5.8 Methoden und Statistische Verfahren zur Analyse der Wirksamkeit alkoholsensitiver Wegfahrsperren

Wolfgang Dau, Rainer Banse

Die Wahl einer adäquaten Auswertungsmethode und deren korrekte Handhabung ist ein integraler Bestandteil jedes Forschungsvorhabens. Dabei steht dem Forscher eine Vielzahl von gut eingeführten Verfahren zur Auswahl zur Verfügung und hochentwickelte Statistiksoftware erleichtert die Datenauswertung. Allerdings ergibt sich bei der Wahl von Auswertungsverfahren auch die Notwendigkeit, Vor- und Nachteile gegeneinander abzuwägen und sich über die inhaltlichen Aussagemöglichkeiten der einzelnen Verfahren Klarheit zu verschaffen. Weiterhin sind die Qualitäten der auszuwertenden Daten bei der Wahl der statistischen Verfahren zu berücksichtigen.

Im folgenden Kapitel werden die wichtigsten Verfahren, die bei der Untersuchung von Effekten alkoholsensitiver Wegfahrsperren in Betracht kommen, anwendungsorientiert beschrieben. Entsprechend dieser Konzeption wird das Grundverständnis über die Beantwortung von Detailfragen gestellt, aber es wird auf weiterführende Literatur verwiesen. Die Darstellung beginnt mit allgemeinen Überlegungen zur statistischen Datenanalyse und einem Überblick zum Teststärke-Konzept (Power-Analyse), die allgemeiner Natur und für alle statistischen Verfahren relevant sind. Anschließend werden die wichtigsten Verfahren der Evaluationsforschung vorgestellt.

5.8.1 Allgemeine Vorüberlegungen und Vorbereitungen – Datenqualität

Gerade bei großen Untersuchungen ist es notwendig bereits vor der eigentlichen Auswertung der Daten eine möglichst hohe Datenqualität sicherzustellen.

Umgang mit fehlenden Werten

Fehlende Werte können entstehen, wenn Versuchspersonen einzelne Fragen oder ganze Teile von Fragebogen bei der Bear-

Dau, W. & Banse, R. (2012). Methoden und Statistische Verfahren zur Analyse der Wirksamkeit alkoholsensitiver Wegfahrsperren. In Nickel, W.-R. & Schubert, W. (Hrsg.) *Best Practice Alkohol-Interlock*, (S. 209-238). Bonn: Kirschbaum.

beitung übersehen oder aus anderen Gründen nicht bearbeiten. Gerade bei Längsschnittuntersuchungen mit zwei oder mehr Messzeitpunkten sind Datensätze oft unvollständig, weil Versuchspersonen an einer Messung nicht teilgenommen haben, oder die Teilnahme an einer noch laufenden Untersuchung beenden.

Von beiden Ursachen für fehlende Daten ist das Problem einzelner, fehlender Antworten in Fragebögen leichter lösbar. In vielen Testmanualen wird angegeben, wie viele fehlende Items in einer Skala toleriert werden können. Das Brief Symptom Inventory (Franke, 2000), als Messinstrument zur Erfassung der psychischen Belastbarkeit, gibt beispielsweise 13 Items (von 53) als Maximalgrenze für fehlende Werte an, die bei der Testauswertung toleriert werden kann. Dabei ist darauf zu achten, ob es systematische Fehlerquellen für fehlende Daten gibt, da hierdurch die Interpretierbarkeit der Untersuchungsergebnisse eingeschränkt werden kann. Dies könnte z. B. der Fall sein, wenn bestimmte Items eines Fragebogens überzufällig häufig nicht verstanden werden. Dies war z. B. in einer Untersuchung von Dau (2012 der Fall: viele jüngere Konsumenten von Cannabis (Altersdurchschnitt: 22.7) haben dabei ein Item des BSI, mit der die Belastung durch „Schwermut" erfasst wird, nicht verstanden.

Die beste Möglichkeit mit komplett fehlenden Datensätzen („drop-outs") umzugehen, besteht darin, diese gar nicht erst entstehen zu lassen. Hierauf sollte bereits während der Vorbereitungsphase einer Untersuchung geachtet werden, z. B. indem möglichst umfassende Kontaktmöglichkeiten mit den Studienteilnehmern vereinbart und erfasst werden. Neben der Telefonnummer der Studienteilnehmer sollten mindestens auch die E-Mail-Adresse und falls möglich Kontaktdaten von Angehörigen erhoben werden. Bei längeren Untersuchungszeiträumen kann daran gedacht werden, auch zwischen den Messzeitpunkten Kontakt zu den Studienteilnehmern zu halten, beispielsweise durch Geburtstagskarten o. ä., um die Erinnerung an die Untersuchung wachzuhalten, die Beziehung zu den Studienteilnehmern zu pflegen und so die Antwortbereitschaft zu erhalten. Ebenfalls sollte darüber nachgedacht werden, welche Anreize (z. B. Büchergutscheine) den Studienteilnehmern geboten werden können. Diese Frage könnte insbesondere für Teilnehmer in den Kontrollgruppen, d. h.,

im Zusammenhang mit Alkohol auffällig gewordene Teilnehmer ohne Alkohol-Interlock-Gerät, wichtig sein. Gute Darstellungen über gelungene Follow-up-Erhebungen, aus denen auch die hier aufgeführten Beispiele entnommen worden, finden sich u. a. bei Cottler, Compton, Ben-Abdallah, Horne und Claverie (1996) und Scott (2004).

Trotz solcher Maßnahmen ist bei Längsschnittstudien immer mit fehlenden Daten zu rechnen. Handelt es sich hierbei um eine relativ kleine Anzahl, besteht die Möglichkeit, Versuchspersonen mit fehlenden Daten ganz von der Untersuchung auszuschließen. Bei diesem Vorgehen fließen nur Datensätze in die Analyse ein, für die zu allen Messzeitpunkten alle Daten vorliegen. Dies ist natürlich ein sehr leicht umzusetzender Ansatz, der allerdings um den Preis eines Verlusts an Studienteilnehmern und damit an Teststärke (siehe den folgenden Abschnitt) erkauft wird. Wie bei den Fragebogendaten besteht aber auch beim Ausschluss von Untersuchungsteilnehmern die Gefahr, dass diese nicht zufällig entstehen und es somit zu einer die Validität der Untersuchung gefährdenden, systematischen Verzerrung der Stichprobe kommt. Die Methode des Ausschlusses von Fällen mit fehlenden Daten für alle Variablen oder pro Variable (listen- oder fallweiser Ausschluss) bietet sich daher eigentlich nur bei einer geringen Anzahl von auszuschließenden Fällen an (Howell, 2010). Es ist bei der Datenanalyse darauf zu achten, dass die meisten Statistikprogramme über Voreinstellungen für den Umgang mit fehlenden Daten verfügen. Bei SPSS© ist beispielsweise der listenweise Fallausschluss bei vielen Verfahren voreingestellt, ein fallweiser Ausschluss muss aktiv gewählt werden. Weitere, häufig überlegenere Methoden zum Umgang mit fehlenden Daten, basieren z. B. auf *Imputation*. Hierbei werden für unvollständige Fälle unterschiedliche oder konstante Werte eingesetzt. Eher abzuraten ist davon, fehlende Werte einfach durch Stichprobenkennwerte (Mittelwert, Median usw.) zu ersetzen. Hieraus ergeben sich bedeutende methodische Nachteile, welche die Vorteile einer solchen Vorgehensweise bei weitem überwiegen. Es existiert eine Reihe von überlegenen Verfahren, die auf Regressions- oder Maximum-Likelihood-Modellen beruhen. Ein sehr guter Überblick über mögliche Verfahren zum Umgang mit fehlenden Werten sowie über die relevante Literatur findet sich bei Toutenburg, Heumann und Nittner (2009).

Vermeidung von Fehlern bei der Dateneingabe

Besonders bei einer großen Stichprobe liegt eine weitere Fehlerquelle bei der Eingabe der Daten. Hier ist an Kontrollprozeduren zu denken, bei denen die eingegebenen Daten immer wieder möglichst durch unabhängige Personen kontrolliert werden. Aus methodischen Überlegungen hinsichtlich der Reliabilität und Validität sollten Eingabefehler ausgeschlossen werden. Es empfiehlt sich in jedem Fall Vorkehrungen zu treffen, durch die Eingabefehler minimiert oder idealerweise ausgeschlossen werden können. Sofern möglich, sollte die Datenerhebung von vornherein computergestützt erfolgen. Dies stellt nicht nur eine immense Zeitersparnis dar. Übertragungsfehler bei der Dateneingabe sind bei diesem Verfahren ausgeschlossen. Darüber hinaus ist es je nach verwendeter Software möglich, die Eingabeprozeduren so zu gestalten, dass auch die Versuchspersonen keine Fragen mehr aus Versehen auslassen können, da das Programm erst nach Beantwortung fortfährt, womit eine weitere Fehlerquelle ausgeschlossen werden kann. Ferner kann bereits bei der Eingabe überprüft werden, ob Daten den richtigen Datentyp aufweisen (z. B. numerisch oder Buchstabenfolgen), das richtige Datenformat (z. B. ein korrektes Datum), oder außerhalb des möglichen Datenbereichs liegen (z. B. Monatsangaben über 12, Werte außerhalb des Skalenbereichs etc.). Inkorrekte Eingaben werden idealer Weise vom Datenerfassungsprogramm nicht zu gelassen.

5.8.2 Planung des Stichprobenumfangs über die Teststärke – Poweranalyse

Als Teststärke oder auch *power* wird die Wahrscheinlichkeit bezeichnet, mit der ein statistischer Signifikanztest einen tatsächlich vorhandenen Unterschied einer bestimmten Größe zwischen verschiedenen Stichproben auch entdeckt. Anders ausgedrückt könnte man die Teststärke als Sensitivität einer Untersuchung bezeichnen, einen tatsächlich vorhandenen Effekt einer bestimmten Größe auch korrekt zu identifizieren (Bortz & Schuster, 2010). Es geht also um die Frage, mit welcher Wahrscheinlichkeit ein auf den Einsatz von alkoholsensitiven Wegfahrsperren beruhender Effekt auf die Häufigkeit späterer alkoholbezogener Verkehrsdelikte mittels einer durchgeführten Studie gezeigt werden kann. Hierbei wäre die Annahme, dass die Wegfahrsperren keinerlei

Effekte haben die statistische Nullhypothese H_0. Demgegenüber würde die Alternativhypothesen H_1 einen Effekt der Wegfahrsperren voraussagen. Ist es aufgrund theoretischer Überlegungen möglich, eine bestimmte Richtung des Effekts vorherzusagen, z. B. die Aussage: der Einsatz von Wegfahrsperren verbessert die Legalprognose bei zuvor wegen Alkoholdelikten im Straßenverkehr auffällig gewordenen Fahrern, wird von einer gerichteten Hypothese gesprochen. Dies ist im Zusammenhang mit der Teststärke wichtig, da gerichtete Hypothesen eine höhere Teststärke aufweisen als ungerichtete (Bortz & Schuster, 2010). Darüber hinaus hängt die Teststärke von der Stichprobengröße und natürlich der Effektstärke ab. Die Effektstärke gibt an, wie stark ein beobachteter oder erwarteter Effekt ist. Bezogen auf die hier in Frage stehenden Untersuchungen wäre die Effektstärke etwas vereinfacht beispielsweise die Größe der Differenz zwischen den Rückfallraten mit oder ohne Einsatz einer alkoholsensitiven Wegfahrsperre relativ zur Streuung innerhalb der Gruppen. Es ist unmittelbar einleuchtend, dass ein starker Effekt leichter entdeckt werden kann als ein schwacher. In großen Stichproben werden Unterschiede schneller signifikant, da sich der Standardfehler mit zunehmendem Stichprobenumfang verringert. Eine Verringerung des Standardfehlers führt dazu, dass die Messwerte sich enger um den Mittelwert verteilen. Die den beiden konkurrierenden Hypothesen zugrundeliegenden Verteilungen um die Mittelwerte von Experimental- und Kontrollgruppe werden demzufolge schmaler und der Überlappungsbereich kleiner. Mit anderen Worten: die Teststärke steigt. Diese zugegebenermaßen vereinfachenden Ausführungen sind notwendig, um zu verstehen, wie vor Beginn einer Untersuchung mittels einer Poweranalyse der optimale Stichprobenumfang a priori festgelegt werden kann. Der „optimale" Stichprobenumfang stellt dabei in aller Regel einen Kompromiss zwischen den ökonomischen Kosten dar, der die Stichprobengröße nach oben limitiert, und der Wahrscheinlichkeit, einen tatsächlich vorhandenen Effekt einer bestimmten Größe auch zu entdecken, die einen minimalen Stichprobenumfang vorgibt.

Es ist jetzt deutlich geworden, dass Teststärke, Stichprobenumfang und Effektstärke zusammenhängen. Als vierte Komponente beeinflussen die Irrtumswahrscheinlichkeiten für die fälschliche Zurückweisung der H_0 (α-Fehler oder Fehler 1. Art)

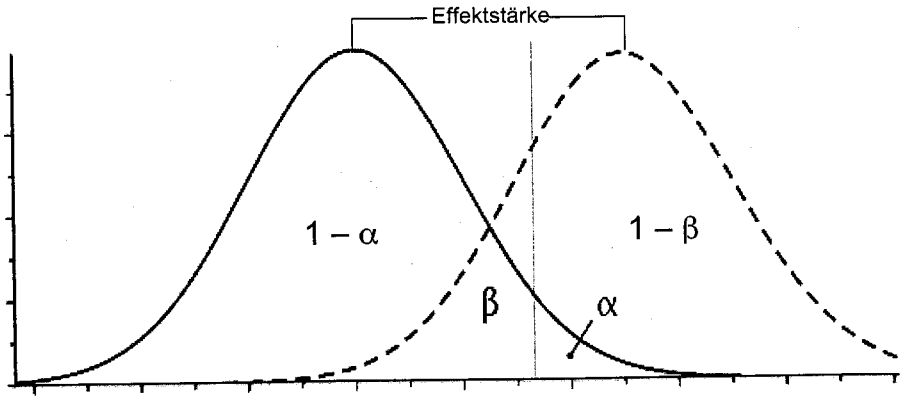

Abb. 5.16: Darstellung des Zusammenhangs zwischen Überlappungsbereich, Standardfehler und Teststärke

und der H_1 (β-Fehler oder Fehler 2. Art) die Teststärke. Dieser Zusammenhang mit der Teststärke ist leicht verständlich. Mit der α-Fehlerwahrscheinlichkeit setzen wir das Risiko einer irrtümlichen Zurückweisung der Nullhypothese auf einen akzeptablen Wert fest. Konventionell wird diese auf 5 % festgelegt, was bedeutet, dass die Wahrscheinlichkeit einer fälschlichen Zurückweisung der korrekten Nullhypothese 5 % beträgt. Je kleiner der α-Fehler gewählt wird, desto geringer ist zwar die Wahrscheinlichkeit eines Irrtums unter Beibehaltung der Nullhypothese, aber gleichzeitig sinkt auch die Wahrscheinlichkeit, sich korrekt für die Annahme der Alternativhypothese zu entscheiden. Der β-Fehler verhält sich komplementär. Je geringer die Irrtumswahrscheinlichkeit für eine falsche Zurückweisung der Alternativhypothese gewählt wird, desto größer ist die Wahrscheinlichkeit, dass ein tatsächlich vorhandener Unterschied nicht signifikant wird. Dementsprechend wird die Teststärke (oder statistische *power*) auch mit der Wahrscheinlichkeit $1 - \beta$ definiert. Konventionell wird der β-Fehler mit 20 % festgelegt, was zu einer Teststärke von 80 % führt. Das heißt, in 80 % oder vier von fünf durchgeführten Untersuchungen würde ein vorhandener Unterschied entdeckt werden. Wichtig ist an dieser Stelle noch die Anmerkung, dass der α- und β-Fehler auch anders bestimmt werden können. Bei der Entscheidung hierüber gilt es jeweils abzuwägen, welche Konsequenzen mit einer leichteren Ablehnung oder Annahme der jeweiligen Hypothese verbunden sind. Eine höhere Irrtumswahrscheinlichkeit für den α-Fehler würde unter Umständen dazu führen, dass

alkoholsensitive Wegfahrsperren eingeführt würden, obwohl die Effekte eher gering ausfallen. Ohne dies an dieser Stelle vertiefen zu können, gibt es natürlich Argumente für die eine und andere Position. Zur Berechnung der Teststärke ist es notwendig, die Effektstärke mit Hilfe der Forschungsliteratur zu ermitteln. Da im Bereich der Effektivität von alkoholsensitiven Wegfahrsperren bereits Studien durchgeführt wurden, kann die Effektstärke aus der Literatur entnommen oder errechnet werden. Liegen dann Werte für α, β und die Effektstärke vor, kann der optimale Stichprobenumfang leicht errechnet werden. Bei der Versuchsplanung ist weiter zu berücksichtigen, dass bei eher schwachen Effekten unter Umständen ein sehr hoher Ressourcenaufwand betrieben werden muss um schwache Effekte statistisch abzusichern. Eine Poweranalyse hilft somit bei der Entscheidung, ob der Aufwand an Ressourcen lohnend ist und somit eine fundierte Abschätzung des Kosten-Nutzen-Verhältnisses vorgenommen werden kann. Ein international etabliertes Standardwerk zu dieser Thematik ist das Buch von Cohen (1988). Eine sehr anschauliche Übersicht findet sich auch in Bühner und Ziegler (2009), die auch beschreiben wie mittels des kostenlosen Programms G*Power (Erdfelder, Faul & Buchner, 1996) Poweranalysen durchgeführt werden können. Daneben finden sich bei Bortz und Schuster (2010) Angaben zu optimalen Stichprobenumfängen für eine Vielzahl von Testverfahren.

5.8.3 Allgemeine Anmerkung zur Durchführung der statistischen Verfahren

Alle statistischen Verfahren sind für ihre Durchführung an Voraussetzungen gebunden, wie z. B. die Annahme normalverteilter Merkmale in der Population, Varianzgleichheit usw. Eine Verletzung dieser Voraussetzung kann unter Umständen dazu führen, dass Ergebnisse stark verfälscht oder nicht interpretierbar sind. Zu jedem in diesem Kapitel vorgestellten Verfahren werden Empfehlungen für weiterführende Literatur gegeben. Dort werden auch die Voraussetzungen für die jeweiligen Verfahren genannt und wie mit Verletzungen der Voraussetzungen umgegangen werden kann. Generell unterscheiden sich statistische Verfahren darin, wie empfindlich sie auf die Verletzung einer bestimmten Voraussetzung, z. B. der Normalverteilungsannahme,

reagieren. Man spricht in diesem Zusammenhang auch von der Robustheit von Testverfahren. Ein statistischer Test wird als robust bezeichnet, wenn sich α- und β-Fehler bei Verletzung statistischer Annahmen wenig ändern. Einige Testverfahren reagieren bei der Verletzung bestimmter Voraussetzung mit Ergebnissen, die eine häufigere Annahme der Nullhypothese nahelegen, d. h., die Verfahren reagieren konservativ. Dies hat allerdings zur Folge, dass die Wahrscheinlichkeit, dass ein tatsächlich existierender Unterschied auch seltener entdeckt wird. Der β-Fehler steigt also an. Umgekehrt verhält es sich, wenn die Verletzung von Vorrausetzungen zu sogenannten progressiven Entscheidungen führen. In diesem Fall wird die Nullhypothese leichter verworfen und die Alternativhypothese akzeptiert. Es besteht also eine höhere Gefahr eines α-Fehlers. Es existieren für viele Situationen, in denen die Voraussetzungen von Testverfahren verletzt sind, Prozeduren und Empfehlungen, wie zu verfahren ist. Die korrekte Darstellung der teilweise komplexen Zusammenhänge und Vorgehensweisen würde allerdings den Rahmen dieser Übersicht sprengen. Aus diesem Grunde wird in der Regel auf die Prüfung der Voraussetzungen und den Umgang mit Verletzungen der statistischen Voraussetzungen hier nicht eingegangen, sondern auf die weiterführende Literatur verwiesen. Generell ist zu empfehlen, bei der Verletzung von Voraussetzungen alternative Verfahren in Betracht zu ziehen, die mit weniger Voraussetzungen auskommen. In vielen Fällen ist die Anwendung möglich und sinnvoll, wenn der Test konservativ reagiert. Dies gilt, wenn eine vorschnelle, fälschliche Annahme der Alternativhypothese(n) größere Nachteile bringt. Reagiert ein Testverfahren progressiv auf die Verletzung der zugrundeliegenden Voraussetzungen, ist von einer Anwendung abzuraten (Nachtigall & Wirtz, 2009). Es ist demzufolge zwingend notwendig, sich vor der Anwendung eines Verfahrens über diese Thematik zu informieren.

5.8.4 Auswertungen von Häufigkeiten

χ^2-Kontingenzanalyse

Bei der Untersuchung von Effekten der alkoholsensitiven Wegfahrsperren dürfte der Forscher oftmals mit der Auswertung von Häufigkeiten konfrontiert sein. Hierzu werden Verfahren eingesetzt, die auf der χ^2-Verteilung (sprich: chi-Quadrat) beruhen.

Mögliche Fragestellungen wären z. B. „Wie viele Versuchspersonen werden während oder nach der Einsatzphase eines Alkohol-Interlock-Gerätes alkoholbedingt auffällig?" oder „Gibt es bei der Wirkung von Alkohol-Interlock-Geräten geschlechtsbezogene Unterschiede?" In beiden Fällen werden die Häufigkeiten erfasst, mit der ein bestimmtes Merkmal in einer bestimmten Kategorie oder Zelle des Versuchsplans auftritt. Es wird daher auch von kategorialen Daten gesprochen. Erfolgt hierbei eine einfache Zuordnung von Häufigkeiten zu anderen Merkmalen, wie z. B. Geschlecht oder Zugehörigkeit zur Experimentalgruppe (mit Alkohol-Interlock-Gerät) oder Kontrollgruppe (ohne Alkohol-Interlock-Gerät), handelt es sich um nominale Daten. Merkmale wie das Geschlecht, die lediglich zwei Ausprägungen aufweisen, werden als dichotom bezeichnet. Die Zusammenhänge zwischen den Merkmalen können dann mit einem χ^2-Test untersucht werden. Die Auswertung zwei dichotomer Merkmale erfolgt mittels einer *2 × 2-Kontingenzanalyse*. Die Kontingenzanalyse überprüft die Nullhypothese, dass die beiden Merkmale unabhängig voneinander sind, dass also beispielsweise zwischen Geschlecht und Wirkung von Alkohol-Interlock-Geräten kein Zusammenhang besteht. Der χ^2-Test beruht auf dem Vergleich zwischen den empirisch beobachteten und den unter der Nullhypothese, bei angenommener Unabhängigkeit, zu erwartenden Häufigkeiten.

Interpretation der Ergebnisse und Effektstärken

Zur Durchführung werden die Daten zunächst in einer Kreuztabelle angeordnet und ausgewertet. Liefert die Kontingenzanalyse ein signifikantes Ergebnis, ist die Nullhypothese zu verwerfen und von einem Zusammenhang zwischen den Merkmalen auszugehen. Vor der Auswertung ist unbedingt zu beachten, dass eine Gewichtung für die einzelnen Häufigkeiten stattfindet, die sich aus den Randsummen der Kreuztabelle ergeben, da ansonsten das Ergebnis verzerrt wird bzw. nicht sinnvoll interpretierbar ist. Eine anschauliche, praxisorientierte Darstellung findet sich bei Backhaus, Plinke, Erichson und Weiber (2006).

Das Verfahren ist ohne weiteres auch für eine höhere Anzahl von Merkmalen durchführbar. Es handelt sich dann um eine *k × l-Kontingenzanalyse* mit *k*-Zeilen und *l*-Spalten. Bei einer gleichzeitigen Auswertung nach Versuchsbedingung (Experimental-