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)

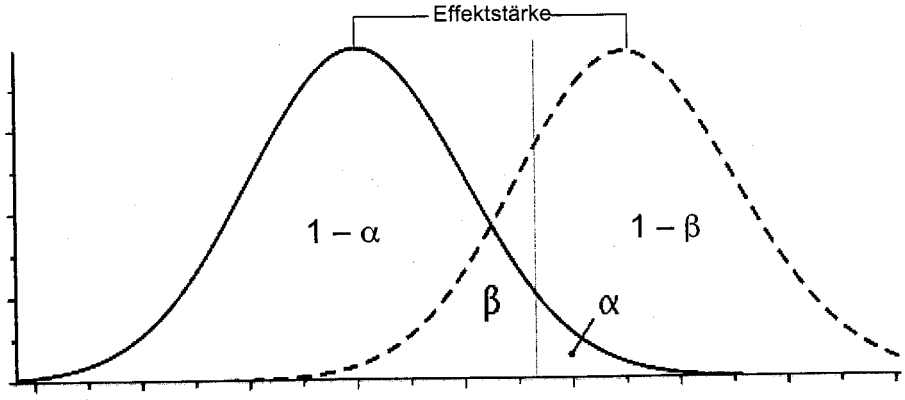

Abb. 5.16: Darstellung des Zusammenhangs zwischen Überlappungsbereich, Standardfehler und Teststärke

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-

versus Kontrollgruppe), Geschlecht und Rückfallrate erhielte man eine 4 × 2-Kontingenzanalyse. Spalten und Zeilen sind beliebig austauschbar. Bei der Interpretation ist unbedingt zu beachten, dass eine Zusammenhangshypothese überprüft wird. Für eine kausale Interpretation ist ein entsprechend aufgebautes experimentelles Design zwingend notwendig. Darüber hinaus verrät der χ^2-Test zunächst nichts über die Stärke oder die Richtung des Zusammenhangs. Allerdings ist es insbesondere im Fall einer 2 × 2-Kontingenzanalyse sehr leicht möglich, die Richtung des Zusammenhangs und die Effektstärke zu ermitteln.

Dies geschieht über den sogenannten Phi-Koeffizienten (φ), der dem Korrelationskoeffizienten bei metrischen Daten entspricht und auch ebenso zu interpretieren ist: je größer φ, desto stärker der Zusammenhang zwischen den Merkmalen. Werte größer 0.3 gelten bereits als bedeutsam (Howell, 2010), was auf Besonderheiten der auf der χ^2-Verteilung beruhenden Verfahren zurückgeht. Bei einer 2 × 2-Kontingenzanalyse entspricht φ darüber hinaus der für χ^2-Verfahren mit dem Parameter ω bezeichneten Effektstärke. Für erweiterte Kontingenzanalysen ist eine andere Berechnung von ω notwendig, die aber nicht aufwendig ist (Cohen, 1988). Mit Hilfe gängiger Softwarepakete lassen sich leicht alle notwendigen Werte für die Berechnung von ω bestimmen.

Bestimmung der Effektstärke über Risiken, Odds und die Odds-Ratio

Eine weitere, einfache Methode zur Bestimmung der Stärke von Zusammenhängen kann über die Risikoschätzung erfolgen. Die hierzu notwendigen Informationen lassen sich ohne weiteres aus einer Kreuztabelle oder 4-Felder-Tafel ablesen (siehe Tabelle 5.5).

Tab. 5.5: Beispiel einer Kreuztabelle für eine fiktive Untersuchung

Gruppe mit Alkohol-Interlock	800	150	950
Gruppe ohne Alkohol-Interlock	770	210	980
			1.930

Das *Risiko* im Untersuchungszeitraum ein Delikt zu begehen, beträgt in dieser fiktiven Untersuchung für die Gruppe mit Alkohol-Interlock 150 : 950 = 16 %. Das *Risiko* ist die Auftretenswahrscheinlichkeit eines Ereignisses in einer Versuchsbedingung und beträgt für die Gruppe ohne Alkohol-Interlock 210 : 980 = 21 %. Es wäre also in diesem Fall zu erwarten, dass etwa jeder fünfte aus der Kontrollgruppe ein alkoholbezogenes Delikt im Untersuchungszeitraum begeht. Das *Relative Risiko (RR)* ist der Quotient aus den Wahrscheinlichkeiten des Ereignisses „Deliktbegehung" in den beiden Bedingungen: $R = \text{Risiko}_{\text{ohne Alkohol-Interlock}} : \text{Risiko}_{\text{mit Alkohol-Interlock}} = 0.21 : 0.16\,\% = 1.31$. Das relative Risiko, ein Delikt während des Untersuchungszeitraums zu begehen ist also für die Kontrollgruppe 1.33 mal höher als für die Gruppe mit Alkohol-Interlock.

Ein ähnlicher Ansatz zur Quantifizierung von Effekten erfolgt über die Ermittlung der Chancen oder englisch: *odds*. Die Chancen ein Delikt zu begehen betragen für ein Mitglied der Alkohol-Interlock-Gruppe 150 : 800 = 0.19 und für die Kontrollgruppe 210 : 770 = 0.27. Die *Odds Ratio (OR)* wird aus dem Quotienten beider „Odds" oder Wahrscheinlichkeiten gebildet: 0.27 : 0.19 = 1.45. Die Chancen für die Begehung eines Delikts sind in der Kontrollgruppe fast anderthalb mal größer als in der Gruppe, die ein Alkohol-Interlock-Gerät erhalten hatte. Es existiert keinerlei Festlegung, welcher Wert bei der Berechnung der *OR* für den Zähler oder den Nenner gewählt wird. Es ist nur darauf zu achten, dass sich bei einem Tausch auch die Aussagerichtung umkehrt. Wären im obigen Beispiel Zähler und Nenner vertauscht, ergebe sich eine *OR* von 0.7. Die Aussage wäre demzufolge, dass die Chancen einer Deliktbegehung für die Alkohol-Interlock-Gruppe um den Faktor 0.7 verringert sind, was weniger anschaulich zu sein scheint. Eine *OR* = 1 bedeutet, dass die Chancen für eine Deliktbegehung in beiden Gruppen gleich ist und die Merkmale unabhängig voneinander sind.

Was sind die Vor- und Nachteile der einzelnen Maße, die in Abbildung 5.16 noch einmal im Überblick dargestellt sind? Das relative Risiko ist anschaulicher als die *Odds Ratio*. Es sollte aber nur berechnet werden, wenn die Randwahrscheinlichkeiten in der Kreuztabelle zufällig und nicht durch das Studiendesign entstehen. Gegen die Verwendung der *Odds Ratio* ist eingewandt wor-

$$Wahrscheinlichkeit\ (p) = \frac{\text{Häufigkeit eines Ereignisses}}{\text{Häufigkeit aller Ereignisse}}$$

$$Risiko = \frac{\text{Häufigkeit Delikt}}{\text{Häufigkeit aller Ereignisse einer Bedingung}} = p\ (\text{Delikt} \mid \text{Versuchsbedingung}$$

$$Relatives\ Risiko = \frac{p_1}{p_2} = \frac{(p_1 \mid \text{kein Interlock})}{(p_2 \mid \text{Interlock})}$$

$$odds = \frac{p}{(1-p)}$$

$$Odds\ Ratio = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}} = \frac{(odds \mid \text{kein Interlock})}{(odds \mid \text{Interlock})}$$

Abb. 5.17: Übersichtsdarstellung der Begriffe Wahrscheinlichkeit, Risiko, Relatives Risiko, Odds und Odds Ratio

den, sie sei unanschaulich, zu leicht mit Wahrscheinlichkeiten zu verwechseln und überschätze die Stärke von Zusammenhängen (Katz, 2006; Sackett, Deeks & Altmann, 1996). Richtig angewandt und interpretiert ist die *Odds Ratio* aber ein einfach zu berechnendes Effektmaß für dichotome Daten, das auch dazu genutzt werden kann, Effekte aus Studiendaten zu vergleichen (Howell, 2010; Rumpf, 2010). Die Berechnung eines Konfidenzintervalls ermöglicht eine Prüfung der *Odds Ratio* auf statistische Signifikanz. Die Berechnung ist über Statistiksoftware, z. B. SPSS©[1] leicht möglich. Eine Formel findet sich bei Morris und Gardner (1988).

Umrechnung der Odds Ratio zu Cohen's *d*

Die *Odds Ratio* kann Werte zwischen 0 und unendlich annehmen. Eine Interpretation im Sinne eines Korrelationskoeffizienten ist daher nur mit einer Transformation möglich, wie zum Beispiel dem *Y*-Koeffizienten von Yule (Yule, 1912), der hier vor allen Dingen wegen seiner Anschaulichkeit vorgestellt wird. Wie weiter unter beschrieben, ist es möglich, eine *Odds Ratio* direkt in Korrelationskoeffizienten zu transformieren. Der *Y*-Koeffizient berechnet sich folgendermaßen:

$$Y = \frac{\sqrt{OR} - 1}{\sqrt{OR} + 1} = \frac{\sqrt{1.45} - 1}{\sqrt{1.45} + 1} = 0.09$$

1 SPSS©-Menüführung: Analysieren – Deskriptive Statistik – Kreuztabellen – Statistik: Risiko

Für das fiktive Beispiel ergibt sich nur ein äußerst schwacher positiver Zusammenhang zwischen den Merkmalen „Häufigkeit Deliktbegehung" und Gruppenbedingung (Alkohol-Interlock/ kein Alkohol-Interlock). Eine einfache Formel zur Umrechnung einer *Odds Ratio* in ein Effektstärkemaß nach Cohen's *d* lautet (Cox & Snell, 1989):

$$d_{cox} = \frac{Ln(OR)}{1.65} = \frac{ln(1.45)}{1.65} = 0.23.$$

„Ln" bezeichnet die Funktion für den natürlichen Logarithmus. Auch nach der von Cohen (1988) vorgeschlagenen Klassifikation für die Einteilung von Effektstärken ergibt sich für die fiktive Untersuchung lediglich ein schwacher Effekt. Aber auch diese Formel wurde hier wegen ihrer relativen Einfachheit gewählt, denn ihr liegt die Annahme einer logarithmischen Verteilung der Daten zugrunde. Bei den meisten Untersuchungen wird aber von einer Normalverteilung ausgegangen. Die verwendete Formel muss jedoch an die den Daten zugrundeliegende Verteilung angepasst sein. Andernfalls kommt es zu einer Verzerrung der Ergebnisse. Aber auch für normalverteilte Daten gibt es Möglichkeiten zur Umrechnung der *Odds Ratio* sowohl in einen Korrelationskoeffizienten als auch in ein an Cohen angelehntes Effektstärkemaß. Einen sehr guten Überblick hierzu findet sich bei Sánchez-Meca, Marín-Martínez und Chacón-Moscoso (2003). Es steht weiterhin eine Reihe von kostenfreien Umrechnungsprogrammen zur Verfügung, die über das Internet bezogen werden können.

5.8.5 Logistische Regression

Die logistische Regression dient dazu, ein Modell zu entwickeln, mit dem Merkmale vorhergesagt werden können. Das Verfahren ist verwandt mit der linearen Regression für metrische (intervallskalierte) Daten, auf die später eingegangen wird. Bei der logistischen Regression weist die vorherzusagende (abhängige) Variable jedoch lediglich zwei Ausprägungen auf. Bei der Wirksamkeitsüberprüfung von alkoholsensitiven Wegfahrsperren sind Fragestellungen interessant, mit welcher Wahrscheinlichkeit Rückfälle auftreten und welche Einflussgrößen hierbei wirken. Das Merkmal „Rückfall" wird hierbei binär mit liegt vor („1") und liegt nicht vor („0") kodiert und bezeichnet im logistischen

Regressionsmodell die abhängige Variable. Soll überprüft werden, welchen Einfluss z. B. die Teilnahme an einem Alkohol-Interlock-Programm, an einer bestimmten, begleitenden psychotherapeutischen Intervention, das Geschlecht, Alter, Vorliegen einer Suchtstörung, die Anzahl allgemeiner Verkehrsdelikte usw. auf die Auftretenswahrscheinlichkeit von Rückfällen nach Abschluss des Interventionsprogramms haben, kann dies durch eine multiple logistische Regression berechnet werden. Merkmale, deren Einfluss auf die abhängige Variable (Rückfall) untersucht wird, werden als unabhängige Variablen oder Prädiktoren bezeichnet. Die Prädiktoren können dichotom oder metrisch sein.

Modellannahmen

Analog zum Vorgehen bei einer linearen Regression wird für die logistische Regression eine Gleichung bestimmt, mit der sich die Werte der abhängigen Variablen optimal vorher sagen lassen. Im hier gewählten Beispiel weist das Merkmal „Rückfall" allerdings nur zwei Werte auf, ohne Berücksichtigung von Zwischenausprägungen. Aus diesem Grunde wird auch von *zensierten* Daten gesprochen. Dieser Begriff spielt auch bei der Survival-Analyse eine wichtige Rolle. Die Grundidee der logistischen Regression ist es, die *Chancen (odds)* für den Eintritt des Ereignisses „Rückfall" unter den Bedingungen oder Bedingungskombinationen der verschiedenen Prädiktorvariablen zu ermitteln. Der Grund, warum diese Berechnung nicht über eine multiple Regression erfolgen kann, besteht darin, dass die Beziehung zwischen den odds und den unabhängigen Variablen nicht linear ist. Dies ist aber eine zentrale Modellannahme der multiplen Regression. Weiterhin sind bei einer dichotomen abhängigen Variablen auch weitere, wichtige Anwendungsvoraussetzungen der multiplen Regression verletzt, auf die hier nicht näher eingegangen werden soll. Eine gut verständliche Erklärung findet sich bei Backhaus, Plinke, Erichson und Weiber (2006). Um das Problem zu beheben, werden die odds mit Hilfe des natürlichen Logarithmus transformiert *(logarithmische Transformation)*. Die logarithmierten Odds werden auch als *logits* bezeichnet. Die logarithmische Transformation ist eine Technik, um den vermuteten nicht-linearen Zusammenhang zwischen den Prädiktoren und der abhängigen Variablen, in einen linearen Zusammenhang zu überführen. Ein nicht-linearer Zusammenhang bedeutet, dass eine Zunahme beim Wert des

Prädiktors nicht zu einer entsprechend höheren Eintrittswahrscheinlichkeit bei der abhängigen Variablen führt. Ist der Prädiktor die „Anzahl früherer Verkehrsdelikte" bedeutet dies, dass die Wahrscheinlichkeit für einen „Rückfall" nach Beendigung des Alkohol-Interlock-Programms nicht linear steigt, sondern ggf. eine andere Beziehung zwischen den beiden Variablen besteht. Denkbar wäre etwa, dass ab einer gewissen Anzahl von früheren Verkehrsdelikten die Rückfallwahrscheinlichkeit weniger steil ansteigt, da die Ausprägung des psychologischen Konstrukts, das für einen Rückfall kausal verantwortlich ist, bei mehr als drei früheren Delikten nicht mehr weiter zunimmt.

Dementsprechend wird die Beziehung zwischen den Prädiktoren und der abhängigen Variablen bei der logistischen Regression nicht durch eine gerade, sondern eine s-förmige (sigmoidale) Kurve wieder gegeben, die sich durch die folgende Regressionsgleichung beschreiben lässt:

$$P(Y) = \frac{1}{1 + e^{-(b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_n X_n)}}$$

Der Exponent der eulerschen Zahl *e* entspricht der Gleichung, wie sie sich auch für die lineare Regression formulieren lässt. Die Parameter „*b*" in der Gleichung geben an, wie sich das Chancenverhältnis bei einer Veränderung der Prädiktorvariablen „*X*" verändert.

Überprüfung der Ergebnisse und Interpretation

Die Prüfstatistik für die logistische Regression ist wie bei der Kontingenzanalyse die χ^2-Verteilung. Es stehen verschiedene Prüfgrößen für die Modellgüte zur Verfügung. Modellgüte bedeutet, wie gut die empirisch durch die Daten ermittelte Regressionsgleichung den tatsächlich vorhandenen Zusammenhang beschreibt. Die *Likelihood* gibt Aufschluss darüber, wie groß die Wahrscheinlichkeit ist, mit den ermittelten Parameterschätzungen die empirisch erhobenen Werte zu erhalten. Es wird mit der logarithmierten Likelihood (*LL*) gerechnet, die sich bei einer Multiplikation mit –2 einer χ^2-Verteilung annähert. Die Größe $-2LL$ bezeichnet man auch als *Devianz*, womit die Abweichung vom Idealwert gemeint ist. Bei einer perfekten Modellanpassung (Modellfit) ist $-2LL = 0$. Das bedeutet, der Anteil der nicht

durch das Modell aufgeklärten Information beträgt Null. Ein hoher –2*LL*-Wert bedeutet daher eine schlechte Modellanpassung. Mit der Wald-Statistik lässt sich ebenfalls prüfen, ob die gefundenen Regressionskoeffizienten signifikant von Null verschieden sind. Da die Wald-Statistik das Ergebnis insbesondere bei kleinen Stichproben verzerrt, ist der Likelihood-Test vorzuziehen (Cohen, Cohen, West & Aiken, 2003). Die sogenannte Pseudo-R^2-Statistik *Nagelkerke-*R^2 überprüft ebenfalls die Güte der Modellanpassung. Sie gibt an, wie viel mehr an Information durch das Hinzufügen der Prädiktoren gewonnen werden kann und kann Werte zwischen 0 (Prädiktoren sind wertlos) und 1 (perfekte Modellanpassung) annehmen. Es existieren noch weitere Pseudo-R^2-Statistiken, die sich von *Nagelkerke-*R^2 dadurch unterscheiden, dass sie den Maximalwert von 1 nicht erreichen können. Letzterer fällt daher immer größer aus als konkurrierende Maße. Backhaus, Plinke, Erichson und Weiber (2006) empfehlen klar die Verwendung von *Nagelkerke-*R^2, wohingegen Cohen, Cohen, West und Aiken (2003) lediglich darauf hinweisen, dass angegeben werden sollte, welches Maß verwendet worden ist. Wie bei der linearen Regressionsanalyse auch, sollte überprüft werden, ob es extreme Ausreißer in den empirischen Werten gibt. Ausreißer können das Ergebnis der Regressionsanalyse stark beeinflussen.

Bei der Interpretation der Regressionskoeffizienten ist darauf zu achten, dass es sich um Odds handelt. Ein b-Koeffizient von 0.37 für die Prädiktorvariable „Anzahl früherer Verkehrsdelikte" bedeutet eine Erhöhung für die Chance einen Rückfall zu begehen um den Faktor 0.37, wenn alle übrigen in der Vorhersage vorhandenen Prädiktoren gleich bleiben.

Interaktionen zwischen Prädiktoren und hierarchische Modellierung

Neben der Frage, welchen Einfluss mehrere Prädiktoren auf die Ausprägung der abhängigen Variablen haben, ist die Frage von Interesse, inwieweit sich die Prädiktoren gegenseitig beeinflussen. Dies wird als *Interaktion* bezeichnet. Die Interaktion zweier Merkmale lässt sich berechnen, indem die Variablen miteinander multipliziert werden und das Ergebnis als neuer Prädiktor in das Modell eingefügt wird. Durch das Hinzufügen und Wegnehmen von Prädiktoren kann weiterhin überprüft werden, welchen Ein-

fluss die einzelnen Prädiktoren und deren Interaktionen auf die Modellanpassung nehmen. Dies beantwortet die Frage, ob durch das Hinzufügen eines Prädiktors, z. B. Alter, mehr Information bezüglich der Wahrscheinlichkeit eines Rückfalls aufgeklärt wird oder nicht. Eine gute Beschreibung der Vorgehensweise bietet Field (2009).

Kodierung der Prädiktorvariablen

Prinzipiell ist die Kodierung von dichotomen Daten beliebig, d. h., normalerweise spielt es keine Rolle, ob das Datum „Merkmal liegt vor" mit 1, 2 oder jedem anderen Wert kodiert wird. Eine Kodierung mit –1 und 1 (Merkmal liegt nicht vor/Merkmal liegt vor) wird als *Effektkodierung*, und mit 0 und 1 (Merkmal liegt nicht vor/Merkmal liegt vor) als *Dummykodierung* bezeichnet. Da die logistische Regression auf Odds beruht, ist zu beachten, dass bei Effektkodierung die Regressionskoeffizienten erst umgerechnet werden müssen, um korrekt interpretiert werden zu können (Cohen, Cohen, West & Aiken, 2003).

5.8.6 Varianzanalyse für kategoriale Daten: Log-lineare Analyse

Liegen nur dichotome Daten vor, kann eine log-lineare Analyse durchgeführt werden. Die log-lineare Analyse basiert darauf, den jeweiligen Anteil der einzelnen Faktoren und Interaktionen am Gesamt-χ^2-Wert zu ermitteln. Wenn nur zwei Variablen untersucht werden, unterscheidet sich eine log-lineare Analyse nicht wesentlich von einer Kontingenz-Analyse. Hauptzweck der log-linearen Analyse ist, die Beziehung von dichotomen Variablen in komplexeren Situationen zu untersuchen, bei denen eine Vielzahl von Variablen involviert ist. Die log-lineare Analyse, die wie die logistische Regression auf der Berechnung von Odds beruht, kann zur Modellfindung *(exploratorisch)* und zur Testung von statistischen Hypothesen *(konfirmatorisch)* eingesetzt werden. Aufgrund der Komplexität der zugrundeliegenden Modelle muss hier auf eine Darstellung verzichtet werden. Eine sehr gute und ausführliche Darstellung über das Verfahren gibt Howell (2010). Eher praxisorientiert beschreiben Rudolf und Müller (2004) das Vorgehen.

5.8.7 Survival-/Ereigniszeit- oder auch Ergebniszeitanalyse

Bei den hier beschriebenen Verfahren, handelt es sich um eine Gruppe von Verfahren, die auf einer Analyse der Zeitspanne beruhen, die vergeht, bevor ein bestimmtes Ereignis eintritt. Die Bezeichnung Survival-Analyse (Überlebensanalyse) oder auch Sterbenszeit-Analyse geht auf Anwendungen dieser Verfahren in der medizinischen Forschung zurück, bei denen die Zeitspanne bis zum Tod innerhalb von Patientengruppen mit unterschiedlicher Behandlung verglichen wird.

Überlebens-Tafeln beschreiben die Zeitspanne bis zum Eintritt eines vorher definierten negativen Ereignisses, die *Überlebenszeit* (survival time). Eine Anwendung zur Überprüfung der Effektivität von Alkohol-Interlock-Programmen bestünde in der Fragestellung, ob die mittlere Zeitspanne bis zu einem Rückfall in einer Versuchsgruppe mit Alkohol-Interlock-Gerät länger ist als in einer unbehandelten Kontrollgruppe. Es wird die „Überlebensrate" (kein Rückfall) als Funktion der Zeit untersucht: *Überlebens-Funktion* (survival function). Diese Vorgehensweise ermöglicht eine sehr anschauliche Präsentation der Daten mit Hilfe von Überlebens-Tafeln. Gruppenunterschiede können mit einem χ^2-Test auf Signifikanz geprüft werden.

Es kann auch präziser untersucht werden, durch welche Merkmale die Überlebens-Zeit beeinflusst wird. Es handelt sich hierbei um die gleiche Fragestellung wie bei der Regressionsanalyse. Im Unterschied zur logistischen Regression ist hier die Zeit die vorherzusagende abhängige Variable. Ein weiterer wichtiger Unterschied liegt in dem Umstand begründet, dass bei Ereigniszeit-Analysen die Überlebenszeit häufig nicht für alle Fälle bekannt ist. Um die Effektivität von Alkohol-Interlock basierten Interventionen zu überprüfen, könnte beispielsweise die Fragestellung untersucht werden, welchen Einfluss Alter, Geschlecht, Teilnahme an einer psychotherapeutischen Intervention, frühere Verkehrsauffälligkeit und Impulsivität auf die Zeitspanne ohne weitere alkoholbedingte Auffälligkeit im Straßenverkehr nehmen. Zu diesem Zweck könnten in einer Studie vor Beginn, unmittelbar nach Beendigung und zu zwei weiteren Messzeitpunkten 1 und 2 Jahre nach Ende des Programms Daten erhoben werden. Eine Analyse der Daten muss berücksichtigen, dass der genaue Zeitpunkt bis das Ereignis (Verkehrsdelikt) eingetreten

ist, zumeist nicht genau oder, aufgrund z. B. von drop-outs, nicht für alle Versuchspersonen bekannt ist. Die Daten sind diesbezüglich *zensiert* (Cohen, Cohen, West & Aiken, 2003). Endet die Studie bevor für alle untersuchten Fälle das in Frage kommende Ereignis, bspw. „Deliktbegehung", eingetreten ist, wird von *rechts zensierten* Daten gesprochen. Bei links zensierten Daten, tritt das Ereignis bereits vor Studienbeginn ein. Dies sei nur der Vollständigkeit halber erwähnt, da es im Zusammenhang mit den hier zu untersuchenden Fragestellungen wahrscheinlich eher selten der Fall sein wird.

Das Problem der zensierten Daten muss bei der Analyse berücksichtigt werden, da andernfalls die Ergebnisse verzerrt würden. Im Gegensatz zur logistischen oder linearen Regression ist dies bei den hier vorgestellten Verfahren der Fall. Eine lesenswerte Studie, bei der ein entsprechendes Verfahren, die Cox-Regression (Cox & Snell, 1989) angewendet wurde, wurde von Rauch, Ahlin, Zador, Howard und Duncan (2011) durchgeführt. Verwirrend kann sein, dass bei den hier beschriebenen Verfahren alle Prädiktoren als Kovariate bezeichnet werden.

Überlebens-Tafeln

Die bekannteste Methode zur Berechnung von Überlebens-Tafeln ist die Kaplan-Meier-Methode oder die Produkt-Limit-Methode (Tabachnick & Fidell, 2008), welche die Überlebenszeit über nur eine einzelne Statistik (Mittelwert, Median) schätzt. Bei der Produkt-Limit-Methode wird nicht das Zeitintervall zwischen den Untersuchungen berücksichtigt, sondern die Kennwerte für die Überlebensrate werden für jeden Ereigniseintritt berechnet. Die Ergebnisse können über die Bestimmung eines Konfidenzintervalls gegen zufällig auftretende Effekte abgesichert werden. Für einen Vergleich von Gruppenunterschieden stehen verschiedene Tests zur Verfügung, von denen der Log-Rank-Test verwendet werden sollte, da er die größte Teststärke besitzt. Dies gilt allerdings nur, solange die Überlebensraten in beiden Gruppen proportional zueinander sind, was zumindest annäherungsweise anhand der grafischen Darstellung abgeschätzt werden kann: kreuzen sich die Überlebens-Kurven, ist die Annahme verletzt. In diesem Fall hat der Wilcoxon-Test die größere Power (Ziegler, Lange & Bender, 2007 b). Beide Tests basieren auf der χ^2-Verteilung.

Die Programmausgabe enthält weiterhin eine Auswertung über die Anzahl der Fälle, die für jeden Analysezeitpunkt zensiert wurde. Es handelt sich dann um die Fälle, für die das untersuchte Merkmal noch nicht eingetreten war. Eine Alternative zur Produkt-Limit-Methode stellt die Aktuarial-Methode dar, die feste Zeitintervalle für die Analyse nutzt. Tabachnick und Fidell (2008) geben eine gute anwendungsorientierte Beschreibung.

Verfahren zur Vorhersage der Überlebenszeit: Cox-Regression

Auch die Cox-Regression nimmt wie die logistische Regression einen logarithmisch-linearen Zusammenhang zwischen den Prädiktoren und der Überlebenszeit an. Dementsprechend lassen sich die Regressionskoeffizienten ähnlich wie Odds interpretieren. Speziell am Ansatz von Cox (Cox & Snell, 1989) ist, dass keine bestimmte Verteilung für die Überlebenszeiten angenommen wird, sondern dass die Effekte der Kovariaten (oder mehrerer Prädiktoren) auf das vorherzusagende Ereignis (das Überleben) über die Zeit konstant sind (Proportionalitätsannahme). Das Cox-Modell nutzt die *Hazard-Funktion*, die als Wahrscheinlichkeit definiert ist, dass das Zielereignis innerhalb eines Zeitintervalls eintritt (Ziegler, Lange & Bender, 2007 a):

$$h(t) = \frac{\text{Anzahl der Personen mit Zielereignis im Intervall beginnend bei } t}{(\text{Anzahl der Personen, ohne Zielereignis bis } t) \cdot \text{Intervallbreite}}.$$

Die Cox-Regression hat somit den Vorteil, dass sie bei unterschiedlich langen Beobachtungszeiträumen eingesetzt werden kann. Wird beispielsweise untersucht, wie viele Teilnehmer an einem Alkohol-Interlock-Programm nach 3, 7 und 8 Monaten erneut auffällig geworden sind, kann diese Wahrscheinlichkeit durch die Hazard-Funktion ermittelt werden. Die Prädiktorvariablen, die hier als Kovariate bezeichnet werden, können dichotom oder metrisch sein. Einschränkungen ergeben sich durch die Proportionalitätsannahme, die eine gleiche Zunahme der Häufigkeiten von Zielereignissen in den zu vergleichenden Gruppen annimmt. Erfolgt beispielsweise ein Vergleich zwischen zwei Gruppen mit und ohne zusätzliche psychotherapeutische Intervention, dürfen sich zwar die Zeiträume bis zum ersten Auftreten von Rückfällen unterscheiden. Die Rate, mit der die Rückfälle zunehmen, muss aber in beiden Gruppen gleich sein. Mit anderen Worten müssen die Treatment-Effekte unabhängig von der Zeit konstant bleiben.

Graphisch wird Proportionalität durch parallel verlaufende Überlebens- bzw. Sterbe-Kurven veranschaulicht. Die Modellannahme der Proportionalität muss vor Durchführung der Cox-Regression statistisch überprüft werden, wozu ein Maximum-Likelihood-Test eingesetzt werden kann. Eine Beschreibung hierzu findet sich bei Tabachnick und Fidell (2008). Ist die Voraussetzung verletzt, darf die Cox-Regression nur für die Zeiträume angewendet werden, für die Proportionalität gegeben ist. Trotz dieser Einschränkung ist die Cox-Regression das Standardverfahren bei der Vorhersage des Einflusses von Prädiktoren auf Überlebenszeiten (Tabachnick & Fidell, 2008).

Die Kaplan-Maier-Methode ist ein nonparametrisches Verfahren, was bedeutet, dass die Anwendung nicht an spezifische Annahmen über die Verteilung der zugrundeliegenden Daten geknüpft ist. Die Cox-Regression wird als semi-parametrisches Verfahren bezeichnet, da auch hier keine Modellannahme über die Verteilung der Hazard-Raten getroffen wird, aber der Einfluss der Kovariaten über einen Parameter geschätzt wird. Ist jedoch die Verteilungsform der Hazard-Funktion bekannt, z. B. normal, exponentiell usw., stehen für die Ereigniszeitanalyse auch verteilungsgebundene, sogenannte parametrische Verfahren zur Verfügung. Einen praxisbezogenen Überblick geben Tabachnick und Fidell (2008). Zwar dürfte es in der Praxis häufig sehr schwer sein, fundierte Annahmen über die vorliegende Verteilungsform zu treffen, aber wenn dies möglich ist, bieten die parametrischen Verfahren eine größere Teststärke und im Regelfall auch eine bessere Modellgüte. An eine Anwendung von parametrischen Verfahren zur Überlebenszeitanalyse sollte auch gedacht werden, wenn die Verletzung der Proportionalitätsannahme eine Analyse unter Verwendung der Cox-Regression ausschließt. In allgemeinen ist die Cox-Regression aber aufgrund ihrer großen Robustheit als Verfahren der Wahl anzusehen (Tabachnick & Fidell, 2008).

5.8.8 Varianzanalyse und (multiple) lineare Regression

Varianzanalyse

Die Varianzanalyse ist das wichtige Verfahren zur Auswertung von Experimenten (Backhaus, Plinke, Erichson & Weiber, 2006). Sie überprüft den Einfluss von unabhängigen Variablen, die

durch das Studiendesign festgelegt werden, auf eine oder mehrere abhängige Variablen. Die unabhängige Variable kann dichotom oder mehrfach abgestuft sein, die abhängige Variable muss mindestens intervallskaliert bzw. metrisch sein. Die unabhängige Variable wird auch als Faktor bezeichnet, der unterschiedliche Stufen oder Ausprägungen aufweisen kann. Die unabhängige Variable „Geschlecht" hat beispielsweise die beiden Stufen „männlich" und „weiblich". Anders als der Name zunächst vermuten lässt, erfolgt die Analyse durch eine Überprüfung von Mittelwertsunterschieden. Der Name des Verfahrens erklärt sich dadurch, dass die Varianz der abhängigen Variablen in verschiedene Bestandteile zerlegt wird. Da dies durch quadrierte Summen geschieht, wird auch von Quadratsummenzerlegung gesprochen. Durch dieses Vorgehen kann ermittelt werden, welcher Anteil an Unterschiedlichkeit (Varianz) der abhängigen Variablen auf das Treatment oder unsystematische Fehler zurückgeht. Über das Verhältnis zwischen der Effekt- und der Fehlervarianz kann als statistische Prüfgröße die *F*-Ratio bestimmt werden. Überschreitet die *F*-Ratio einen kritischen Wert, der anhand einer Tabelle zur *F*-Verteilung überprüft wird, liegt ein signifikanter Effekt vor. Diese kurze Zusammenfassung gibt die grundlegende Idee des Verfahrens wieder. Da eine Vielzahl von verschiedenen Einflussgrößen bei der Durchführung beachtet werden muss, kann es bei falscher Anwendung der Varianzanalyse zu fehlerhaften Ergebnissen kommen. Es existiert eine Reihe von guten Darstellungen des Verfahrens. Es wird empfohlen, sich vor Anwendung des Verfahrens, mit der Literatur vertraut zu machen. An erster Stelle ist die Zusammenfassung von Backhaus, Plinke, Erichson und Weiber (2006) zu nennen, aber auch bei Bortz und Schuster (2010) und Bühner und Ziegler (2009) finden sich wichtige Hinweise. Eine sehr umfassende Darstellung findet sich bei Howell (2010), die auch in die Tiefe geht. Etwas anwendungsorientierter wird das Verfahren von Tabachnick und Fidell (2008) dargestellt.

Überblick über Varianten und spezielle Anwendungen der Varianzanalyse

Es gibt eine Vielzahl von Anwendungsfeldern für die Varianzanalyse. Die Grundidee des Verfahrens bleibt dabei aber stets gleich. Wird nur eine abhängige Variable untersucht, wird von einer univariaten Varianzanalyse (ANOVA) gesprochen. Das

Verfahren lässt sich aber problemlos auf mehrere abhängige Variable erweitern. Es handelt sich dann um eine multivariate Varianzanalyse (MANOVA).

Messwiederholungen

Erfolgt die Untersuchung der gleichen Versuchspersonen über mehrere Zeiträume, muss eine Varianzanalyse für Messwiederholungen durchgeführt werden, da die Prüfverfahren der ‚normalen' Varianzanalyse aufgrund der Abhängigkeit der Daten zu falschen Ergebnissen führen würde.

Interaktionen

Bei mehreren Faktoren kann und sollte die Interaktion zwischen den Faktoren analysiert werden. Bei signifikanten Haupt- und Interaktionseffekten dürfen die unabhängigen Variablen nicht mehr isoliert interpretiert werden, da Wechselwirkungen zwischen den Variablen Haupteffekte qualifizieren oder diese verursachen können.

Kovarianzanalyse

Soll abgesehen von einer dichotomen unabhängigen Variablen der Einfluss einer weiteren, metrischen Variable auf die abhängige Variable untersucht werden, kann eine Kovarianzanalyse durchgeführt werden. Die zentrale Idee der Kovarianzanalyse ist es, den Einfluss einer metrischen Variablen, z. B. Alter, durch eine Regressionsrechnung statistisch zu kontrollieren und damit gleichsam aus der abhängigen Variable zu eliminieren.

Feste und zufällige Faktoren

Werden für die Varianzanalyse nur bestimmte von mehreren Ausprägungen eines Faktors ausgewählt, spricht man von einem *fixed model*. Handelt es sich um eine zufällige Auswahl von Faktorstufen, wird von einem *random model* gesprochen. Das gleiche gilt, wenn bei der Untersuchung alle möglichen Ausprägungen einer unabhängigen Variablen berücksichtigt werden, z. B. männliches oder weibliches Geschlecht. Enthält eine Varianzanalyse sowohl feste als auch zufällige Faktorstufen, liegt ein *mixed model* vor. Die Berechnung zwischen den Modellen unterscheidet sich

bei mehrfaktoriellen Versuchsplänen, daher ist eine Berücksichtigung der Einteilung vor der Auswertung wichtig. Weiterhin ist die Übertragbarkeit der Ergebnisse bei einem fixed model eingeschränkt. Sie haben nur für die festen Faktorstufen Gültigkeit und können nicht ohne weiteres auf andere Ausprägungen der unabhängigen Variablen übertragen werden.

Ungleiche Stichprobengröße

Sind die Gruppen nicht gleich groß, kann es möglich sein, dass die Quadratsummenzerlegung nicht nach dem Standardverfahren durchgeführt werden kann. Es muss dann eine Korrektur erfolgen, die von vielen computergestützten Auswertungsprogrammen automatisch durchgeführt wird. Da es aber auch andere Methoden als die in den Programmen voreingestellten gibt, die in besonderen Fällen zu besseren Ergebnissen führen können, empfiehlt es sich, sich bei ungleichen Stichprobengrößen mit der Thematik vertraut zu machen. Howell (2010) setzt sich hiermit intensiv auseinander und beschreibt die Vor- und Nachteile der einzelnen Vorgehensweisen.

Verletzung der Voraussetzungen

Sind eine oder mehrere Voraussetzungen für die Durchführung einer Varianzanalyse verletzt, bedeutet dies nicht, dass auf eine Anwendung verzichtet werden muss. Zum einen reagiert die Varianzanalyse bei der Verletzung bestimmter Voraussetzung robust, d. h., die Ergebnisse bleiben unverändert interpretierbar, und zum anderen existieren eine Reihe von Korrekturmöglichkeiten. Auch hier gilt, dass deren Darstellung den Rahmen dieses Überblicks sprengen würde und es sei auf die entsprechende Literatur verwiesen. Als Faustregel kann gelten, dass die Varianzanalyse bei gleicher Zellenbesetzung robust gegen die Verletzung ihrer Voraussetzungen ist (Backhaus, Plinke, Erichson und Eiber, 2006; Howell, 2010).

(Multiple) lineare Regressionsanalyse

Ein etabliertes, internationales Standardwerk zu dieser Thematik ist das Buch von Cohen, Cohen, West und Aiken (2003), auf das in diesem Abschnitt im Wesentlichen Bezug genommen wird. Ebenfalls sehr gut und anwendungsorientiert befassen sich

Backhaus, Plinke, Erichson und Weiber (2006) mit der Thematik. Ein sehr flexibles und daher vielseitig einsetzbares Verfahren ist die lineare Regressionsanalyse bei der die Ausprägung einer abhängigen Variablen durch andere Variablen (Prädiktoren) vorhergesagt werden kann. Liegen mehrere Prädiktoren vor, wird von einer multiplen Regressionsanalyse gesprochen. Im Gegensatz zur logistischen Regression muss die abhängige Variable metrisch sein, wohingegen die Prädiktoren dichotom oder metrisch sein können. Eine mögliche Fragestellung könnte etwa lauten „Lässt sich Häufigkeit von Fehlstartversuchen bei Einsatz von Alkohol-Interlock-Geräten durch andere Merkmale vorhersagen/beschreiben?" Eine sehr einfache, praktische Möglichkeit zur Analyse der Beziehung zwischen den Variablen wäre die graphische Analyse, wobei der Prädiktor auf der X-Achse und die abhängige Variable auf der Y-Achse eingezeichnet werden. Jeder Messwert erhält einen Wert für die X- und Y-Achse und kann so in den Graphen eingezeichnet werden. Soll etwa die Anzahl der Fehlstartversuche durch das Alter vorhergesagt werden, wird für jede Versuchsperson dem erhobenen Alter die entsprechende Anzahl von Fehlstartversuchen zugeordnet. Auf diese Weise erhält man ein Streuungsdiagramm für alle Messwerte. Aufgabe der Regressionsanalyse ist nun, eine Gerade zu finden, die möglichst gut zur Punktwolke „passt", d. h., bei der die Abstände zu den einzelnen Punkten minimal sind. Ist dies gelungen, ließe sich der Zusammenhang zwischen der Häufigkeit von Fehlstartversuchen und dem Alter als Funktionsgleichung dieser Geraden darstellen. Es ist dann möglich, anhand des Alters eine Schätzung der zu erwartenden Fehlstartversuche vorzunehmen. Der Versuch, eine solche Gerade per Hand zu zeichnen, zeigt schnell, dass es eine Vielzahl von möglichen Geraden gibt (tatsächlich sogar unendlich viele). Die Durchführung einer Regressionsanalyse ermöglicht es, die am besten zu den Daten passende Gerade zu bestimmen. Dies geschieht im Prinzip dadurch, dass für jeden Messwert eines jeden Prädiktors die Abweichung der quadrierten Differenzen zur Geraden minimiert wird. Natürlich ist eine perfekte Anpassung in den seltensten Fällen möglich. Die nicht zu beseitigende Abweichung zwischen der Geraden und den einzelnen Punkte wird als *Residuum* bezeichnet. Formalisiert lässt sich die Regressionsgleichung als

$$Y = (b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_n X_n)$$

schreiben. Hierbei bezeichnet Y die abhängige Variable, X_n die Prädiktoren und b die Regressionskoeffizienten. Mit b_0 wird das konstante Glied bezeichnet. Das konstante Glied liefert den Wert auf der Y-Achse, den die abhängige Variable annimmt, wenn alle Prädiktoren gleich Null sind. Graphisch betrachtet schneidet die Regressionsgerade bei diesem Wert die Y-Achse.

Überprüfung der Ergebnisse und Interpretation

Da es wie bereits gesagt unendlich viele Regressionsgerade gibt, muss überprüft werden, wie gut die anhand der empirischen Daten ermittelte Gleichung die abhängige Variable vorhersagen kann. Dies geschieht über die Berechnung des Bestimmtheitsmaßes R^2. R^2 kann Werte zwischen 0 und 1 annehmen und gibt an, wie groß der Anteil an der Varianz der abhängigen Variablen ist, der durch die Prädiktoren vorhergesagt wird. Hierzu wird das Verhältnis aus erklärter Streuung und der Gesamtstreuung gebildet. Nimmt in einer Regression der Fehlstartversuche auf das Alter R^2 zum Beispiel den Wert 0.14 an, bedeutet dies, dass durch das Alter 14 % der Varianz (Unterschiedlichkeit) der Fehlstartversuche erklärt werden kann. Das Bestimmtheitsmaß gibt Auskunft über die Stärke des gefundenen Zusammenhangs und ist ein Parameter für die Modellgüte. Weiterhin muss überprüft werden, ob die ermittelte Regressionsbeziehung zufällig zustande gekommen ist oder tatsächlich vorhanden ist. Dies kann mittels einer F-Statistik überprüft werden. Die einzelnen Regressionskoeffizienten werden durch die Errechnung von Konfidenzintervallen gegen den Zufall abgesichert. Wie können bei signifikanten Ergebnissen die Regressionskoeffizienten interpretiert werden? Allgemein gesagt, geben die Regressionskoeffizienten an, um welchen Betrag die abhängige Variable zunimmt, wenn der Prädiktor um einen Punkt erhöht wird. Ein Regressionskoeffizient der Variable Alter von b = –0.47 bedeutet, dass sich die Anzahl der Fehlversuche um 0.47 verringert, wenn das Alter um 1 Jahr höher ist. Das Vorzeichen bestimmt die Richtung des Zusammenhangs. Enthält das Regressionsmodell mehrere Prädiktoren, empfiehlt es sich eine Standardisierung der Ausgangswerte vor der Regressionsanalyse vorzunehmen, um die Vergleichbarkeit der Regressionskoeffizienten zu gewährleisten. Man spricht in diesem Zusammenhang auch von einer z-Standardisierung. Die meisten Statistikprogramme enthalten hierzu Prozeduren. Bei der

z-Transformation wird von jedem Messwert der Mittelwert der Variable subtrahiert und dieser Differenzwert durch die Streuung der Variablen geteilt. Nach dieser linearen Transformation wird der Mittewert aller z-standardisierten Variablen 0 und ihre Standardabweichung 1. Mit diesem Verfahren werden Variablen mit ganz unterschiedlichen Skalenwerten und auch deren Regressionskoeffizienten direkt vergleichbar. Beispielsweise könnte man neben dem Lebensalter auch die Testwerte einer Fragebogenskala zur Impulsivität als Prädiktor heranziehen. Eine Zunahme des Testwertes um 10 Punkte entspräche nicht einer Zunahme des Alters um 10 Jahre. Nach erfolgter Standardisierung ist eine vergleichende Interpretation möglich. Eine z-Standardisierung bedeutet, dass die Variablen einen Mittelwert von 0 und eine Streuung von 1 aufweisen. Als Folge nimmt das konstante Glied den Wert Null an und der Zusammenhang zwischen Prädiktoren und der abhängigen Variablen lässt sich jetzt in Standardabweichungseinheiten ausdrücken. Weist ein z-standardisierter Regressionskoeffizient den Wert von 0.5 auf, bedeutet die Zunahme des Prädiktors um eine Standardabweichung eine Steigerung der abhängigen Variablen um eine halbe Standardabweichung.

Interaktionen zwischen Prädiktoren und hierarchische Modellierung

Wie bei der logistischen Regression können auch für die lineare Regression Interaktionseffekte bestimmt und in die Regressionsanalyse einbezogen werden. Das gleiche gilt für die Durchführung einer hierarchischen Modellierung. Zur Berechnung von Interaktionen empfiehlt es sich, wie bei der logistischen Regression beschrieben, eine z-Transformation durchzuführen. Ziel ist es, einer Interkorrelation zwischen den Prädiktorvariablen und den Interaktionen entgegenzuwirken. Die Modellgüte bzw. der Erklärungswert des Regressionsmodells sinkt allgemein, wenn die Prädiktoren einen hohen Zusammenhang aufweisen, also stark miteinander korrelieren. Dieses Phänomen wird als *Multikollinearität* bezeichnet und diese sollte so gering wie möglich ausfallen. Eine alternative Methode zur Standardisierung ist das *Zentrieren* der Variablen. Hierzu wird von jedem individuellen Wert der Mittelwert der Prädiktorvariablen subtrahiert. Dies führt dazu, dass anschließend eine mittlere Ausprägung eines Merkmales den Wert 0 erhält. Allerdings behalten die Variablen dann ihre

Tab. 5.6: Entscheidungshilfe zur Auswahl eines geeigneten Verfahrens in Anlehnung an Backhaus, Plinke, Erichson & Weiber (2006, S. 430)

		Unabhängige Variable	
		Metrisch	nominal
Abhängige Variable	Metrisches Skalenniveau	Regressionsanalyse	Varianzanalyse Regressionsanalyse
	Nominales Skalenniveau	Logistische Regression	Logistische Regression Log-lineare Analyse

unterschiedliche Metrik. Das Verfahren wird hier beschrieben, da es in manchen Lehrbüchern Erwähnung findet.

Kodierung der Prädiktorvariablen

Auch hinsichtlich der Kodierung der Prädiktorvariablen gilt das gleiche Vorgehen wie für die logistische Regression (vgl. Kapitel 5.8.5).

Die in Frage kommenden statistischen Verfahren hängen natürlich auch immer mit der Forschungsstrategie und den Hypothesen, die anhand einer übergeordneten Fragestellung gebildet werden, zusammen. Eine Entscheidungshilfe zur Auswahl des Testverfahrens ist in Tabelle 5.6 dargestellt. Ein wesentlicher Unterschied zwischen der vorgestellten Kaplan-Maier-Methode und der Cox-Regression zur Ereigniszeit-Analyse besteht darin, dass bei letzter mehrere Prädiktoren (Kovariaten) in die Analyse einbezogen werden können, die zudem nominales oder metrisches Skalenniveau aufweisen dürfen. Die abhängige Variable ist in beiden Fällen die Zeit, die vergeht, bis ein bestimmtes Ereignis eingetreten ist. Die Auswahl des statistischen Verfahrens ist für die Datenauswertung von entscheidender Bedeutung, da sich bei unterschiedlichen statistischen Testverfahren verschiedene Interpretationen der Daten ergeben, wobei aber die vorgestellten Voraussetzungen für die statistischen Verfahren beachtet werden müssen. Es ist empfehlenswert, bereits während der Planungsphase einer Untersuchung Überlegungen zu den zu verwendenden Auswertungsverfahren anzustellen und sich mit der Materie vor Erhebungsbeginn hinreichend vertraut zu machen. Liegen die Daten bereits vor und es stellt sich heraus, dass die Erhebung methodische Mängel aufweist, lässt sich dies später eventuell nicht mehr korrigieren.

Literatur

Backhaus, K., Plinke, W., Erichson, B. & Weiber, R. (2006). Multivariate Analysemethoden: Eine anwendungsorientierte Einführung (11., überarb.). Berlin: Springer

Bortz, J. & Schuster, C. (2010). Statistik für Human- und Sozialwissenschaftler. (7., vollst. überarb. u. erw). Berlin: Springer.

Bühner, M. & Ziegler, M. (2009). Statistik für Psychologen und Sozialwissenschaftler. München: Pearson.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2. rev. edition.). Hillsdale N.J: Lawrence Erlbaum Associates.

Cohen, J., Cohen, P., West, S. G. & Aiken, L. S. (2003). Applied multiple regression/correlation analysis for the behavioral sciences (3. ed.). Mahwah, NJ: Erlbaum.

Cottler, L. B., Compton, W. M., Ben-Abdallah, A., Horne, M. & Claverie, D. (1996). Achieving a 96.6 percent follow-up rate in a longitudinal study of drug abusers. Drug and Alcohol Dependence, (41), 209–217.

Cox, D. R. & Snell, E. J. (1989). Analysis of binary data (2nd). London: Chapman and Hall.

Dau, W. (2012, 19. Mai). Kompass – eine stationäre Kurzintervention für junge Cannabis-/Partydrogenpatienten. Vortrag beim 30. Symposium der Fachgruppe Klinische Psychologie und Psychotherapie. Abtei Neumünster: Luxemburg. [unveröffentlichtes Vortragsmanuskript].

Erdfelder, E., Faul, F. & Buchner, A. (1996). GPOWER: A general power analysis program. Behavior Research Methods, Instruments and Computers, (28), 1–11.

Field, A. (op. 2009). Discovering statistics using SPSS (and sex and drugs and rock ‚n' roll) (3rd). Los Angeles, Calif.: Sage.

Franke, G. H. (2000). Brief Symptom Inventory von L. R. Derogatis.: Kurzform der SCL-90-R. Deutsche Version. Göttingen: Hogrefe.

Howell, D. C. (2010). Statistical methods for psychology (7th ed.). Belmont, Calif.: Wadsworth Cengage Learning.

Katz, K. A. (2006). The (Relative) Risks of Using Odds Ratios. Archives of Dermatology, 142(6), 761–764. doi:10.1001/archderm.142.6.761.

Nachtigall, C. & Wirtz, M. (2009). Wahrscheinlichkeitsrechnung und Inferenzstatistik. Statistische Methoden für Psychologen. Teil 2. Weinheim: Juventa.

Morris, J. A. & Gardner, M. J. (1988). Calculating confidence intervals for relative risks (odds ratios) and standardised ratios and rates. British Medical Journal, (296), 1313–1316.

Rauch, W. J., Ahlin, E. M., Zador, P. L., Howard, J. M. & Duncan, G. D. (2011). Effects of administrative ignition interlock license restrictions on drivers with multiple alcohol offenses. Journal of Experimental Criminology, 7(2), 127–148. doi:10.1007/s11292-010-9118-0.

Rudolf, M. & Müller, J. (2004). Multivariate Verfahren: Eine praxisorientierte Einführung mit Anwendungsbeispielen in SPSS. Göttingen: Hogrefe.

Rumpf, H.-J. (2010). Was ist ein Odds Ratio? PPmP – Psychotherapie Psychosomatik Medizinische Psychologie, 60(05), 190–191. doi:10.1055/s-0030-1248432

Sackett, D. L., Deeks, J. J. & Altmann, D. G. (1996). Down with odds ratios! Evidence Based Medicine, (1), 164–166.

Sánchez-Meca, J., Marín-Martínez, F. & Chacón-Moscoso, S. (2003). Effect-Size Indices for Dichotomized Outcomes in Meta-Analysis. Psychological Methods, 8(4), 448–467. doi:10.1037/1082-989x.8.4.448.

Scott, C. K. (2004). A replicable model for achieving over 90% follow-up rates in longitudinal studies of substance abusers. Drug and Alcohol Dependence, 74(1), 21–36. doi:10.1016/j.drugalcdep.2003.11.007.

Tabachnick, B. G. & Fidell, L. S. (2008). Using multivariate statistics (5. ed.). Boston: Pearson/Allyn & Bacon.

Toutenburg, H., Heumann, C. & Nittner, T. (2009). Statistische Methoden bei unvollständigen Daten. In H. Holling (Ed.), Grundlagen und statistische Methoden der Evaluationsforschung (pp. 649–692). Göttingen: Hogrefe.

Yule, G. U. (1912). On the Methods of Measuring Association Between Two Attributes. Journal of the Royal Statistical Society, (75), 579–652.

Ziegler, A., Lange, S. & Bender, R. (2007a). Überlebenszeitanalyse: Die Cox-Regression. DMW – Deutsche Medizinische Wochenschrift, 132(S 01), e42-e44. doi:10.1055/s-2007-959039.

Ziegler, A., Lange, S. & Bender, R. (2007b). Überlebenszeitanalyse: Der Log-Rang-Test. DMW – Deutsche Medizinische Wochenschrift, 132(S 01), e39-e41. doi:10.1055/s-2007-959040.