinlichkeit, der Vertrauenswahrscheinlichkeit, nicht rein zufällig ist, oder andersherum gesagt: Wenn die Wahrscheinlichkeit, dass wir fälschlicherweise von einem Unterschied ausgehen, sehr klein ist. Diese Irrtumswahrscheinlichkeit für die Annahme der Nullhypothese ist der **Alpha-Fehler**, der auch **Fehler 1. Grades** genannt wird.

Wissen | Konvention für Signifikanz
Vertrauenswahrscheinlichkeit 95 % $(1 - \alpha)$
Signifikanzniveau (Alpha-Fehler) $\alpha = 5\ \%$

Die Vorgehensweis ist also eine indirekte: Statt direkt nach Indizien für die Alternativhypothese zu suchen, was leider nicht geht, versuchen wir die Nullhypothese als unwahrscheinlich zu verwerfen, um damit die Annahme der Alternativhypothese zu begründen. Das konventionelle Verfahren geht über die Ermittlung des p-Wertes.

Wissen | p-Wert
Der p-Wert ist die bedingte Wahrscheinlichkeit dafür, dass der Unterschied, den wir in der Stichprobe sehen (oder ein höherer) noch mit der Nullhypothese vereinbar ist. Ist der p-Wert kleiner als der akzeptierte Alpha-Fehler, auch Signifikanzniveau genannt, liegt Signifikanz vor.

> Zu beachten ist, dass der wissenschaftlich-statistische Begriff „signifikant" nicht die gleiche Bedeutung hat wie im allgemeinen Sprachgebrauch. Dort steht das Wort *signifikant*, seinem lateinischen Wortursprung *significans* gemäß, oft für etwas Großes oder Bedeutendes. Das wird aber in der quantitativen Forschung als *Relevanz* bezeichnet und über Effektstärken ausgedrückt. Ein statistisch signifikanter Unterschied in der Forschung kann durchaus unbedeutend klein sein und ein großer Unterschied wird in einer kleinen Stichprobe oft nicht signifikant sein.

Ist der Unterschied signifikant, dann meint man damit das Umgangssprachliche: „Das kann doch wohl kein Zufall mehr sein!", die Nullhypothese wird als unwahrscheinlich abgelehnt und damit die Alternativhypothese angenommen. Bewiesen im Sinne eines logischen Schlusses ist die Alternativhypothese damit natürlich nicht, dies ist auch gar nicht möglich; sie kann ja wegen der tolerierten Irrtumswahrscheinlichkeit immer noch falsch sein. Man sagt deshalb vorsichtig, „die Alternativhypothese habe sich im Test bewährt" oder „ihre Annahme ist plausibel".

Dabei ist die Wahl eines Signifikanzniveau von 5 % natürlich willkürlich, darauf hat man sich geeinigt. Prinzipiell ist aber ein p-Wert von 5,01 % oder 4,99 % so ziemlich das Gleiche, nur würde man in dem einen Fall das Ergebnis als nicht-signifikant und im anderen als signifikant einordnen. Man sollte deshalb bei der Interpretation von Werten nahe der 5-Prozentmarke etwas zurückhaltender formulieren und von einem „knappen Ergebnis" sprechen.

Signifikanztest können einseitig (mit gerichteter Hypothese) oder zweiseitig (ungerichtete Hypothese) durchgeführt werden:

- Der **zweiseitige Test** ist der Normalfall. Die H_A wird wie im Eingangsbeispiel ungerichtet aufgestellt (es gibt einen Unterschied); man weiß nicht vorher schon in welcher Gruppe höhere Werte liegen. Die Richtung kann erst nach dem Signifikanztest angegeben werden.

- Beim **einseitigen Test** wird H_A gerichtet aufgestellt; z. B. in Gruppe 1 gibt es höhere Werte. Der gegenteilige Fall wird vorab ausgeschlossen. Das kann sachlogisch begründet sein, weil dieser Fall unplausibel ist. Wenn Sie testen, welche Auswirkung eine Werbung auf den Bekanntheitsgrad eines Produktes hat, kann dieser nach der Werbung nur höher, aber nicht niedriger sein. Es können auch empirische Gründe vorliegen, so beispielsweise, wenn in anderen Studien noch nie ein Unterschied in der anderen Richtung festgestellt wurde.

Einseitige Tests sind ökonomischer, diese werden schneller signifikant und kommen so mit kleineren Stichproben aus. Jedoch sollte die Entscheidung für den einseitigen Test immer gut begründet sein. Im Zweifel gilt; zweiseitig testen.

Wirft man einen Blick in die meisten deutschsprachigen Statistiklehrbücher, entsteht schnell der Eindruck, dass der Test mit der Untersuchung auf Signifikanz abgeschlossen ist. Das ist nicht der Fall und wird bei Veröffentlichungen nach internationalen Standards auch nicht akzeptiert![52] Der Signifikanztest ist nur der erste Teil des Hypothesentest. Zu den Konventionen guter wissenschaftlicher Arbeiten gehört es, (mindestens bei) signifikanten Unterschieden eine Effektstärke anzugeben und insignifikante Unterschiede mit Hilfe der Teststärke zu interpretieren.

Diese Prüfschritte sind hier noch einmal im Diagramm dargestellt:

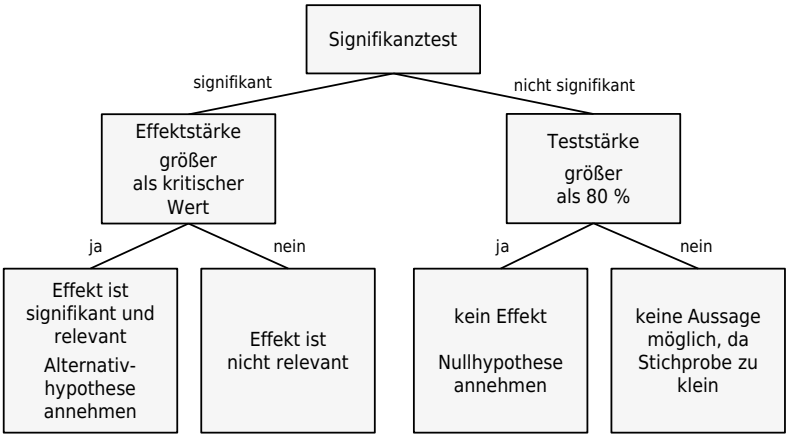

Abbildung 2: Vorgehen bei Hypothesentests mit vorgegebener Effektstärke

52 vgl.: Dempster, Hanna, 2019; S.:306 und 471

Idealerweise sollten Sie beim Hypothesentesten vorab folgende Überlegungen anstellen und ihre Festlegungen dazu in ihrer Arbeit berichten:

- Bestimmen Sie, mit welchem Test Sie ihre Daten auswerten wollen.
- Formulieren Sie zwei sich gegenseitig ausschließende Hypothesen über das Ergebnis von denen genau eine wahr sein muss.
- Begründen Sie, ob der Test einseitig oder zweiseitig ausgeführt wird.
- Legen Sie das Signifikanzniveau (den akzeptierten Alpha-Fehler) fest (i. d. R. nehmen Sie hier 5 %).
- Wählen Sie ein geeignetes Maß für die zu berichtenden Effektstärke und definieren Sie die Mindestgröße für einen relevanten Effekt in ihrer Forschung.
- Bestimmen Sie die angestrebte Teststärke (i. d. R. nehmen Sie hier 80 %).
- Ermitteln Sie den optimalen Stichprobenumfang.

Die Auflistung orientiert sich an dem Hypothesentest nach *Neyman* und *Pearson*.[53]

Die folgenden Ausführungen zeigen ihnen, was es mit diesen Größen auf sich hat und wie man diese in den jeweiligen Tests berechnet und interpretiert.

Wenn Sie sich nicht auf einen für ihre Studie relevanten Mindesteffekt festlegen können oder wollen, sollten der Ablauf wie folgt aussehen:

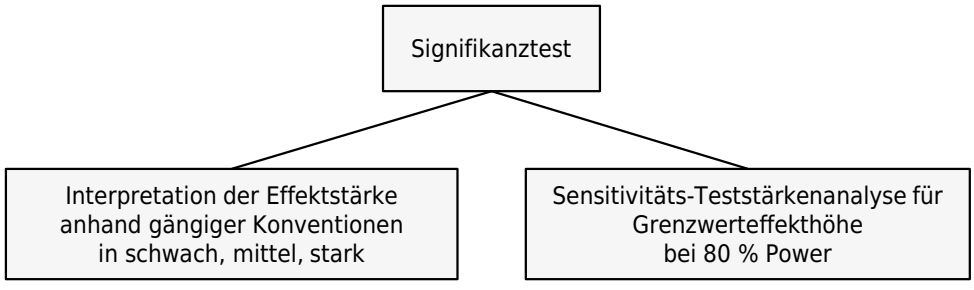

Abbildung 3: Vorgehen bei Hypothesentest ohne vorgegebene Effektstärke

53 vgl.: z. B.: Bühner, Ziegler, 2017, S. 221ff. oder Sedlmeier, Renkewitz, 2018, S. 393ff.

4.3 Bedeutung Effektstärke

> Wenn ein Unterschied signifikant ist, sagt das nichts über seine Größe und praktische Relevanz aus. Bei großen Stichproben wird so ziemlich jeder, noch so kleine Unterschied, statistisch signifikant. Deshalb gehört es zu den Konventionen empirischer Forschung ein signifikantes Ergebnis durch die Angabe einer **Effektstärke** zu ergänzen.

Ein einheitliches **Effektstärkemaß** gibt es jedoch nicht, fast jeder Test hat sein eigenes Maß. Man kann aber zwei „Familien" von Effektgrößen unterscheiden,

- den Abstandsmaßen (Unterschiedsmaßen) und
- den Zusammenhangmaßen.

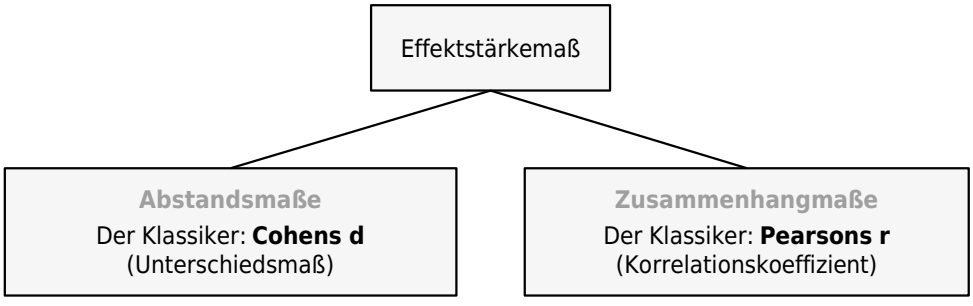

Abbildung 4: Effektstärkemaße

Cohens d wird insbesondere bei Mittelwertunterschiede benutzt. Dabei wird die Mittelwertdifferenz zur Standardabweichung ins Verhältnis gesetzt. Je höher der Wert, desto stärker der Effekt. Der **Korrelationskoeffizient r** nimmt die betragsmäßigen Werte zwischen 0 (kein Zusammenhang) und 1(vollständiger Zusammenhang) an. Bei metrischen Merkmalen gibt zudem das Vorzeichen an, in welche Richtung der Zusammenhang geht.

Die Wahl der Effektgröße wird durch die Forschungsfrage bestimmt. Lautet diese beispielsweise: „Verdienen Frauen und Männer bei gleicher Qualifikation unterschiedlich?" sollte d angegeben werden. Ist die Frage so formuliert: „Gibt es einen Zusammenhangzwischen Geschlecht und der Lohnhöhe", ist eher die Interpretation mit Hilfe von r angebracht. Wenn Sie Studien vergleichen wollen, in denen mal d und mal r berichtet wird, können Sie das gegebene Maß mit Hilfe folgender **Konversionsformeln** in das andere Maß überführen:

Formel | Konversion von Effektgrößen

$d = \frac{2r}{\sqrt{1 - r^2}}$ bzw. $r = \frac{d}{\sqrt{d^2 + \frac{N^2}{n_1 \times n_2}}}$

Oder als Faustformel: d ist doppelt so groß wie r.

Man sollte nun idealerweise **vor** Beginn der Forschung definieren, mit welchem Maß man arbeitet und ab welcher Höhe ein Effekt als relevant gilt. Nur wenn man die Effekthöhe kennt, auf die man testen will, kann man den nötigen Stichprobenumfang bestimmen. Aber wie kommt man nun zu der „richtigen" Höhe der Effektstärke, die man ansetzen sollte? Man muss darüber nachdenken, ab welcher Höhe ein Unterschied anfängt, praktisch relevant zu werden. Es sollte, wie es im Englischen heißt, eine **Minimum Important Difference (MID)**[54] bestimmt werden. Dazu können oft Kosten-Nutzen-Überlegungen herangezogen werden. Wenn beispielsweise die unterschiedlichen Umsatzzahlen eines Verkäufers nach und vor einer Schulung getestet werden sollen, wird man bei bekannten Schulungskosten auch eine notwendige Umsatzsteigerung angeben können, damit sich die Schulung rechnet. Diese kann man dann in eine MID-Effekthöhe umwandeln.

Bei **Replikationsstudien** kann man sich auch an der in der Ursprungsstudie festgestellten Effektstärke orientieren. Oder man hält sich an den Konventionen der jeweiligen Effektgröße, dann brauch man sich nur noch dafür zu entscheiden, ob man auf einen schwachen, mittleren oder großen Effekt untersucht. Gängige Interpretationsvorschläge[55] zu diesen beiden Effektmaßen sind[56]:

Effektgröße	d	r
schwacher Effekt	> 0,2	> 0,1
mittlerer Effekt	> 0,5	> 0,3
starker Effekt	> 0,8	> 0,5

Tabelle 4

54 vgl.: Dempster, Hanna, 2019, S. 311
55 vgl.: Döring, Bortz, 2016, S. 820
56 Da d (näherungsweise) doppelt so groß ist wie r, sind die Konventionen für mittlere und starke Effekte allerdings nicht genau kompatibel.

Die Tabellenwerte hier sind aber keinesfalls als unabänderliche Standards zu verstehen. Besser man macht sich selbst Gedanken; die hier vorgestellten Vorschläge sind eine Hilfe, wenn einem weiter nichts Eigenes einfällt.

Ein Unterschied muss also, damit wir ihn annehmen, nicht nur (statistisch) signifikant, sondern auch (praktisch) relevant sein, also größer sein als eine vorab festgelegte Mindesteffektstärke. Haben wir keinen Mindesteffekt vorab definiert, können und sollten wir nach Abschluss der Forschung den Stichprobenunterschied wenigstens anhand der hier vorgestellten Konventionen interpretieren.

4.4 Bedeutung Teststärke

In den Statistikbücher kann man oft lesen, dass man bei einem nicht-signifikanten Ergebnis die Nullhypothese „beibehält“. Nur stimmt das leider so nicht!

> Ist ein Testunterschied nicht signifikant, so bedeutet dies nicht zwangsläufig, dass die Nullhypothese angenommen werden kann.

Es kann zwei Gründe dafür geben:

- Es gibt tatsächlich in der Grundgesamtheit keinen (relevanten) Unterschied; in diesem Fall wäre die Annahme der Nullhypothese natürlich die richtige Entscheidung.
- Es gibt in der Grundgesamtheit einen Unterschied; die Stichprobe war aber zu klein, um den Unterschied zu entdecken, also ihn signifikant werden zu lassen; eine Annahme der Nullhypothese wäre nun falsch. Wir würden einen **Beta-Fehler**, auch **Fehler der 2. Art** genannt, begehen.

So haben wir es beim Testen mit zwei potenziellen Fehlerarten zu tun:

	es besteht in der Realität ein Unterschied	es besteht in der Realität kein Unterschied
Alternativhypothese wird angenommen	richtig Entscheidung	Alpha-Fehler (Irrtumswahrscheinlichkeit über einen nicht existierenden Unterschied)
Nullhypothese wird angenommen	Beta-Fehler (Übersehenswahrscheinlichkeit eines existierenden Unterschieds)	richtig Entscheidung

Tabelle 5

> Der Beta-Fehler ist die Wahrscheinlichkeit den Unterschied übersehen zu haben. Die **Teststärke,** auch **Power** genannt und mit großem P abgekürzt, ist nun das Gegenstück zum Beta-Fehler, nämlich P = 1 – Beta-Fehler. Sie gibt die Wahrscheinlichkeit an, dass ein definierter Effekt, sofern er denn in der Grundgesamtheit existiert, im Test signifikant wird, sie ist also, anschaulicher gesagt, die **Entdeckungswahrscheinlichkeit**.

Stellen wir uns vor, wir würden im Wohnzimmer nach einem verlegten Haustürschlüssel suchen. Schauen wir nur kurz ins Wohnzimmer rein, wäre die Aussage „Hier ist er nicht!“ (die Nullhypothese) ziemlich voreilig. Bei einem kurzen Blick ist die Entdeckungswahrscheinlichkeit gering. Haben wir aber alles gründlich abgesucht, ist die Aussage viel eher wahr

Bei einem nicht-signifikanten Testunterschied kann vernünftigerweise die Nullhypothese („Da ist nichts!“) nur angenommen werden, wenn wir „gründlich“ nach einem Effekt gesucht haben und die Entdeckungswahrscheinlichkeit hoch war. Gründlich bedeutet hier, dass die Stichprobe groß genug sein muss. Um den nötigen Stichprobenumfang ermitteln zu können, muss die Höhe der angestrebten Teststärke vorab festgelegt werden; auch dafür gibt es eine Konvention:[57]

57 vgl.: Döring, Bortz, 2016, S. 809

Wissen | Konvention für Teststärken
Beta-Fehler β = 20 % n
Teststärke (Power P) 1 – β = 80 %

Allerdings ist es diskussionswürdig, ob eine Teststärke von 80 % immer genug ist.[58] Immerhin akzeptiert man damit einen viel höheren Beta-Fehler gegenüber dem Alpha-Fehler. Dahinter verbirgt sich die Auffassung, dass eine fälschliche Annahme der Nullhypothese weniger schlimm ist als die falsche Annahme der Alternativhypothese. Auch wird argumentiert, dass die Alternativhypothese die „Wunschhypothese" des bzw. der Forscher:in ist und dieses besonders streng geprüft werden soll, im Gegensatz zur gegenteiligen Nullhypothese. Beide Argumente müssen aber nicht immer zutreffen. Es kann auch gerechtfertigt sein, beispielsweise Alpha- und Beta-Fehler gleich anzusetzen, was dann aber auch schnell die nötige Stichprobengröße ansteigen lässt. Wir können hier die Diskussion um den angemessenen Wert für die Power nicht abschließend diskutieren und verwenden im Weiteren die gängige Konvention.

Die Teststärke ist abhängig vom Stichprobenumfang, Signifikanzniveau und der Effektgröße. Aber welcher? Manche Statistikbücher ignorieren die Teststärkenberechnung völlig oder erklären diese für nicht bestimmbar. Dies beruht jedoch auf einem Missverständnis: Nicht berechenbar ist die Teststärke für den „wahren" Effekt der Population, schlicht deshalb, weil wir ihn nicht kennen. Auch ist es falsch (aber leider verbreitet) ersatzweise die Teststärkeberechnung mit der im Test festgestellten empirischen Effekthöhe zu berechnen. Dies ist eine aussagenlose Rechnung und sollte tunlichst unterbleiben![59]

> Was in der Forschungspraxis unbedingt interessiert, ist die Entdeckungswahrscheinlichkeit für die vom Forschenden **vorab** definierte, auf praxisrelevanten oder theoretischen Überlegungen basierende, Mindesteffektgröße.

Die Power dafür ist für jeden Test ohne größere Probleme berechenbar!

Für die nachträgliche (post hoc) Teststärkeberechnung wird die Power also auf Basis der festgelegte Mindesteffektgröße, und nur für diese, inter-

58 vgl.: Döring, Bortz, 2016, S. 809
59 vgl.: Döring, Bortz, 2016, S. 813

pretiert. Wenn wir einen solchen MID-Effekt nicht vorgegeben haben, gibt es auch keine sinnvolle Teststärkeangabe. Man kann keine Entdeckungswahrscheinlichkeit angeben, wenn man nicht weiß, was man entdecken will.

> Aber: Wenn wir keinen MID-Effekt definiert haben, kann – und sollte – bei nicht-signifikanten Ergebnissen jene Effektgröße angegeben werden, die zu einer Teststärke von 80 % führt (Sensitivitätsteststärkeanalyse). Effekte in dieser Stärke (und größere) wären mit hoher Wahrscheinlichkeit entdeckt worden. Dass ein Effekt oberhalb dieses Grenzwertes existiert, kann bei nicht-signifikante Ergebnissen dann als unplausibel gelten.

Die Annahme der Nullhypothese ist dabei nicht wörtlich zu verstehen; es geht nicht darum, dass Null, also gar kein Unterschied besteht. Das wäre wohl auch unrealistisch. Das zwei Gruppen auf die letzte Nachkommastelle den gleichen Wert liefern, kommt praktisch so gut wie nie vor. Eine absolute Nullhypothese ist so betrachtet immer falsch. Es geht darum, ob ein Unterschied in der Höhe eines relevanten Effekts – oder bei der Sensitivitätsanalyse in Höhe eines Grenzwertes – vernünftigerweise ausgeschlossen werden kann oder nicht.

5 Test bei zwei Mittelwerte (t-Test)

5.1 Signifikanztest

Wenn zwei Stichprobengruppen auf Unterschiede in den Mittelwerten eines Merkmales getestet werden sollen, nimmt man dafür den **t-Test für unabhängige Stichproben**, auch bekannt als **Student-t-Test**.[60] Die Bestimmung von Mittelwerten geht streng genommen nur für metrische (intervallskalierte) Daten, also „echte" Zahlen. Bei hinreichender Differenzierung von Nominaldaten, wie sie bei einer Ratingskala mit mindestens fünf Kategorien erhoben werden, ist es aber Praxis, wie schon erwähnt, die Daten in Zahlen zu kodieren und daraus Mittelwerte zu berechnen und zu vergleichen.

Beispiel | Mathenoten
Forschungsfrage: Sind die Noten der Matheklausuren unterschiedlich in den Studiengänge Betriebswirtschaft (BW) und Ingenieurwissenschaften (IG)?

- H_A: die Durchschnittsnoten unterscheiden sich
- H_N: die Durchschnittsnoten unterscheiden sich nicht.

Auswertung:

	Durchschnittsnote	Standard abweichung	Stichprobengröße (n)
BW	3,5	1,3	100
IG	3,0	1,2	80

Tabelle 6

60 Der Erfinder des 1908 publizierten t-Test war aber kein Student, sondern Mitarbeiter der Guinness-Brauerei in Dublin, und hieß *William Seayl Gosset* (1876–1937). Er wählte „Student" als Pseudonym, da die Brauerei die Veröffentlichung verboten hatte. Nachzulesen bei Wikipedia.

Für den t-Test ist es unabdingbar, auch die Streuung um den Mittelwert in Form der Standardabweichung σ zu berücksichtigen. Die Standardabweichung ist die Wurzel aus den quadrierten Abweichungen der einzelnen Noten (x_i) vom Durchschnitt der Gruppe ($\bar{x}$)

$$\sigma = \sqrt{\frac{1}{n-1} \times \sum \left(x_i - \bar{x}\right)^2}.$$

In Excel steht dafür die Funktion +**STABW.S** zur Verfügung. Für den Durchschnitt können Sie in Excel die Funktion +**MITTELWERT** nutzen.

Ist die absolute Abweichung von einer halben Note ($|\text{ABW}| = 3,5 - 3,0 = 0,5$) nun signifikant, also wahrscheinlich nicht zufällig? Zunächst die **einfache Methode**, die man mit einem simplen Taschenrechner durchführen kann. Demnach ist der Unterschied signifikant, wenn der **empirische t-Wert** ($t_{emp.}$) der Stichprobe größer ist als der **kritische t-Wert** ($t_{krit.}$). Für einen akzeptieren Alpha-Fehler von 5 % ist der Wert $t_{krit.}$ = 1,96 im zweiseitigen Test und 1,65 im einseitigen Test.[61] Bei der hier vorliegenden Forschungsfrage ist zweiseitig zu testen, weil man nicht sicher im Voraus sagen kann, welcher Studiengang die besseren Noten erreichen wird.

Formel | t-Wert bei zwei Mittelwerten

$$t_{emp.} = \frac{|ABW|}{\sqrt{\frac{\sigma_1^2 \times n_1 + \sigma_2^2 \times n_2}{n_1 \times n_2}}}$$

Hier also: $t_{emp.} = \frac{0,5}{\sqrt{\frac{169 + 115,2}{8000}}} = 2,65$

Da 2,65 > 1,96 ist der Unterschied signifikant, also mit mehr als 95 % Wahrscheinlichkeit nicht zufällig.

61 Streng genommen gelten die Werte nur für große Stichproben ab 240 aufwärts. Für kleinere Werte gelten geringfügig höhere kritische Werte (bei je 30 Teilnehmer:innen z. B. im zweiseitigen Test 2,00 statt 1,96); dieses können wir bei der einfachen Methode aber vernachlässigen. Genaue Werte liefert die *Excel*-Funktion T.INV.2S für zweiseitige und T.INV für einseitige Tests wobei $\alpha und f\ddot{u}r Freiheitsgrade f = N - 2$ einzugeben sind.

> Die „klassische“ Methode ist genauer und sollte deshalb in Veröffentlichungen verwendet werden. Sie besteht in der Angabe des p-Wertes. Der **p-Wert** ist die bedingte Wahrscheinlichkeit, dass sich die empirische ABW (oder mehr) einstellt, obwohl die Nullhypothese gilt. Ist dieser Wert kleiner als der vorgegeben Alpha-Fehler (von hier 5 %), ist das Ergebnis signifikant. Die einfache Methode zeigt nur, ob der p-Wert unter 5 % ist, aber eben nicht seine genaue Größe.

Leider kann der p-Wert jetzt nicht einfach so mit dem Taschenrechner ermittelt werden. Es handelt sich bei p-Werten um Flächeninhalte unter statistischen Verteilungen. Er kann aber aus Tabellen – wie der im → Anhang – zu jedem t-Wert abgelesen werden. Aus der Tabelle ergibt sich für t = 2,65 im zweiseitigen Test ein p-Wert zwischen 1 % und 0,7 %.

Ganz genau geht es in Excel mit der Funktion **+T.VERT.2S**. Dabei ist neben dem t-Wert noch die Zahl der Freiheitsgrade (f)[62] mit anzugeben. Dies ist der Gesamtstichprobenumfang ($N = n_1 + n_2$) minus 2 ($f = N - 2$). Excel liefert T.VERT.2S (2,65;178): 0,9 %. In wissenschaftlichen Arbeiten schreibt man dies meistens als Dezimalzahl, also ohne Prozentzeichen, mit drei Nachkommastellen: p = 0,009. Bei internationalen Publikationen ist folgende Schreibweise üblich: p = .009[63]

Die Wahrscheinlichkeit, dass die Abweichung von 0,5 noch mit dem Zufall zu vereinbaren ist, liegt also nur bei 0,9 %. Bei p-Werten < 0.01 spricht man auch von hochsignifikanten Unterschieden. Bei einem einseitigen Test wäre der p-Wert zu halbieren oder man nimmt dazu die Excelfunktion +**T.VERT.RE**.

Zum Rechnen können Sie auch folgende Exceldatei nutzen:

◆ **Download 1** | Unterschiedstest 2 Mittelwerte (Service_53241_01.xls)

Hier noch ein weiteres Beispiel zum selbst nachrechnen:

62 In der Statistikliteratur werden die Freiheitsgrade oft auch mit df = degrees of freedom abgekürzt. Auf das Konzept der Freiheitsgrade können wir hier nicht näher eingehen, tiefere Kenntnis darüber ist aber für die praktische Anwendung auch nicht erforderlich.

63 Ist der p-Wert kleiner 0,1 % schreibt man mit dem Ungleichheitszeichen p < .000, niemals p =.000, da dieser nicht Null werden kann.

Beispiel | Einkommen
Forschungsfrage: Verdienen Frauen weniger als Männer bei gleicher Qualifikation?

- H_A: Die Gehälter der Frauen sind niedriger
- H_N: Die Gehälter unterscheiden sich nicht

Auswertung:

	Durchschnittseinkommen	Standardabweichung	Befragte (n)
Männer	2.800 €	1.000 €	100
Frauen	2600 €	800 €	100

Tabelle 7

Bei der Formulierung der Forschungsfrage ist hier ein einseitiger Test angebracht. Offenbar ist klar (z. B. durch vergleichbare andere Studien), dass Frauen nirgendwo mehr verdienen als Männer, es kann also nur in eine Richtung gehen. Wir können damit für $t_{krit.} = 1{,}65$ heranziehen.

Der empirische t-Wert ist nun hier $t = \frac{200}{\sqrt{16400}} = 1,56$. Da dies kleiner als 1,65 ist, ist der Unterschied nicht signifikant. Excel liefert über TVERT.RE(1,56; 198) den p-Wert = 0.06; höher als 5 %, zu hoch, um die Alternativhypothese annehmen zu können. Wenn Sie den p-Wert angeben, und das ist die übliche Methodik, ist die zusätzliche Angabe von kritischen t-Werten nicht erforderlich.

5.2 Effektstärke

Im Beispiel „Mathenoten“ hatten wir einen signifikanten Effekt festgestellt. Da dieses noch nichts über die Relevanz aussagt, ist nun zusätzlich ein Effektstärkemaß anzugeben und zu interpretieren. Bei Mittelwerten nimmt man das Maß von Cohen:[64] **Cohens d**. Er schlug vor, die Abweichung

64 Nach dem Statistiker *Jacob Cohen* (1923–1998)

durch die Streuung zu dividieren. Bei unterschiedlichen Streuungen und Gruppengröße wie hier, kann man folgende Formel[65] verwenden:

Formel | Cohens d

$$d = \frac{|\text{ABW}|}{\sqrt{\frac{\sigma_1^2 \times n_1 + \sigma_2^2 \times n_2}{n_1 + n_2}}}$$

Hier ergibt sich[66]

$$d = \frac{0,5}{\sqrt{\frac{1,3^2 \times 100 + 1,2^2 \times 80}{180}}} = 0,4.$$

> Was sagt das nun aus? Ein Unterschied ist relevant, wenn der empirische d-Wert ($d_{emp.}$) über einen vorab definierten Mindest-d-Wert ($d_{krit.}$) liegt. Wenn wir das versäumt haben, kann man sich an den Interpretationsvorschlag von Cohen aus ⟶ Kapitel 4.3 halten: Demnach haben wir es hier mit einem schwachen Effekt zu tun.

Haben Sie mit d_{krit} vorab eine interessierende Mindesteffektgröße definiert, sollten Sie ihre Hypothesen gleich spezifischer formulieren, z. B. für $d_{krit.}$ =0,2:

- H_A: Es gibt einen relevanten Unterschied mit d > 0,2
- H_N: Es gibt keinen relevanten Unterscheid mit d > 0,2

Man kann mit Hilfe von d auch einen alternativen Signifikanztest durchführen. Dazu bildet man das Konfidenzintervall über den empirischen d-Wert von hier 0,4. Wenn auch die untere Grenze des Intervalls noch > 0 ist, liegt ein signifikanter Effekt vor. Das 95-Prozent-Konfidenzintervall ist $d \pm 1,96 \times \sigma_d$.

Die Standardabweichung von d[67] lässt sich gut schätzen mit:

---

65 Häufig wird die Berechnung von d mit der „gepoolten" Standardabweichung auch als „Hedges g" bezeichnet, wir bleiben hier bei der Bezeichnung d.

66 Alternativ und wenn die Stichprobengrößen nicht zu unterschiedlich sind, kann man d auch über t einfacher ermitteln mit: $d = \frac{2t}{\sqrt{N}}$. Wenn die Korrelation als Effektgröße gewünscht ist, ist diese: $r = \sqrt{\frac{t^2}{t^2 + f}}$.

67 vgl.: Westermann, 2000, S. 355

Formel | Standardabweichung von d

$$\sigma_d = \sqrt{\frac{N}{n_1 \times n_2} + \frac{d^2}{2N}}$$

Im Beispiel „Mathenoten“ ist $\sigma_d = \sqrt{0,0225 + 0,0004} = 0,15$ und damit liegt das Intervall zwischen $0,4 \times \pm 1,96 \times -1,5$, also etwa zwischen 0,1 und 0,7. Auch die untere Grenze ist noch positiv, damit ist der Effekt auch mit dieser Methode signifikant. Wir sehen hier aber auch, dass der Effekt zwischen unbedeutend und mittel sein kann. Diese Rechenmethode ist zwar (noch) nicht besonders verbreitet, aber mindestens eine empfehlenswerte Ergänzung. Die Betrachtung von Konfidenzintervallen einer Effektgröße gilt als moderne Alternative zur p-Wert- Methode.[68] Die *Deutsche Gesellschaft für Psychologie* schreibt die Angabe des Konfidenzintervalls zu d in ihren Richtlinien für Veröffentlichungen sogar explizit vor.[69]

Mit Hilfe der unteren Intervallgrenze (UG), der oberen Intervallgrenze (OG) und $d_{krit.}$ können wir nun signifikante Effekte (UG > 0) wie folgt näher interpretieren:

- Ist OG < $d_{krit.}$ liegt ein signifikanter aber praktisch bedeutungsloser Effekt vor.

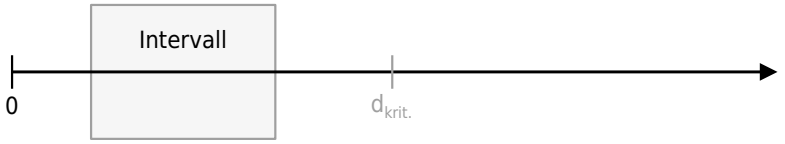

Abbildung 5: Bedeutungslose Effektstärke

- Liegt $d_{krit.}$ im Intervall, also OG > $d_{krit.}$ und gleichzeitig UG < $d_{krit.}$ liegt ein signifikanter Effekt mit unklarer praktischer Bedeutung vor.

68 Bei der Analyse metrischer Merkmale geht man im Grunde genommen schon immer so vor, man ermittelt hier zuerst mit der Korrelation die Effektgröße und erst dann, wenn überhaupt, deren Signifikanz über Konfidenzintervalle.

69 DGP: Richtlinien der Manuskriptgestaltung, 2019, S. 39

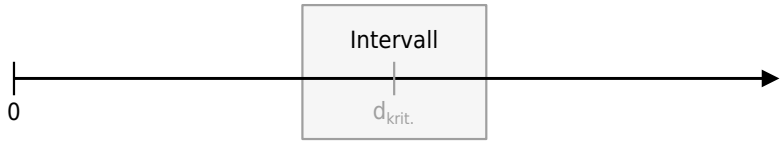

Abbildung 6: Unklare Effektstärke

- Ist auch UG > $d_{krit.}$ haben wir einen signifikanten und praktisch relevanten Effekt gefunden.

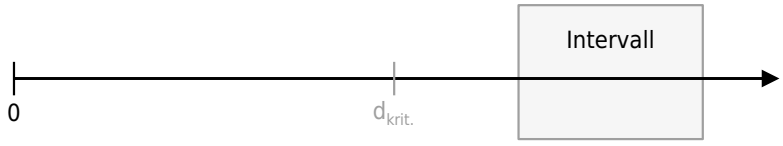

Abbildung 7: Relevante Effektstärke

Je größer die Stichprobe, desto enger liegen die Intervallgrenzen beieinander. Der zweite, unklare Fall wird dadurch unwahrscheinlicher.

Wie würden wir das nun in einer Studie berichten? Wir zeigen die Tabelle mit den Mittelwerten, Standardabweichungen und der Stichprobegröße. Dazu berichten wir den t-Wert mit Freiheitsgaden, den p-Wert und die Effektgröße, z. B. so:

> „Im Studiengang IG erreichten die Teilnehmer:innen eine signifikant bessere Note mit t(178)= 2,65; p = 0, 009.
> Es liegt ein schwacher Effekt mit unklarer praktischer Bedeutung vor mit d=0,4 [0,1;0,7] im 95 % KI."

> Geben Sie immer an, wie Sie die Werte ermittelt haben, also welche Statistiksoftware oder welche *Excel*-Funktion benutzt wurde bzw. die Literaturstelle der benutzen Tabelle. Keinesfalls sollten Sie den Output einer Statistiksoftware einfach in ihre Arbeit reinkopieren! Die Tabellen zeigen oft zu viele unnötige oder auch zu wenige Daten, haben andere Überschriften und sehen unschön aus. Erstellen Sie alle Tabellen deswegen mit ihrer Textverarbeitungssoftware.

5.3 Teststärke

Im Beispiel „Einkommen“ hatten wir keinen signifikanten Effekt festgestellt. Hier müssen wir nun mit der Teststärke prüfen, ob der Schluss erlaubt ist, dass (mit hoher Wahrscheinlichkeit) kein relevanter Unterschied in der Grundgesamtheit besteht. Die Teststärke P (Power) ist die Wahrscheinlichkeit, dass ein Unterschied in Höhe eines definierten Effekts, sofern er denn existiert, erkannt worden wäre. Die Power sollte mindestens 80 % betragen, um die Nullhypothese anzunehmen.

Die Power kann nicht direkt ermittelt werden. Es kann jedoch ein Zwischenwert (z-Wert) errechnet werden, aus dem sich dann aus Tabellen zur Standardnormalverteilung der Wert für P ablesen lässt. Der kritische z-Wert ist $z_{krit} = 0{,}84$; dies entspricht etwa einem P = 0,8. Kleinere z-Werte stehen für kleinere P-Werte.

Formel | Zwischenwert zur Teststärkenbestimmung

$$z = 0,5 \times d_{krit} \sqrt{\frac{4n_1 \times n_2}{n_1 + n_2}} - t_{krit}$$

> Für die post hoc (nachträgliche) Teststärke sind die **vorab** festgelegten kritischen Werte von d und t einzusetzen. Die Teststärke kann man also nur bestimmen, wenn man weiß, auf welche Mindesteffektgröße man untersucht und welcher Alpha-Fehler akzeptiert wurde. Es sei noch einmal daran erinnert, dass man hier nicht mit der ermittelten Effektstärke der Stichproben rechnen soll, auch wenn das tatsächlich manchmal von der ein oder anderen Statistiksoftware gemacht wird.

Nehmen wir für unser Beispiel „Einkommen“, dass wir auf einen schwachen Effekt der Stärke $d_{krit} = 0{,}2$ testen. Der z-Wert ergibt sich nun zu

$z = 0,1\sqrt{200} - 1,65 = -0,24$.

Damit ist z deutlich kleiner als 0,84; die Teststärke ist also geringer als 80 %. In der Tabelle am Ende dieses Buches kann man ablesen, dass bei z = −0,24 die Teststärke bei etwa 40 % liegt. Excel ermittelt die Teststärke genau

über die Funktion **+STANDNORMVERT**[70](−0,24) zu 40,6 %. Ein Effekt in der für relevant angesehen Größe von d = 0,2 hatte also nur eine geringe Wahrscheinlichkeit, um entdeckt zu werden. Die „Nullhypothese" für d = 0,2 kann nicht angenommen werden.

Wir können hier keine statistisch belastbaren Aussagen tätigen. Die Stichprobe war einfach zu klein, um damit etwas anfangen zu können. Das ist in einer Bachelorarbeit auch nicht schlimm, wenn Sie das klar darstellen. Im Rahmen einer anspruchsvollen Forschung wäre es hingegen nicht tragbar und damit der GAU: Der ganze Aufwand ohne belastbares Ergebnis und folglich ohne Erkenntnisgewinn.

Im Bericht könnte dann stehen:

> „Der Einkommensunterschied war (knapp) nicht signifikant mit t(198) = 1,56; p = 0,06. Die Teststärke war jedoch mit 40,6 % (für d = 0,2) zu gering, um einen schwachen Effekt ausschließen zu können."

Wenn Sie sich nicht vorab auf eine Mindesteffektstärke festgelegt haben, sollten Sie stattdessen nun eine **Sensitivitätsteststärkeanalyse**[71] durchführe, d. h. die Effektgröße angeben, bei der die Teststärke gleich 80 % wird. Berechnen kann man diese Grenzeffektstärke d_{Grenz} wie folgt:

Formel | Sensitivität

$$d_{\text{Grenz}} = \frac{2\left(t_{\text{krit}} + z_{\text{krit}}\right)}{\sqrt{\frac{4 \times n_1 \times n_2}{n_1 + n_2}}}$$

Und es ergibt in unserem Fall für den einseitigen Test:

$$d_{\text{Grenz}} = \frac{2\left(1,65 + 0,84\right)}{\sqrt{\frac{4 \times 100 \times 100}{100 + 100}}} = 0,35$$

Sie können also berichten, dass die Existenz einer Effekthöhe von d > 0,35 für die Grundgesamtheit als unplausibel anzusehen ist, da diese mit hoher

70 In neueren *Excel*-Versionen verwendet man NORM.S.VERT mit der Eingabe (−0,24;1). Die Ergebnisse sind identisch.

71 vgl.: Döring, Bortz, 2016, S. 815

Wahrscheinlichkeit aufgedeckt worden wäre. Spezifiziert für d > 0,35 kann die „Nullhypothese“ akzeptiert werden.

> Sensitivitätsanalysen können Sie zu vielen Tests mit dem kostenlosen Programm ***G*Power*** durchführen.

5.4 Optimaler Stichprobenumfang

Der **Stichprobenumfang** kann so geplant werden, dass der GAU verhindert wird, also auf jeden Fall nach der Datenerhebung eine plausible Aussage zu Gunsten der Alternativhypothese oder der Nullhypothese möglich wird. Dazu muss die Stichprobe pro Gruppe nur groß genug sein; es muss ein optimaler Stichprobenumfang eingehalten werden.

Der **optimale Stichprobenumfang** $n_{opt.}$ pro Gruppe errechnet sich im t-Test wie folgt:

Formel | Optimaler Stichprobenumfang

$$n_{\text{opt}} = \frac{2\left(t_{\text{krit}} + z_{\text{krit}}\right)^2}{d_{\text{krit}}^{\ 2}}$$

Entscheidend sind also folgenden Größen:

- Der akzeptierte Alpha-Fehler und ob wir einseitigen oder zweiseitig testen. Dadurch wird t_{krit} bestimmt:

t_{krit}	Einseitiger Test (gerichtete Hypothese)	Zweiseitiger Test (ungerichtete Hypothese)
90 % Quantil (Alpha 10 %)	1,28	1,65
95 % Quantil (Alpha 5 %)	**1,65**	**1,96**
99 % Quantil (Alpha 1 %)	2,33	2,58

Tabelle 8

Wir entscheiden uns im Beispiel „Einkommen“ für den einseitigen Test und den „normalen“ Standard einen Alpha-Fehler von 5 % zu akzeptieren und wählen damit $t_{krit} = 1{,}65$.

- Die verlangte Teststärke, die den Wert z_{krit} bestimmt:

Teststärke	**80 % (Beta 20 %)**	90 % (Beta 10 %)	95 % (Beta 5 %)	99 % (Beta 1 %)
z_{krit}-Wert	**0,84**	1,28	1,65	2,33

Tabelle 9

Wir nehmen die konventionell üblichen 80 % und wählen $z_{krit} = 0{,}84$

- Die Effektstärke, die mindestens erreicht sein muss, damit wir von einem relevanten Effekt ausgehen. Hier entscheiden wir uns, schon einen schwachen Effekt nach Cohen von $d_{krit} = 0{,}2$ für relevant zu halten.

In unserem Beispiel wäre also $n_{opt} = \frac{2(1,65 + 0,84)^2}{0,2^2} = 310$. Statt 100 pro Gruppe hätten also mehr als dreimal so viele Personen befragt werden müssen. Da die Effektstärke quadriert in die Berechnung einfließt, kommt ihr eine hohe Bedeutung zu. Würde man hier nur auf einen mittleren Effekt von $d = 0{,}5$ testen, so reduziert sich die Stichrobbengröße deutlich auf nur noch 50 pro Gruppe!

Für die Berechnung kann auch folgende *Excel*-Datei benutzt werden:

◆ **Download 2** | Stichprobenumfangsplanung (Service_53241_02.xls)

Die beiden Gruppen müssen nicht exakt gleich groß sein, dies ist aber zu empfehlen. Eine geringere Anzahl in einer Gruppe kann nämlich nur durch eine überproportional größere Anzahl in der anderen Gruppe ausgeglichen werden. Bei gegeben n_{opt} und gegebener Gruppengröße n_1 kann die nötige Gruppengröße von n_2 berechnet werden mit:[72]

72 vgl.: Döring, Bortz, 2016, S. 843

$$n_2 = \frac{n_1 \times n_{opt}}{2n_1 - n_{opt}}$$

Wenn die Anzahl der Proband:innen größer ist als der optimale Stichprobenumfang, ist dies für die Studie zunächst egal. Ist die Anzahl der Proband:innen mit Kosten verbunden, dann wäre eine größere Stichprobe aber unökonomisch. Ist die Forschung, wie es oft in medizinischen Studien der Fall ist, mit Unannehmlichkeiten oder gar Risiken für die Teilnehmer:innen verbunden, ist es schon rein forschungsethisch nicht vertretbar, über die optimale Stichprobengröße hinauszugehen.

> Bei Fragebögen sollten Sie bedenken, dass nie alle Bögen auswertbar sind. Deshalb sollten Sie die Stichprobengröße großzügig nach oben aufrunden.

Bei einer typischen Onlineumfrage, verteilt in den sozialen Netzwerken, wird man eher keine Stichprobenumfangsplanung vorab vornehmen; denn man hat ohnehin keinen direkten Einfluss auf die Zahl der Teilnehmer:innen bzw. man nimmt was man kriegen kann und das ist oft deutlich weniger als hier berechnet. Wie erwähnt, ist das nicht schlimm; der Unterschied kann bei entsprechender Effektgröße trotzdem signifikant werden. Lediglich bei insignifikanten Ergebnissen lässt sich dann eben kein belastbarer Schluss ziehen. Darauf muss gegebenenfalls hingewiesen werden.

> Kann der optimale Stichprobeumfang realisiert werden, können wir sicher sein, dass
>
> - bei einem Effekt von mindestens $d_{krit.}$ in der Stichprobe dieser signifikant wird und die Alternativhypothese angenommen werde kann.
> - bei einem nicht signifikanten Unterschied die Teststäke für $d_{krit.}$ hoch genug ist, um die Nullhypothese anzunehmen.

Wenn Sie eine solche Stichprobenumfangsplanung vorgenommen haben, kann die Post-hoc-Berechnung der Teststärke auch entfallen. Wurde die ermittelte Anzahl an Proband:innen erreicht oder übertroffen, so kann bei einem nicht signifikanten Unterschied die Nullhypothese für die vorgegebene Effektgröße ohne weitere Prüfung angenommen werden. Ist das nicht der Fall, ist bei einem nicht signifikanten Wert die Teststärke für eine Aussage zu gering.

5.5 Voraussetzungen für den t-Test

Der t-Test gehört zu den parametrischen Tests und ist als solcher an bestimmte Voraussetzungen gebunden.[73] Dazu gehört die Annahme der **Normalverteilung,** das ist die symmetrische Verteilung der Werte um ihren Mittelwert, die in der berühmten **Glockenkurve** auch optisch zum Ausdruck kommt.[74] Dies kann man zwar nun mit speziellen Verteilungstests prüfen, wir empfehlen aber dies nicht zu tun. Der t-Test gilt als robust bei Verletzung der Normalverteilungsannahme, d. h. er liefert auch ohne normalverteilte Daten brauchbare Ergebnisse. Hinzu kommt, dass bei Stichproben über 30 die verwendete t-Verteilung ohnehin immer mehr zu einer Normalverteilung wird.

Etwas kniffliger ist die Annahme der **Varianzhomogenität**. Demnach müssen die Standardabweichungen der Grundgesamtheiten der beiden Gruppen gleich sein. Im Regelfall wird man gar nicht wissen können, ob dies zutrifft. Ein **F-Test** oder der **Levene-Test**, wie ihn *SPSS* ausführt, ist so ein Test, der die Varianzhomogenität anhand der Streuung der Stichprobenwert überprüft. Auch hier können wir zumindest teilweise Entwarnung geben: Der t-Test funktioniert nämlich auch bei völlig unterschiedlichen Standardabweichungen in den Gruppen, wenn die Gruppengrößen einigermaßen gleich groß sind. Nur wenn sowohl die Standardabweichungen **und** die Gruppengröße deutlich unterschiedlich sind, wird der t-Test ungenau. In diesem Ausnahmefall kann man auf den **Welch-Test**[75] ausweichen, der aber eine geringere Teststärke hat.

Die Statistiksoftware *SPSS* liefert immer sowohl die Ergebnisse des „original" t-Test (unter „bei gleichen Varianzen") als auch des Welch-Test (unter „bei ungleichen Varianzen"), ohne den Begriff Welch-Test zu verwenden. Wenn ihre Stichproben über jeweils mehr als 30 Werte verfügen und einigermaßen gleich groß sind, können Sie das ignorieren, Sie brauchen sich um die Voraussetzungen dann nicht weiter zu kümmern.

Die hier vorgestellten Formeln gelten aber nur für den t-Test mit zwei **unabhängigen Stichproben**. Bei einem **Vorher-nachher-Vergleich** mit denselben Proband:innen spricht man von abhängigen Stichproben oder von Messwiederholung. Dann sind andere Formeln, nämlich die für abhängige Stichproben[76] zu verwenden. Unbedingt nötig sind diese aber nicht, wir

73 vgl.: Westermann, 2000, S. 331ff.
74 vgl.: Blasius, Thiessen, 2021, S. 95ff.
75 vgl.: Janczyk, Pfister, 2009, S. 55
76 vgl.: z. B.: Sedlmeier, Renkewitz, 2018, S. 414ff.

werden sehen, dass man bei Messwiederholung auch den t-Test für den 1-Stichprobenfall aus dem nächsten Kapitel verwenden kann.

5.6 Der 1-Stichprobenfall

Es ist auch möglich, einen Stichprobenmittelwert aus nur einer Stichprobe mit einem vorgegebenen Mittelwert, etwa aus der bekannten Gesamtpopulation, zu vergleichen, z. B. wenn Sie testen wollen, ob Studierende einen höheren IQ haben als die Durchschnittsbevölkerung, von der bekannt ist, dass die Mitte bei 100 Punkten und die Standardabweichung bei 15 Punkten liegt. Bei solchen Fragstellungen spricht man auch von einem **Anpassungstest**. Die Formeln verändern sich jetzt zum Grenzwert hin, bei der die Vergleichsgröße aus einer unendlich großen „Stichprobe“ stammt.

Der empirische t- Wert ermittelt sich im Anpassungstest wie folgt:

Formel | t-Wert bei einem Mittelwert

$$t = \frac{|\text{ABW}| \times \sqrt{n}}{\sigma}$$

Wenn bekannt (wie im Beispiel), kann man als Standardabweichung, die des Populationswertes verwenden, dann sprich man auch vom **Gauß-Test**. Andernfalls nimmt man die empirisch ermittelte Standardabweichung der Stichprobe, wie das der t-Test für den 1-Stichprobenfall vorsieht.

Nehmen wir nun an, wir haben n = 60 Studierende befragt und als Mittelwert 104 erhalten, also eine Abweichung von 4. Der t-Wert ist:

$$t = \frac{4 \times \sqrt{60}}{15} = 2,1$$

Den p-Wert erhalten wir wieder über *Excel*, wobei wir hier einseitig testen. Die Freiheitgrade im 1-Stichprobenfall betragen f = n − 1. *Excel* liefert über **T.VERT.RE**.(2,1;59): p = .022; der Unterschied ist signifikant.

Ob die 4 Punkte aber auch relevant sind, ist zu diskutieren. Das Effektstärkemaß d[77] errechnet sich beim Anpassungstest wie folgt:

Formel | Effektstärke bei einem Mittelwert

$$d = \frac{|\text{ABW}|}{\sigma}$$

Es beträgt hier 4/15 = 0,27 was nach Cohen einen schwachen Effekt entspricht. Cohen hat für den 1-Stichprobenfall allerdings eine korrigierte Berechnung bzw. abweichende Interpretation von d vorgeschlagen. So sollen die Interpretationsschwellen durch $\sqrt{2}$ geteilt werden und es ergibt sich dann für d > 0,14 bereits ein schwacher, für d > 0,35 ein mittlerer und ab d > 0,57 ein großer Effekt.[78]

Statt sich diese krummen Zahlen zu merken, empfehlen wir, gleich die Konventionen für die Korrelation r zu übernehmen, also mit > 0,1 als schwachen Effekt, > 0,3 als mittlerer Effekt und > 0,5 als starker Effekt.

Ist die Teststärke gefragt, kann diese über die Standardnormalverteilung von z beim Anpassungstest wie folgt bestimmt werden:

Formel | Zwischenwert zur Teststärkenberechnung bei einem Mittelwert (Anpassungstest)

$$z = d_{\text{krit}} \times \sqrt{n} - t_{\text{krit}}$$

77 Die Effektgröße d ist auch wie folgt aus t berechenbar: $d = \frac{t}{\sqrt{n}}$

78 vgl.: Eil et al, 2017, S. 230

Formel | Optimale Stichprobenumfang bei einem Mittelwert (Anpassungstest)

$$n_{opt} = \frac{\left(t_{krit} + z_{krit}\right)^2}{d_{krit}^{\ 2}}$$

Formel | Sensitivität bei einem Mittelwert (Anpassungstest)

$$d_{Grenz} = \frac{\left(t_{krit} + z_{krit}\right)}{\sqrt{n}}$$

Nutzen Sie dazu auch gerne in *Excel*:

◆ **Download 3** | Unterschiedstest 1Mittelwert (Service_53241_03.xls)
◆ **Download 4** | Stichprobenumfangsplanung (Service_53241_02.xls)

Auch für zwei abhängige Stichproben können wir diesen Test verwenden. Stellen wir uns vor, wir fragen eine Gruppe von Proband:innen nach ihrer Zahlungsbereitschaft für ein Produkt, einmal vor und dann nochmal nach dem diese einen Werbefilm zu diesem Produkt gesehen haben. Das wäre ein typisches Beispiel für eine abhängige Stichprobe bzw. eine Messwiederholung. Die Größe um die es jetzt geht, ist die jeweilige **Differenz** der Zahlungsbereitschaft vorher und nachher. Aus diesen Differenzgrößen lassen sich nun wieder Mittelwert und Standardabweichung errechnen. Getestet wird dann gegen eine Differenz von Null als Populationsgröße.

5.7 Mehr als zwei Mittelwerte

Im Beispiel „Mathenoten“ im → Kapitel 5.1 haben wir die Notenmittelwerte zweier Studiengänge auf Signifikanz geprüft. Wie gehen wir nun vor, wenn

wir mehr als zwei Studiengänge und damit auch mehr als zwei Mittelwerte haben?

Die naheliegende Möglichkeit wäre es nun, paarweise Mittelwertvergleiche zu bilden. Bei 3 Gruppen können wir dann 3 Paare bilden, bei 5 Gruppen wären es schon 10 Paare. Das wird überwiegend kritisch gesehen, da es zu einer Fehlerakkumulation kommt. Wenn ein einziger t-Test eine Vertrauenswahrscheinlichkeit von 95 % hat, so haben 10 t-Tests dann nur noch $0{,}95^{10}$ = 60 % Vertrauenswahrscheinlichkeit. Man müsste, um im Gesamtergebnis auf 95 % zu kommen, die Alpha-Fehler für die einzelnen Tests kleiner wählen (**Bonferroni-Korrektur**), so klein, dass dabei oft kein signifikantes Ergebnis mehr herauskommt.

Allerdings wird in der Literatur dabei fast immer übersehen, dass die Fehlerakkumulation in dieser Höhe nur für stochastisch unabhängig Tests gilt. Genau das dürfte im Regelfall bei einer Studie zu einem zusammenhängenden Thema gar nicht zutreffen. Schließlich verlangt beispielsweise schon die Anforderung nach „interne Konsistenz" eines Fragebogens eine starke Korrelation der Items. Außerdem sind die Vertrauenswahrscheinlichkeiten bei Signifikanz nach dem Test oft höher. Das Argument der Fehlerakkumulation dürfte deshalb wohl deutlich überschätzt sein.

> Es ist üblich, die Signifikanz mehrere Mittelwerte über eine **Varianzanalyse**, auch **ANOVA** (Analysis of Variance) genannt, zu testen da diese die Fehlerakkumulation vermeidet. Das Effektmaß ist hier Eta-Quadrat, das anders zu interpretieren ist als Cohens d. Dabei gibt es aber nicht die eine Varianzanalyse, sondern verschiedene Varianten, die je nach Aufgabenstellung anzuwenden sind. Die Datenauswertung erfolgt hierbei am besten über Statistiksoftware, wie *SPSS* oder *R*.

Die Interpretation der Daten ist knifflig. Ein signifikantes Ergebnis bedeutet nur, dass irgendwo ein Unterschied besteht. Das ist aber nur eine sehr unspezifische Alternativhypothese. Was soll signifikant bei mehr als zwei Mittelwerten eigentlich bedeuten? Das sich alle Gruppen unterscheiden? Nur eine Gruppe von der Summe aller anderen oder nur die Extremgruppen am Rand?

Es müssen nun immer noch zusätzliche Post-hoc-Analysen oder paarweise Kontrasttests erfolgen, die alle ihre Besonderheiten haben. Wobei ein paarweiser Kontrasttest auch nichts anderes ist als eine Art modifizierter t-Test und dieser eine vorhergehend ANOVA eigentlich überflüssig macht.

Auch ist es sogar möglich, dass die ANOVA kein signifikantes Ergebnis zeigt, obwohl es im Paarvergleich dann doch signifikante Unterschiede gibt.[79]

Letztlich ist es zweifelhaft, ob es überhaupt einen Erkenntnisgewinn bringt auf Unterschiede mehrerer Mittelwerte zu untersuchen. Es ist eher so, dass der Begriff „Unterschied" nur im Vergleich zweier Wertepaare sinnvoll interpretierbar ist.

Ungeachtet dieser kritischen Punkte aber hat sich die **Varianzanalyse** in der Literatur etabliert, wozu die oft kostenlos zur Verfügung stehenden Statistiksoftwarepakete sicher stark beigetragen habe, denn diese erledigen für uns die notwendigen komplexen Rechenoperationen. Eine Interpretation liefern sie ab nicht mit. Deshalb setzt ihre Benutzung ein hohes Maß an Einarbeitung und Kenntnisse voraus. Das ist dann schon – wenn überhaupt – eher etwas für Masterarbeiten. Für eine Bachelorarbeit sollten Sie ihre Forschungsfrage so stellen, dass Sie mit einem oder wenigen t-Test auskommen oder die metrischen Daten in Kategorien einteilen und einen Chi-Quadrat-Test nutzen, dazu später mehr.

79 vgl.: Westermann, 2000, S. 402

6 Test bei zwei Anteilswerte (Prozentzahlen)

6.1 Signifikanztest

Ein Test auf unterschiedliche Mittelwerte ist nur bei metrischen Merkmalen (also „echten“ Zahlen) oder Werten, die man methodisch großzügig als metrisch definieren kann, durchführbar. Manchmal sind die abgefragten Merkmale aber rein kategorial (also Wörter wie ja, nein). Dann kann man prozentuale Anteile der Antwortkategorien auf Unterschiede testen.

Beispiel | Bekanntheitsgrad
Forschungsfrage: Ist der Bekanntheitsgrad einer Biermarke in Nord- und Süddeutschland unterschiedlich?

- H_A: Der Bekanntheitsgrad ist unterschiedlich
- H_N: Der Bekanntheitsgrad ist gleich

Eine Befragung brachte folgende Ergebnisse:

	Marke bekannt	Marke unbekannt	Summe Teilnehmer:innen	Bekanntheitsgrad
Nord	30	70	$n_1 = 100$	$p_1 = 0{,}30$
Süd	45	55	$n_2 = 100$	$p_2 = 0{,}45$
Summe	75	125	$N = 200$	$p = 0{,}375$

Tabelle 10

Wir verwenden dafür einen Test auf unterschiedliche Anteilswerte, der für Stichprobengrößen n > 30[80] benutzt werden kann. Dieser Test wird in den Statistikbücher auch als **approximativer Binomialtest**[81] bezeichnet.

Wir ermitteln auch hier wieder einen t-Wert. Es wird damit getestet, ob der Unterschied (ABW) von 15 Prozentpunkten (45 % – 30 %) signifikant ist. Der Test ist zweiseitig mit $t_{krit} = 1{,}96$. Der empirische t-Wert errechnet sich bei Anteilswerten wie folgt:

Formel | t-Wert für 2 Anteilswerte

$$t_{emp} = \frac{|ABW|}{\sqrt{\pi(1-\pi) \times \frac{N}{n_1 \times n_2}}}$$

Hier also:

$$t_{emp} = \frac{0,15}{\sqrt{0,375(1-0,375) \times \frac{200}{10000}}} = 2,19 > 1,96$$

Der p-Wert liegt laut Tabelle zwischen 2,5 und 3 %. Mit *Excel* lässt er sich wieder genau angeben. Bei Anteilswerten sind die Freiheitsgrade unendlich, hier können Sie deshalb eine beliebig hohe Zahl, z. B.1000, vorgeben. Wir erhalten bei T.VERT.2S.(2,19;1000): p = . 029 < 5 %.

Der Unterschied von 15 Prozentpunkten ist signifikant, die Alternativhypothese kann angenommen werden.

6.2 Effektstärke

Obwohl Cohen sein Maß „Cohens d“ ausdrücklich für den Effekt von Mittelwertunterschieden erdacht hatte, lässt es sich auch für Anteilswerte

80 In einigen Lehrbücher wird hingegen $n = 9/(\pi(1-\pi))$. als Faustformel für die nötige Stichproben zur Anwendung des Tests vorgeschlagen, vgl.: z. B. Bortz, Lienert, 2008, S.:42

81 Nicht zu verwechseln mit dem eigentlichen (exakten) Binomialtest, der für kleinere Stichproben benutzt werden soll.

in einer dafür modifizierten Form anwenden. Dabei empfiehlt sich folgende Formel:[82]

Formel | Adaptiertes d für 2 Anteilswerte

$$d = \frac{|\text{ABW}|}{\sqrt{\frac{\pi_1(1-\pi_1) \times n_1 + \pi_2(1-\pi_2) \times n_2}{n_1 + n_2}}}$$

Damit ist im Beispiel:

$d = \frac{0,15}{\sqrt{\frac{0,3 \times 0,7 \times 100 + 0,45 \times 0,55 \times 100}{200}}} = 0,31$, also nach Cohen ein schwacher Effekt.

Das Konfidenzintervall ist KI95%[0,03;0,59].

Es ist jedoch auch hier zu empfehlen, sich selbst Gedanken zu machen, ab welcher Effektstärke ein für die konkrete Fragestellung relevanter Effekt anfängt. Dabei werden in diesem Beispiel wohl die Kosten der Werbung und der Zusammenhang zwischen Bekanntheitsgrad und Umsatz eine Rolle spielen und vielleicht kommt man dann nach solchen Wirtschaftlichkeitsüberlegungen zu einem anderen Mindesteffekt.

Hier ein paar Beispiele für Unterschiede in den Anteilswerten und die dazugehörigen Effektstärken d:

d = 0,2	d = 0,35	d = 0,5	d = 0,8
10%/17%	10%/23%	10%/29%	10%/43%
30%/40%	30%/47%	30%/54%	30%/67%
50%/60%	50%/67%	50%/74%	50%/85%
70%/79%	70%/84%	70%/90%	70%/97%

Tabelle 11

82 In der Literatur, z. B.: Döring, Bortz, 2016, S. 827, wird stattdessen eine Transformation in eine Arcussinusfunktion und die Bildung der Effektstärke $h = \left|2 \times ARCSIN\sqrt{\pi_1} - 2 \times ARCSIN\sqrt{\pi_2}\right|$ empfohlen. Dies führt bei aufwändigerer Rechnung in den Bereichen in denen Effekte praktisch vorkommen aber zu den gleichen Ergebnissen wie die vorgestellte modifizierte d Formel und kann deshalb unterbleiben.

6.3 Teststärke und optimaler Stichprobenumfang

Wenn wir auch für den Anteilswerttest die (adaptierte) Effektgröße d verwenden, brauchen wir keine neuen Formeln für Power und die Stichprobenumfangsplanung. Sowohl für die Teststärkenermittlung als auch für die Berechnung eines optimalen Stichprobenumfangs können wir die Formeln aus den Mittelwertvergleichen in den ⟶ Kapiteln 5.3 und 5.4 verwenden. Das gilt auch für die Sensitivität. Die Formeln finden Sie auch am Ende des Buches wieder. Hier die *Excel*-Tools dazu:

◆ **Download 5** | Unterschiedstest 2 Anteilswerte (Service_53241_04.xls)
◆ **Download 6** | Stichprobenumfangsplanung (Service_53241_02.xls)

6.4 Auswertung als 4-Felder-Matrix

Ein Test auf Unterschiede zweier Anteilswerte kann alternativ und ohne Umrechnung der Häufigkeiten in Anteilswerte durch eine 4-Felder-Matrix durchgeführt werden. Der Grundtyp der 4-Felder-Matrix ist:

	mit Merkmal	ohne Merkmal
Gruppe 1	**a**	**b**
Gruppe 2	**c**	**d**

Tabelle 12

Der empirische t-Wert lässt sich hierbei einfach aus den 4-Feldern a, b, c und d errechnen:

$$t_{\text{emp}} = \frac{\sqrt{N} \times |a \times d - b \times c|}{\sqrt{(a+b) \times (c+d) \times (a+c) \times (b+d)}}$$

Für das Beispiel „Bekanntheitsgrad“ aus dem ⟶ Kapitel 6.1 ergibt sich die 4-Felder-Matrix zu:

	Marke bekannt	Marke unbekannt
Nord	30	70
Süd	45	55

Tabelle 13

Die Summe der 4-Felder ist N = 200 und es errechnet sich der schon bekannte t-Wert nun wie folgt:

$$t_{\text{emp}} = \frac{\sqrt{200} \times |30 \times 55 - 70 \times 45|}{\sqrt{(30+70) \times (45+55) \times (30+45) \times (70+45)}} = 2,19$$

> Die Höhe des t-Wertes in der 4-Felder-Matrix und damit auch von p ist identisch mit der Berechnung im Anteilstest. Es handelt sich hier also nur um einen anderen Rechenweg, und nicht um eine andere Methode.

Da hier der relative Unterschied der Anteile in Prozent nicht berechnet ist, kann das Effektmaß d nicht direkt berechnet werden. Tatsächlich ermittelt und interpretiert man bei der 4-Felder-Maritx ein anderes Maß für die Effektstärke, den Korrelationskoeffizienten r, in der 4-Felder-Matrix auch **Phi-Korrelation** genannt:

$$r = \frac{t}{\sqrt{N}} = \frac{2,19}{\sqrt{200}} = 0,15$$

Der Wert steht für einen **schwachen Zusammenhang**.

Mit Hilfe der Konversionsformel aus ⟶ Kapitel 4.2 können wir die Korrelation bei Bedarf näherungsweise in d umwandeln:

$$d = \frac{2r}{\sqrt{1-r^2}} = \frac{0,3}{\sqrt{1-0,0225}} = 0,3$$

6.5 Der 1-Stichprobenfall

Es kann auch hier ein Anpassungstest eines Anteilwertes an einer Vorgabe getestet werden.

Beispiel | Studierende
Forschungsfrage: Studieren mehr Frauen als Männer BWL?

- H_A: Der Anteil der Frauen ist größer als der der Männer
- H_N: Der Anteil der Frauen ist gleich oder kleiner als der der Männer

Ergebnis einer Befragung von n = 50 BWL-Studierende war, dass davon 60 % Frauen waren.

Wir bezeichne mit π_S den empirischen Stichprobenanteil von hier 60 % und mit p den Populationsanteil bei Gültigkeit der Nullhypothese der hier 50 % beträgt und ermitteln eine absolute Abweichung von ABW von $|\pi_S - \pi| = |0,6 - 0,5| = 0,1$.

Der empirische t-Wert ist im Anpassungstest:

Formel | t-Wert bei einem Anteilswert (Anpassungstest)

$$t = \frac{ABW \times \sqrt{n}}{\sqrt{\pi(1-\pi)}}$$

Hier kommen wir auf $t = \frac{0,1 \times \sqrt{50}}{\sqrt{0,5(1-0,5)}} = 1,41$. Dies ist auch für den einseitigen Test zu niedrig, der p-Wert ist: nach T.VERT.RE(1,41;1000) p = 0.079. Der Unterschied ist nicht signifikant.

Alternativ dazu kann die Signifikanz auch über das Konfidenzintervall, wie in → Kapitel 3.1 beschrieben, ermittelt werden. Wenn die Untergrenze des Intervalls den Populationswert von hier 50 % überdeckt, ist der Unterschied nicht signifikant. Hier ist die Intervalluntergrenze: $0,6 - 1,65 \times \sqrt{\frac{0,6 \times 0,4}{50}} = 48,6$ % und damit ist der Test insignifikant, da der Zufallsbereich auch noch die 50 % umfasst. Beide Methoden führen widerspruchsfrei zu gleichen Ergebnissen.

Die Effektstärke d entspricht im 1-Stichprobenfall der Korrelation r; hier sind diese:

Formel | Effektgrößen bei einem Anteilswert (Anpassungstest)

$$d = \frac{\text{ABW}}{\sqrt{\pi(1-\pi)}}$$

$$r = \frac{t}{\sqrt{n}}$$

$$d = \frac{\text{ABW}}{\sqrt{\pi(1-\pi)}} = \frac{0,1}{0,5} = 0,2$$

$$r = \frac{t}{\sqrt{n}} = \frac{1,41}{7,07} = 0,2$$

Für die Teststärke, den optimalen Stichprobenumfang und die Sensitivität können die gleichen Formeln verwenden werden wie im → Kapitel 5.6 beim 1-Stichprobenfall für einen Mittelwertvergleich. Die Formeln finden Sie auch nochmal am Ende des Buches.

Hier geht es zu den *Excel*-Tools:

◆ **Download 7** | Unterschiedstest 1 Anteilswert (Service_53241_05.xls)
◆ **Download 8** | Stichprobenumfangsplanung (Service_53241_02.xls)

7 Häufigkeitsverteilungen im Chi-Quadrat-Tests (χ2-Test)

7.1 Der Chi-Quadrat-Unabhängigkeitstest

Mit dem vorgestellten Anteilstest können wir die Häufigkeitsverteilung eines Merkmals mit zwei Ausprägungen bei zwei Gruppen testen. Auch mit dem **Chi-Quadrat-Unabhängigkeitstest** werden Häufigkeitsverteilungen mit kategorialen Merkmalen von unabhängigen Gruppen getestet. Aber: Die Anzahl der Gruppen und die Anzahl der Merkmale, die man auswerten kann, sind hierbei theoretisch unbegrenzt.

Die Häufigkeiten der Nennungen pro Gruppe werden in einer **Kontingenztafel** (auch **Kreuztabelle** genannt) erfasst. Der Test unterliegt zwar keinen besonderen Voraussetzungen, als „asymptotischer Test" funktioniert er aber umso besser, je größer die Stichprobe ist.

> Der Gesamtstichprobenumfang sollte deshalb mindestens N = 60 sein und in keinem Feld der Tabelle die Null stehen. Es werden immer ungerichtete Hypothesen getestet, außer die Kontingenztafel besteht nur aus 4 Feldern, dann kann auch eine gerichtete Hypothese getestet werden, aber für diesen Fall können wir auch auf den einfacheren 4-Felder-Test aus ⟶ Kapitel 6.4 oder den Anteilstest, ⟶ Kapitel 6.1 zurückgreifen.[83]

83 In einem 4-Felder-Chi-Quadrat-Test ist $\chi 2 = t^2$ und der kritische Chi-Quadrat-Wert für das Signifikanzniveau 5 % ist nichts anderes als der quadrierte kritische t-Wert $t^2 = 1{,}96^2 = 3{,}84$. Damit sind beide Methoden ergebnisgleich.

Beispiel | Jobmerkmale
Forschungsfrage: Ist das wichtigste Merkmal[84] bei der Jobsuche abhängig von den gewählten Studiengängen Betriebswirtschaft (BW), Ingenieurwesen (ING) und Maschinenbau (MA)?

- H_A: Die Häufigkeitsverteilungen unterscheiden sich
- H_N: Die Häufigkeitsverteilungen unterscheiden sich nicht

Das Häufigkeitsverteilung einer Umfrage ist in folgender Kontingenztabelle erfasst:

	Gehalt	Arbeitsklima	lokale Nähe	Summe
BW	40	10	30	80
ING	20	20	25	65
MA	20	10	30	60
Summe	80	40	85	205

Tabelle 14

Die Nullhypothese eines $\chi 2$-Testes lautet stets, dass die Häufigkeitsverteilungen gleich sind, also hier: dass die Jobmerkmale unabhängig vom Studiengang gewählt werden. Entsprechend lautet die Alternativhypothese, dass es (irgendwo) einen nicht mehr zufälligen, also signifikanten Unterschied gibt. Die Vorgehensweise ist jetzt, analog zu anderen Tests: Signifikanz lieg vor, wenn ein empirischer $\chi 2$-Wert größer ist als ein kritischer Chi-Quadrat-Wert $\chi 2$(krit).

Um den empirischen $\chi 2$-Wert zu ermitteln, wird zunächst eine Tabelle der Erwartungswerte bei Gültigkeit der Nullhypothese ermittelt. Dazu bildet man für jedes Feld das Produkt aus Zeilensumme mal Spaltensumme und teilt dieses durch die Gesamtstichprobengröße. Für das Feld links oben

84 Wenn Sie Mehrfachantworten zulassen, die Befragten also mehrere für sie wichtige Merkmale nennen können, ist der Chi-Quadrat-Test nicht geeignet. Dann wäre eine Bewertung der Merkmale über eine Ratingskala und die Auswertung der Ratingmittelwerte über einen t-Test besser.

(BW/Gehalt) also: $\frac{80 \times 80}{205} = 31,2$; für das Feld BW/Arbeitsklima $\frac{40 \times 80}{205} = 15,6$ usw.:

Erwartungswerte	Gehalt	Arbeitsklima	lokale Nähe
BW	31,22	15,61	33,17
ING	25,36	12,68	26,95
MA	23,41	11,71	24,88

Tabelle 15

An dieser Stelle empfehlen die Lehrbücher oft eine Voraussetzung zu prüfen[85]: Es soll nämlich in keinem Feld die Zahl 0 stehen und in 80 % der Felder sollen Werte über 5 stehen, anderenfalls kann man Zeilen oder Spalten zusammenfassen. Es ist aber unter Statistiker:innen durchaus strittig, ob der Test nicht auch ohne diese Voraussetzung gültig ist.[86] Hier gibt es dieses Problem nicht, die Voraussetzungen sind erfüllt.

Nun bilden wir die Prüfgrößen (PG) indem wir für jedes Feld die Differenz zwischen dem Wert der Kontingenztafel und dem Erwartungswert quadrieren und durch den Erwartungswert dividieren. Also für links oben: $\frac{(40 - 31,22)^2}{31,22} = 2,47$ .Damit ergeben sich die PG wie folgt:

Prüfgrößen PG	Gehalt	Arbeitsklima	lokale Nähe
BW	2,47	2,02	0,30
ING	1,13	4,23	0,14
MA	0,50	0,25	1,05

Tabelle 16

Der empirische Chi-Quadrat-Wert ist nun die Summe aller dieser 9 Prüfgrößen:

$\chi 2 = 2,47 + 2,02 + 0,30 + 1,13 + 4,23 + 0,14 + 0,5 + 0,25 + 1,05 = 12,09.$

85 vgl.: Bortz, Lienert, 2008, S. 106
86 vgl.: Sedlmeier et al; 2018, S 568f.

Wenn es gar keinen Unterschied gibt, wäre der χ2-Wert Null, je höher er also ist, desto größer wird die Wahrscheinlichkeit, dass sich die Gruppen nicht nur zufällig unterscheiden.

Der **kritische Chi-Quadrat-Wert** ist neben dem akzeptierten Alpha-Fehler abhängig von der Spalten- und Zeilenanzahl, genauer von den Freiheitsgraden f. Bezeichnen wir mit k die Zeilenanzahl und mit m die Spaltenanzahl gilt:

$$f = (k - 1) \times (m - 1)$$

Hier also haben wir $f = (3 - 1) \times (3 - 1) = 4$ Freiheitsgrade. Für das übliche Alpha-Fehlerniveau von 5 % gelten nun folgende Tabellenwerte:

f	χ2 kritisch
1	3,84
2	5,99
3	7,81
4	**9,49**
5	11,07
6	12,59
7	14,07
8	15,51
9	16,93
10	18,31

Tabelle 17

Wir müssen also hier 9,49 wählen, noch genauere Werte kann man für einen Alpha-Fehler von 5 % über *Excel* ermitteln mit: **+CHIQU.INV(0,95;f)**

Damit kann von einem signifikanten Unterschied ausgegangen werden, da der empirische Chi-Quadrat-Wert größer ist als der kritische bei 4 Freiheitsgraden:

χ2 > χ2(krit): 12,09 > 9,49

Auch hier sollte man aber am besten den p-Wert angeben, der kleiner 5 % sein sollte. Genau lässt er sich über spezielle Tabellen ablesen oder einfach

in *Excel* mit +**CHIQU.VERT.RE** und Eingabe des empirischen Chi-Quadrat-Wertes und der Freiheitsgrade. Es ergibt sich CHIQU.VERT.RE (12,09;4) = 0,017 (alternativ können Sie den p-Wert mit CHIQU.TEST bestimmen und dabei die Kontingenztafel und die Tafel der Erwartungswerte markieren). Der p-Wert des Chi-Quadrat-Test wird auch oft, z. B. in Statistikprogramm *SPSS*, mit **asymptotischer Signifikanz** bezeichnet.

Wir können nun zwar die Alternativhypothese annehmen, doch sagt diese nur, dass es irgendwo einen Unterschied gibt, nicht aber wo.

Bei einem signifikanten Unterschied ist ein weitere Prüfschritt erforderlich, um die Felder zu identifizieren, die den Unterschied ausmachen. Dazu empfiehlt sich der **Fuchs-Kenett-Ausreißertest**.[87]

Dessen Tabellenwerte (FKW) berechnen sich aus den beobachteten Häufigkeiten (H), den erwarteten Häufigkeiten (E), Zeilensumme (Z), Spaltensumme (S) und Stichprobengröße (N) für jedes Feld wie folgt:

Formel | FK-Werte (standardisierte Residuen)

$$FKW = \frac{H - E}{\sqrt{E \times \left(1 - \frac{Z}{N}\right) \times \left(1 - \frac{S}{N}\right)}}$$

Für das Feld BW/Gehalt z. B.: $FKW = \frac{40 - 31,22}{\sqrt{31,22 \times \left(1 - \frac{80}{205}\right) \times \left(1 - \frac{80}{205}\right)}} = 2,58$

Damit ergibt sich die Tabelle der Fuchs-Kenett-Werte (FKW)[88] wie folgt:

Residuen	Gehalt	Arbeitsklima	lokale Nähe
BW	**2,58**	**-2,03**	-0,92
ING	-1,65	**2,77**	-0,59
MA	-1,07	-0,66	1,60

Tabelle 18

87 vgl.: Bortz, Lienert, 2008, S. 113ff.

88 In der Literatur werden die FKW auch als standardisierte Residuen bezeichnet, in *SPSS* als korrigierte Residuen.

Die einfachste Interpretation[89] ist nun:

> Werte mit einem Betrag größer 1,96 deuten auf signifikante Unterschiede hin, bei positiven Vorzeichen auf überproportional hoch und bei negativen Vorzeichen auf unterproportional niedrig.

Hier zeigt sich eine überproportionale Bedeutung des Gehaltes für BW-Studierende und eine überproportionale Bedeutung des Arbeitsklimas bei ING-Studierende, während dieses für BW Studierende signifikant weniger wichtig ist.

7.2 Ergänzungen: Effektstärke/Teststärke

Auch bei einem Chi-Quadrat-Test sagt die Signifikanz nichts über die praktische Bedeutsamkeit der Unterschiede aus, so dass eine ergänzende Betrachtung mit Hilfe einer Effektstärke notwendig wird. In der Literatur werden mehrere Effektstärkemaße für den Chi-Quadrat-Test diskutiert. Das Statistikprogramm *G*Power*[90] rechnet mit **Cohens** ω (sprich Omega). Diese Effektstärke entspricht genau der Phi-Korrelation des 4-Felder-Test und ist:[91]

Formel | Cohens Omega

$$\omega = \sqrt{\frac{\chi^2}{N}}$$

Im Beispiel „Jobmerkmale“ ergibt sich $\omega = \sqrt{\frac{12,09}{205}} = 0,24$.

89 Auf die sehr konservative Interpretation mit Hilfe der Bonferroni-Korrektur verzichten wir, da diese bei signifikanten Unterschieden im Test nicht immer einen Ausreißer identifizieren kann.

90 Das überaus hilfreiche und international anerkannte Tool *G*Power* der Uni Düsseldorf können Sie kostenlos runterladen unter: *www.gpower.hhu.de*

91 Eigentlich ist damit die Phi-Korrelation als Effektgröße überflüssig, da sie Cohens Omega immer entspricht, einige Autor:innen machen es umgekehrt und bezeichnen Omega immer mit Phi. Wie auch immer, benötigt wird nur eines davon.

Die Effektstärke ω wird üblicherweise wie eine Korrelation interpretiert (obwohl sie keine ist), also mit > 0,1 schwacher Zusammenhang, > 0,3 mittlerer Zusammenhang und > 0,5 starker Zusammenhang.

Wir sollten in einem Bericht die Kontingenztabelle und bei Signifikanz die Tabelle des Fuchs-Kennet-Tests angeben sowie den $\chi 2$- Wert mit Freiheitsgraden, den p-Wert und die Effektgröße. Dann kann so berichtet werden:

> „Es liegt ein signifikanter Unterschied der Studiengänge bei der Merkmalswahl vor mit $\chi 2(4) = 12{,}09$; p =.017 und ein schwacher Effekt mit $\omega = 0{,}24$. Im Studiengang BW wird das Merkmal Gehalt überproportional häufig und das Merkmal Arbeitsklima unterproportional häufig gewählt. Letzteres ist im Studiengang ING überproportional vertreten. Im Studiengang MA gibt es keine statistischen Ausreißer."

Ist die Angabe einer Teststärke erforderlich, weil der Testunterschied insignifikant ist, so kann diese für ein vorgegebenes ω_{krit} über das schon erwähnte Tool *G*Power* ermittelt werden. Wenn *G*Power* nicht zur Verfügung steht, kann die Standardnormalverteilung über einen z-Wert verwendet werden. Der z-Wert eines Chi-Quadrat-Testes bei Alpha-Fehler von 5 % kann mit Hilfe der Gesamtstichprobengröße (N) und der Freiheitsgrade (f) wie folgt geschätzt werden:

Formel | Zwischenwerte zur Teststärkenberechnung im Chi-Quadrat Test

$$z = \omega_{krit} \times \sqrt{N} \times f^{-0,16} - 1,96.$$

Der optimale Gesamtstichprobenumfang im Chi-Quadrat-Test mit Alpha-Fehler 5 % und Teststärke 80 % lässt sich mit folgender Formel schätzten:

Formel | Optimaler Stichrobenumfang Chi-Quadrat-Test

$$N = \frac{7,85 \times f^{0,31}}{\omega^2}$$

Die Tabelle zeigt einige optimale Stichprobengrößen des Chi-Quadrat-Test für Alpha 5 % ermittelt mit *G*Power*:

	ω = 0,1	ω = 0,3	ω = 0,5
f = 1	785	88	32
f = 2	964	108	39
f = 3	1091	122	44
f = 4	1194	133	48
f = 5	1283	143	52
f = 6	1363	152	55
f = 7	1436	160	58
f = 8	1503	167	61
f = 9	1565	174	63
f = 1	1625	181	65

Tabelle 19

Die Tabelle macht deutlich, dass man im Rahmen einfacher Studien eher auf mittlere Effekte testen sollte, da die notwendigen Stichprobenumfänge, um auch noch kleine Effekt zu entdecken, sehr hoch sind.

Für die Sensitivität bei einer Teststärke von 80 % und 5 % Signifikanzniveau kann man folgende Formel nutzen:

Formel | Sensitivität im Chi-Quadrat-Test

$$\omega_{\text{Grenz}} = \frac{2,8}{\sqrt{N} \times f^{-0,16}}$$

Cohens ω hat in den Augen mancher Forscher:innen den „Schönheitsfehler", dass es im Unabhängigkeitstest mit mehr als 4 Feldern auch Werte über 1 annehmen kann, was mit einer Interpretation als Korrelation nicht vereinbar ist. In der Literatur[92] wird deshalb oft als Alternative das Effektmaß **Cramers v** vorgeschlagen[93]. Cramers v ist eine Art korrigiertes ω, dass nur Werte zwischen 0 und 1 annehmen kann. Die Formel dazu ist:

Formel | Cramers v

$$v = \sqrt{\frac{\chi^2}{N(L-1)}}$$

Dabei ist L (Low) der kleinere Werte aus der Zeilenanzahl (k) oder der Spaltenanzahl (m). So ist L = k, wenn k < m und L = m, wenn m < k. Im Beispiel „Jobmerkmale" ist es egal, da hier m = k = 3 ist und somit ist v:

$$v = \sqrt{\frac{12,09}{200(3-1)}} = 0,17$$

Aus der Formel erkennt man: Wenn es in der Kontingenztabelle nur zwei Zeilen oder zwei Spalten gibt, ist Cramers v mit Cohens ω identisch. Auch Cramers v ist somit im 4-Felder-Fall gleich der dort benutzen Phi-Korrelation. Man sollte aber nicht übersehen, dass Cramers v zwar den gleichen Wertebereich wie eine Korrelation umfasst, aber dass es keine „echte" Korrelation ist, es sieht nur so aus. Es ist schon definitionsgemäß bei kategorialen Daten mit mehr als zwei Kategorien gar nicht möglich eine

92 vgl.: z. B.: Quatember, 2017, S 71

93 Nach dem schwedischen Statistiker *Harald Cramér* (1893–1985)

Korrelation anzugeben![94] Insofern ist Cramers v nicht das bessere Maß gegenüber Cohens ω, es ist einfach nur ein anderes Maß. Wenn man akzeptiert, dass ein Effektmaß auch Werte über 1 annehmen kann,[95] gibt es eigentlich keinen Grund Cramers v vorzuziehen, zumal es für letzteres keine einheitlichen Konventionen zur Interpretation gibt. Es empfiehlt sich also eher ω zu verwenden.

In älteren Lehrbüchern finden man noch eine weitere Effektgröße, den Kontingenzkoeffizienten. Dieser ist aber zu Recht aus der Mode gekommen und braucht uns hier nicht weiter zu interessieren.

Die Effektgrößen ω und v können leicht ineinander überführt werden. Dabei ist dann

$\omega = v\sqrt{L-1}$ oder andersrum: $v = \omega/\sqrt{L-1}$.

Für die Berechnungen stehen ihnen die *Excel*-Tools zur Verfügung:

◆ **Download 9** | Chi-Quadrat-Unabhängigkeitstest (Service_53241_06.xls)
◆ **Download 10** | Stichprobenumfangsplanung (Service_53241_02.xls)

7.3 Der Chi-Quadrat-Anpassungstest

Wenn die Abweichung einer Häufigkeitsverteilung von einer theoretischen Vorgabe für die Nullhypothese getestet werden soll, kommt der **Chi-Quadrat-Anpassungstest** zur Anwendung. Der Stichprobenumfang sollte hier größer 40 sein.

Beispiel | Design
Forschungsfrage: Werden unterschiedliche Designs für Smartphones unterschiedlich bewertet?

- H_A: Unterschiedliche Designs werden unterschiedlich bewertet
- H_N: Das Design hat keinen Einfluss auf die Bewertung

94 Heimsch et al, 2018, S. 166
95 Cohens d kann auch Werte über 1 annehmen, ohne dass sich jemand daran stört.

Es wurden n =200 Proband:innen befragt, welches Smartphone-Design ihnen am besten gefällt:

	Design A	Design B	Design C	Design D	Summe
Stichprobenverteilung	42	38	55	65	200

Tabelle 20

Wenn die Nullhypothese gilt, gibt es keinen Unterschied und alle Designs werden gleich oft gewählt, also würden dann in jedem der 4 Felder bei N = 200 als Erwartungswert die Zahl 50 stehen. Die Prüfgrößen ergeben sich aus der quadrierten Differenz der Stichprobenwerte mit dem Erwartungswert der Nullhypothese, geteilt durch den Erwartungswert, also beispielsweise im ersten Feld zu $\frac{(42-50)^2}{50} = 1,28$.

Wir ergänzen die Tabelle um Erwartungswerte und Prüfgrößen:

	Design A	Design B	Design C	Design D	Summe
Stichprobenverteilung	42	38	55	65	200
Erwartungswert der Nullhypothese	50	50	50	50	200
Prüfgrößen	**1,28**	**2,88**	**0,50**	**4,50**	

Tabelle 21

Der Chi-Quadrat-Wert ist wiederum die Summe der Prüfgrößen:

$\chi^2 = 1,28 + 2,88 + 0,5 + 4,5 = 9,16$.

Da es im Anpassungstest immer nur eine Zeile gibt, werden die Freiheitsgrade nur durch die Spaltenanzahl bestimmt und sind f = m − 1 = 4 − 1 = 3. Laut Tabelle ist für f = 3 der kritische Chi-Quadrat-Wert 7,81 und damit ist 9,16 > 7,81, die Alternativhypothese, dass ein nicht zufälliger Unterschied in

der Häufigkeitsverteilung besteht, kann angenommen werden. Der genaue p-Wert, laut *Excel* CHIQU.VERT.RE (9,16;3), ist p = .027.

Wo nun der Unterschied liegt, kann durch auffällig hohe Prüfwerte bestimmt werden. Zu beachten ist, dass die Prüfgrößen immer positiv sind. Ob ein hoher Wert für einen Ausreißer nach oben oder nach unten steht, ist anhand der Daten zu prüfen. Die Betrachtung der einzelnen Prüfgrößen zeigt hier einen auffällig hohen Wert für das Design D, dieses wird signifikant häufiger bevorzugt. Der hohe Wert bei Design B steht hingegen für eine geringe Häufigkeit.

Da es nur eine echte Stichprobenzeile gibt, ist hier Cramers v nicht definiert, so dass wir mit Cohens ω bzw. der identischen Phi-Korrelation für die Effektstärke arbeiten:

$$\omega = \sqrt{\frac{\chi^2}{N}}$$

Hier beträgt $\omega = \sqrt{\frac{9,16}{200}} = 0,21$ (schwacher Zusammenhang).

Der z-Wert zur Bestimmung einer Teststärke kann wie folgt geschätzt werden:

$$z = \omega_{\text{krit}} \times \sqrt{n} \times f^{-0,16} - 1,96$$

Der optimale Stichprobenumfang ist (für Alpha 5 % und Power 80 %):

$$n_{\text{opt}} = \frac{7,85 \times f^{0,31}}{\omega^2}$$

und die Sensitivität ist:

$$\omega_{\text{Grenz}} = \frac{2,8}{\sqrt{n} \times f^{-0,16}}$$

Die Berechnungstools finden Sie hier:

◆ **Download 11** | Chi-Quadrat-Anpassungstest (Service_53241_07.xls)
◆ **Download 12** | Stichprobenumfangsplanung (Service_53241_02.xls)

Auch der vorgestellte 1-Stichproben-Anteilstest zum Anteil der Frauen im BWL-Studium lässt sich nun alternativ und mit gleichem Resultat als Chi-Quadrat-Anpassungstest durchführen:

	Frauen	Männer	Summe
Stichprobenverteilung	30	20	50
Erwartungswert der Nullhypothese	25	25	50
Prüfgrößen	**1,00**	**1,00**	

Tabelle 22

Der Chi-Quadrat-Wert ist $\chi^2 = 1 + 1 = 2$ und *Excel* liefert für den einseitigen Test mit einem Freiheitsgrad den schon bekannten p-Wert: CHIQU.VERT.RE (2;1)/2 = 0,079.

Die Effektstärke ist $\omega = \sqrt{\frac{2}{50}} = 0,2$. Cohens ω entspricht im Anpassungstest mit 4 Feldern das dort benutzte Cohens d bzw. der Korrelation.

8 Korrelationsanalyse bei metrischen Merkmalen

8.1 Punktdiagramm und Trendlinie

Geht es darum, zwei Zahlenreihen, also metrisch messbare Merkmale, auf einen Zusammenhang zu testen, kommt die **Korrelationsanalyse** zum Einsatz. Bei Auswertungen aus Umfragen ist es eher selten, dass beide Merkmale metrisch sind: Bei der Analyse quantitativen Datenmaterials ist es aber ein häufig anzutreffender Fall.

Beispiele solcher Zusammenhangshypothesen sind: „Je höher der Preis, desto geringer die Nachfrage“ oder: „Steigt der *Dow Jones*, so steigt auch der *DAX*.“ Dabei müssen die Einheiten der Zahlenreihen nicht identisch sein. So können Sie auch beispielsweise den Punktestand von Fußballvereinen einer Liga auf den Zusammenhang mit dem Sportetat in Euro untersuchen.

Die Zusammenhänge der Werte kann man sich optisch schon mal vorab in einem **Punktdiagramm** anschauen. Das geht mit allen gängigen Statistikprogrammen oder in *Excel*. In *Excel* kann man sich auch eine dazugehörigen „Trendlinie“ (Regressionsgerade) dazu anschauen.

Wir betrachten als Beispiele die n = 4 Wertepaare:

x = 1; y = 4
x = 2; y = 5
x = 3; y = 4
x = 4; y = 8

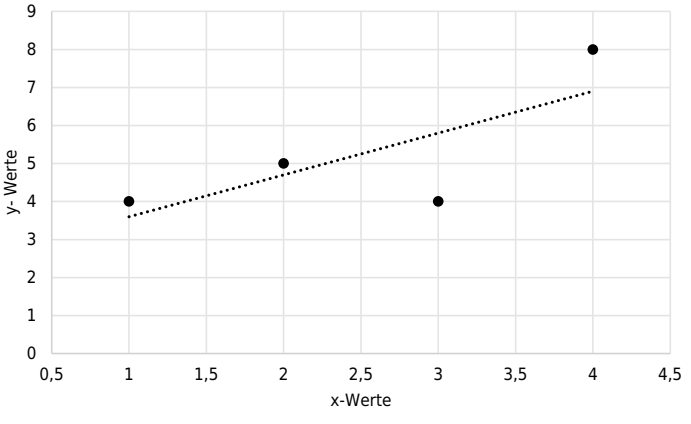

Abbildung 8: Regressionsgrade

Die genaue Gleichung der linearen Trendlinie, die Regressionsgleichung der Form y = a + bx, kann man sich in *Excel* ebenfalls anzeigen lassen. Sie wird so bestimmt, dass die quadrierten Abweichungen der Werte aus der Gleichung von den tatsächlichen Werten minimal werden. Ihre Parameter errechnen sich wie folgt:

$$b = \frac{n \times \sum xy - \sum x \times \sum y}{n \times \sum x^2 - \sum (x)^2} \text{ und}$$

$$a = \bar{y} - b \times \bar{x}$$

wobei $\bar{y}$und $\bar{x}$die Mittelwerte sind, also $\bar{y} = \frac{\sum y}{n}$ und $\bar{x} = \frac{\sum x}{n}$ und führen hier zu y = 2,5 + 1,1x.

Damit kann man nun auch y-Werte schätzen, die zwischen den gegebenen x-Werten liegen. Man sollte aber berücksichtigen, dass die **Regressionsgerade** von den tatsächlichen Werten abweicht. Es gibt also einen Schätzfehler, der mehr oder weniger groß sein kann. Die Gerade ist also nur brauchbar, wenn sie eng an den empirischen Zahlen dran liegt und das heißt, dass sie eine hohe Korrelation haben sollte, dazu später mehr.

Monoton steigende oder fallen Trends, die eher einer progressiven oder degressiven Kurve folgen, können durch Logarithmieren mit dem natürlichen Logarithmus linearisiert werden. Man wandelt die empirischen Werte mit Hilfe der „ln-Taste“ des Taschenrechners in logarithmierte Werte um, der Zusammenhang wird jetzt oft linear sein. Die Trendlinie lautet dann:

$$\ln y = a + b \times \ln x$$

Diese kann dann auf einen linearen Zusammenhang getestet werden und dann bezogen auf die ursprünglichen Werte in eine nicht-lineare Trendkurve des Typus

$$y = a \times x^b$$

umgewandelt werden.[96] Bei einem nicht-monotonen Trend ist eine lineare Regression nicht mehr sinnvoll. Verfahren der nicht-linearen Regression und der multiplen Regression, also Zusammenhänge von mehr als zwei Variablen werden hier nicht behandelt.

96 vgl.: Heimsch et al, 2018, S. 193

8.2 Der Pearson-Korrelationskoeffizient[97]

Bei der Korrelationsanalyse geht man in einer anderen Reihenfolge vor als bei den bisher behandelten Hypothesentests. Es wird hier zunächst mit dem Korrelationskoeffizienten r die Effektgröße bestimmt und erst im Nachhinein auf Signifikanz überprüft.

Für lineare Zusammenhänge und normalverteilte[98] Daten verwendet man üblicherweise den **Pearson-Korrelationskoeffizient**[99]**,** der zwischen –1 und +1 liegt. Er kann wie folgt aus den Wertepaaren berechnet werden:

Formel | Korrelation nach Pearson

$$r = \frac{n \times \sum xy - \sum x \times \sum y}{\sqrt{\left(n \times \sum x^2 - \left(\sum x\right)^2\right) \times \left(n \times \sum y^2 - \left(\sum y\right)^2\right)}}$$

Die Korrelation zeigt, wie nah die lineare Trendlinie an den tatsächlichen Werten liegt. Bei einer Korrelation von 1 liegen alle Wertepaare genau auf einer ansteigenden Trendlinie, bei –1 auf einer fallenden Trendlinie. Eine Korrelation von 0 bedeutet, dass es keinen Zusammenhang zwischen den Variablen gibt.

Im Beispiel mit n = 4 ist die Rechnung dann:

y	x	xy	x^2	y^2
4	1	4	1	16
5	2	10	4	25
4	3	12	9	16
8	4	24	16	64
Σ = 21	Σ = 10	Σ = 58	Σ = 30	Σ = 121

Tabelle 23

97 Auch Bravais-Pearson-Korrelation nach *Auguste Bravais* (1811–1863) und *Karl Pearson* (1857–1936)

98 Tests auf Normalverteilung findet man beispielsweis bei: Cleff, 2019, S. 217ff.

99 vgl.: Heimsch et al, 2018, S. 171

$$r = \frac{4 \times 58 - 10 \times 30}{\sqrt{(4 \times 30 - 100) \times (4 \times 121 - 441)}} = 0,75$$

Dies steht für einen starken positiven linearen Zusammenhang zwischen x und y. Ganz einfach geht die Berechnung in *Excel* mit der Funktion **+KORELL**. Bei der Interpretation ist zu beachten, dass die Korrelation selbst kein metrisches Merkmal ist, dass also eine Korrelation von 0,8 nicht doppelt so hoch ist wie eine Korrelation von 0,4; sie ist lediglich höher!

Das gängige, auf Cohen zurückgehende Interpretationsschema für Korrelationen nach schwachen, mittleren und starken Effekten haben wir schon im Kapitel „Effektstärken" vorgestellt. Es hat sich vorrangig in den Sozialwissenschaften und in der Psychologie bewährt, wo man es eher mit „weichen" Daten, wie auf Rankingskalen abgefragten Einstellungen, zu tun hat. Bei anderen Anwendungsgebieten mit „harten" Zahlenmaterial wie Preisen oder Kursen sind die Korrelation generell höher und die Einteilung erscheint dann unangemessen niedrig. Alternativ empfiehlt es sich hier, erst ab 0,25 von einem kleinen, ab 0,5 von einem mittleren und ab 0,75 von einem großen Effekt zu sprechen.

> Anschaulicher zu interpretieren ist das **Bestimmtheitsmaß B**, das durch Quadrieren von r entsteht: $B = r^2$. Es gibt an, zu welchem Anteil x an der Veränderung von y beteiligt ist. Hier beträgt es $0,75^2 = 0,5625$. Die Veränderung von y ist zu 56,25 % durch die Veränderung von x erklärbar. Wenn z. B. der Punktestand einer Bundesligamannschaft zum Etat ein Bestimmtheitsmaß von 0,7 aufweist, dann bedeutet dies, dass 70 % der Punktausbeute durch den eingesetzten Etat erklärt werden kann.

8.3 Signifikanz von Korrelationen

Rein intuitiv haben Sie sicher im vorangestellten Beispiel ihre Zweifel gehabt, ob eine Korrelation von r = 0,75, die auf nur 4 Wertepaare, beruht signifikant sein kann. Die Signifikanz können wir auf zwei Arten prüfen:

Signifikanztest durch Konfidenzintervalle

Man gibt das Intervall für ein 95-prozentiges Vertrauensniveau an. Man benötig dazu den Korrelationskoeffizienten der Stichprobe r und die Anzahl

der Wertepaar n. Ein Zusammenhang ist signifikant, wenn beide Intervallgrenzen das gleiche Vorzeichen haben, der Wert 0 also nicht dazugehört. Die Berechnung geht in folgenden Schritten:[100]

Schritt 1: Die Umwandlung von r in einen Z-Wert (Fishers Z-Transformation[101])

Dabei ist:

$$\mathbf{Z} = 0,5\ln\left(\frac{1+\mathbf{r}}{1-\mathbf{r}}\right)$$

Im Beispiel ist das:

$$Z = 0,5\ln\left(\frac{1,75}{0,25}\right) = 0,97$$

Schritt 2: Es wird der Standardfehler für Fishers Z berechnet

$$\boldsymbol{\sigma z = 1n - 3\sigma z = \sqrt{\frac{1}{n-3}}}$$

Im Beispiel ist das:

$$\sigma z = \sqrt{\frac{1}{4-3}} = 1$$

Schritt 3: Konfidenzintervall von Z bestimmen

$$\mathbf{INV(Z) = Z \pm 1,96 \times \sigma z}$$

Im Beispiel ist das:
0,97±1,96:[−0,99;2,93]

100 vgl.: Heimsch et al, 2018, S. 187ff.
101 Nach *Ronald Almer Fisher* (1890–1962)

Schritt 4: Rücktransformation der beiden Z Werte in r Werte

$$\mathbf{r} = \frac{e^{2Z} - 1}{e^{2Z} + 1}$$

Im Beispiel ist das:

$$-\frac{0,86}{1,14} = -0,75 \text{ und } \frac{349,7}{351,7} = 0,99 : [-0,75; 0,99]$$

Die „wahre" Korrelation liegt demnach mit hoher Wahrscheinlichkeit im Intervall zwischen **–0,75 und 0,99**, kann also praktisch fast überall und auch 0 sein. Der Zusammenhang ist nicht signifikant, da die Intervallgrenzen unterschiedliche Vorzeichen haben und die enthaltene 0 ja für keinen Zusammenhang steht.

Signifikanztest als t-Test

Man kann die Signifikanz auch über einen t-Test herausfinden, der empirische t-Wert muss über den kritischen t-Wert1,96 liegen. Er berechnet sich:

Formel | t-Wert bei Korrelationsanalysen

$$t_{emp} = \frac{r \times \sqrt{n-2}}{\sqrt{1-r^2}}$$

Im Beispiel ist das:

$$t_{emp} = \frac{0,75 \times \sqrt{4-2}}{\sqrt{1-0,75^2}} = 1,6 < 1,96$$

Über die schon bekannte *Excel*-Funktion kann bei Bedarf der t-Wert auch wieder in einen p-Wert umgewandelt werden. Die Freiheitsgrade dazu sind n − 2. Der p-Wert liegt dann bei p = 0,25 im zweiseitigen Test. Wenn von der Forschungsfragen her klar ist, dass ein Zusammenhang nur positiv oder negativ sein kann, so kann auch mit 1,65 statt 1,96 einseitig getestet werden.

8.4 Teststärke und optimaler Stichprobenumfang

Bei nicht signifikanten Korrelationen kann die Nullhypothese nur angenommen werden, wenn für eine vorgegebene Mindestkorrelation die Teststärke größer 80 % ist. Die Teststärke (Power) ist wieder die Standardnormalverteilung über einen z-Wert mit $z_{krit} = 0{,}84$:

Formel | Zwischenwert zur Teststärkeberechnung bei Korrelationsanalysen

$$z = \sqrt{n \times Z^2 - Z^2} - 1{,}96$$

Daraus lässt sich auch ein optimaler Stichprobenumfang für einen zweiseitigen Test mit Alpha-Fehler 5 % und 80 % Teststärke berechnen:

Formel | Optimaler Stichprobenumfang bei Korrelationsanalysen

$$n_{opt} = \frac{(1{,}96 + 0{,}84)^2 + Z^2}{Z^2}$$

Wenn Sie es nachrechnen, sehen Sie, dass für eine erwartete Korrelation von 0,75 (Z = 0,97) schon 10 Werte ausgereicht hätten, aber eben nicht 4. Allerdings haben in der Praxis die meisten Zusammenhänge niedrigere Korrelationen. Da sie dies vorab nicht wissen, sollten Sie eher konservativ, also niedrig, schätzen.

Für einen einseitigen Test kann in beiden Formeln wieder 1,96 durch 1,65 ersetzt werden.

Beispiele für optimale Stichprobengrößen für Alpha 5 % und Teststärke 80 % sind:

R	Z	$n_{opt.}$ zweiseitig	$n_{opt.}$ Einseitig
0,1	0,1	785	621
0,3	0,31	83	66
0,5	0,55	27	22

Tabelle 24

8.5 Mittelwertunterschiede von Korrelationen

Nehmen wir an, Sie haben ermittelt, dass die BMW-Aktie im letzten Jahr (n_1 = 250) mit 0,8 zum DAX korreliert. In den ersten Monaten des neuen Jahres (n_2 = 90) nur noch mit 0,7. Um die durchschnittliche Korrelation des gesamten Untersuchungszeitraumes zu ermitteln, muss zunächst die Umwandlung der Korrelation in die Z-Werte erfolgen. Diese betragen hier 1,10 und 0,86. Daraus wird dann, gewichtet mit den Stichprobengrößen, zunächst der Mittelwert der Z-Werte ermittelt:

$$Z_{\text{Mittel}} = \frac{1,1 \times 250 + 0,86 \times 90}{340} = 1,036$$

Die mittlere Korrelation ergibt sich dann durch Rücktransformation des mittleren Z-Werts und führt zu

$$r = \frac{6,94}{8,94} = 0,776.$$

Auch lässt sich überprüfen, ob der Unterschied der Korrelationen signifikant ist. Dazu ist wieder ein t-Test möglich mit folgendem t-Wert.

$$t = \frac{|Z_1 - Z_2|}{\sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}}$$

Der führt hier zu 0,24/0,1246 = 1,93 und da dies kleiner ist als 1,96 ist der Korrelationsunterschied der beiden Zeiträume (knapp) nicht signifikant. Es kann also auch nur zufällig anders sein.

Nutzen Sie zu den Berechnungen gerne auch die *Excel*-Tools:

◆ **Download 13** | Korrelationsanalyse (Service_53241_08.xls)
◆ **Download 14** | Korrelationsvergleich (Service_53241_09.xls)

8.6 Korrelation und Kausalität

Korrelationen werden benutzt, um kausale Zusammenhänge aufzuspüren. Der Korrelationskoeffizient sagt jedoch nichts über Ursache und Wirkung aus. Bildet man die Korrelation der Variablen A und B, so ist es rein rechnerisch gleich, ob A von B oder B von A abhängt. Für den Korrelationskoeffizienten gilt:

$r_{AB} = r_{BA}$

> Der bzw. die Forscher:in muss also theoretisch begründen, ob A aus B folgt oder umgekehrt. Häufig ergibt sich dies aus zeitlichen Abläufen oder aus dem „gesunden Menschenverstand".

Manchmal bedingen sich auch beide Variabel gegenseitig. Wenn es eine positive Korrelation zwischen Zufriedenheit der Mitarbeiter:innen und Umsatz gibt, führt dann steigende Zufriedenheit zu mehr Umsatz oder führt ein höherer Umsatz zu höherer Zufriedenheit der Mitarbeiter:innen? Oder kann nicht auch beides gleichzeitig richtig sein? Solche Kausalfragen können durch die Korrelationen nicht beatwortet werden.

> Korrelationen lassen sich immer berechnen, aber nicht jede statistische Korrelation steht zwingend für Kausalität. Eine dritte Variable z, die ursächlich für Veränderungen von x **und** y ist, kann zu einer Korrelation von x und y führen, ohne dass für x mit y eine kausale Beziehung vorliegt.

So kann der Umsatz von Speiseeis (x) mit der Hautkrebsrate (y) korrelieren, obwohl man vom Eis essen sicher keinen Hautkrebs bekommt, es also keine kausale Beziehung gibt. Hier gibt es eine dritte Variable z, die Anzahl der Sonnentage, und diese ist ursächlich für beide Beobachtungen.

Durch die Berechnung von **Partialkorrelationen** lassen sich bisweilen solche Scheinkausalitäten aufdecken. Dazu ein Beispiel:

Die Korrelation Eis/Hautkrebs sei: $r_{xy} = 0{,}75$

Die Korrelation Eis/Sonne sei: $r_{xz} = 0{,}9$

Die Korrelation Hautkrebs/Sonne sei: $r_{yz} = 0{,}8$

Die Partialkorrelation von Eis/Hautkrebs ohne Einfluss der Sonne $r_{xy/Z}$, also bei unveränderten Sonnentagen ist nun:[102]

Formel | Partialkorrelation

$$r_{xy/z} = \frac{r_{xy} - r_{yz} \times r_{xz}}{\sqrt{\left(1 - r_{yz}^2\right) \times \left(1 - r_{xz}^2\right)}}$$

Daraus errechnet sich im Beispiel:

$$r_{xy/z} = \frac{0,75 - 0,8 \times 0,9}{\sqrt{(1 - 0,64) \times (1 - 0,81)}} = 0,11$$

Die Korrelation ist also ohne Einfluss der Sonnentage nur noch sehr gering und je nach Anzahl der beobachteten Werte auch nicht signifikant.

> Es liegen oft nicht die nötigen Korrelationen vor, um Partialkorrelationen berechnen zu können. Jedoch sollte man dann trotzdem auf mögliche **Drittvariablen** hinweisen, um Fehlschlüsse zu vermeiden. Ein kausaler Zusammenhang sollte nur angenommen werden, wenn neben einer relevanten und statistisch signifikanten Korrelation auch eine **theoretische Begründung** dafür gegeben werden kann, bzw. eine Korrelationsanalyse sollte nicht „ins Blaue" erfolgen, sondern nur, wenn diese theoretisch begründbar ist.

102 vgl.: Heimsch et al, 2018, S. 180

8.7 Rangkorrelation nach Spearman

Wenn die Annahme der Normalverteilung der Daten nicht gegeben ist oder in mindestens einer Datenreihe nur ordinale Werte, also Rangfolgen vorliegen, ist der **Rangkorrelationskoeffizient** nach Spearman[103], genannt **Spearmans Rho**, hier R geschrieben, zu verwenden. Das wäre beispielsweise der Fall, wenn wir den Sportetat einer Fußballmannschaft (metrisches Merkmal) nicht mit den erzielten Punkten, sondern mit dem Tabellenstand am Saisonende (ordinales Merkmal) in Bezug setzten wollen. Dabei werden die Daten beider Datenreihen in Ränge umgewandelt, (sofern sie nicht ohnehin als Rangdaten vorliegen) und die quadrierten Differenzen aus den Rängen (d^2) dienen zur Berechnung von R nach folgender Formel:

Formel | Rangkorrelation Rho nach Spearman

$$R = 1 - \frac{6\sum d^2}{n(n^2 - 1)}$$

Wir nehmen dazu das schon bekannte Zahlenbeispiel. Bei gleichen Werten nehmen wir den mittleren Rang für beide Werte:

x	y	Rang x	Rang y	d Rang x– Rang y	d^2
1	4	4	3,5	0,5	0,25
2	5	3	2	1	1
3	4	2	3,5	–1,5	2,25
4	8	1	1	0	0

Tabelle 25

103 nach *Charles Spearman* (1863–1945)

Mit

$$\sum d^2 = 0,25 + 1 + 2,25 = 3,5\,.$$

Spearmans Rho ist dann:

$$R = 1 - \frac{6 \times 3,5}{n(16-1)} = 0,65$$

Spearmans Rho ist „gröber“ als die Pearson-Korrelation. Sie sollten diese nur dann benutzen, wenn die Voraussetzungen für die Pearson-Korrelation nicht erfüllbar sind. Im Beispiel mit der Fußballliga wäre es besser, statt des Tabellenranges das metrische Merkmal der Punktzahl heranzuziehen, da es zwischen zwei Tabellenplätze große oder kleine Leistungsunterschiede bei den Punkten geben kann, das würde aber bei der Rangkorrelation vernachlässigt werden.

Die Signifikanz kann hier genau wie bei der Pearson-Korrelation getestet werden. Auch die Berechnung von Teststärke und optimaler Stichprobengröße ist identisch.

9 Spezielle Befragungen

9.1 Expertenbefragung und Delphi-Befragung

> Eine Besonderheit der Befragung ist die **Expertenbefragung**. Hierbei werden Expert:innen zu der Fragestellung befragt, um z. B. Einschätzungen, Entwicklungen oder Trends zu erfassen. Die Ergebnisse spiegeln dann die Meinung vieler Expert:innen wider und können von Unternehmen bei der Prognose von Entwicklungen genutzt werden. Denkbar ist auch die Durchführung in einem Unternehmen, um die Perspektiven und Erfahrungen unterschiedlicher Fachabteilungen hier zu nutzen.

In dem Kontext wird häufig die sogenannte **Delphi-Befragung** angewendet. Das *Orakel von Delphi* war ein Ort des antiken Griechenlands, an dem Ratsuchende an wenigen Tagen im Jahr einen Orakelspruch bekamen, um eine zukünftige Entscheidung besser zu treffen.

> Die Delphi-Methode ist eine **strukturierte Befragung in mehreren Stufen**. So werden mit dem jeweiligen Thema affine Personen bzw. Expert:innen befragt, um zu Themen ein breites Meinungsbild zu erfassen und dieses Meinungsbild dann wiederum bei den Befragten zu spiegeln. Dadurch kann ein gesamtheitliches, eventuell auch konsensorientiertes Meinungsbild abgeleitet werden. Jedoch wird hier die Kritik angeführt, dass sich bei einer Konsensbildung eher die Einschätzung der Expert:innen durchsetzt, die besonders von ihrer Meinung überzeugt sind.[104]

Delphi-Befragungen eignen sich für unterschiedliche Fragestellungen. So können mit dieser Methode Entwicklungen und Trends in Wissenschaft, Forschung und Zukunftsszenarien untersucht werden. Eine häufige Anwendung ist konkret die Abschätzung von Technologien und deren Wirkung.[105]

---

104 vgl.: Döring & Bortz, 2016, S. 420f.; Häder, 2014, S. 24f.

105 vgl.: Häder, 2021; Niederberger & Deckert, 2022; Göpfert, 2022

Die Delphi-Methode wird auch im **Projektmanagement** eingesetzt, um Aufwände und Entwicklungszeiten richtig einzuschätzen. So werden Personen aus verschiedenen Disziplinen befragt, um möglichst ein komplementäres Bild zu den Fragestellungen zu bekommen.[106]

Auch in der **Medizin** wird die Delphi-Methode eingesetzt, u. a. bei Erstellung von Leitlinien zur Behandlung von bestimmten Krankheiten. Hier wird damit die Diversität der Behandlungsmöglichkeiten berücksichtigt.[107]

Die Befragung mehrerer themenaffinen Personen und Experten aus unterschiedlichen Disziplinen führt zu einem umfassenden Bild und damit auch zu realistischen Einschätzungen, z. B. bei der Entwicklung neuer Produkte oder Aufwandsschätzung bei Projekten.

> Eine durchgehende digitale Durchführung der Befragung ermöglicht, eine zeitnahe Ergebnispräsentation gegenüber den Befragten sowie einer entsprechenden schnellen Rückkopplung. Wenn Delphi-Befragungen bzw. die Stufen in einem bestimmten Zeitfenster durchgeführt werden, kann von einer **Realtime-Delphi-Befragung** gesprochen werden.[108]

Ein Problem könnte bei der Befragung sein, dass Personen bewusst die eigene Einschätzung nicht bzw. falsch abgeben, da diese einen Wissensvorsprung behalten wollen. Es wird auch die Frage diskutiert, ob Expert:innen überhaupt bessere Prognosen abgeben können, als dies Nicht-Expert:innen tun.

Auch konzentrieren sich Expert:innen auf ihr Fachgebiet bzw. auf ihre Perspektive und damit werden andere entfernte Fachgebiete und deren Verbindung unzureichend beachtet. Gerade bei gesellschaftlichen Entwicklungen im Bereich Technologie wird oft die Geschwindigkeit, die sogenannte Diffusionsgeschwindigkeit von Innovationen[109], als zu schnell eingeschätzt.

Häufig werden Delphi-Befragungen per E-Mail bzw. online durchgeführt. Bei der Durchführung in Präsenz, also in einer Gruppe, entsteht auch eine soziale Gruppe. Hier können Autoritäten oder Dominanzen einzelner Personen die Einschätzungen und eine Konsensorientierung entsprechend beeinflussen.[110]

---

106 vgl.: von der Gracht & Kisgen, 2022
107 vgl.: Sforzini et al, 2022
108 vgl.: Gerhold, 2019
109 vgl.: Rogers, E. M. 2003
110 Häder, 2014, S. 171

Wissen | Ablauf einer Delphi-Befragung
Der Ablauf einer Delphi-Befragung kann grob wie folgt beschrieben werden:[111]

1. **Festlegung des Themenbereichs:** Wichtig hierbei ist, den Themenbereich eher größer zu fassen und dementsprechende Fragen abzuleiten, damit möglichst viele Aspekte, Meinungen und Expertenwissen erfasst werden können. Das Ziel der Befragung sollte möglichst viele Sichtweisen und Fachdisziplinen berücksichtigen.
2. **Ableitung der Fragen:** Aus dem Befragungsziel sind entsprechende Fragen abzuleiten. Die Fragen müssen kurz, prägnant und eindeutig sein. Dabei können sowohl geschlossene als auch offene Fragen zum Einsatz kommen. Es bedarf einer Einschätzung in Abhängigkeit der Auskunftsfähigkeit der Befragten und der Fragen, ob offene Fragen zu qualitativen Antworten führen. Bei geschlossenen Fragen ist es sinnvoll, eine Struktur wie eine Likert-Skala zu verwenden. Generell sollten Enthaltungen bzw. Auslassungen von Fragen möglich sein.
3. **Identifikation der Expert:innen bzw. themenaffinen Personen sowie Verteilung der Befragung:** Für die Befragung sind mögliche themenaffine Personen bzw. Expert:innen zu identifizieren. Hierbei ist zu beachten, dass möglichst ein breites Spektrum an Personen mit unterschiedlichem Wissen, Kompetenzen und Ansichten identifiziert wird. Die Diversität der Personen ist ein entscheidender Erfolgsfaktor, um möglichst ein breites und damit vollständiges Bild an Einschätzungen zu bekommen. Eine vorherige Kontaktaufnahme und Avisierung der Delphi-Befragung ist sinnvoll.
   Bei der Verteilung der Befragung an diese Personen ist zu beachten, dass die Personen einen passenden Zeitraum zur Beantwortung haben. Somit kann es sinnvoll sein, den Zeitpunkt der Befragung adressatengerecht zu gestalten.
4. **Auswertung:** Nach dieser ersten Befragung werden die Antworten ausgewertet und aufbereitet. Dabei wird bewusst die Diversität der Antworten aufgezeigt. Bei quantitativen Fragen ist es wichtig, nicht nur den Mittelwert anzugeben, sondern auch die Varianz und die Extremwerte. Bei offenen Fragen werden in der Regel die Antworten

111 Häder & Häder, 2022, S. 922; Häder, 2014, S. 81–84; Döring & Bortz, 2016, S. 420f.

verdichtet, aber es müssen dabei sämtliche Einschätzungen deutlich werden.

5. **Erneute Befragung:** Diese anonymisierten Ergebnisse werden an die Teilnehmer:innen der Befragung gesandt. Das Ziel dieser erneuten Befragung ist eine Einschätzung der Ergebnisse aus der ersten Befragung. Die Befragten kennen die Gruppenergebnisse und können ihre Einschätzung abgeben.
6. **Ergebnis:** Durch diesen mehrstufigen Prozess steigen die Qualität und die Güte der Antworten der Befragten, da sie die Einschätzung aus der ersten Befragung kennen und bei der finalen Antwort berücksichtigen.
7. **Iterativer Prozess:** Häufig wird eine Delphi-Befragung als zweistufige Befragung durchgeführt, um dadurch eine iterative Schleife mit den Gruppenergebnisse aus der ersten Befragung zu ermöglichen. Denkbar ist auch, dass diese Iterationen wiederholt werden, um eine möglichst hohe Konvergenz der Einschätzung zu erreichen oder dass sich die Einschätzungen der einzelnen Befragten nicht mehr ändert.

Mit der Delphi-Methode können Einschätzungen und Prognosen von verschiedenen Expert:innen zusammengeführt werden. Durch den mehrstufigen Charakter ist es möglich, die Aussagegüte der Befragungsergebnisse zu erhöhen. Dadurch können Schätzungen besser vorgenommen werden. Bei mehreren Stufen ist eine Konsensfindung denkbar.[112]

112 Häder & Häder, 2022, S. 922

9.2 Conjoint-Analyse

Definition | Conjoint-Analyse
Die Conjoint-Analyse ist ein Verfahren der multivariaten Analysemethoden und stellt ein dekompositionelles Verfahren dar, welches die Präferenzen von Personen über verschiedene Alternativen (Stimuli) erfasst. Diese Stimuli setzen sich aus verschiedenen Merkmalen zusammen. Jede Person nimmt hierbei eine individuelle Bewertung vor, somit wird dieses Verfahren auch **Individualanalyse** genannt.

Gerade bei der Beurteilung von Varianten von Produkten einhergehend mit deren Preisgestaltung wird die Conjoint-Analyse genutzt, um den Nutzen der verschiedenen Varianten von Produkten zu beurteilen. Dieser setzt sich additiv aus den Nutzenbeiträgen der einzelnen Leistungskomponenten zusammen. Diese Komponenten sollten weitestgehend unabhängig voneinander sein, um in einer „kompensatorischen Beziehung" zueinander zu stehen. Die Wichtigkeit der einzelnen Leistungskomponenten wird dabei in Abhängigkeit zueinander bewertet. Diese Komponenten bzw. Eigenschaften müssen präferenzrelevant sein, d. h. diese müssen die Kaufentscheidung beeinflussen, und müssen von der bzw. dem Durchführenden der Befragung beeinflussbar sein.[113]

> Somit wird ein Set verschiedener Produktfunktionalitäten bzw. Leistungskomponenten quantitativ bewertet, um zu einer Beurteilung des Gesamtnutzens eines Produkts zu kommen. Nicht die Beurteilung einzelner Funktionalitäten ist entscheidend, sondern die Zusammensetzung der Funktionalitäten, die das Produkt ausmachen. Somit führt die Abwägung des Gesamtnutzens zur Kaufentscheidung und nicht die Beurteilung einziger Leistungskomponenten.[114]

Um den Nutzen zu beurteilen, werden den Proband:innen Produkte mit sämtlichen Ausprägungen dargestellt, z. B. auch Bestandteile im Bereich Service, und der bzw. die Proband:in hat dann diese Produkte mit unterschiedlichen Ausprägungen in eine Rangfolge zu bringen.[115]

113 Keller, S. 94ff.
114 Keller, S. 94ff.
115 Fiedler et al, 2017, S. 1–2

Anwendung

Ein wesentlicher Vorteil der Conjoint-Analyse ist, dass die Entscheidungen der Proband:innen einer realen Entscheidungssituation nachgestellt sind und so der Realität sehr nahekommen. Wie in der Realität müssen vollständige Produkte bewertet werden. Eine Bewertung der einzelnen Funktionalitäten bzw. Komponenten erfolgt nicht.[116]

Im Bereich der **Produktentwicklung** kann die Conjoint-Analyse angewandt werden, um neue oder weiterentwickelte Produkte von der anvisierten Zielgruppe beurteilen zu lassen. Hier wird untersucht, welche Produkteigenschaften welchen Nutzen für die Kund:innen generieren und damit einhergehend auch, welche Funktionalitäten nicht nachgefragt werden. Somit kann schon während der Produktentwicklung der Fokus auf die Eigenschaften mit entsprechenden Nutzen für die Kundinnen fokussiert werden und ggfs. werden dementsprechend auch bestimmte Funktionalitäten nicht entwickelt. [117]

> Analog zu dem Nutzen für die Kund:innen kann somit auch die Preisbereitschaft der verschiedenen Produkte analysiert werden. Mit den verschiedenen **Preisbereitschaften** kann dann eine Preis-Absatz-Funktion abgeleitet werden. Diese kann dann wiederum für die betriebliche Entscheidungen, z. B. Gewinnung von Kund:innen oder Gewinnmaximierung, genutzt werden.

Ergänzend können auch Produkte von Wettbewerbern sowie eigene Produkte durch die Conjoint-Analyse getestet werden. Dadurch wird eine Rangfolge der Produkte in dem Marktsegment generiert und das Unternehmen kann dadurch Aussagen zu Marktanteilen der Produkte treffen.

Vorgehensweise und Beispiel

Anhand des folgenden Beispiels wird die Funktionsweise der Conjoint-Analyse vorgestellt.

116 Krämer, Burgartz, 2022, S.90f.
117 Krämer, Burgartz, 2022, S. 93

Beispiel | Abo-Modelle eines Fitness-Studios

Eigenschaften	Eigenschaftsausprägungen
Nutzungszeit	vormittag, abends, ganztägig, ein Tag pro Woche
Nutzungsbereich	nur Kurse, nur Geräte, ohne Einschränkung
Laufzeit	6 Monate, 12 Monate, 24 Monate
Preis	20 Euro, 40 Euro, 50 Euro, 60 Euro

Tabelle 26

Es stehen also 4 Eigenschaften zur Verfügung, 2 davon mit je 3 Ausprägungen, 2 davon mit 4 Ausprägungen. Insgesamt gibt es also $3 \times 3 \times 4 \times 4 = 144$ mögliche Produkte. Diese zusammengesetzten Produkte werden durch die Proband:innen in eine Reihenfolge gebracht.

Bei der Durchführung dieser Conjoint-Analyse werden den Proband:innen nur die möglichen und relevanten Kombinationen präsentiert. In dem obigen Beispiel würde damit ganztägig, ohne Einschränkung, 6 Monate für 20 € kein sinnvolles Produkt sein. Ein Erfahrungswert ist, dass maximal 10–20 Alternativen getestet werden. Je weniger Alternativen in Frage kommen, umso geringer ist der Aufwand bei der Durchführung und der Vergleiche auf Seite der Proband:innen. Es sind nur die Alternativen (Stimuli) zu testen, die auch existieren bzw. geplant sind.

In dem obigen Beispiel sollen nun die möglichen Alternativen getestet werden, nachfolgend ein Ausschnitt:

Produkt	Nutzungszeit	Nutzungsbereich	Laufzeit	Preis
A	vormittags	ohne Einschränkung	6	60
B	vormittags	nur Kurse	24	40
C	ein Tag	nur Geräte	24	20
D	abends	ohne Einschränkung	12	60
E	ganztägig	ohne Einschränkung	24	50
…	…	…	…	…

Tabelle 27: Beispiel Conjoint-Analyse

In einem weiteren Schritt werden die Alternativen meist paarweise gegenübergestellt, wie „Welches Produkt präferieren Sie: Produkt A oder Produkt C?“ Es liegt auf der Hand, dass möglichst viele Vergleichspaare von den Proband:innen verglichen werden, damit eine Präferenz zwischen den verschiedenen Alternativen abzulesen ist. Die Alternativen müssen nach der Bewertung in eine Reihung gebracht werden.

In dem Kontext ist die **Nutzenfunktion** der Proband:innen entscheidend. So gibt es beispielsweise Eigenschaften, die eine nutzenmaximierende Eigenschaftsausprägung (**Idealpunktmodell**) haben, z. B. Kofferraumgröße bei einem Auto oder Größe einer Verpackung eines Lebensmittels.

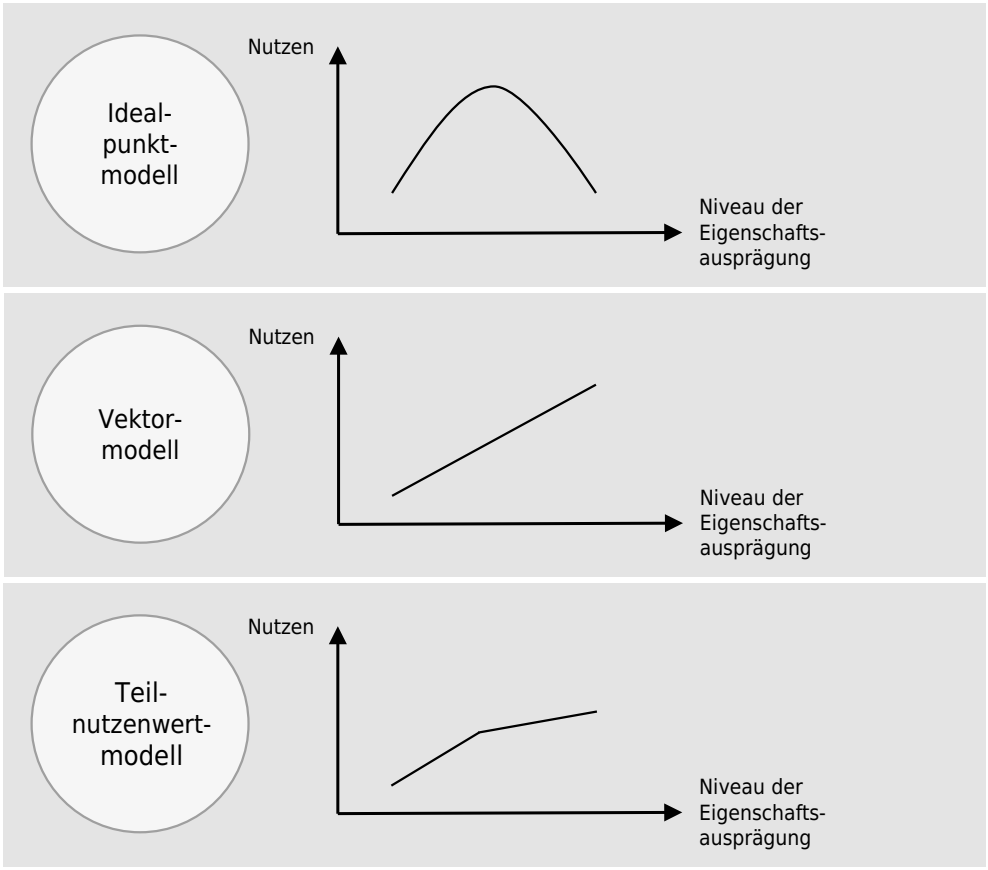

Abbildung 9: Modelle der Nutzenfunktion

Der Nutzen einer Eigenschaftsausprägung steigt bzw. fällt proportional mit der Menge bzw. Intensität dieser Ausprägung (**Vektormodell**). Ein typisches Beispiel für diese vektormodellbasierte Nutzenfunktion ist der Preis.

> Das **Teilnutzenwertmodell** ermittelt für die jeweilige Eigenschaftsausprägung einen sogenannten Teilnutzenwert. Aufgrund der Unterschiedlichkeit der Ausprägungen der Eigenschaft und damit des Nutzens ist dieses Modell bei Conjoint-Analysen am häufigsten anzutreffen.[118]

In den einzelnen Teilnutzenmodellen ist dann die Ausprägung des Nutzens zu erkennen. Für die Errechnung dieser Teilnutzenwerte gibt es verschie-

118 vgl.: Fiedler et al, 2017, S. 36ff. und 74–76

dene multivariate statistische Methoden. In dem Kontext kann eine Regressionsanalyse angewandt werden, in der Regel werden eher Schätzverfahren wie Hierarchical Bayes[119] oder die Maximum-Likelihood-Methode[120] angewendet. So sind diese u. a. im Rahmen der Berechnung der Conjoint-Analyse in einigen Statistik-Anwendungssystemen integriert und es existieren spezielle Software-Produkte in dem Bereich. An dieser Stelle wird auf diese konkrete Umsetzung dieser Berechnungsmöglichkeiten nicht näher eingegangen.

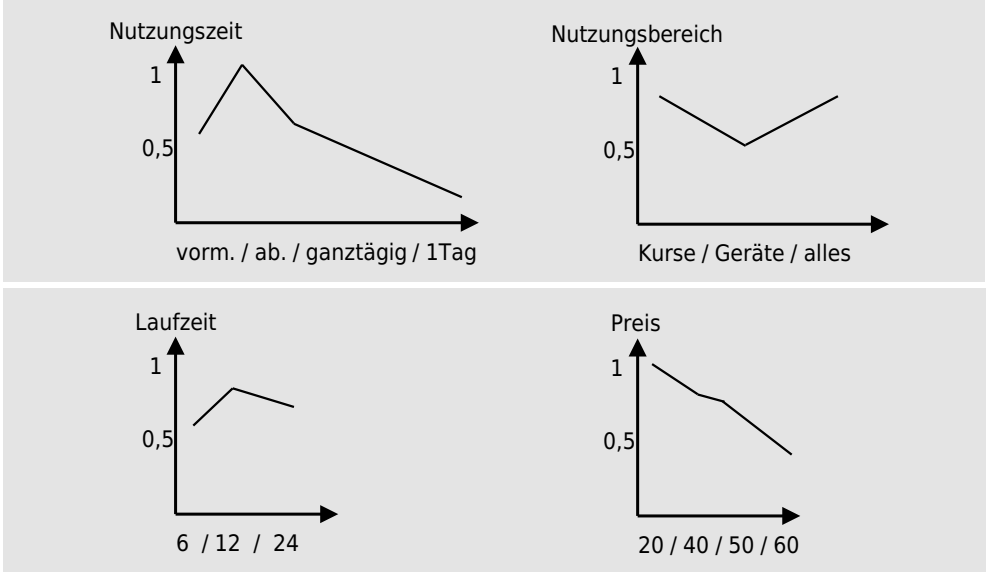

Abbildung 10: Beispiel Nutzen pro Eigenschaft des Fitnessstudios

Hieraus leitet sich dann der Bereich des Nutzens pro Eigenschaftsausprägung ab. Der Anteil des Nutzens wird prozentual angegeben.

119 vgl.: Fiedler et al, 2017, S. 44
120 vgl.: Döring & Bortz, 2016, S. 637ff.

Eigenschaftsausprägung	Nutzenbereich	Nutzenanteil
Nutzungszeit	0,31–0,98 = 0,67	44,30 %
Nutzungsbereich	0,69–0,97 = 0,28	18,50 %
Laufzeit	0,53–0,79 = 0,26	17,20 %
Preis	0,98–0,68 = 0,30	19,90 %

Tabelle 28: Nutzenbereich und Nutzenanteil pro Eigenschaft des Fitnessstudios

Die **Gesamtnutzenanteil** ist 100 % und setzt sich aus dem Beitrag der Teilnutzenwerte zusammen. Die Eigenschaftsausprägung mit dem höchsten Nutzenanteil hat die größte Differenzierungsrelevanz.

Damit ist es möglich, auf Basis der Eigenschaftsausprägungen und deren Gewichtung am **Gesamtnutzen** (basierend auf den individuellen Teilnutzenfunktionen) bestehende potenzielle Produkte im Rahmen der Produktweiterentwicklung zu bewerten. Dies kann zu Eigenschaftsbündeln für den Nutzen für die Zielgruppen konzipiert bzw. angepasst werden.

Ausgehend von den obigen möglichen Alternativen ergeben sich somit folgende Gesamtnutzenwerte für die einzelnen Produktvarianten A, B und C in unserem Beispiel.

Produkt	Nutzungszeit	Nutzungs-bereich	Lauf-zeit	Preis	Gesamt-nutzen
	44,30 %	*18,50 %*	*17,20 %*	*19,90 %*	
A	vormittags	ohne Ein-schränkung	6	60	
	0,51	0,97	0,53	0,68	**0,63186**
B	vormittags	nur Kurse	24	40	
	0,51	0,8	0,61	0,8	**0,63805**
C	ein Tag	nur Geräte	24	20	
	0,31	0,69	0,61	0,98	**0,56492**

Tabelle 29: Nutzenwerte für Produktvarianten des Beispiels Fitnessstudio

Bei dem Vergleich hat das Produkt B den größten Gesamtnutzen gegenüber den anderen Produkten.

Diese Ergebnisse können für die obigen Anwendungsbereiche genutzt werden. Bei der Entscheidung beispielsweise für die Produktauswahl und Preisfestsetzung sind auch gleiche oder vergleichbare Produkte der Wettbewerber miteinzubeziehen. Die bzw. der Konsument:in vergleicht bei der Kaufentscheidung das Eigenschaftsbündel und entsprechende Preise im gesamten Markt.

9.3 Erfolgsfaktorenanalyse

Eine eindimensionale Betrachtung des Wirkungszusammenhanges zwischen einem Faktor und dem Unternehmenserfolg ist nicht ausreichend. Es beeinflussen mehrere Faktoren den Unternehmenserfolg. Ein Faktor muss auch einen relevanten, nachweisbaren, direkten oder indirekten Einfluss auf den Erfolg besitzen. Ferner muss ein Erfolgsfaktor vom Unternehmen gestaltbar sein. Externe **Erfolgsfaktoren**, die einen Einfluss auf den Unternehmenserfolg haben, aber nicht durch das Unternehmen beeinflussbar sind, werden in der vorliegenden Erfolgsfaktorenanalyse nicht betrachtet. Hier können jedoch aufgrund der Wirkung externer Faktoren unternehmensinterne und damit gestaltbare Erfolgsfaktoren abgeleitet werden.

Das Verständnis über den Begriff „**kritische Erfolgsfaktoren**“ variiert in der Literatur. Erstmalig wurde der Begriff Erfolgsfaktor von *Daniel* (1961) verwendet. Nach Daniel können für die meisten Branchen wenige Faktoren identifiziert werden, die den Unternehmenserfolg determinieren (Daniel, 1961).[121] Der Ansatz von Daniel wurde später von *Rockart* mit folgender Definition zu „kritischen Erfolgsfaktoren“ weiterentwickelt:

> *„Critical success factors thus are, for any business, the limited number of areas in which results, if they are satisfactory, will ensure successful competitive performance for the organization”* (Rockart, 1979).[122]

Durch die Identifikation der kritischen Erfolgsfaktoren wird erreicht, dass unternehmerische Entscheidungen sowie Maßnahmen an den richtigen Stellen, d. h. den erfolgskritischen Punkten, erfolgen.[123]

---

121 vgl.: Daniel, 1961, S. 111–121
122 vgl.: Rockart,1979, S. 81–93
123 vgl.: Geier, 1999

Wissen | Kernelement der Erfolgsfaktorenanalyse
Kernelement dieser kritischen Erfolgsfaktorenanalyse ist die Beurteilung eines Erfolgsfaktors nach Wichtigkeit (Priorität) und Qualität der Umsetzung (Leistung). Erfolgsfaktoren, deren Wichtigkeit hoch, die unternehmensinterne Qualität der Umsetzung allerdings niedrig ist, werden als kritische Erfolgsfaktoren bezeichnet.

Grabowski und *Geiger* definieren kritische Erfolgsfaktoren als solche Faktoren, denen ausgesprochene Schwächen auf Unternehmensseite gegenüberstehen und bei denen sofortiger Handlungsbedarf besteht.[124] Überdies bestehen auch wechselseitigen Interdependenzen der Erfolgsfaktoren. Sie sind gerade bei der Interpretation der Ergebnisse und der Ableitung von Handlungsempfehlungen wichtig.

Zusammenfassung | Erfolgsfaktorenanalyse
Die Bewertung der einzelnen Erfolgsfaktoren findet anhand von Fragebögen von themenaffinen Personen wie Expert:innen oder Mitarbeiter:innen (T) statt. Dabei wird die Priorität P(K), die Leistung L(K) und der Gesamterfolg E(K) des jeweiligen Erfolgsfaktors bewertet. In der Regel wird eine 7-stufige Skala genutzt. So wird die Priorität P(K) von 1 „irrelevant“ bis 7 „sehr wichtig“ sowie die Leistungserfüllung L(K) von 1 „sehr schlecht“ bis 7 „sehr gut“ bewertet. Die Werte werden über die Antworten der Befragten kumuliert.[125]

Die kumulierten Werte drücken aus, ob die kritischen Erfolgsfaktoren, auch oft mit KEF abgekürzt, ausreichend unterstützt sind.

Der Erfolg E(K) eines kritischen Erfolgsfaktors ist somit:

$$E(K) = \frac{\sum_{T=1}^{t}(P(K,T) \times L(K,T))}{\sum_{T=1}^{t} P(K,T)}$$

Der Erfolg ist umso größer, je höher die Priorität und die Leistungserfüllung von den Befragten eingeschätzt wurde.

124 vgl.: Grabowski & Geiger, 1997
125 vgl.: Heinrich & Lehner, 2005, S. 344ff.

Für den bzw. die jeweilige:n Befragte:n (T) kann der Erfolg hinsichtlich der Erfolgsfaktoren wie folgt ausgedrückt werden:

$$E(K) = \frac{\sum_{K=A}^{Z}(P(K,T) \times L(K,T))}{\sum_{K=A}^{Z} P(K,T)}$$

Für die Klassifikation der Erfolgsfaktoren ist die sogenannte **Leistungsdifferenz** zwischen Priorität und Leistungserfüllung relevant.

$$D(K) = \frac{1}{t}\sum_{T=1}^{t} P(K,T) - \frac{1}{t}\sum_{T=1}^{t} L(K,T)$$

> Mit der Leistungsdifferenz wird ausgedrückt, wie sehr ein Erfolgsfaktor hinsichtlich der Priorität und adäquater Leistungserfüllung umgesetzt ist. D(K) liegt bei einer 7-stufigen Skala dementsprechend zwischen –6 und +6.

Das Ergebnis mündet in einer Erfolgsfaktoren-Matrix, die die Erfolgsfaktoren nach Relevanz und Ausprägung bzw. Leistungserfüllung klassifiziert. Somit werden die Faktoren als kritisch, maßgeblich, überbewertet und unwesentlich klassifiziert.

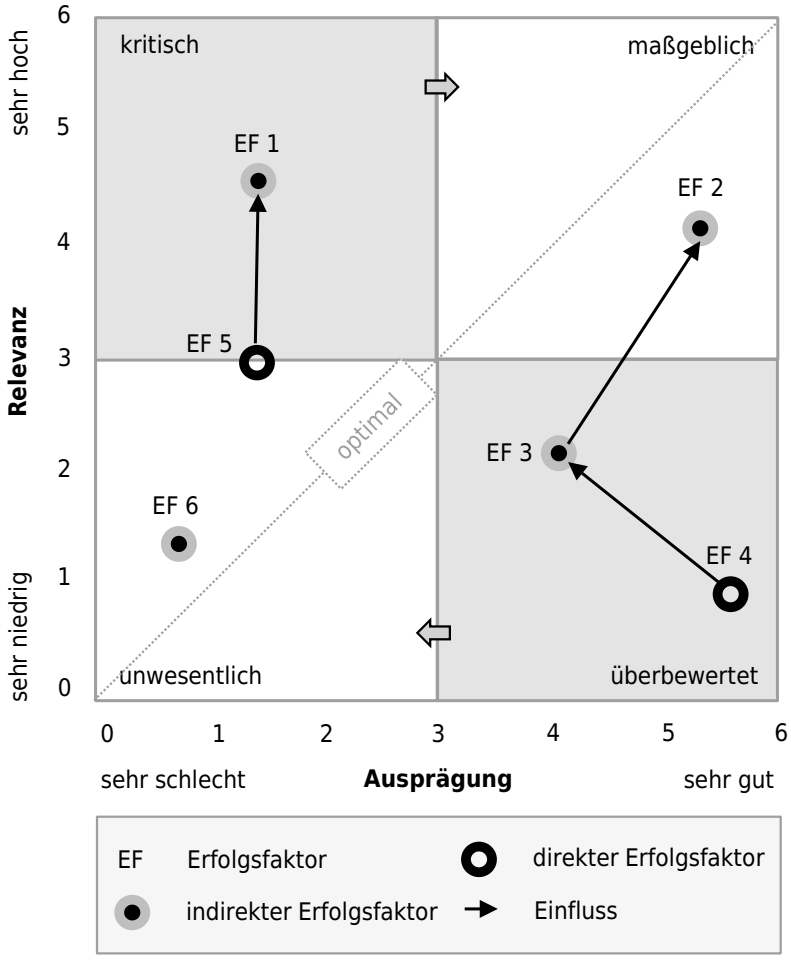

Abbildung 11: Erfolgsfaktoren-Matrix, in Anlehnung an Gausemeier et al.[126]

Wissen | Arten von Erfolgsfaktoren

- **Kritische Erfolgsfaktoren** sind schlecht ausgeprägt und haben eine hohe Relevanz für das Unternehmen. Bei solchen identifizierten Faktoren besteht Handlungsbedarf, da dort die größten Risiken für das Unternehmen zu vermuten sind.
- **Maßgebliche Erfolgsfaktoren** sind gut ausgeprägt und haben eine hohe Relevanz für das Unternehmen. Solche Faktoren sind die

126 vgl.: Gausemeier et al, 2009, S. 138,140

eigentlichen Erfolgsträger für das Unternehmen und sollten in Unternehmen verstetigt werden.
- **Überbewertete Erfolgsfaktoren** sind zwar gut ausgeprägt, aber nicht für das Unternehmen relevant. In diesem Zusammenhang werden möglicherweise interne Ressourcen verschwendet. Ressourcen, die für überbewertete Erfolgsfaktoren beansprucht sind, können anderweitig verwendet werden.
- **Unwesentliche Erfolgsfaktoren** sind weder gut ausgeprägt noch für das Unternehmen relevant. Solche Faktoren erfordern keine besondere Beachtung durch das Unternehmen.

Vorgehen

Für das Vorgehen der Erfolgsfaktorenanalyse orientiert man sich an Methodik der kritischen Erfolgsfaktorenanalyse nach *Heinrich* und *Lehner*.[127]

I. Identifikation der Erfolgsfaktoren

Zur Identifizierung der Erfolgsfaktoren ist es notwendig, die unterschiedlichen Dimensionen des zu untersuchenden Bereichs abzubilden und eine Auflistung von möglichen Faktoren vorzunehmen. Am Beispiel des Vertriebs stellen Organisation, Prozesse, Instrumente, Service sowie Controlling die dimensionalen Ausprägungen einer Vertriebspolitik dar. Eine Zusammenstellung möglicher Faktoren und damit Erfolgsfaktoren erfolgt durch die Analyse in dem Bereich und Marktumfeld. So können in der Regel diese Faktoren aus Studien, Marktberichten und Literaturquellen zusammengetragen werden. Der nächste Schritt ist eine Verdichtung dieser Faktoren. Das Ziel liegt nicht in einem vollständigen Abbild sämtlicher Erfolgsfaktoren, sondern in der Auflistung der relevanten Faktoren.

Ergänzt werden können diese Faktoren auch durch unternehmens- und branchenspezifische Faktoren, um ggfs. Spezifika der Branche abzubilden und die Faktoren zu komplementieren.

> Eine Empfehlung für die Anzahl der Faktoren gibt es nicht. Wenn eine befragte Person 50 oder mehr Faktoren beurteilen soll, kann die

127 vgl.: Heinrich & Lehner, 2005, S 350ff.

> Motivation, Konzentration und damit die Antwortqualität leiden. Es hängt auch davon ab, wie viele Personen für die Befragung gewonnen werden können. So können beispielsweise 50 Faktoren zufällig aufgeteilt werden, so dass nicht jede befragte Person alle Faktoren beantwortet, sondern jeweils nur einen Teil.

II. Festlegen der Teilnehmer:innen an der Befragung

Im zweiten Schritt werden die Teilnehmer:innen der Befragung festgelegt. Heinrich und Lehner empfehlen eine Totalerhebung oder bei größeren Befragungen eine Begrenzung auf ca. 200 Teilnehmer:innen.

> Entscheidend ist, dass die Teilnehmer:innen entsprechendes Vorwissen bzw. Affinität zu dem Thema der Befragungen besitzen.

So sollten bei der Beurteilung der Faktoren im Bereich Vertrieb die Personen befragt werden, die in dem betreffenden Unternehmen entsprechende Vertriebserfahrung haben oder eng mit dem Vertrieb zusammenarbeiten. Wird unternehmensübergreifend die Befragung durchgeführt, so wären der Vertriebsleiter pro Unternehmen in einer Region bzw. Branche zu befragen.

III. Formulieren des Fragebogens

Der Fragebogen gliedert sich in verschiedene Teile. Die Erfolgsfaktoren werden nach der Wichtigkeit und nach der Umsetzung beurteilt. Dabei ist es wichtig, dass diese Fragen getrennt voneinander in dem Fragebogen kommen, damit ein direkter Vergleich oder eine Erinnerung zwischen der Beurteilung nach Priorität und nach Umsetzung nicht bzw. schwer möglich ist.

Wissen | Fragebogenaufbau Erfolgsfaktorenanalyse
Im ersten Teil werden die Erfolgsfaktoren hinsichtlich ihrer Priorität anhand einer Skala bewertet. Die Fragestellung in diesem Teil lautet: „Welche Priorität bzw. Wichtigkeit haben Ihrer Erfahrung nach die folgenden Erfolgsfaktoren für das Ziel …?“. Die Antworten werden mit P(K) bezeichnet und geben die Prioritätseinschätzung des Erfolgsfaktors K wieder.

In einem weiteren Teil des Fragebogens werden die Erfolgsfaktoren in zufälliger Reihenfolge hinsichtlich ihrer Leistung bzw. der Qualität der Umsetzung bewertet. Die Fragestellung kann hier lauten: „Welche Ausprägung bzw. Umsetzung haben Ihrer Erfahrung nach die folgenden Erfolgsfaktoren für das Ziel …?“. Die Antworten werden mit L(K) bezeichnet.
Ein dritter inhaltlicher Teil beinhaltet ergänzende Fragen, die nicht direkt für die Erfolgsfaktorenanalyse notwendig sind. So können die Befragten nach ihrer beruflichen Stellung, der Branchenzugehörigkeit und deren Affinität zu dem Thema, in dem Beispiel Vertrieb, befragt werden.
Im letzten Teil wird um Einschätzung des Gesamterfolgs des Untersuchungsbereichs, in dem Beispiel Vertrieb, anhand der Bewertungsskala L(K) gebeten. Zweck hier ist die Überprüfung der Plausibilität des über die Leistung errechneten Erfolgs.

Um möglichst sicherzustellen, dass der erste und der zweite Teil der Erfolgsfaktorenanalyse so in dem Fragebogen abgebildet ist, damit die Befragten keine Rückschlüsse zwischen den Fragen ziehen können, bietet es sich an, den dritten Teil des Fragebogens zwischen den beiden Teilen zu platzieren.

Als Skala bei den Fragen nach Bedeutung und Leistungserfüllung wird häufig eine 7-stufige-Skala eingesetzt. Vor dem Hintergrund der Benutzerfreundlichkeit und Einfachheit, die Antworten zu markieren, kann auch eine 5-stufige-Skala eingesetzt werden, wenn die Erfolgsfaktoren hinsichtlich Bedeutung und Umsetzung gut differenziert beurteilt werden können.

IV. Durchführen der Datenerhebung

Die Erfolgsfaktorenanalyse kann in Präsenz oder online durchgeführt werden. Die Proband:innen sind über das Ziel, die Methode und den Ablauf zu informieren, damit die Relevanz der Antworten verstanden wird und dadurch auch die Güte der Antworten hoch ist. Ebenso sind Fachbegriffe und das Verständnis zu den Erfolgsfaktoren zu erläutern. Die Befragung sollte bei allen Teilnehmer:innen in einem gleichen Zeitraum durchgeführt werden, damit z. B. bei einer Befragung in einem Unternehmen keine Absprachen möglich sind.

V. Auswerten der Erhebungsdaten und Darstellen der Erhebungsergebnisse

Auf Basis der Antworten werden pro Erfolgsfaktor die Leistungsdifferenzen D(K) gebildet sowie der Gesamterfolg E(K) der jeweiligen Erfolgsfaktoren.

$$E(K) = \frac{\sum_{T=1}^{t}(P(K,T) \times L(K,T))}{\sum_{T=1}^{t} P(K,T)}$$

$$D(K) = \frac{1}{t}\sum_{T=1}^{t} P(K,T) - \frac{1}{t}\sum_{T=1}^{t} L(K,T)$$

„Erfolgsorientierte Manager versuchen, bei allen Erfolgsfaktoren eine Leistung zu erbringen, die ihrer Priorität entspricht."[128]

Die **Leistungsdifferenz** drückt die Notwendigkeit von leistungsverbessernden Maßnahmen aus. Bei negativen Werten wird empfohlen, zu desinvestieren. Bei einer positiven Leistungsdifferenz, d. h. die Wichtigkeit ist höher als die aktuelle Umsetzung, sollte in den Erfolgsfaktor investiert werden.

Die Erfolgsfaktoren werden hinsichtlich Wichtigkeit und Leistungserfüllung in ein Diagramm eingezeichnet. Der Durchschnitt der Wichtigkeiten sowie der Leistungserfüllungen bildet die jeweiligen Achsen.

So gibt es folgende Quadranten:

- **Erfolg:** wichtiger Erfolgsfaktor mit guter Leistung
- **Killer:** wichtiger Erfolgsfaktor mit schlechter Leistung
- **Verschwendung:** unwichtiger Erfolgsfaktor mit guter Leistung
- **OK:** unwichtiger Erfolgsfaktor mit schlechter Leistung

128 vgl.: Heinrich & Lehnert, 2005, S. 349

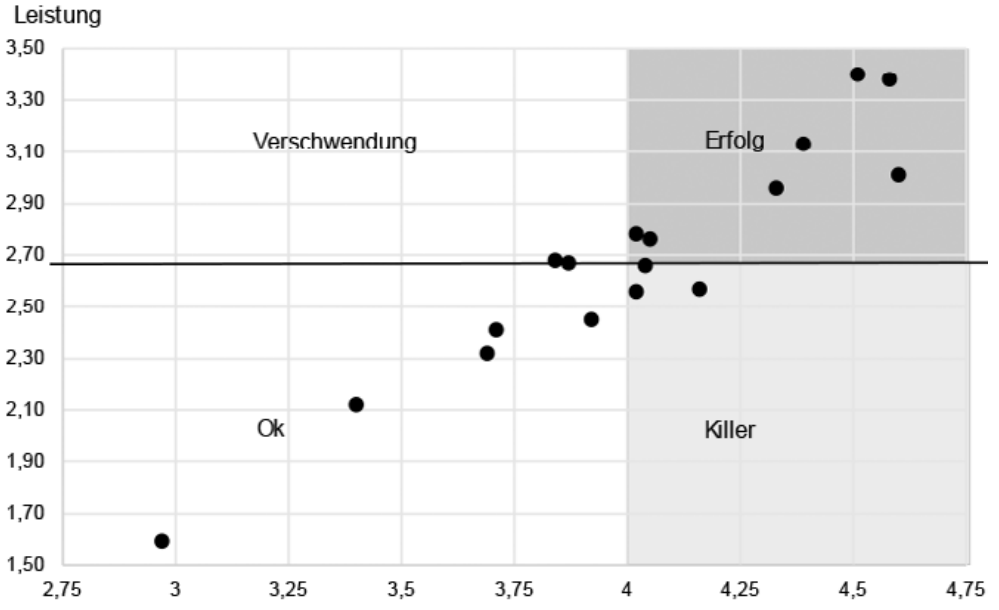

Abbildung 12: Exemplarisches Ergebnis der Erfolgsfaktorenanalyse[129]

Die Achsen können die Skala der Befragung widerspiegeln oder diese können auch nur den Ausschnitt wiedergeben, der relevant ist. Für einen Überblick über alle Erfolgsfaktoren und deren Lage und entsprechende Leistungsdifferenzen ist eine Skala analog der Befragungsskala hilfreich. Für eine Detailbetrachtung jedes Quadranten kann die Skala auf den relevanten Bereich verkürzt werden.

VI. Interpretieren der Erhebungsergebnisse

Die Erfolgsfaktoren, die in dem Bereich „Killer" liegen, haben eine hohe Priorität und eine nicht adäquate Umsetzung. Diese Erfolgsfaktoren sollten prioritär betrachtet und verstärkt umgesetzt werden.

Durch diese Methode wird die Mehrdimensionalität eines Themas aufgezeigt. Sie konkretisiert, welche Faktoren für eine erfolgreiche Umsetzung notwendig sind und gibt konkret Hinweise für Investitions- und Deinvestitionsentscheidungen. So wird eine regelmäßige Wiederholung der Analyse empfohlen (alle zwei bis drei Jahre), um die Wirkungen der erfolgsverbessernden Maßnahmen festzustellen und ggfs. neue und modifizierte Faktoren in die Analyse miteinzubeziehen.

129 vgl.: Heinrich, Riedl, & Stelzer, 2014, S. 375

9.4 Semantisches Differential

Das **semantische Differenzial** ist eine Befragung, um ein **Polaritätenprofil** zu bestimmten Eigenschaften und Wahrnehmungen bei den Befragten zu erheben. Das Verfahren wurde in der Psychologie entwickelt, um die Vorstellungen der Proband:innen mit bestimmten Produkten, Sachverhalten oder Planungen zu verbinden.[130]

> Das semantische Differential eignet sich gut, um die Wahrnehmung von Eigenschaften eines Produktes bei der Zielgruppe herauszufinden. So können die Wahrnehmungen zwischen Produkten, z. B. gegenüber Wettbewerbern, und von Produkten bei verschiedenen Zielgruppen und im Zeitverlauf analysiert werden.

So kann gerade auch das semantische Differential als Befragung vor und nach einer Marketingmaßnahme, z. B. Werbekampagne, angewendet werden, um die Wirkung bei der **Wahrnehmung** differenziert bei der Zielgruppe zu analysieren.

Die Proband:innen beurteilen auf einer Skala die Wahrnehmung eines Produkts. Das Besondere ist, dass für die Beurteilung gegensätzliche Begriffe genutzt werden, z. B. „groß" und „klein", „schön" und „nicht schön". Als **Skala** wird meist eine 5- oder 7-studige Skala verwendet. Die ungeraden Stufen sind hier angebracht, damit der bzw. die Proband:in auch ihre bzw. seine Indifferenz zwischen diesen beiden bipolaren Begriffen ausdrücken kann. Durch die Antworten entstehen dann für jede Frage ein Mittelwert und Streuungsmaß.

An dem Beispiel eines Protein-Riegels soll die Anwendung des semantischen Differenzials verdeutlicht werden. So werden hier Vereinsmitglieder eines Sportvereins vor und nach einer Werbekampagne befragt. Das Ziel dieser affektiv ausgeprägten Kampagne liegt in der Erhöhung der Wahrnehmung und der Steigerung der Affinität zu dem Produkt. Die Fragen können wie folgt lauten:

130 vgl.: Theobald, 2017, S. 58; Döring & Bortz, 2016, S. 276f; Raab, Unger, & Unger, 2009, S. 86–88

Wie finden Sie den Protein-Riegel?	angenehm	unangenehm
Was assoziieren Sie mit dem Protein-Riegel?	dominant	unterlegen
Wie fühlt sich der Protein-Riegel für Sie an?	dynamisch	ruhig

Tabelle 30: Beispiel semantisches Differential

Wissen | Semantisches Differential
Diese drei Arten von Fragen fokussieren auf die drei Ausprägungen beim semantischen Differenzial. So wird in der Valenzdimension nach dem Gefühl bei dem Produkt bzw. Sachverhalt gefragt. Hier wird der Sympathiewert des Produkts herausgefunden. Die Potenzdimension bei der zweiten Frage zielt darauf ab, die Stärke und die Macht, die von dem Begriff bzw. Produkt ausgeht, zu analysieren. Bei der letzten Frage wird die Aktivierungsdimension beschrieben, um die Dynamik und Aktivitäten, die mit dem Produkt verbunden sind, herauszufinden.

Ergänzend zu diesen Valenz-, Potenz- sowie Aktivierungsdimensionen können andere Eigenschaften durch bipolare Eigenschaften abgefragt werden. Das Ergebnis wird dann in der Regel in einem **Polaritätenprofil** dargestellt.

In dem Beispiel kann so erkannt werden, welche Eigenschaften nach der Marketingmaßnahme mehr von den Proband:innen wahrgenommen werden. So stiegen in dem Fall die Sympathiewerte und das Produkt wird eher ruhiger wahrgenommen.

> Diese drei Dimensionen (Valenz, Potenz und Aktivierung) sind wichtig, da diese die unterschiedlichen Affektivitätsausprägungen bzw. Emotionen abbilden. Diese drei Dimensionen werden als **sozioemotionale Grundausstattung**[131] des Menschen beschrieben, da diese unabhängig von Kultur und Sprache sind. Je nach Befragungsobjekt sind die bipolaren Eigenschaften anzupassen. Bei der Anwendung des semantischen Differentials bei der Erhebung vor und nach einer Marketingmaßnahme ist es essenziell, dass die Eigenschaftspaare gleich sind.

131 vgl.: Osgood, 1952; Lewin, 1986, S. 171–172

So existieren auch Produkte, bei denen ein Eigenschaftspaar in den jeweiligen Dimensionen zu finden ist, z. B. eine Valenz- und Aktivierungsdimension bei einer Tablette gegen Kopfschmerzen. In solchen Fällen kann auf allgemeine bipolare Eigenschaften zurückgegriffen werden:

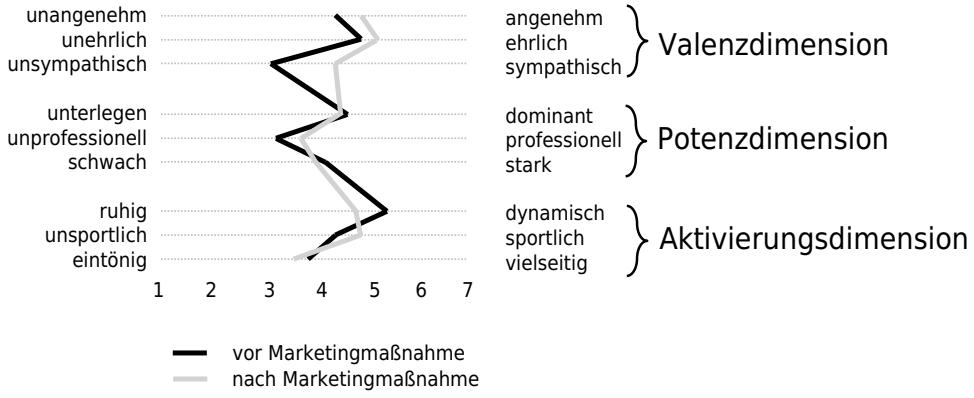

Abbildung 13: Beispiel semantisches Differential

Die Sympathie, die Stärke und auch die Dynamik sind beeinflussende Faktoren für die Wahrnehmung und damit auch die Kaufbereitschaft von Produkten. So kann wie in dem Beispiel analysiert werden, welche Eigenschaften dazu beitragen, die Kaufbereitschaft zu erhöhen. So ist es bei solchen Fragenstellungen sinnvoll, die Korrelation der Valenz-, Potenz- und Aktivierungsdimension mit Fragestellungen der **Kaufbereitschaft** oder der **Weiterempfehlung** zu bilden, um eine Basis für weitergehende betriebswirtschaftliche Entscheidungen zu haben.

Die Vorteile der Methode des semantischen Differentials liegen in der Einfachheit der Formulierung der Fragen sowie der damit verbundenen einfacheren Auswertung der Fragen im Vergleich zu offenen Fragen. Dabei sind die Antworten differenzierter als Fragetypen mit vorgegebenen Antworten oder Ja/Nein-Antworten.

Nachteile entstehen primär durch die bipolaren Eigenschaften, da die Proband:innen die Eigenschaften eventuell nicht richtig verstehen. Auch ist es möglich, dass die Unterschiede zwischen den Antwortmöglichkeiten nicht klar sind.

Anhang

Statistiktabelle für Hypothesentests

t-Wert	p-Wert	p-Wert		z-Wert	Teststärke
					(Power)
	einseitiger Test	zweiseitiger Test			
0,68	25,0 %	50,0 %		−3,00	0 %
0,84	20,0 %	40,0 %		−2,00	2 %
1,04	15,0 %	30,0 %		−1,60	5 %
1,15	12,5 %	25,0 %		−1,30	10 %
1,28	10,0 %	20,0 %		−1,05	15 %
1,44	7,5 %	15,0 %		−0,85	20 %
1,48	7,0 %	14,0 %		−0,52	30 %
1,52	6,5 %	13,0 %		−0,38	35 %
1,56	6,0 %	12,0 %		−0,25	40 %
1,60	5,5 %	11,0 %		0,00	50 %
1,65	**5,0 %**	10,0 %		0,13	55 %
1,70	**4,5 %**	9,0 %		0,26	60 %
1,75	**4,0 %**	8,0 %		0,39	65 %
1,82	**3,5 %**	7,0 %		0,53	70 %
1,88	**3,0 %**	6,0 %		0,69	75 %
1,96	**2,5 %**	**5,0 %**		0,84	**80 %**
2,01	**2,2 %**	**4,5 %**		0,88	**81 %**
2,06	**2,0 %**	**4,0 %**		0,90	**82 %**
2,11	**1,8 %**	**3,5 %**		0,95	**83 %**

2,17	**1,5 %**	**3,0 %**		1,00	**84 %**
2,25	**1,2 %**	**2,5 %**		1,10	**86 %**
2,33	**1,0 %**	**2,0 %**		1,20	**88 %**
2,44	**0,7 %**	**1,5 %**		1,30	**90 %**
2,58	**0,5 %**	**1,0 %**		1,40	**92 %**
2,70	**0,4 %**	**0,7 %**		1,50	**93 %**
2,80	**0,3 %**	**0,5 %**		1,60	**95 %**
2,90	**0,2 %**	**0,4 %**		1,80	**96 %**
3,00	**0,1 %**	**0,3 %**		1,90	**97 %**
3,10	**0,1 %**	**0,2 %**		2,00	**98 %**
3,20	**0,1 %**	**0,1 %**		2,20	**99 %**

Tabelle 31

2-Stichproben-fall	Mittelwerte	Anteilswerte
Signifikanz	$t = \frac{\text{ABW}}{\sqrt{\frac{\sigma_1^2 \times n_1 + \sigma_2^2 \times n_2}{n_1 \times n_2}}}$	$t = \frac{\text{ABW}}{\sqrt{\pi(1-\pi) \times \frac{N}{n_1 \times n_2}}}$
Effektstärke	$d = \frac{\text{ABW}}{\sqrt{\frac{\sigma_1^2 \times n_1 + \sigma_2^2 \times n_2}{N}}}$	$d = \frac{\text{ABW}}{\sqrt{\frac{\pi_1(1-\pi_1) \times n_1 + \pi_2(1-\pi_2) \times n_2}{N}}}$
Teststärke	$z = 0,5 \times d_{\text{krit}} \times \sqrt{\frac{4n_1 \times n_2}{N}} - t_{\text{krit}}$	
optimale Stichproben-größe	$n_{\text{opt}} = \frac{2 \times (t_{\text{krit}} + z_{\text{krit}})^2}{d_{\text{krit}}^2}$	
Sensitivität	$d_{\text{Grenz}} = \frac{2(t_{\text{krit}} + z_{\text{krit}})}{\sqrt{\frac{4 \times n_1 \times n_2}{n_1 + n_2}}}$	

Tabelle 32

1-Stichprobenfall	Mittelwert	Anteilswert
Signifikanz	$t = \frac{ABW \times \sqrt{n}}{\sigma}$	$t = \frac{ABW \times \sqrt{n}}{\sqrt{\pi(1-\pi)}}$
Effektstärke	$d = \frac{\text{ABW}}{\sigma}$	$d = \frac{\text{ABW}}{\sqrt{\pi(1-\pi)}}$
Teststärke	$z = d_{krit} \times \sqrt{n} - t_{krit}$	
optimale Stich-probengröße	$n_{\text{opt}} = \frac{(t_{\text{krit}} + z_{\text{krit}})^2}{d_{\text{krit}}^2}$	
Sensitivität	$d_{\text{Grenz}} = \frac{(t_{\text{krit}} + z_{\text{krit}})}{\sqrt{n}}$	

Tabelle 33

Literatur

Baier, D., & Brusch, M. (2021). *Conjointanalyse Methoden – Anwendungen – Praxisbeispiele.* Berlin: Springer Gabler.

Beckmann, K., Glemser, A., Heckel, C., & al., e. (2016). *Demographische Standards.* Eine gemeinsame Empfehlung des ADM, Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V., der Arbeitsgemeinschaft Sozialwissenschaftilcher Institute e.V. (ASI) und des Statistischen Bundesamtes.

Blasius, Thiessen, (2021), *Argumentieren mit Statistik*, Opladen und Toronto, utb

Böhler, H., Germelmann, C. C., Baier, D., & Woratschek, H. (2021). *Marktforschung.* Kohlhammer Verlag.

Bortz, Lienert, (2008), *Kurzgefasste Statistik für die Klinische Forschung*, Heidelberg, Springer

Bruhn, M. (2022). *Marketing. Grundlagen für Studium und Praxis.* Springer Gabler.

Bühner, Ziegler, (2017), *Statistik und Forschungsmethoden für Psychologen und Sozialwissenschaftler*, Halbergmoss, Pearson

Cleff, (2019), *Angewandte Induktive Statistik und Statistische Testverfahren*, Wiesbaden, Springer Gabler

Daniel, R. D. (1961). Management Information Crisis. *Harvard Business Review*(39 (5)), S. 111–121.

Dempser, Hanna, (2019), *Statistik und Forschungsmethoden für Psychologen und Sozialwissenschaftler für Dummies*, Weinheim, Wiley-VCH

DGP: *Richtlinien der Manuskriptgestaltung*, 2019, Göttingen, Hogrebe,

Döring, N., & Bortz, J. (2016). *Forschungs- methoden und Evaluation* (5. vollständig überarbeitete, aktualisierte und erweiterte Ausg.). Berlin, Heidelberg: Springer.

Eid, Gollwitzer, Schmitt, 2017, *Statistik und Forschungsmethoden*, Basel, Beltz

Föhl, U., & Friedrich, C. (2022). *Quick Guide Onlinefragebogen.* Wiesbaden: Springer Gabler.

Fiedler, H., Kaltenborn, T., Lanwehr, R., & Melles, T. (2017). *Conjoint-analyse.* Rainer Hampp Verlag.

Gausemeier, J., Plass, C., & Wenzelmann, C. (2009). *Zukunftsorientierte unternehmensgestaltung: Strategien, geschäftsprozesse und it-systeme für die produktion von morgen.* Hanser Verlag.

Göpfert, I. (2022). Zukunftsforschung. In *In Logistik der Zukunft-Logistics for the Future* (S. 1–35). Wiesbaden: Springer Gabler.

Geier, C. (1999). Kritische Erfolgsfaktoren und Meßgrößen. *In Optimierung der Informationstechnologie bei BPR-Projekten*, S. 157–164.

Grabowski, H., & Geiger, K. (1997). *Neue Wege zur Produktentwicklung*. Stuttgart u.a: Raabe.

Häder, M. (2014). *Delphi-Befragungen. Ein Arbeitsbuch.* Wiesbaden: Springer Fachmedien.

Häder, M. (2021). Delphi-Analyse. In C. Zerres, *Handbuch Marketing-Controlling* (S. 205–222). Berlin, Heidelberg: Springer Gabler.

Häder, M., & Häder, S. (2022). Delphi-Befragung. In N. Baur, & J. Blasius, *Handbuch Methoden der empirischen Sozialforschung* (S. 921–928). Wiesbaden: Springer VS.

Heimsch, Niederer, Zöfel, (2018), *Statistik im Klartext*, Halbergmoss, Pearson

Heinrich, L. J., Riedl, R., & Stelzer, D. (2014). *Informationsmanagement: Grundlagen, Aufgaben, Methoden.* München: De Gruyter Oldenbourg.

Heinrich, L., & Lehner, F. (2005). *Informationsmanagement Planung, Überwachung und Steuerung der Informationsinfrastruktur* (Bde. 8., vollst. überarb. und erg. Aufl.). München: Oldenbourg.

Hoffmann, S., Franck, A., Schwarz, U., Soyez, K., & Wünschmann, S. (2018). *Marketing-Forschung.* München: Verlag Franz Vahlen.

Homburg, C. (2017). *Marktingmanagement. Strategie – Instrumente – Umsetzung – Unternehmensführung* (Bd. 6). Wiesbaden: Springer Gabler.

Jacob, R., Heinz, A., Décieux, J., & Eirmbter, W. (2012). *Umfrage. Einführung in die Methoden der Umfrageforschung.* München: Oldenbourg Wissenschaftsverlag.

Janczyk, Pfister, (2013), *Inferenzstatistik verstehen*, Berlin, Springer

Keller, K. (kein Datum). Anwendungsbeispiel: Conjoint-Analyse – 10 Mehr als die Summe seiner Teile. In B. Abdel-Karim, *Data Science.* Vieweg, Wiesbaden: Springer.

Krämer, A., & Burgartz, T. (2022). Kernfunktion Kundennutzen: Was sind die tatsächlichen Kundenbedürfnisse? In *Kundenwertzentriertes Management.* Wiesbaden: Springer Gabler.

Kreis, H., Wildner, R., & Kuß, A. (2021). *Marktforschung Datenerhebung und Datenanalyse.* Wiesbaden: Springer Gabler.

Lewin, M. (1986). *Psychologische Forschung im Umriss.* Berlin, Heidelberg: Springerverlag.

Mayer, (2013), *Interview und schriftliche Befragung*, München, Oldenbourg

Meffert, H. (2013). *Marktforschung. Grundriss mit Fallstudien.* Springer-Verlag.

Meffert, H., Burmann, C., & Kirchgeorg, M. (2015). *Marketing. Grundlagen marktorientierter Unternehmensführung. Konzepte – Instrumente – Praxisbeispiele* (Bd. 12). Wiesbaden: Springer Gabler.

Meffert, H., Burmann, C., Kirchgeorg, M., & Eisenbeiß, M. (2019). *Grundlagen marktorientierter Unternehmensführung. Konzepte – Instrumente – Praxisbeispiele.* Wiesbaden : Springer Gabler.

Meyer, H. O. (2013). *Interview und schriftliche Befragung*. München: Oldenbourg Verlag.

Naderer, G., & Balzer, E. (2011). *Qualitative Marktforschung in Theorie und Praxis: Grundlagen, Methoden und Anwendungen*. Springer-Verlag.

Niederberger, M., & Deckert, S. (2022). Das Delphi-Verfahren: Methodik, Varianten und Anwendungsbeispiele. *Zeitschrift für Evidenz, Fortbildung und QUalität im Gesundheitswesen*.

Osgood, C. (1952). The nature and measurement of meaning. *Psychological Bulletin*(49(3)), S. 197–237.

Porst, R. (2014). *Fragebogen. Ein Arbeitsbuch*. Halle (Saale), Nürnberg: Springer VS.

Pusler, M. (2019). *Dem Konsumenten auf der Spur. Erfolgreiches Marketing durch zeitgemäße Marktforschung*. Freiburg, München, Stuttgard: Haufe Group.

Quatember, (2017), *Statistik ohne Angst vor Formeln*, Halbergmoss, Pearson

Raab, G., Unger, A., & Unger, F. (2009). *Methoden der Marketing-Forschung*. Wiesbaden: Gabler.

Raab, G., Unger, A., & Unger, F. (2018). *Methoden der Marketing-Forschung. Grundlagen und Praxisbeispiele*. Wiesbaden: Springer Gabler.

Reichheld, F. (2011). *The ultimate question 2.0 (revised and expanded edition): How net promoter companies thrive in a customer-driven world*. Harvard Business Review Press.

Rockart, J. (Mar-Apr 1979). Chief executives define their own data needs. *Harvard Business Review*(57(2)), S. 81–93.

Rogers, E. M. (2003). *Diffusion of innovations* (5. Aufl.). New York: Free Press.

Schmidt-Atzert, L., & Ameland, M. (2012). *Psychologische Diagnostik*. Berlin, Heidelberg: Springer .

Sedlmeier, Renkewitz, (2018), *Forschungsmethoden und Statistik*, Halbergmoss, Pearson

Sforzini, L., Worrell, C., Kose, M., Anderson, I. M., Aouizerate, B., Arolt, V., & Pariante, C. M. (2022). A Delphi-method-based consensus guideline for definition of treatment-resistant depression for clinical trials. *Molecular psychiatry*(27(3)), S. 128.

Steiner, E., & Benesch, M. (2018). *Der Fragebogen. Von der Forschungsidee zur SPSS-Auswertung*. utb GmbH.

Steiner, E., & Benesch, M. (2021). *Der Fragebogen. Von der Forschungsidee zur SPSS-Auswertung*. facultas bei UTB.

Theobald, A. (2017). *Praxis Online-Marktforschung Grundlagen – Anwendungsbereiche – Durchführung*. Wiesbaden: Springer Gabler.

von der Lippe, P., & Kladroba, A. (2002). Repräsentativität von Stichproben ", 24 (2002), S. 227–238. *Marketing*(24 (1)), S. 227–238.

von der Gracht, H., & Kisgen, S. (2022). METHODEN DER STRATEGISCHEN VORAUSSCHAU. In *Management der Zukunft. SIBE-Edition* (S. 31–78). Berlin, Heidelberg: Springer Gabler.

Westermann, (2000); *Wissenschaftstheorie und Experimentalmethodik*, Göttingen, Hogrebe

Register

Idealer Ratgeber für Haus-, Bachelor- und Masterarbeiten

Bei Haus-, Bachelor- und Masterarbeiten ist die Umfrage eine beliebte Forschungsmethode. Wolfgang Ortmanns und Ralph Sonntag vermitteln dazu alles Wissenswerte – angefangen von den Rahmenbedingungen, den Fragetypen bis hin zum Umfrageaufbau und der Stichprobenauswahl. Wichtiges statistisches Know-how vermitteln sie zudem, u.a. wichtige Testverfahren und die Korrelationsanalyse.

Das Buch richtet sich an Studierende und junge Forschende aus den Bereichen der Wirtschafts- und Sozialwissenschaften.

Gefördert vom Konsortium der sächsischen Hochschulbibliotheken.

Prof. Dr. Wolfgang Ortmanns lehrt Volkswirtschaftslehre und Finanzmärkte an der Hochschule für Technik und Wirtschaft (HTW) in Dresden.

Prof. Dr. Ralph Sonntag lehrt im Bereich Marketing und Existenzgründung und ist an der Hochschule Stralsund tätig.

ISBN 978-3-7398-3241-8

9 783739 832418

www.uvk.de

UVK