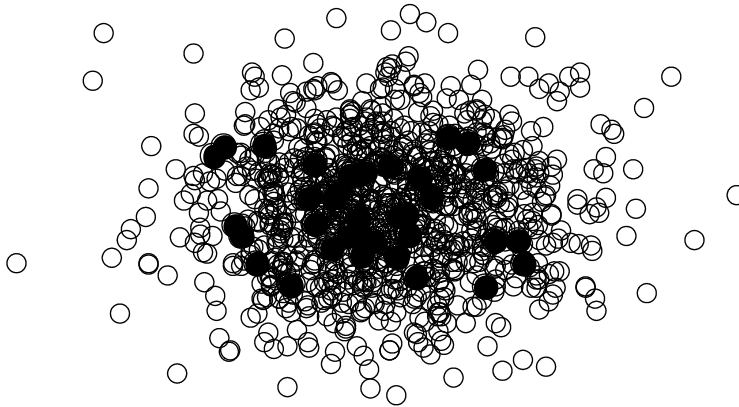


Statistische Methoden der empirischen Sozialforschung



Norbert Nothbaum

2008

Nothbaum GmbH

© 2007 Dr. Norbert Nothbaum

Autor:

Norbert Nothbaum, Jahrgang 1962, Diplom-Psychologe, Dr. rer. nat. Seit 1989 Mitarbeiter der Gesellschaft für Organisation und Entscheidung, Bielefeld, von 1999 bis 2004 auch Gesellschafter. Dort Durchführung verschiedener Projekte im Gesundheitsbereich (Entwicklung von Dokumentationssystemen, Evaluation von Behandlungen und Therapien). Lehraufträge an der Universität Bielefeld zu statistischen Methoden und zu umweltpsychologischen Themen. Wissenschaftlicher Mitarbeiter im Institut für Didaktik der Mathematik (zwischen 1990 und 1994) und an der Fakultät für Gesundheitswissenschaften (1998). Seit 2001 mit der Nothbaum GmbH Entwicklung und Moderation von Internetplattformen.

Inhaltsverzeichnis

1 Vorwort.....	10
1.1 Einführung.....	10
1.2 Aufbau dieses Studientextes.....	12
1.3 Wozu Methoden, wozu Statistik?.....	12
2 Gütekriterien.....	15
Wissensabschnitt 1: Kriterien für eine gute Messung.....	15
Wissensabschnitt 2: Repräsentativität.....	16
Wissensabschnitt 3: Suffizienz.....	18
Wissensabschnitt 4: Indikativität.....	19
2.1 Wie verlässlich ist das Ergebnis der Untersuchung?.....	20
Wissensabschnitt 5: Reliabilität.....	20
Wissensabschnitt 6: Objektivität.....	20
Wissensabschnitt 7: Validität.....	21
2.2 Kann das angestrebte Wissen auch mit geringerem Aufwand gewonnen werden?.....	22
Wissensabschnitt 8: Ökonomie.....	22
2.3 Kriterien für eine gute Studiendurchführung.....	23
Wissensabschnitt 9: Evaluationsstandards.....	23
Übungsaufgabe 1: Gütekriterien.....	26
Übungsaufgabe 2: Studienplanung.....	28
3 Studienplanung.....	29
3.1 Population und Stichprobe.....	29
Wissensabschnitt 10: Population.....	29
Wissensabschnitt 11: Die Größe der Stichprobe.....	30
3.2 Arten wissenschaftlicher Studien.....	30
Wissensabschnitt 12: Experimentelle Studien - Hypothesentestung....	30
Wissensabschnitt 13: Experimentelle Studien - experimentelle Manipulation.....	32

Wissensabschnitt 14: Experimentelle Studien - experimentelle Kontrolle und Störvariablen.....	33
Wissensabschnitt 15: Quasiexperiment.....	34
Wissensabschnitt 16: Korrelative Studien.....	37
Wissensabschnitt 17: Unterschied zwischen korrelativem Zusammenhang und Kausalbeziehung.....	38
Wissensabschnitt 18: Deskriptive und hypothesengenerierende Studien.....	39
Wissensabschnitt 19: Querschnitts- und Längsschnittsstudien.....	40
3.3 Arten epidemiologischer Studien.....	41
Wissensabschnitt 20: Deskriptive epidemiologische Studien.....	41
Wissensabschnitt 21: Analytische epidemiologische Studien.....	42
Wissensabschnitt 22: Interventionsstudien.....	42
Wissensabschnitt 23: Kohortenstudien.....	43
Wissensabschnitt 24: Fall-Kontroll-Studien.....	45
3.4 Evaluation.....	46
Wissensabschnitt 25: Methodische Anforderungen an Evaluationen..	46
Wissensabschnitt 26: Ziele von Evaluation.....	47
Wissensabschnitt 27: Was wird evaluiert?.....	48
Wissensabschnitt 28: Wann wird evaluiert?.....	48
Wissensabschnitt 29: Wo ist die Evaluation angesiedelt?.....	48
Wissensabschnitt 30: Hinweise für eine erfolgreiche Evaluation.....	48
Übungsaufgabe 3: Studientypen.....	50
Übungsaufgabe 4: Kausalität.....	51
Übungsaufgabe 5: Interpretation von Gruppenunterschieden.....	52
Übungsaufgabe 6: Fall-Kontroll-Studien.....	52
Übungsaufgabe 7: Evaluation 1.....	52
Übungsaufgabe 8: Evaluation 2.....	53
4 Datengewinnung: Messen, Fragen, Beobachten.....	54
4.1 Datenquellen.....	54
Wissensabschnitt 31: Konzeptspezifikation.....	54
Wissensabschnitt 32: Operationalisierung.....	55

Wissensabschnitt 33: Indexbildung.....	57
Wissensabschnitt 34: Die Operationalisierung einer Variablen be- stimmt das Skalenniveau der gewonnenen Messwerte.....	57
Wissensabschnitt 35: Objektive Untersuchungsergebnisse.....	58
Wissensabschnitt 36: Befragung von Experten.....	58
Wissensabschnitt 37: Befragung der Teilnehmenden.....	59
Wissensabschnitt 38: Prozessproduzierte Daten.....	60
Wissensabschnitt 39: Überschneidungen der Datenquellen, Nutzung mehrerer Quellen.....	60
4.2 Die Entwicklung eines Messinstrumentes.....	61
Wissensabschnitt 40: Selbst entwickelte Messinstrumente - Überblick	61
Wissensabschnitt 41: Fragebogenentwicklung und Fragebogentestung	62
Wissensabschnitt 42: Item-Arten.....	63
Wissensabschnitt 43: Fragebogaufbau.....	65
Wissensabschnitt 44: Beispielfragebogen SOC-Skala.....	66
Wissensabschnitt 45: Polung von Items.....	67
Wissensabschnitt 46: Fragebogentestung.....	68
Wissensabschnitt 47: Strukturierte Interviews.....	68
Wissensabschnitt 48: Beobachtungsprotokolle, Auswertung von Beobachtungen.....	69
Wissensabschnitt 49: Untersuchungsprotokolle, Experimentenprotokolle.....	70
4.3 Skalierung und Skalenniveau.....	70
Wissensabschnitt 50: Skalierung und Skalenniveau - Überblick.....	70
Wissensabschnitt 51: Rationalskala.....	71
Wissensabschnitt 52: Intervallskala.....	74
Wissensabschnitt 53: Ordinalskala.....	77
Wissensabschnitt 54: Nominalskala.....	80
Wissensabschnitt 55: Das richtige Skalenniveau.....	81
Übungsaufgabe 9: Fragebogenentwicklung.....	81
Übungsaufgabe 10: Item-Polung.....	82

5 Beschreibende Statistik.....	83
Wissensabschnitt 56: Zielsetzung der statistischen Kapitel dieses Studientextes.....	83
Wissensabschnitt 57: Häufigkeiten.....	83
Wissensabschnitt 58: Grafische Darstellung von Häufigkeiten.....	84
Wissensabschnitt 59: Maße der zentralen Tendenz.....	86
Wissensabschnitt 60: Maße für die Streuung.....	89
Wissensabschnitt 61: Kreuztabellen.....	93
Wissensabschnitt 62: Streudiagramme und Korrelationen.....	98
Übungsaufgabe 11: Interpretation von Korrelationen.....	100
Übungsaufgabe 12: Deskriptive Kennzahlen.....	101
6 Wahrscheinlichkeiten.....	102
Wissensabschnitt 63: Wahrscheinlichkeiten, Einführung.....	102
Wissensabschnitt 64: Logische Wahrscheinlichkeiten.....	102
Wissensabschnitt 65: Frequentistische Wahrscheinlichkeiten.....	103
Wissensabschnitt 66: Wahrscheinlichkeiten von komplexen Ereignissen.....	104
Wissensabschnitt 67: Konjunktive Wahrscheinlichkeiten.....	105
Wissensabschnitt 68: Disjunktive Wahrscheinlichkeiten.....	105
Wissensabschnitt 69: Errechnete disjunktive Wahrscheinlichkeiten bei stochastischer Unabhängigkeit.....	107
Wissensabschnitt 70: Bedingte Wahrscheinlichkeiten.....	109
Wissensabschnitt 71: Prävalenz.....	112
Wissensabschnitt 72: Kumulative Inzidenz.....	113
Wissensabschnitt 73: Inzidenzdichte.....	113
Wissensabschnitt 74: Das Basis-Raten-Problem.....	114
Wissensabschnitt 75: Relatives Risiko.....	118
Wissensabschnitt 76: Odds-Ratio.....	119
Übungsaufgabe 13: Wahrscheinlichkeitsrechnung.....	120
Übungsaufgabe 14: Bedingte Wahrscheinlichkeiten.....	120
Übungsaufgabe 15: Prävalenz.....	121
Übungsaufgabe 16: Kumulative Inzidenz.....	123

Übungsaufgabe 17: Relatives Risiko.....	123
Übungsaufgabe 18: Odds-Ratio.....	123
7 Andere beschreibende Parameter für Populationen.....	125
Wissensabschnitt 77: Wahrscheinlichkeitsverteilungen.....	125
Wissensabschnitt 78: Die Normalverteilung.....	126
8 Schließende Statistik.....	129
Wissensabschnitt 79: Von den Stichprobendaten auf die Population schließen.....	129
Wissensabschnitt 80: Die Verteilung von Stichprobenparametern....	131
Wissensabschnitt 81: Konfidenzintervalle.....	136
Wissensabschnitt 82: Die Hypothesen des statistischen Hypothesen- tests.....	139
Wissensabschnitt 83: Beispiel für einen Hypothesentest.....	140
Wissensabschnitt 84: Die Irrtumswahrscheinlichkeit (alpha-Fehler).	141
Wissensabschnitt 85: Der Rechenweg für den Beispiel-Hypothesentest	142
Wissensabschnitt 86: Stichprobenergebnisse und Signifikanzentschei- dung.....	143
Wissensabschnitt 87: alpha-Fehler, beta-Fehler und Power.....	144
Wissensabschnitt 88: Rechenbeispiel für den beta-Fehler.....	148
Wissensabschnitt 89: Effektgrößen.....	149
Wissensabschnitt 90: Beispiel für die Berechnung einer Effektstärke aufgrund der Stichprobendaten.....	149
Übungsaufgabe 19: Hypothesentest.....	150
9 Die Interpretation der Ergebnisse.....	152
Wissensabschnitt 91: Dateninterpretation bereits bei der Auswertung	152
Wissensabschnitt 92: Abschließende Ergebnisbewertung.....	152
Wissensabschnitt 93: Einbeziehung der Beteiligten in die Ergebnisin- terpretation und Information über die Studienergebnisse.....	153

10 Literaturverzeichnis..... 155

11 Glossar..... 157

1 Vorwort

1.1 Einführung

Sir Francis Bacon war englischer Lordkanzler und Verkünder eines wissenschaftlichen Programms zur sorgfältigen Prüfung des Sonderbaren. Er schrieb 1620 in seinem Hauptwerk *Novum Organum*, man müsse den Verstand mit Gewichten beschweren, sonst ziehe der Mensch in seiner Eile allzu leicht die falschen Schlüsse. Dieser Studententext will ein solches Gewicht für Ihren Verstand darstellen.

Weshalb sollen Sie sich überhaupt mit einem methodischen Verstandesgewicht belasten? In einem Forschungsbericht, einem Buch oder einer Zeitschriftenveröffentlichung kann man stets leicht darlegen, warum man die Untersuchung auf eine bestimmte Art durchgeführt hat oder warum man seine Daten auf eine bestimmte Art ausgewertet hat. Falls einem die Ergebnisse von Signifikanztests nicht passen, lässt man sie einfach weg oder berichtet nur den Teil der Daten, welche die Meinung der Verfasser bestmöglich unterstützen. Solche Vorgehensweisen und Tricks sollen Sie in Zukunft besser erkennen können und bei Ihrer Meinungsbildung über den Wert der berichteten Ergebnisse berücksichtigen können.

Für die kritische Lektüre anderer Studien sollten Sie lernen, welche Fragestellungen Sie aus methodischer Sicht an die Darstellung richten sollen, welche Kennzahlen und statistischen Auswertungen Sie erwarten können und welche alternativen Vorgehensweisen möglich und sinnvoll gewesen wären.

Für Ihre eigenen Studien sollen Sie wissen, welche Vorgehensweisen möglich und welche davon für Ihre Fragestellung sinnvoll sind. Welche

Auswertungen Sie von Ihrer Statistikerin¹ anfordern und wie Sie die Kennzahlen verstehen sollten, die Sie erhalten.

Dieser Text will Ihnen einen Überblick über die Methoden der empirischen Sozialforschung geben. Das primäre Ziel dieses Textes ist es, Ihnen eine Vielzahl von Kriterien und Ansatzpunkten an die Hand zu geben, um fremde Studien und Veröffentlichungen kritisch zu lesen und um eigene Studien in Zusammenarbeit mit Methodikerinnen optimal zu planen, durchzuführen und auswerten zu können.

Dieser Text wird Ihnen keine Methodik- oder Statistikausbildung ersetzen, sondern denjenigen das methodische Grundwissen vermitteln, die empirische Texte kritisch lesen und eigene Studien in Kooperation mit Methodikern durchführen wollen. Was Sie hier lernen können reicht aber keinesfalls aus, um selbst eine gute methodische Planung oder eine angemessene statistische Datenauswertung durchführen zu können. Es gibt eine Vielzahl von guten Statistikbüchern, denen wir kein weiteres hinzufügen wollten. Hinweise auf weiterführende Literatur finden Sie in Kapitel 10 („Literaturverzeichnis“, S. 155).

Um diese Grundrichtung konsequent umzusetzen, haben wir in diesem Studententext fast vollständig auf die Wiedergabe von Formeln verzichtet. Vielleicht hilft dies auch denjenigen, die aus ihrer bisherigen Schul- und Studienerfahrung auf Mathematik und Formeln mit instinktiver Abwehr reagieren, den vorliegenden Text interessant und anregend zu finden. Methoden und Statistik gelten vielen Studierenden als notwendiges Übel bei der Ausbildung. Es wäre schön, wenn dieser Text zu der Erkenntnis beitragen könnte, dass eine gute methodische Grundlage entscheidend dabei hilft, alle

¹ Es ist inzwischen selbstverständlich, einen Text geschlechtsneutral zu formulieren, wenn beide Geschlechter gemeint sind. Dies in der Praxis umzusetzen ist allerdings manchmal schwierig, weil sich in der deutschen Sprache nur unter Verwendung langatmiger Wiederholungen („Die Teilnehmerinnen und Teilnehmer eines Experiments bezeichnet man als Versuchspersonen“) oder durch Kunstgriffe wie dem Binnen-I („Die TeilnehmerInnen eines Experiments bezeichnet man als Versuchspersonen“) geschlechtsneutral formulieren lässt. Da in einem Statistik-Buch der Zufall durchaus eine gewisse Rolle spielen darf, habe ich mich in diesem Text aber für ein anderes Vorgehen entschieden: Bei allen Sätzen, bei denen eine geschlechtsneutrale Formulierung notwendig erschien, weil mit einem männlichen Substantiv auf eine Gruppe verwiesen wurde, die sowohl weibliche als auch männliche Personen umfassen kann, habe ich eine Münze geworfen. Je nach Ergebnis des Münzwurfs habe ich im Text dann die weibliche oder die männliche Form gewählt. In der Summe sollte dies zu einer ausgewogenen Nennung von Frauen oder Männern führen, so dass ich hoffe, diesen Text insgesamt geschlechtsneutral formuliert zu haben.

weiteren Studieninhalte besser zu verstehen und kritisch einordnen zu können.

Dieser Text wurde 2002/2003 geschrieben und 2008 sorgfältig überarbeitet. Es ist also bereits einiges an Feedback von früheren Leserinnen eingeflossen. Dennoch bleibt für den Autor noch einiges zu lernen:

- x Welche Teile sind unverständlich?
- x Wo sind noch unnötig komplizierte Erklärungen?
- x Wo sind inhaltliche Fehler oder unzulässige Vereinfachungen?
- x Was finden Sie zu ausführlich?
- x Was fehlt Ihnen?

Es wäre schön, wenn wir im Arbeitsforum zum Studientext in der Worksphere eine rege Diskussion führen könnten, einerseits, um Sie beim Verständnis dieses Studientextes zu unterstützen, andererseits, um mir zu helfen, den Studientext für die folgenden Jahrgänge zu verbessern.

1.2 Aufbau dieses Studientextes

Der Studientext hat fünf zentrale Kapitel, die jeweils unterschiedliche Inhalte behandeln. Die Kapitelreihenfolge spiegelt die Reihenfolge wieder, in der man sich über diese Themen klar werden sollte, wenn man selbst eine Studie durchführt. Beim Lesen müssen Sie sich nicht unbedingt an die Reihenfolge im Text halten, sondern können entsprechend Ihren Interessen oder Ihrem Vorwissen springen oder sich eine eigene Reihenfolge überlegen.

Wir haben die Wissensinhalte in kurze Wissensabschnitte untergliedert und diese durchgehend nummeriert. Vielleicht erleichtert Ihnen dies die Orientierung im Text. Falls Sie sich für eine eigene Lesereihenfolge entscheiden, können Sie im Inhaltsverzeichnis diejenigen Wissensabschnitte markieren, die Sie bereits gelesen (und verstanden) haben.

1.3 Wozu Methoden, wozu Statistik?

Methodische Fragen durchziehen Ihre gesamte Tätigkeit in den Sozialwissenschaften. In Ihrem weiteren Studium und auch in Ihrer praktischen Tätigkeit sind Sie ständig damit befasst, empirische Informationen aufzunehmen und in Ihr eigenes Fachwissen einzuordnen. Auch kommt es vor, dass Sie selbst empirische Informationen über Ihre eigenen Projekte und Ihre eigenen Themen berichten und veröffentlichen.

Die Wissenschaft funktioniert als geregelter Diskurs zwischen den Beteiligten. Was eine hinreichende Mehrheit (oder auch eine einflussreiche Minderheit) für wahr halten, gilt als gesicherte wissenschaftliche Erkenntnis.

Auf diese Art ist die wissenschaftliche Wahrheit in ständiger Entwicklung. Alte Wahrheiten werden hinterfragt und neu bewertet, neue Erkenntnisse in dieses System integriert.

Es hat sich als allgemein anerkanntes Vorgehen etabliert, dass neue empirische Informationen nur dann in diesen Diskurs aufgenommen werden, wenn Sie entsprechend der formalen Kriterien der Methodik gewonnen und ausgewertet wurden. Dabei gilt, dass die empirischen Methoden ebenso wie alle wissenschaftlichen Erkenntnisse in einem ständigen Diskurs und einer kontinuierlichen Weiterentwicklung stehen.

Wollen Sie einen Artikel oder ein Buch veröffentlichen, ist es notwendig, die Gutachterinnen (und später auch die Leserinnen) davon zu überzeugen, dass die vorgestellten empirischen Angaben auf eine methodisch angemessene Art gewonnen und überprüft wurden. Wenn Sie einen Vortrag halten, gilt das gleiche. Aber auch, wenn Sie in einer Organisation eine Neuerung einführen wollen, ist es sinnvoll, Ihre Anregungen mit empirischen Daten zu begründen. Können Sie diese auf eine methodisch überzeugende Art darstellen, so gewinnt Ihre Argumentation das Vertrauen, dass unsere Gesellschaft allgemein wissenschaftlichen Erkenntnissen beimisst².

Wenn Sie selbst Informationen aus der Literatur benötigen, können Sie Ihr methodisches Wissen einsetzen, um die angebotenen empirischen Informationen zu bewerten und zu gewichten. Nicht alle wissenschaftlichen Veröffentlichungen begründen Ihre Schlussfolgerungen auf eine methodisch überzeugende Art. Wenn Sie ein solides methodisches Grundwissen haben, können Sie dies dazu verwenden, zwischen vertrauenswürdiger und unzureichend belegter Information zu unterscheiden.

Daten oder Veröffentlichungen, die diesem methodischen Kanon *nicht* unterworfen wurden, werden nicht als wissenschaftliche Veröffentlichungen angesehen. So kann man bereits vieles aus dem wissenschaftlichen Diskurs ausschließen³. Dies bedeutet nicht, dass alles, was methodisch einwandfrei ist, auch relevant und von wissenschaftlicher Bedeutung ist. Diese Bewertung kann Ihnen die Methodik nicht abnehmen.

Es gibt wertvolle und überzeugende Vorgehensweisen, die nicht der wissenschaftlichen Methodik entsprechen. Eine intuitive Datensammlung oder

² Entscheiden Sie sich für andere (nicht wissenschaftliche) Bezugssysteme in Ihrer Argumentation und verwenden deren methodische Ansätze, so können Sie nur auf das Vertrauen setzen, dass Ihre Rezipientinnen diesen Systemen entgegenbringen.

³ Solche Daten oder Veröffentlichungen können natürlich Anlass einer wissenschaftlichen Studie sein, die es dann unternimmt, diese neuen Ansätze in den wissenschaftlichen Diskurs einzubringen.

ein individueller, subjektiver und möglicherweise auch emotional gefärbter Erfahrungsbericht können für die eigene Tätigkeit sehr wichtig sein. Aber sie sind nicht wissenschaftlich.

Die These, die diesem Studientext zugrunde liegt, lautet, dass wissenschaftliche Arbeit notwendigerweise eine angemessene Kenntnis der Methodik voraussetzt und dass auch die Nutzung der wissenschaftlichen Arbeit anderer ohne methodische Grundkenntnisse nicht möglich ist.

Wie alles in der Wissenschaft, muss auch die Methodik inhaltlich selbst überzeugen und sollte von Ihnen nicht als ein Dogma angesehen werden. Ein jeder übernehme für die eigene Praxis nur das, was ihm überzeugend und plausibel erscheint. So entwickelt sich auch die Methodik weiter.

Eine sorgfältige Beschäftigung mit den Grundkonzepten der Methodik sollte jeder inhaltlichen Arbeit vorangehen, weil nur diese Sie befähigt, in den wissenschaftlichen Diskurs einzutreten. Möglicherweise entwickelt sich so sogar ein fortdauerndes Interesse an methodischen Fragen. Gute methodische Kenntnisse führen dazu, dass man einen schärferen und kritischeren Blick auf die Inhalte eines Fachgebietes haben kann und häufig sogar zu neuen Ideen und Fragestellungen findet, die einem ohne dieses methodische Rüstzeug verschlossen geblieben wären.

2 Gütekriterien

In diesem Kapitel werden wir eine Vielzahl von Aspekten vorstellen, die einen Einfluss auf die Güte einer Studie, einer Messung oder einer wissenschaftlichen Angabe haben. Um fremde Studien kritisch einordnen zu können und um bei eigenen Studien eine hohe wissenschaftliche Qualität sicherstellen zu können, ist es notwendig, die verschiedenen Aspekte wissenschaftlicher Güte zu kennen. Sie sollten in der Lage sein, vorgegebene Studien oder Studienpläne hinsichtlich dieser Aspekte kritisch einschätzen zu können.

In diesem Kapitel stellen wir die Gütekriterien überblicksartig vor. Im Rest des Studientextes werden wir uns dann damit befassen, was man machen kann, um Studien durchzuführen, die diesen Kriterien entsprechen. Zunächst geht es um sieben klassische Gütekriterien, die in der empirischen Methodenlehre entwickelt und diskutiert wurden. Dies sind

1. Repräsentativität, Suffizienz, Objektivität, Indikativität
(Kann mit der geplanten Untersuchung das angestrebte Wissen überhaupt gewonnen werden?)
2. Reliabilität, Validität
(Wie verlässlich ist das Ergebnis der Untersuchung) und
3. Ökonomie
(Kann das angestrebte Wissen auch mit geringerem Aufwand gewonnen werden?)

Daran anschließend befassen wir uns mit einer deutlich umfangreicheren Liste von Kriterien, die für die Durchführung von Evaluationen entwickelt wurden. Diese Kriterien legen einen stärkeren Schwerpunkt auf die Qualität der *Studiendurchführung* als die klassischen Gütekriterien, die primär auf die Qualität der gewonnenen *Daten* ausgerichtet sind. Insofern erscheinen die Program Evaluation Standards als eine gute und praxisorientierte Ergänzung zu den klassischen Maßen.

Wissensabschnitt 1: Kriterien für eine gute Messung

Die Güte einer wissenschaftlichen Angabe kann nicht auf einer einzigen Dimension beschrieben werden, etwa, indem man Schulnoten vergibt. Vielmehr gibt es unterschiedliche Güteaspekte, die von einer wissenschaftliche Angabe mehr oder weniger gut erfüllt werden können. Zum Teil hängen diese Aspekte miteinander zusammen. Beispielsweise führt eine nicht reprä-

sentative Stichprobe zu einer nicht validen Aussage über die Population. Zum Teil stehen die Kriterien aber auch zueinander in Widerspruch. So führt beispielsweise die Erhöhung des Stichprobenumfangs zu einer größeren Suffizienz (die Studie lässt eine ausreichende Genauigkeit der Messung erwarten), aber gleichzeitig auch zu einer geringeren Ökonomie der Studie (sie wird zu teuer). Es ist deshalb nicht zu erwarten, dass eine Studie alle Gütekriterien optimal erfüllt. Vielmehr muss deutlich werden,

- dass einerseits angesichts der jeweiligen Anforderungen und Schwierigkeiten, die sich aus der Fragestellung und den praktischen Rahmenbedingungen der Studie ergeben, ein bestmögliches Ergebnis erzielt wurde, und
- andererseits bei den Güteaspekten, die zueinander in Konkurrenz stehen, ein sinnvoller Ausgleich gefunden wurde.

Hinzu kommt, dass die hier vorgestellten Gütekriterien aus verschiedenen Bereichen wissenschaftlicher Forschung zusammengestellt wurden. So stammen Objektivität, Reliabilität und Validität aus der klassischen Testtheorie. Repräsentativität und Suffizienz sind primär Konstrukte zu Beschreibung der Güte einer Stichprobenziehung. Ökonomie bezieht sich auf wirtschaftliche und andere Effizienzaspekte einer Studie.

Wie häufig, wenn man Konzepte aus verschiedenen Bereichen zusammenstellt, kommt es auch hier zu Überschneidungen und Ähnlichkeiten zwischen verschiedenen Konzepten. Es geht im Folgenden also weniger darum, die verschiedenen Konzepte scharf voneinander zu trennen, als vielmehr, die unterschiedlichen Kernbedeutungen zu erfassen und einen Eindruck über das gesamte Feld der wissenschaftlichen Gütekriterien zu erhalten, das von allen Konzepten zusammen hoffentlich weitgehend vollständig abgedeckt wird.

Wissensabschnitt 2: Repräsentativität

Liefert die Studie einen guten Überblick über die Grundgesamtheit (Population), oder ist das Bild verzerrt? Da in einer Studie gewöhnlich nur ein sehr geringer Teil der Population untersucht wird (Stichprobe), ist es wichtig, dass die Stichprobe ein gutes Abbild der Population ist. Wurde dies erreicht, spricht man von einer repräsentativen Stichprobe. Schwierig wird die Beurteilung der Repräsentativität besonders dadurch, dass man niemals erschöpfend weiß, hinsichtlich welcher Aspekte (Variablen) sich die Stichprobe von der Population unterscheiden könnte. (Reichen Alter, Bildung und Familienstand, oder muss man bis zu Hobbys und politischen Einstellungen auf Repräsentativität achten?) Dabei kann man auf die Überprüfung von Aspekten,

die offensichtlich von Irrelevanz sind (und die nicht mit anderen, relevanten Kriterien in einem Zusammenhang stehen), wie etwa die Vorliebe für bestimmte Inneneinrichtungen, verzichten.

Die Repräsentativität einer Stichprobe in Bezug auf eine Anzahl von Kriterien lässt sich nachprüfen, indem man vorhandenes Wissen über die Verteilung der Kriterien in der Population (Alter, Familienstand etc.) mit der Verteilung in der Stichprobe vergleicht. Weicht die Verteilung in der Stichprobe nur unwesentlich von der Verteilung in der Population ab (dies wird mit einem statistischen Hypothesentest überprüft), so kann man hinsichtlich dieser Kriterien von einer Repräsentativität ausgehen.

Das Vorgehen ist problematisch, weil man diese Prüfung nur für die Kriterien vornehmen kann, bei denen man Daten über die Populationsverteilung besitzt. Fehlen diese, so ist eine Prüfung nicht möglich.

Ein anderes und übliches Vorgehen besteht darin, dass man bei der Auswahl der Stichprobe auf eine Zufallsziehung setzt. Ist die Stichprobe ausreichend groß und wurde sie per Zufall aus der Population gezogen, so kann man aufgrund der Wahrscheinlichkeitsgesetze davon ausgehen, dass sie in Bezug auf alle Kriterien (bekannte und unbekannt) weitgehend der Population gleicht. Größere Abweichungen sind extrem unwahrscheinlich und können deshalb ignoriert werden.

Eine echte Zufallsziehung aus einer Population durchzuführen ist allerdings keine einfache Aufgabe. Um Repräsentativität zu erreichen, muss sichergestellt werden, dass jedes Mitglied der Population die gleiche Wahrscheinlichkeit hat, in die Stichprobe aufgenommen zu werden. Dies ist bei vielen Studien aber nicht gegeben:

- Wenn Sie eine repräsentative Stichprobe aus einer Stadtbevölkerung erheben wollen, können Sie diese nicht durch Zufallsziehung aus dem Telefonbuch ermitteln, da hierdurch die Wahrscheinlichkeit, in die Stichprobe zu gelangen, für Personen ohne Telefon, für Personen mit Geheimnummer sowie für alle Familienangehörigen mit gemeinsamem Anschluss gleich null wäre. Single-Haushalte wären überrepräsentiert, Personen in Familien und Wohngemeinschaften wären unterrepräsentiert.
- Wenn Sie einen deutschsprachigen Fragebogen einsetzen, haben Populationsmitglieder, die der deutschen Schriftsprache nicht mächtig sind (Personen mit anderen Sprachen, Personen, die Leseschwierigkeiten haben oder einen Fragebogen nicht bearbeiten können) eine geringere Wahrscheinlichkeit, in die Stichprobe zu gelangen.

Wenn die Repräsentativität einer Stichprobe also mit einer Zufallsziehung gewährleistet werden soll, so sollten Sie sehr kritisch überlegen bzw. nachfragen, ob die Realisierung dieser Ziehung auch zu einer gleichen Aufnahmewahrscheinlichkeit für alle Populationsmitglieder führt. Gibt es hier berechtigte Zweifel, so sollte man eine möglichst umfassende kriterienorientierte Prüfung der Stichprobe durchführen.

Wird deutlich, dass bestimmte Teilgruppe komplett in der Stichprobe fehlen, so kann man sich damit behelfen, dass man die Population auf die sich die Studie bezieht redefiniert: Alle Bürger der Stadt, die über ausreichende deutsche Sprachkenntnisse verfügen, beispielsweise. Häufig führen solche Populationseinschränkungen aber auch zu erheblichen Einschränkungen im Erkenntniswert der Studie.

Häufig wird die Frage nach der Repräsentativität einer Stichprobe mit der Rücklaufquote bei einer Fragebogenbefragung oder dem Stichprobenumfang verknüpft: „Ist eine Befragung repräsentativ, wenn nur 24% der verteilten Fragebögen ausgefüllt zurückgesendet wurden?“ oder „wenn Ihre Stichprobe nur 1,2% der Population umfasst.“ Aus dem bisher Gesagten sollte Ihnen deutlich geworden sein, dass Stichprobenumfang und Rücklaufquote keine Aussage über die Repräsentativität erlauben. Sogar eine Studie, die 95% eine Population umfasst, kann unter Umständen nicht repräsentativ sein: Wenn Sie in einer Kleinstadt alle Kinder auf eine bestimmte Erkrankung untersuchen, indem Sie die Schulen besuchen, aber diejenigen vergessen, die (vielleicht sogar wegen dieser Erkrankung) zu Hause geblieben sind oder im Krankenhaus liegen, hat Ihre Studie ein ernstes Repräsentativitätsproblem. Haben Sie hingegen aus einer sehr großen Population eine Stichprobe von vielleicht 250 Studienteilnehmerinnen mit hoher Sorgfalt per Zufall gezogen, können Sie von einer entsprechend hohen Repräsentativität ausgehen.

Wissensabschnitt 3: Suffizienz

Das Suffizienzkriterium verlangt, dass die Studie einen ausreichenden Umfang hat, um die Fragestellung angemessen beantworten zu können. Dies bedeutet einerseits, dass die Stichprobe ausreichend groß sein muss, um die Fragestellung beantworten zu können. Auch eine repräsentative Stichprobe von 100 Teilnehmerinnen kann nicht ausreichen, um etwa die Prävalenz (das Auftreten) einer sehr seltenen Krankheit zu bestimmen. Um eine Einschätzung über die notwendige Größe eines suffizienten Stichprobenumfangs zu gewinnen, können Fallzahlberechnungen durchgeführt werden. Dies sind mathematisch-statistische Verfahren, die es erlauben, zu ermitteln,

wie groß eine Stichprobe sein muss, damit ein Effekt, der eine bestimmte Stärke hat, oder ein Ereignis, das eine bestimmte Auftretenswahrscheinlichkeit hat, mit hinreichender Sicherheit auch gefunden werden kann.

Zusätzlich bedeutet Suffizienz aber auch, dass die Datenmenge ausreichend sein muss. Werden alle für das angestrebte Wissen relevanten Variablen erfasst? Lassen sich gewünschte Aussagen über räumliche Verteilung der Werte oder über zeitliche Veränderungen treffen? Hierzu muss die Studie der Fragestellung angemessen geplant sein. Auf diese Frage kommen wir später im Studientext noch zurück. Eine hinreichend große (suffiziente) Stichprobe kann also vergeudet sein, wenn Ihr Untersuchungsinstrument wichtige Informationen nicht erfasst (also insuffizient ist).

Wissensabschnitt 4: Indikativität

Ist die Messung so genau und so umfassend, dass die gesuchten Unterschiede oder Risiken überhaupt entdeckt werden können?

Einige Beispiele für Studienfragestellungen, bei denen die Indikativität kritisch sein könnte:

- Sie interessieren sich für das Risiko, dass durch eine Schadexposition eine nur geringe Erhöhung einer Krankheitsanfälligkeit resultiert. Um diese Erhöhung aber nachweisen zu können, ist vielleicht ein bestimmtes Screening-Verfahren zu ungenau (dass heißt: nicht indikativ) und es muss eine genauere Diagnostik durchgeführt werden.
- Ein Fragebogen ist so summarisch und unpräzise, dass die gewonnenen Ergebnisse nicht ausreichen, den Effekt einer Therapie oder einer anderen Maßnahme nachzuweisen. Angenommen, Sie wollen den Effekt eines Anti-Raucher-Seminars evaluieren, fragen im Fragebogen aber nur ab, ob die Teilnehmerinnen das Rauchen komplett aufgegeben haben. Möglicherweise würden Sie es ja als ein Erfolg werten, wenn die Teilnehmerinnen im Durchschnitt den Konsum deutlich reduzierten. Der genannte Fragebogen wäre für ein solches Erfolgskriterium aber nicht indikativ.

Indikativität kann aber nicht nur durch eine zu ungenaue Messung verfehlt werden. Ein ebenfalls relevantes Problem besteht darin, dass man die falsche Variable erfasst oder aber falsche Aspekte der richtigen Variable misst. In diesem Fall sagt man, dass die erfasste Variable nicht indikativ ist für die angestrebte Aussage. Will man beispielsweise bei sehr höflichen Patienten die Zufriedenheit mit einer Therapie erfassen, so könnte es sein, dass diese auf die direkte Frage „Wie zufrieden waren Sie?“ alle eher positiv antworten. Diese Frage wäre dann nicht indikativ für die Zufriedenheit. Eine

eher indirekte Frage wie „Würden Sie diese Therapie noch einmal wählen?“ könnte möglicherweise indikativer sein.

2.1 Wie verlässlich ist das Ergebnis der Untersuchung?

Wissensabschnitt 5: Reliabilität

Die Reliabilität bezieht sich auf die Zuverlässigkeit einer Angabe oder einer Messung.

Angenommen Sie möchten Reaktionszeiten messen, etwa um eine leichte Verlangsamung aufgrund einer Medikamentengabe nachzuweisen. Wenn Sie zur Erfassung der Reaktionszeiten, die unterhalb von einer Sekunde liegen können, keine geeigneten Apparate einsetzen, sondern den Sekundenzeiger einer Armbanduhr verwenden, so werden sie unreliable, das heißt, nicht verlässliche Messungen erhalten. Würden Sie eine solche Messung wiederholen (vielleicht, indem Sie die Reaktion der Untersuchten auf Video aufnehmen und mehrfach mit der Armbanduhr eine Messung vornehmen), werden Sie unterschiedliche Ergebnisse erhalten.

Bei jeder Messung müssen Sie damit rechnen, bei Wiederholung ein etwas anderes Ergebnis zu erhalten, sei es auch ab der fünften Nachkommastelle. Deshalb geht man in der Messtheorie davon aus, dass eine jede Messung aus der Erfassung des wahren Wertes plus einem (zufälligen) Messfehler besteht. Ist dieser Messfehler gering, so bezeichnet man die Messung als reliabel, ist er aber hoch, so ist die Messung unreliabel.

Bei einem guten Messinstrument oder einer guten Untersuchungsmethode wird der Messfehler angegeben. Man kann ihn bestimmen, indem man die gleiche Messung mehrfach wiederholt, also etwa die gleiche Probe mehrfach im Labor untersuchen lässt.

Eine geringere Reliabilität kann auch auftreten, wenn es Besonderheiten bei der Untersuchung oder Befragung gibt, welche die Messwerte beeinflussen. Dies könnte etwa störender Lärm sein oder eine fehlerhafte Lagerung oder Verarbeitung von Probenmaterial (z.B. Blutproben).

Wissensabschnitt 6: Objektivität

Die Objektivität bezieht sich auf die Unabhängigkeit einer Angabe oder einer Messung von demjenigen, der die Messung durchführt.

Die Objektivität ist bei Laboruntersuchungen oder bei der Verwendung von Fragebögen im allgemeinen gegeben. Wenn Sie dieselbe Probe im Labor von verschiedenen Mitarbeiterinnen analysieren lassen, oder wenn Sie

8 Schließende Statistik

Wissensabschnitt 79: Von den Stichprobendaten auf die Population schließen

Das Grundproblem, mit dem sich dieses Kapitel befasst, ist die Frage, wie wir Informationen über Populationen erhalten. Da es im allgemeinen unmöglich ist, die gesamte Population zu untersuchen oder zu befragen, bleibt aus praktischen Gründen nichts anderes, als eine repräsentative Stichprobe zu ziehen und diese zu untersuchen oder zu befragen.

Wie wir gesehen haben, ist es möglich, mit Verfahren der deskriptiven Statistik verständliche und übersichtliche Aussagen über Stichproben abzuleiten. Aber normalerweise sind Aussagen über eine Stichprobe wissenschaftlich uninteressant. Die Stichprobe ist eine einmalig zusammengestellte Gruppe, die in dieser Konstellation niemals wieder auftritt.

Stichproben sind Mittel zum Zweck, sie sollen uns möglichst genaue Hinweise auf die zugehörige Population geben. Wie zieht man nun den Schluss von der Stichprobe zur Population? Betrachten wir als Beispiel eine Stichprobe von 50 zufällig ausgesuchten Frauen und Männern zwischen 25 und 69 Jahren aus NRW. Bei diesen Probanden wurde untersucht, ob sie unter Hypertonie oder Grenzwerthypertonie leiden oder ob sie keinen Bluthochdruck haben. Es zeigt sich, dass 11 StudienteilnehmerInnen (22%) unter Hypertonie leiden, ebenfalls 11 TeilnehmerInnen unter Grenzwerthypertonie und 28 TeilnehmerInnen (56%) keinen Bluthochdruck aufweisen¹¹.

Die Kenntnisse über die Population, die wir zu diesem Beispiel im vorigen Kapitel diskutiert haben, sind für unser Beispiel hilfreich, um zu sehen, wie sich die Stichprobenwerte und die wahren Werte der Population unterscheiden. Sie sollten sich stets bewusst sein, dass bei einer echten Studie die wahren Werte der Population natürlich unbekannt sind, und die Stichprobenwerte die einzigen Informationen sind, über die man verfügt.

Um unter diesen Umständen etwas über die Population aussagen zu können, bleibt also nichts übrig, als die Werte, die man in der Stichprobe gemessen hat, als beste Schätzung der Populationswerte zu nehmen. Man muss also annehmen, dass in der Population, aus der die Stichprobe gezogen

¹¹ Diese und alle folgenden Stichproben, die in diesem Abschnitt diskutiert werden, wurden mit einer Computersimulation zufällig aus der Population gezogen.

wurde, die Wahrscheinlichkeit von Hypertonie 0,22 und die Wahrscheinlichkeit von Grenzwerthypertonie ebenfalls 0,22 beträgt.

Wir wissen aus dem vorigen Kapitel, dass diese Stichprobe den Anteil der Hypertoniker in der Population überschätzt, weil in Wirklichkeit die Wahrscheinlichkeit für eine Hypertonie in der Population nur 0,177 beträgt. Dieses Wissen hat man aber nicht, wenn man nur die Daten der Stichprobe vorliegen hat.

Wie sind diese Abweichungen der Stichprobenhäufigkeiten von den wahren Wahrscheinlichkeiten zu erklären? Wenn die Stichprobe wirklich repräsentativ aus der Population gezogen wurde, können alle systematischen Fehler ausgeschlossen werden, denn wenn ein systematischer Fehler vorliegt, wäre die Stichprobe eben nicht repräsentativ. Ist beispielsweise Übergewicht ein solcher Einflussfaktor, und wir haben unsere Probanden aus den TeilnehmerInnen eines Diätkurses rekrutiert, so könnte es sein, dass wir zu viele übergewichtige TeilnehmerInnen in unserer Stichprobe haben und so das Ergebnis erklärlich wäre. In diesem Fall wäre die Stichprobe nicht repräsentativ.

Die einzige Erklärung, die für die Abweichungen in unserem Beispiel bleibt, ist, dass wir es nicht mit einem systematischen Fehler zu tun haben, sondern mit einem Zufallsfehler. Wir haben unter unseren 50 TeilnehmerInnen, die ja zufällig aus der Population ausgewählt wurden, einige mehr erwischt, die unter Hypertonie leiden, als wir erwartet hätten. Würden wir die gleiche Untersuchung mit einer anderen Stichprobe, die wir auf gleich Art rekrutieren, wiederholen, wäre zu erwarten, dass wir einen anderen Zufallsfehler machen, entweder die wahren Werte unterschätzen, genauer treffen oder wiederum überschätzen.

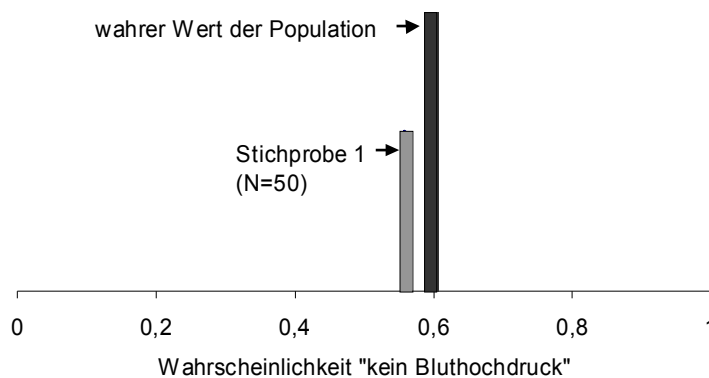
Wir werden uns in diesem Kapitel mit zwei wichtigen Methoden befassen, die geeignet sind, diesen Zufallsfehler beim Schluss auf die Populationswerte zu berücksichtigen. Das eine Verfahren sind die Konfidenzintervalle, das zweite die statistischen Signifikanztests.

Möglicherweise haben Sie ja auch schon überlegt, dass man mit einer größeren Stichprobe genauere Schätzungen bekommen würde. Aber auch hier kann stets ein Zufallsfehler auftreten. Es ist allerdings zu erwarten, dass der Zufallsfehler kleiner ist, je größer die Stichprobe ist. Aus diesem Grund hat man im allgemeinen ein höheres Vertrauen in Studien, die große Stichproben untersucht haben. Wie die Statistik mit dieser Frage der Stichprobengröße umgeht, und warum manchmal auch Studien mit kleinen Stichproben aussagekräftiger sein können, als Studien mit großen Stichproben, werden wir in diesem Kapitel ebenfalls diskutieren.

Wissensabschnitt 80: Die Verteilung von Stichprobenparametern

In Abbildung 29 ist die wahre (aber normalerweise unbekannte) Populationswahrscheinlichkeit für das Ereignis "kein Bluthochdruck" als schwarzer Balken dargestellt. Die Wahrscheinlichkeit in der Population beträgt 0,596. Als grauer Balken ist das Ergebnis unserer Stichprobe von 50 Teilnehmerinnen und Teilnehmern dargestellt. Hier ergab sich eine relative Häufigkeit von 0,560, da 28 der 50 Untersuchten keinen Bluthochdruck aufwiesen. Wie man sieht, wären wir nicht ganz schlecht, wenn wir die Populationswahrscheinlichkeit aufgrund unserer Stichprobe zu 0,560 schätzen würden, hätten uns aber um 0,036, also um 3,6% vertan.

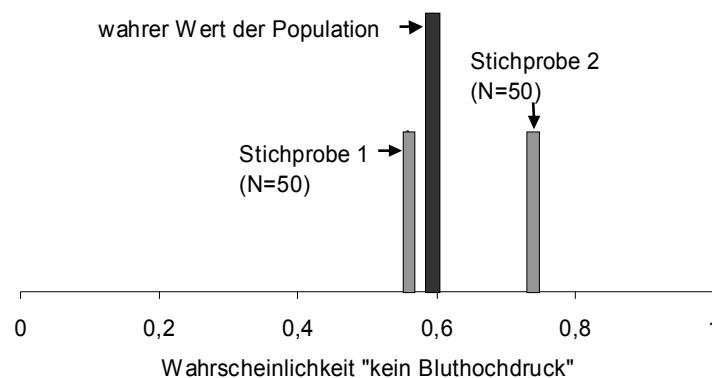
Abbildung 29: Populationswahrscheinlichkeit und relative Häufigkeit der Stichprobe



Nun beginnen wir mit einem Gedankenspiel, das verdeutlichen soll, wie wir den Zufallsfehler abschätzen können, das wir aber in der Realität so nicht verwirklichen würden: Stellen Sie sich vor, Sie legen Ihre Daten zur Seite und untersuchen eine neue Stichprobe von 50 Frauen und Männern auf genau die gleiche Weise wie die erste Stichprobe. Hier beobachten wir nun, dass 37 der 50 Befragten keinen Bluthochdruck aufweisen¹². Würden wir diese zweite Stichprobe also verwenden, um die Populationswahrscheinlichkeit zu schätzen, würden wir sie mit 0,740 deutlich überschätzen.

¹² Dieses Ergebnis ist wieder eine Zufallsziehung, die am Computer simuliert wurde.

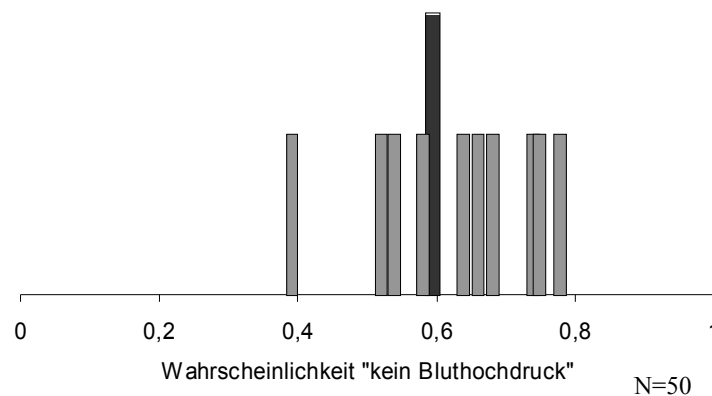
Abbildung 30: Populationswahrscheinlichkeit und relative Häufigkeit von zwei Stichproben



Der Unterschied zwischen beiden Ergebnissen ist ausschließlich auf den Zufallsfehler zurückzuführen, alles andere haben wir exakt identisch gemacht.

In der folgenden Abbildung sehen Sie die zehn verschiedenen relativen Häufigkeiten, die sich ergeben könnten, wenn wir zehnmal eine neue Stichprobe von jeweils 50 zufällig ausgewählten Teilnehmerinnen und Teilnehmern untersuchen.

Abbildung 31: Populationswahrscheinlichkeit und relative Häufigkeit von zehn Stichproben

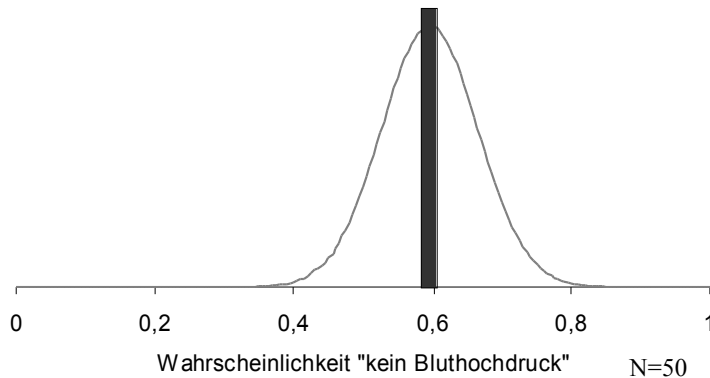


Es fällt auf, dass die ermittelte relative Häufigkeit manchmal geringer ist als die wahre Wahrscheinlichkeit, manchmal größer. Die meisten Werte liegen recht nahe bei der wahren Wahrscheinlichkeit, aber einige Werte liegen auch recht weit entfernt.

Wir können deshalb die ermittelte relative Häufigkeit bei einer Stichprobe mit einer bestimmten Anzahl Befragter wiederum als eine Zufallsvariable betrachten. Jedes mal, wenn wir eine solche relative Häufigkeit errechnen, ist man mehr oder weniger weit vom wahren Wert entfernt. Ziehen wir in

unserem Gedankenspiel also unendlich viele Stichproben vom Umfang 50 Personen. Das Ergebnis ist in der folgenden Abbildung zu sehen.

Abbildung 32: Populationswahrscheinlichkeit und relative Häufigkeit von unendlich vielen Stichproben



Es lässt sich mathematisch zeigen, dass die relativen Häufigkeiten von diesen Stichproben sich annähernd wie eine Normalverteilung verteilen, und dass der Mittelwert dieser unendlich vielen Stichproben der wahre Wert der Populationsverteilung ist. Es ist mathematisch sogar möglich, die Standardabweichung dieser Normalverteilung zu errechnen. In unserem Fall beträgt sie 0,0694.¹³

Wir nennen im folgenden eine solche Verteilung von unendlich vielen Stichprobenergebnissen eine Stichprobenparameterverteilung. Bitte denken Sie stets daran, dass wir eine solche Stichprobenparameterverteilung nur als Gedankenspiel benötigen, um eine Grundlage für die folgenden Methoden zu haben, die es erlauben werden, von einer Stichprobe auf die Population zu schließen. Es gibt keine Studie, welche eine Stichprobenparameterverteilung erhebt.

Nun sollten Sie sich an die Eigenschaften der Normalverteilung erinnern. Wenn man den Mittelwert und die Standardabweichung einer Normalverteilung kennt, kann man für jeden Wert sagen kann, wie wahrscheinlich es ist, diesen oder einen niedrigeren Wert, bzw. einen höheren Wert zu erhalten.

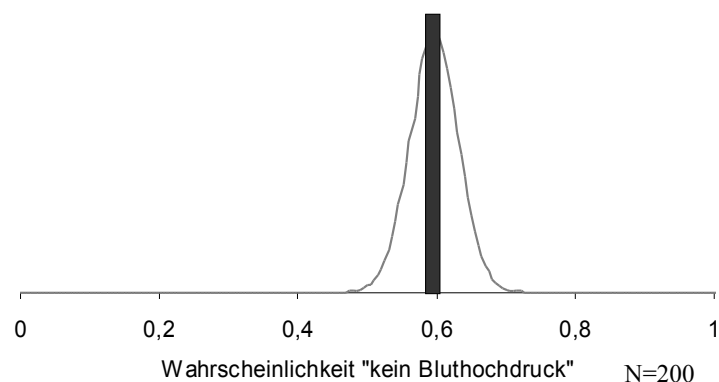
Somit kann man auch für die Stichprobenparameterverteilung sagen, wie wahrscheinlich es ist, eine Stichprobe zu bekommen, die eine relative Häu-

¹³ Sie wird errechnet, indem man die Wahrscheinlichkeit nicht unter Bluthochdruck zu leiden (0,596) mit ihrer Gegenwahrscheinlichkeit (Hypertonie oder Grenzwerthypertonie zu haben, also 0,404) multipliziert, das Ergebnis durch die Stichprobengröße dividiert und zum Abschluss aus dem Ergebnis die Wurzel zieht. Mehr dazu finden Sie in Lehrbüchern der Statistik.

figkeit von 0,560 oder weniger aufweist. 0,560 liegt ungefähr eine halbe Standardabweichung unter dem Mittelwert. Von minus Unendlich bis zu dieser Stelle befindet sich eine Wahrscheinlichkeit von 0,302. Damit wissen wir, dass wir in ca. 30% aller Stichproben aus unserer Population eine relative Häufigkeit von 0,560 oder weniger erhalten werden, alleine aufgrund des Zufallsfehlers. Sehr unwahrscheinlich ist ein solches Ergebnis also nicht.

Was würde nun geschehen, wenn wir eine größere Stichprobe untersucht hätten, beispielsweise 200 Teilnehmerinnen und Teilnehmer? Auch hier führen wir wieder unser Gedankenspiel durch, bei dem wir uns vorstellen, wir würden unendlich viele solcher Stichproben ziehen und die ermittelte relative Häufigkeit von "kein Bluthochdruck" wäre wieder eine Zufallsvariable. Die folgende Abbildung zeigt die Mittelwertsverteilung, die sich bei diesem Stichprobenumfang ergeben würde.

Abbildung 33: Populationswahrscheinlichkeit und relative Häufigkeit von unendlich vielen Stichproben mit jeweils 200 Untersuchten



Wieder resultiert eine Normalverteilung und wieder ist der Mittelwert dieser Normalverteilung die wahre Wahrscheinlichkeit der Population. Allerdings ergibt sich nun eine kleinere Standardabweichung. Sie beträgt nun nur noch 0,0347. Dies geschieht deshalb, weil in der Formel für diese Standardabweichung die Größe der Stichprobe eine Rolle spielt, denn Wahrscheinlichkeit mal Gegenwahrscheinlichkeit werden ja durch die Stichprobengröße dividiert. Je größer also die Stichprobe, desto geringer die Standardabweichung und damit liegen die relativen Häufigkeiten der Stichproben mit größerer Wahrscheinlichkeit näher am wahren Populationswert als bei kleinen Stichproben. Unser Gedankenspiel verhält sich somit genauso, wie wir das auch aus naiver Sicht erwarten würden: Große Stichproben liefern bessere Schätzung für die Population als kleine Stichproben.

Eine relative Häufigkeit von 0,560 würde bei einer 200er-Stichprobe nun etwas mehr als eine Standardabweichung unterhalb des Mittelwertes liegen. Die Wahrscheinlichkeit, dass ein solches Ergebnis auftritt, beträgt bei einem Stichprobenumfang von 200 Untersuchten nur noch 0,159. Etwas weniger als 16% aller Stichproben von 200 Personen ergeben also eine relative Häufigkeit von 0,560 für "kein Bluthochdruck", wenn sie repräsentativ aus der bekannten Population gezogen werden.

Auch hier können wir wieder die beiden Punkte, die 1,64 Standardabweichungen ober- bzw. unterhalb des Mittelwertes liegen als Grenze eines 90%-Intervalls betrachten: 5% unserer Stichproben liegen jeweils außerhalb, also liegen 90% unserer Stichprobe zwischen diesen Grenzen. Bei der Stichprobenparameterverteilung für Stichproben mit einem Umfang von 50 Teilnehmerinnen und Teilnehmern errechnet man dieses Intervall, indem man 1,64 mal die Standardabweichung von 0,0694 zum Mittelwert addiert bzw. von ihm subtrahiert. Somit liegen also 90% der Stichprobenergebnisse mit einem Umfang von 50 zwischen 0,482 und 0,710. Bei einem Stichprobenumfang von 200 dagegen liegen 90% der Ergebnisse zwischen 0,539 und 0,653. Bitte versuchen Sie, diese Zahlen mit einem Taschenrechner nachzuvollziehen.

Es wird also durch die geringere Spannweite dieses Intervalls deutlich, dass eine größere Stichprobe ein genaueres Ergebnis bei der Schätzung der wahren Populationsparameter erwarten lässt.

Solche Stichprobenparameterverteilung lassen sich nicht nur für relative Häufigkeiten als Schätzungen von Populationswahrscheinlichkeiten berechnen. Auf gleiche Art lässt sich unser Gedankenspiel auch für alle anderen deskriptiven Parameter, mit denen sich eine Stichprobe beschreiben lässt, durchführen, beispielsweise für Mittelwerte, Standardabweichungen, Mediane oder Korrelationen.

Für Mittelwerte geht der Gedankengang etwa folgendermaßen: Wir berechnen für unsere Stichprobe einen Mittelwert (z.B. das Durchschnittsalter der Untersuchten). Dieser Mittelwert ist der beste Schätzer für den wahren Mittelwert der Population, den wir besitzen. Hätten wir unendlich viele Stichproben mit dem gleichen Umfang gezogen, hätten wir unendlich viele verschiedene Durchschnittsalter als Stichprobenmittelwerte ermittelt. Diese Mittelwerte bilden eine Zufallsvariable, die ebenfalls normalverteilt ist. Der Mittelwert dieser Normalverteilung ist das wahre Durchschnittsalter der Population. Die Standardabweichung dieser Normalverteilung lässt sich wiederum berechnen (hier: Standardabweichung der Population dividiert durch die Wurzel aus dem Stichprobenumfang). Somit können wir angeben, zwi-

schen welchen Grenzen 90% aller Stichprobenmittelwerte liegen werden, d.h. in welchen Grenzen wir unseren Stichprobenmittelwert mit 90% Wahrscheinlichkeit erwarten würden. Beliebige andere Werte als 90% sind natürlich auch möglich. Wenn wir allerdings 100% wählen, ergeben sich als Grenzen stets minus Unendlich und plus Unendlich, so dass dieses Ergebnis uns nichts nutzt.

Diese Gedanken zu Stichprobenparameterverteilungen sind theoretischer Natur. Wir werden niemals unendlich viele Stichproben ziehen können. Es macht auch keinen Sinn, zehn Stichproben mit einem Umfang von 50 Teilnehmerinnen und Teilnehmern zu ziehen um anschließend die jeweiligen Ergebnisse miteinander zu vergleichen. Stattdessen würden wir eine Stichprobe mit 500 Teilnehmerinnen und Teilnehmern ziehen und aus dieser einen Stichprobe eine recht genaue Parameterschätzung errechnen.

Auch können wir die oben beschriebenen Intervalle um den wahren Populationswert, innerhalb denen jeweils ein fester Anteil der Stichprobenergebnisse liegt, nicht wirklich berechnen, weil wir ja hierzu den wahren Populationswert als Mittelwert der Stichprobenparameterverteilung kennen müssen. Dieser Wert ist aber leider genau derjenige, der unbekannt ist, und dervon uns gesucht wird.

Dennoch hat der Gedanke einer Stichprobenparameterverteilung einen enormen Wert, weil diese Verteilung die Grundlage für die beiden folgenden Verfahren darstellt, die keine theoretischen Gedankenspiele sind, sondern praktisch einsetzbar, und die es erlauben werden, von einer Stichprobe ausgehend Aussagen über die Population zu treffen.

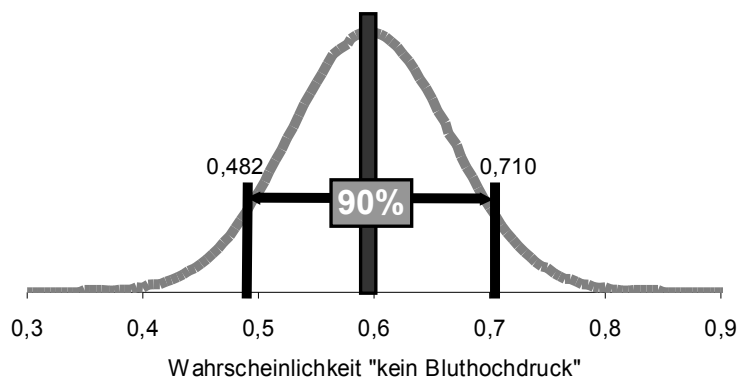
Wissensabschnitt 81: Konfidenzintervalle

Mit dem Intervall, das wir in einer Stichprobenparameterverteilung symmetrisch um den wahren Wert der Population herum aufspannen, können wir den Stichprobenmittelwert mit einer festen (vorher festgelegten) Wahrscheinlichkeit erfassen. Wenn wir nun das gleiche Intervall (mit der gleichen Spannweite) symmetrisch um den Stichprobenwert aufspannen, dann werden wir mit der gleichen Wahrscheinlichkeit den wahren Wert der Population erfassen.

Das dies so ist, ergibt sich aus Symmetriegründen. Wir können uns dies an folgenden Abbildungen verdeutlichen. Abbildung 34 zeigt das im vorigen Abschnitt diskutierte Beispiel eine 90%-Intervalls um die wahre Populationswahrscheinlichkeit von 0,596 bei einer Stichprobengröße von 50

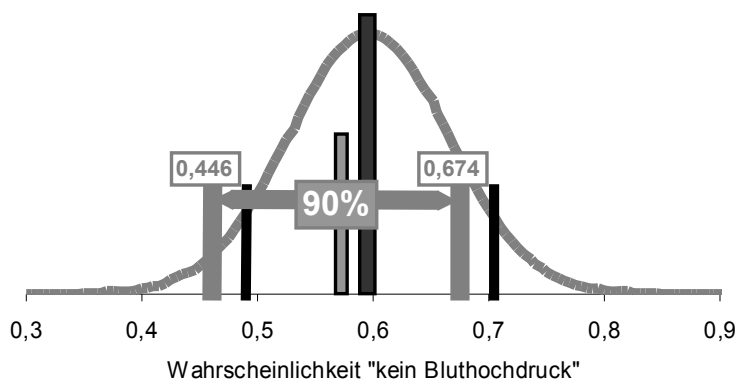
Teilnehmerinnen und Teilnehmern¹⁴. Mit 90%-Wahrscheinlichkeit liegt die in der Stichprobe ermittelte relative Häufigkeit für "kein Bluthochdruck" zwischen 0,482 und 0,710. Die Spannweite dieses Intervalls ist $0,710 - 0,482$, also 0,228.

Abbildung 34: 90%-Intervall für die relative Häufigkeit „kein Bluthochdruck“ in einer Stichprobe mit 50 Untersuchten



Da wir aber die wahre Populationswahrscheinlichkeit nicht kennen, schlagen wir das gleiche Intervall um unsere gemessene relative Wahrscheinlichkeit herum. Nun liegt der wahre Wert mit 90% Wahrscheinlichkeit innerhalb dieses Intervalls. Dies ist in Abbildung 35 dargestellt. Das Konfidenzintervall, innerhalb dessen die wahre Populationswahrscheinlichkeit von "keinem Bluthochdruck" liegt, reicht, aufgrund des Wissens, dass wir mit unserer Stichprobe von 50 Befragten und einem Stichprobenergebnis von 0,560 gewonnen haben, von 0,446 bis 0,674. Es hat also eine Spannweite von 0,228

Abbildung 35: 90%-Konfidenzintervall (grau)



¹⁴ Die Achse geht in dieser Abbildung nur von 0,3 bis 0,9, nicht von 0 bis 1 wie vorher, so dass die Normalverteilung etwas anders aussieht.

Als Konfidenzwahrscheinlichkeit bezeichnet man die Sicherheit, mit welcher der wahre Wert innerhalb dieses Intervalls liegt. In unserem Beispiel beträgt die Konfidenzwahrscheinlichkeit 90%. Selbstverständlich sind Konfidenzintervalle für beliebige Wahrscheinlichkeiten berechenbar. Je höher die Wahrscheinlichkeit, desto sicherer liegt der wahre Populationswert innerhalb dieses Intervalls, desto breiter wird aber auch das Intervall. Ein Konfidenzintervall mit einer Konfidenzwahrscheinlichkeit von 100% würde von minus Unendlich bis plus Unendlich gehen, es wäre also ohne Erkenntniswert. Üblich sind Konfidenzintervalle von 90%, 95% oder auch 99%. Höhere Konfidenzwahrscheinlichkeiten sind sehr selten und werden nur in Fällen verwendet, bei denen ein Irrtum zu erheblichen Schäden führen würde.

Die Breite des Konfidenzintervalls hängt nicht nur von der Konfidenzwahrscheinlichkeit, sondern auch von der Größe der Stichprobe ab. Je größer die Stichprobe ist, desto kleiner ist die Standardabweichung der Stichprobenparameterverteilung. Da bei 90%-Konfidenzintervallen die Grenzen jeweils 1,64 Standardabweichungen ober- bzw. unterhalb der ermittelten relativen Häufigkeit liegen, sind sie also bei kleinen Standardabweichungen näher beieinander als bei großen.

Es gibt noch eine mathematische Ergänzung zu dem bisher Dargestellten nachzutragen: Zur Berechnung des Konfidenzintervalls haben wir auch die Standardabweichung der Normalverteilung benötigt. Wenn Sie zurückblättern, werden sie feststellen, dass wir in dieser Formel die wahre Populationswahrscheinlichkeit (0,596) verwendet haben. Aber diese ist uns ja gerade unbekannt. Deshalb verwendet man bei realen Anwendungen auch für die Berechnung der Standardabweichung die relative Häufigkeit der Stichprobe (hier: 0,560) als beste Schätzung der wahren Populationswahrscheinlichkeit. Bei Konfidenzintervallen für Wahrscheinlichkeit macht dieser Fehler, den man dabei begeht, normalerweise nur einen so geringen Unterschied, dass man ihn vernachlässigt. Bei anderen Konfidenzintervallen, etwa dem für Mittelwerte, führt diese Ungenauigkeit dazu, dass man sie durch bestimmte mathematische Verfahren kompensieren muss.

Zur Übung sollten Sie das 90%-Konfidenzintervall unseres Beispiels neu berechnen, indem Sie die relative Wahrscheinlichkeit der Stichprobe auch zur Berechnung der Standardabweichung der Stichprobenparameterverteilung verwenden.

Wissensabschnitt 82: Die Hypothesen des statistischen Hypothesentests

Der statistische Hypothesentest prüft stets zwei Aussagen über die Population. Die erste Aussage lautet im Allgemeinen, dass es zwischen zwei Gruppen keinen Unterschied gibt, dass es von vorher zu nachher keine Veränderung gegeben hat, oder dass zwei Variablen nicht miteinander korreliert sind. Die ist in den allermeisten Studien das Ergebnis, dass man nicht gerne hätte: Ein Medikament zeigt keine bessere Wirkung als das Placebo, ein Projekt hat im Vergleich zum Vorher-Zustand keine Verbesserung gebracht. Diese Aussage nennt man die Nullhypothese (H_0)

Die zweite Aussage ist die verbleibende Gegenaussage: Dass es zwischen den zwei Gruppen einen Unterschied gibt, dass es von vorher zu nachher eine Veränderung gegeben hat, oder dass zwei Variablen miteinander korreliert sind. Die ist im allgemeinen das gewünschte Ergebnis einer Studie. Man nennt diese Hypothese die Alternativhypothese (H_1).

Häufig ist es so, dass man für die Alternativhypothese nur Veränderung in einer bestimmten Richtung zulassen will: Das Medikament hat die Gesundheit unterstützt und nicht behindert, die Nachher-Befragung zeigte bessere Ergebnisse als die Vorher-Befragung und keine schlechteren. In diesem Fall spricht man von einer einseitigen Testung. Die Nullhypothese umfasst in diesem Fall alle ungünstigen Ergebnisse (schädliche Wirkung des Medikaments, Verschlechterungen) bis einschließlich der fehlenden Wirkung bzw. des fehlenden Effekts und die Alternativhypothese nur die positiven, gesuchten Ergebnisse.

Bei einer zweiseitigen Testung hingegen umfasst die Nullhypothese nur den ausbleibenden Effekt bzw. den fehlenden Gruppenunterschied, während die Alternativhypothese irgendeinen Effekt bzw. Gruppenunterschied behauptet, gleich ob positiv oder negativ.

Der statische Hypothesentest versucht nicht, nachzuweisen, dass die Alternativhypothese gilt, indem er ihre Plausibilität angesichts der erhobenen Stichprobendaten untersucht. Er versucht vielmehr zu zeigen, dass die Nullhypothese nur mit sehr geringer Wahrscheinlichkeit mit den gefundenen Daten in Übereinstimmung zu bringen ist. Wenn aber dies zutrifft (Nullhypothese gilt mit hoher Sicherheit nicht), dann bleibt nur noch die Alternativhypothese übrig. So hat man einen Nachweis für die Alternativhypothese ex negativo erbracht.