REINER KURZHALS

WILEY-SCHNELLKURS

STATISTIK

- Die Grundlagen auf einen Blick
- Vom statistischen Messen bis zu multivariaten Modellen
- Schnelltest: Mit Übungsaufgaben und Lösungen

WILEY

Der Einstiegstest

Zum Einstieg können Sie mit diesen Aufgaben testen, wo Sie in der Statistik noch Lücken haben, die Sie schließen sollten. Dabei steht jede Aufgabe für ein Kapitel im Buch. Sollten Sie feststellen, dass Sie eine Aufgabe nicht lösen können, ist es vielleicht schlau, zuerst dieses Kapitel durchzuarbeiten, wenn Sie nur noch wenig Zeit haben.

Aufgabe 1: Bei welchen Merkmalen gruppieren und bei welchen klassieren wir? Was ist eigentlich der Unterschied zwischen gruppieren und klassieren? Wenn Sie Probleme mit der Aufgabe hatten, lesen Sie bitte Kapitel 1.

Aufgabe 2: Warum reicht die Angabe eines Mittelwerts bei einem Vergleich von zwei Stichproben oftmals nicht aus bzw. kann sogar bei der Anwendung zu Fehlern führen, wenn Sie nicht andere Kennzahlen hinzufügen? Wenn Sie Probleme mit der Aufgabe hatten, lesen Sie bitte Kapitel 2.

Aufgabe 3: Was ist der Unterschied zwischen einer linearen Korrelation und einer linearen Regression? Wenn Sie Probleme mit der Aufgabe hatten, lesen Sie bitte Kapitel 3.

Aufgabe 4: Was ist eine saisonbereinigte Zeitreihe, z. B. eine saisonbereinigte Arbeitslosenzahl? Wenn Sie Probleme mit der Aufgabe hatten, lesen Sie bitte Kapitel 4.

Aufgabe 5: Was unterscheidet eine Wahrscheinlichkeit von einer bedingten Wahrscheinlichkeit? Wenn Sie Probleme mit der Aufgabe hatten, lesen Sie bitte Kapitel 5.

Aufgabe 6: Wozu brauchen wir eigentlich so viele unterschiedliche Wahrscheinlichkeitsmodelle – hat dieser Zentrale Grenzwertsatz damit etwas zu tun? Wenn Sie Probleme mit der Aufgabe hatten, lesen Sie bitte Kapitel 6.

Aufgabe 7: Was ist der Unterschied zwischen einem einseitigen und einem zweiseitigen Konfidenzintervall? Wenn Sie Probleme mit der Aufgabe hatten, lesen Sie bitte Kapitel 7.

Aufgabe 8: Was ist das bessere statistische Verfahren, ein 95 % Konfidenzintervall oder ein statistischer Signifikanztest zum 5 % Niveau? Wenn Sie Probleme mit der Aufgabe hatten, lesen Sie bitte Kapitel 8.

Antworten Für Den Einstiegstest

Aufgabe 1: Das Gruppieren ist bei allen Merkmalstypen möglich, besonders sinnvoll aber bei qualitativen (also nominal und ordinal skaliert) und diskreten quantitativen Merkmalen. Klassieren kommt bei stetigen quantitativen Merkmalen vor. Man gruppiert, wenn man übersichtlich viele Merkmalswerte hat, die man inhaltlich und sachlogisch voneinander trennen und so sinnvoll Gruppen bilden kann. Liegen zu viele Merkmalsausprägungen vor und muss man so erst nach einem festzulegenden Prinzip Trennwerte und somit Klassen definieren, dann klassiert man.

Aufgabe 2: Als Erklärungsbeispiel wähle ich den Vergleich des mittleren Baumalters von zwei Waldaufforstungsgebieten, um anhand des Alters Maßnahmen wie z. B. Abholzung oder Renaturierung einleiten zu können. Beide Waldgebiete könnten anhand des mittleren Baumalters gleich sein, z. B. im Mittel 50 Jahre alt. Beide Waldgebiete könnten jedoch extrem unterschiedliche Streuwerte haben. Das heißt z. B., im ersten Waldgebiet könnten alle Bäume zu einer Zeit vor ca. 50 Jahren gepflanzt worden sein. Im zweiten Waldgebiet könnte ein gutes Drittel vor ca. 5 Jahren angepflanzt, ein gutes Drittel vor ca. 100 Jahren, der Rest irgendwann ohne System in den Jahren dazwischen angepflanzt worden sein. Beide Waldgebiete sind also bezüglich des mittleren Alters gleich, bezüglich der Streuung der Alterswerte jedoch extrem unterschiedlich. Dies gilt es dann beim Einleiten von Maßnahmen wie z. B. Renaturierung oder Abholzung, auf Grundlage von Mittelwerten und Streuwerten, zu berücksichtigen, sodass nicht fälschlicherweise beide

Waldgebiete bezüglich des Alters als homogen zu betrachten sind, sondern als heterogen.

Aufgabe 3: Die Korrelation beschreibt die Stärke des linearen Zusammenhangs (ob es einen gerichteten linearen Zusammenhang gibt), während die lineare Regression eine Ursache-Wirkungsbeziehung des Zusammenhangs unterstellt. Konkret heißt das bei der Regression, dass Sie ein Merkmal als abhängiges Merkmal (Y) und ein Merkmal als unabhängiges Merkmal (X) einteilen und so die Regressionsgleichung formulieren: die Merkmalswerte von Y sind in einer gewissen Weise abhängig von den Merkmalswerten von X. Diese gewisse Weise wird ausgedrückt durch die Regressionskoeffizienten.

Aufgabe 4: Eine Zeitreihe besteht aus mehreren Komponenten, die den Verlauf der Zeitreihe beschreiben. Der Trend zeigt die allgemeine Richtung an. In der Regel wird der Trend von periodisch wiederkehrenden Signalen, den sogenannten saisonalen Effekten begleitet, die in der Regel bekannt sind (z. B. gibt es im Winterquartal in vielen Branchen deutlich höhere Arbeitslosenzahlen als im Sommerquartal). Um einen zukünftigen Trend besser vorauszusagen, werden diese periodischen, saisonalen Schwankungen systematisch erfasst, meist über Saisondurchschnittswerte herausgerechnet und damit der Trend der Zeitreihe durch das Subtrahieren dieser Saisondurchschnittswerte geglättet. Das nennt man Saisonbereinigung. Eine saisonbereinigte Arbeitslosenzahl ist demnach eine Arbeitslosenzahl, von der wir den saisonalen Effekt abziehen. Wenn es zum Beispiel im 1. Quartal 2014 4,4 Millionen Arbeitslose in Deutschland gab, im 4. Quartal 2013 aber 4 Millionen Arbeitslose, dann mag das auf den ersten Blick wie eine 10 %ige Verschlechterung der Arbeitslosensituation

erscheinen. Wenn allerdings in den letzten 10 Jahren die Verschlechterung der Arbeitsmarktsituation vom 4. Quartal zum 1. Quartal bei durchschnittlich 12 % lag, dann lag dieses Mal keine Verschlechterung sondern saisonbereinigt eine Verbesserung vor.

Aufgabe 5: Die Wahrscheinlichkeit, mit einem sechsseitigen Würfel eine Sechs zu würfeln (A = eine Sechs würfeln) beträgt $P(A) = \frac{1}{6}$. Wenn man aber die Information erhält, dass zuvor eine gerade Augenzahl gewürfelt wurde (B = es wurde eine 2 oder 4 oder 6 gewürfelt), ist die Wahrscheinlichkeit eine Sechs zu würfeln unter dieser Bedingung neu zu berechnen und zwar dieses Mal nach den Gesetzen der bedingten Wahrscheinlichkeit:

$$P(A|B) = \frac{P(A \cap B)}{P(A)} = \frac{P(A) + P(B) - P(A \cup B)}{P(B)} = \frac{\frac{1}{6} + \frac{3}{6} - \frac{3}{6}}{\frac{1}{2}} = \frac{\frac{1}{6}}{\frac{1}{2}} = \frac{2}{6} = \frac{1}{3}.$$

Eine bedingte Wahrscheinlichkeit verändert durch eine formulierte Bedingung die Wahrscheinlichkeit eines bestimmten Ereignisses. Diese Veränderung lässt sich über die Formel für die bedingte Wahrscheinlichkeit berechnen.

Aufgabe 6: In der Praxis gibt es unterschiedlichste Probleme, die man mit Hilfe von intelligenten Datenanalysen lösen kann. In manchen Fällen passen wir Daten an ein schon bekanntes Wahrscheinlichkeitsmodell an. Diese Wahrscheinlichkeitsmodelle sind unsere speziellen Wahrscheinlichkeitsverteilungen. Je nachdem, um was für eine Problemkategorie es sich handelt, wählt man eine dazu passende Wahrscheinlichkeitsmodellart. Ein Zeitmodell beantwortet z. B. die Fragestellung nach der

Wahrscheinlichkeit des Eintretens eines Ereignisses in einem festen Zeitabschnitt, ein Urnenmodell beantwortet die Fragestellung nach der Wahrscheinlichkeit eine bestimmte Kugel zu ziehen, wenn ein gewisser Anteil an solchen Kugel in der Urne vorhanden ist.

Der zentrale Grenzwertsatz besagt, dass die Verteilung von Mittelwerten einer Zufallsvariablen unter bestimmten Bedingungen näherungsweise normalverteilt ist, egal wie die Zufallsvariable selbst verteilt ist. Er sorgt also für eine Vereinfachung in dem Sinne, dass er viele der speziellen Wahrscheinlichkeitsmodelle „überflüssig macht“, indem man auf die Normalverteilung approximiert. Der zentrale Grenzwertsatz hat also nichts mit diesen vielen, unterschiedlichen Modellen zu tun, im Gegenteil, er reduziert diese vielen Modelle auf das wichtigste Modell, die Normalverteilung.

Aufgabe 7: Konfidenzintervalle benutzt man, wenn man einen Parameter aus der Grundgesamtheit, zum Beispiel den Mittelwert μ, schätzen möchte und einem ein Punktschätzer, wie zum Beispiel der arithmetische Mittelwert $\bar{x}$, dafür nicht ausreicht. Stattdessen umgibt man diesen Punktschätzer mit dem Stichprobenfehler $\pm z_{1-\alpha} \cdot \sigma_{\bar{x}}$ als symmetrisches Intervall, in dem sich der unbekannte aber wahre Parameter μ mit einer Wahrscheinlichkeit von 1 – α befindet. Anhand der zu beantwortenden Frage erkennt man, ob ein Intervall zu beiden Seiten des wahren aber unbekannten Parameters errechnet werden soll oder nur zu einer Seite. Diese Fragestellungen zielen darauf ab, dass man an einer unteren oder einer oberen Grenze der Berechnung eines Konfidenzintervalls interessiert ist. Gibt es keine explizite Benennung einer solchen Einseitigkeit, liegt eine zweiseitige Fragestellung vor. Eine einseitige

Fragestellung beinhaltet immer Signalwörter, wie zum Beispiel: „Wie hoch ist die mittlere Lebensdauer *mindestens?“* oder „Wie häufig wurde diese Partei *höchstens* genannt?“ oder „Wie hoch ist die *maximale* Sterblichkeitsrate?“ Im Gegensatz dazu beinhalten Fragestellungen zu zweiseitigen Konfidenzintervallen keine Signalwörter, wie zum Beispiel: „Wie hoch ist die mittlere Lebensdauer?“ oder „Wie häufig wurde diese Partei genannt?“ oder „Wie hoch ist die Sterblichkeitsrate?“ Der Unterschied in der Berechnung eines einseitigen gegenüber einem zweiseitigen Konfidenzintervall ist neben der Berechnung von entweder nur $(+z_{1-\alpha} \cdot \sigma_{\bar{x}})$ oder nur $(-z_{1-\alpha} \cdot \sigma_{\bar{x}})$, das heißt nur einer Seite des zweiseitigen Konfidenzintervalls, noch die Tatsache, dass die Irrtumswahrscheinlichkeit a nicht mehr durch 2 geteilt wird.

Aufgabe 8: Hier gibt es kein besser oder schlechter. Allgemein befasst sich die Schließende Statistik mit der Fragestellung, möglichst gesichert aus Stichproben Informationen über das Verhalten eines Merkmals in der Grundgesamtheit zu erhalten. Da wir meist nicht genügend Zeit und Geld haben, die Grundgesamtheit zu betrachten, sind wir darauf angewiesen, uns auf die Ergebnisse einer Stichprobe zu verlassen. Wir unterscheiden dabei zwei mögliche Ansätze, dieses Problem zu lösen. Der eine Ansatz geht über die Konfidenzintervalle, wobei wir zum Beispiel einen 95 % Wahrscheinlichkeitsbereich abschätzen, in dem die uns interessierende Größe liegen könnte. Der andere Ansatz geht über statistische Testverfahren, wobei wir zuerst eine a priori Hypothese für die uns interessierende Größe aufstellen, um dann mit einer Entscheidungsregel ein Stichprobenergebnis mit einem theoretisch unter Gültigkeit der Nullhypothese zu erwartendem Ergebnis zu vergleichen. So sind wir in der Lage, Ergebnisse aus

der Stichprobe als zufällig oder als signifikant (erheblich) zu erklären und das mit einer Irrtumswahrscheinlichkeit von maximal 5 %.

Reiner Kurzhals

Wiley-Schnellkurs Statistik

WILEY

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie;

detaillierte bibliografische Daten sind im Internet über http://dnb.d-nb.de abrufbar.

1. Auflage 2015

Print ISBN: 978-3-527-53017-5

ePub ISBN: 978-3-527-69583-6

mobi ISBN: 978-3-527-69584-3

Inhalt

1
Einführung

Haben Sie sich diese Frage schon einmal gestellt:

Warum muss ich eigentlich eine Statistikvorlesung hören und eine Statistikprüfung schreiben?

Der größte Teil aller Studenten, ich schätze etwa 75 %, an allen Hochschulen auf dieser Welt, muss eine Statistikvorlesung besuchen und eine Statistikprüfung ablegen, in allen möglichen Fachrichtungen (Psychologie, Sozialwissenschaften, Betriebswirtschaft, Medizin, Informatik, usw.). Sie müssen das wahrscheinlich auch – *warum eigentlich*?

Der Grund ist ganz einfach, nur wird er in wissenschaftlichen Büchern gerne komplett ignoriert oder kompliziert umschrieben. Die Antwort ist schnell geschrieben:

- Sie haben eine Vermutung (Hypothese) und möchten diese gerne beweisen; oftmals treffen Sie damit eine Entscheidung. Eine Vermutung könnte beispielsweise lauten: ein neues Medikament heilt besser als ein altes Medikament.
- Sie können diese Vermutung in der Regel wegen fehlender Zeit und mangelndem Budget nicht an allen Objekten einer Grundgesamtheit belegen, also müssen Sie Ihre Vermutung an einer Teilmenge, meist eine Stichprobe, überprüfen.
- Das Ergebnis einer Stichprobenuntersuchung, in der Regel ein Mittelwert (z. B. mittlerer Blutdruck), beinhaltet eine gewisse Schwankungsbreite, da jede mögliche Stichprobe ja auch ein wenig anders aussieht, den Stichprobenfehler. Damit das Ergebnis und die daraus resultierende Entscheidung mit größtmöglicher Sicherheit auch für die Grundgesamtheit aller Patienten gilt, liefert die Statistik Methoden, mit denen Sie mit einer vertretbaren Irrtumswahrscheinlichkeit eine Entscheidung für oder gegen Ihre anfängliche Vermutung treffen können.

Fast alle Wissenschaftler und Anwender haben das gleiche Problem mit der Überprüfung einer Vermutung auf Grundlage einer Stichprobe. Weil der Weg über ein statistisches Testverfahren der

einzige ist, solch eine Vermutung zu belegen, unter Einhaltung einer vertretbaren Irrtumswahrscheinlichkeit, müssen Sie eine Statistikvorlesung hören und eine Statistikprüfung ablegen. Deshalb sollen Studenten fast aller Fachrichtungen die Methodenlehre der Statistik in den Grundzügen beherrschen.

Um Ihnen den Zugang zu dieser statistischen Methodenlehre so einfach wie möglich zu machen, habe ich versucht, so viel Wissenschaft wie möglich heraus zu lassen und so viel wie möglich an Anschauung und Beispielen mit aufzunehmen. Dazu habe ich versucht aus 15 Jahren Lehrerfahrung und vielen Statistikseminaren und Beratungen in der freien Wirtschaft, das nötigste Wissen für die einfachste Vermittlung von Statistik zu extrahieren und mit vielen Beispielen zu beschreiben.

Am Ende möchten Sie nur die Klausur bestehen, das habe ich versucht im Auge zu behalten. Darüber hinaus möchte ich aber dennoch versuchen, Ihnen die Welt der Statistik einfach vertrauter zu machen.

Was haben Amazon und Google mit Statistik zu tun?

Beide Internet-Unternehmen gehören zu den weltweit erfolgreichsten Unternehmen unserer Zeit (bezogen auf das Jahr 2014). Das Geschäftsmodell beider Unternehmen basiert auf simplen, statistischen Datenanalysen, die ich Ihnen in diesem Buch auch näher bringen möchte. Ich verwende hier einige Fachbegriffe, die an dieser Stelle teilweise schwer verständlich sind, aber sehr viel klarer werden, sobald Sie dieses Buch durchgearbeitet haben.

- **Anwendungsfall Amazon**: Geschickte Berechnungen von statistischen Abhängigkeiten zwischen den Datensätzen eines Datenbestandes mit Hilfe von Assoziationsanalysen, die mit „Wenn-dann-Regeln“ beschrieben werden. Ein Beispiel für eine einfache Regel wäre: „Wenn ein Kunde das Produkt Barilla Tagliatelle kauft, dann kauft er in 85 Prozent der Fälle auch einen Chianti“. Mit diesem einfachen *Empfehlungsdienst (Recommender System)* konnte Amazon seine Umsätze, z. B. im Buchgeschäft, um bis zu 30 % erhöhen. Bei einem Milliardenumsatz macht das einen Unterschied.
- **Anwendungsfall Google**: Geschickte Verbindung dreier verschiedener Zielkunden (Suchende, Webseitenbetreiber, Werbetreibende) machen das Geschäftsmodell von Google aus.

Dabei ist das Suchen immer noch das Kerngeschäft von Google. Die Suchmaschine basiert auf einen Algorithmus, der einen zufällig durch das Netz surfenden Benutzer nachbildet. Die Wahrscheinlichkeit, mit der dieser auf eine Webseite stößt, korreliert mit dem PageRank, einer Methode, um die Linkpopularität einer Web-Seite zu berechnen. So werden die Seiten entsprechend ihrer Wertigkeit sortiert, um so eine Ergebnisreihenfolge bei einer Suchabfrage herzustellen.

Diese beiden Anwendungsfälle zeigen, wie Sie mit intelligenter, statistischer Datenanalyse erfolgreich Geschäfte machen können. Dazu passt es gut, dass der größte deutsche IT, Telekommunikations- und Neue Medien-Branche Verband „BITKOM“ Daten als “**vierten Produktionsfaktor**“ in der digitalen Welt bezeichnet (neben Arbeitskraft, Kapital und Rohstoffen). Diese neue Macht im Wirtschaftsleben ist sehr nah unter uns. Fast jeder Leser ist mit Google oder Amazon schon in Berührung gekommen.

Viele Bereiche der Wissenschaft, Wirtschaft und der Verwaltung haben mit Statistik direkt oder indirekt zu tun. Google und Amazon sind nur zwei Aufsehen erregende Beispiele. In Wahrheit ist die Anwendung der Statistik schon längst im Berufsleben verankert. Durch die steigende Möglichkeit der Datenspeicherung und Datenanalysegeschwindigkeit wird die intelligente Datenanalyse eine immer größere Bedeutung in vielfältigen Bereichen annehmen.

Beschreibende und Schließende Statistik

Die vielen Methoden und zugehörigen Anwendungsbereiche der Statistik sind teilweise recht komplex und unübersichtlich, manchmal selbst für erfahrene Statistiker. Dazu kommen noch moderne Begriffe wie Data Mining, Big Data, Smart Data, Data Science und vieles mehr, von denen wir meist nur eine grobe Ahnung haben, was sie bedeuten und wie wir sie in den schon existierenden Methoden und Anwendungen einzuordnen haben.

Hier werden wir uns mit der Basis der statistischen Methoden und Anwendungen auseinandersetzen. Dazu gehört die Unterscheidung der Statistik in die zwei übergeordneten methodischen Bereiche Beschreibende und Schließende Statistik.

Zur Vereinfachung hier eine Kurzdefinition der wesentlichen Inhalte dieser Bereiche anhand ihrer Hauptanwendungen in zwei Sätzen:

- **Beschreibende Statistik**. Hauptanwendung: Messwerte eines Merkmals, wie zum Beispiel Blutdruckmessungen, kompakt und übersichtlich mit Kenngrößen zusammengefasst darstellen, zum Beispiel über einen mittleren Blutdruckwert. Wir nennen sie manchmal auch deskriptive Statistik.
- **Schließende Statistik**. Hauptanwendung: Vermutungen über einen Parameter einer Grundgesamtheit, zum Beispiel eine Blutdruckverbesserung nach Medikamentengabe, anhand einer Stichprobe zu prüfen und das Stichprobenergebnis verallgemeinernd auf die Grundgesamtheit zu übertragen, mit Angabe einer vertretbaren, statistischen Irrtumswahrscheinlichkeit. Wir nennen sie manchmal auch induktive Statistik.

Die Verbindung zwischen diesen beiden Bereichen der Statistik sind Kennzahlen, die eine Datenmenge, in der Regel eine Stichprobe, übersichtlich und zusammenfassend charakterisieren, wie zum Beispiel der Stichprobenmittelwert. Mit solchen Kennzahlen versuchen wir dann mithilfe der statistischen Testmethoden Aussagen über die entsprechenden wahren, aber unbekannten Kennzahlen der Grundgesamtheit zu treffen. Ein Beispiel wäre, dass eine neues Medikament durchschnittlich den Blutdruck erheblich (Statistiker sagen: signifikant) stärker senkt, als das jetzige Standardprodukt auf dem Markt, und zwar für alle Patienten. Solche Aussagen basieren letztlich auf Stichprobendaten und sind daher mit einer vertretbaren Irrtumswahrscheinlichkeit zu beurteilen. Sie können diese Unsicherheiten aber mithilfe von den noch in den folgenden Kapiteln zu erklärenden statistischen Testmethoden kontrollieren und quantifizieren, also in Zahlen ausdrücken.

Das könnte dann zu folgenden Aussagen führen: Mit einer Irrtumswahrscheinlichkeit von höchstens 2% können wir die Vermutung der signifikanten Überlegenheit des neuen Medikaments gegenüber dem bisher gebräuchlichen Medikament bestätigen. Mit anderen Worten, Sie werden mit einer großen Sicherheit davon ausgehen können, dass Sie sich hier richtig entschieden haben.

Das Ziegenproblem

An dieser Stelle möchte ich Ihnen gerne das berühmte Ziegenproblem (Monty-Hall-Dilemma) vorstellen. Dieses Problem möchte ich Ihnen näher bringen, weil es zum einen sehr gerne von

Unternehmen in Assessment Centern zur Selektion von Bewerbern genutzt wird und zum anderen für diejenigen, die Spaß an Hochleistungsdenken haben oder gerne auch mal ihre Grenzen erfahren möchten; das spornt bekanntlich zu mehr Leistung an.

Warnung

Das Ziegenproblem

Dieses Beispiel ist meiner Erfahrung nach für einen Teil der Leser sehr schwer verständlich und führt zu heftigen Diskussionen. Also bitte mit Vorsicht lesen! Sie können das hier erwähnte Ziegenproblem und die Lösung im Kapitel 6 auch komplett ignorieren. Es kommt in einer Statistikklausur quasi nicht vor, weil es zu kontrovers und schwierig ist. Es ist aber auch anregend.

Das Ziegenproblem ist ein weltweit bekanntes Beispiel für die erstaunliche Wirksamkeit von statistischen Analysen, manchmal auch gegen das eigene Bauchgefühl (Intuition). Das Ziegenproblem entstammt einer US Fernsehshow „Let's make a Deal", vergleichbar mit „Geh aufs Ganze" in einem der privaten Bildungskanäle Deutschlands, und bezieht sich inhaltlich auf die Gesetze der bedingten Wahrscheinlichkeit, die wir in Kapitel 6 behandeln werden.

Kurze Zusammenfassung des Ziegenproblems:

Ein Kandidat wird vor drei verschlossene Türen geführt. Hinter Zweien befindet sich eine Ziege, also eine Niete. Hinter einer dritten ein teurer Sportwagen.

Der Kandidat wählt nun eine Tür aus (1. Wahl). Sie wird jedoch nicht geöffnet; stattdessen öffnet der Moderator, der genau weiß, wo das Auto steht, eine andere Tür, hinter der sich eine Ziege befindet.

Die allen Teilnehmern bekannten Spielregeln zwingen den Moderator, dies in jedem Fall zu tun, unabhängig von der 1. Wahl des Kandidaten.

Nun fragt er den Kandidaten, ob dieser vielleicht seine 1. Wahl noch einmal überdenken und zu der anderen noch verschlossenen Tür wechseln möchte (2. Wahl mit Wechseln) oder ob er bei seiner 1. Wahl bleibt (2. Wahl ohne Wechseln).

Frage: Wie verhält sich der Kandidat bei der 2. Wahl? Führt eine „2. Wahl mit Wechseln“ zu einer Erhöhung der Gewinnwahrscheinlichkeit?

Aktion	**Tor A**	**Tor B**	**Tor C**
1. Wahl	Entscheidung für Tor A	Entscheidung für Tor B	Entscheidung für Tor C
Gewinnwahrscheinlichkeit	*1/3*	*1/3*	*1/3*
Moderator öffnet Tor...	B (oder C)	C	B
2. Wahl ohne Wechseln	**Auto**	**Ziege**	**Ziege**
Gewinnwahrscheinlichkeit	*1/3*	*0*	*0*
2. Wahl mit Wechseln	Wechsel zu C (oder B) **Ziege**	Wechsel zu A **Auto**	Wechsel zu A **Auto**
Gewinnwahrscheinlichkeit	*0*	*1/3*	*1/3*

Tabelle 1 Das Ziegenproblem – Entscheidungsbaum

Die Annahme bei dem folgenden Entscheidungsbaum zur Illustration der Lösung ist, dass das Auto sich exemplarisch hinter dem Tor A befindet (die Lösung funktioniert natürlich auch, wenn sich das Auto hinter Tor B oder C befindet).

Antwort: Der Kandidat verdoppelt seine Wahrscheinlichkeit zu gewinnen wenn er sich für die „2. Wahl mit Wechseln“ entscheidet.

Ergebnis: Ein Kandidat hat bei der „2. Wahl mit Wechseln“ eine (durchschnittliche) Gewinnchance von 2/3, bei der „2. Wahl ohne Wechseln“ von 1/3. Also ist es eindeutig besser zu wechseln. Dieses Ergebnis ist für viele Menschen nicht gerade das, was Sie erwartet hätten.

Das Ziegenproblem überspitzt formuliert

Das Ziegenproblem lässt sich auch erklären, indem wir die Situation überspitzt darstellen. Es gibt dann eine Million Tore und hinter genau

einem befindet sich das Auto. Nachdem der Kandidat ein Tor gewählt hat, öffnet der Moderator alle anderen Tore bis auf eines. Hier ist es sofort einsichtig, dass der Kandidat wechseln sollte: Die Wahrscheinlichkeit, mit dem zuerst gewählten Tor richtig zu liegen, ist sehr gering. Die Gewinnchancen bei der „2. Wahl ohne Wechseln“ liegen bei 1 zu 1.000.000, die Gewinnchancen bei der „2. Wahl mit Wechseln“ liegen bei 999.999/1.000.000. Wenn Sie die Zahl der Tore verringern, ändert sich nichts daran, dass der Kandidat das Tor wechseln sollte, nachdem der Moderator alle bis auf eine Niete entfernt hat. Insbesondere gilt dies auch für den Fall mit drei Toren.

Wenn Sie das Ziegenproblem nicht ganz verstanden haben, dann ist das erst einmal gar kein Problem, denn dann befinden Sie sich in bester Gesellschaft. Sogar hochrangige Professoren haben sich in wissenschaftlichen Fachmagazinen einen Streit um die richtige Interpretation geliefert und damit gezeigt, wie schwer diese Thematik zu verdauen ist. Nach Lektüre des Kapitels über bedingte Wahrscheinlichkeiten, in dem Sie den übersichtlichen Beweis für die Lösung dieses Problems nach dem Satz von Bayes zu sehen bekommen, werden Sie dieses Ziegenproblem nicht mehr als Problem ansehen, sondern als amüsante Anekdote und dürfen in Ihrem Bekanntenkreis oder beim nächsten Bewerbungsgespräch Fachwissen an den Tag legen.

Spickzettel

Weil ich gerne Wissen überprüfen möchte und nicht auswendig Gelerntes, fertigt jeder meiner Studenten mehr oder weniger kunstvoll den eigenen Spicker an, den ich offiziell zu meinen Klausuren zulasse. Hier sehen Sie die Vorderseite eines ausgesuchten Spickzettels von der Studentin Pauline G. aus dem Statistikkurs, den ich im Wintersemester 2013/14 an der FH Münster gehalten habe und den ich hier mit Zustimmung der Studentin präsentiere. Bei der Klausurkorrektur hatte ich gemerkt, dass viele der Klausurteilnehmer diesen Spicker benutzt haben, das spricht für ihn und deswegen wird er hier gewürdigt. Allerdings ergibt meiner Meinung nach ein Spickzettel nur dann Sinn, wenn er eigenhändig angefertigt wird.

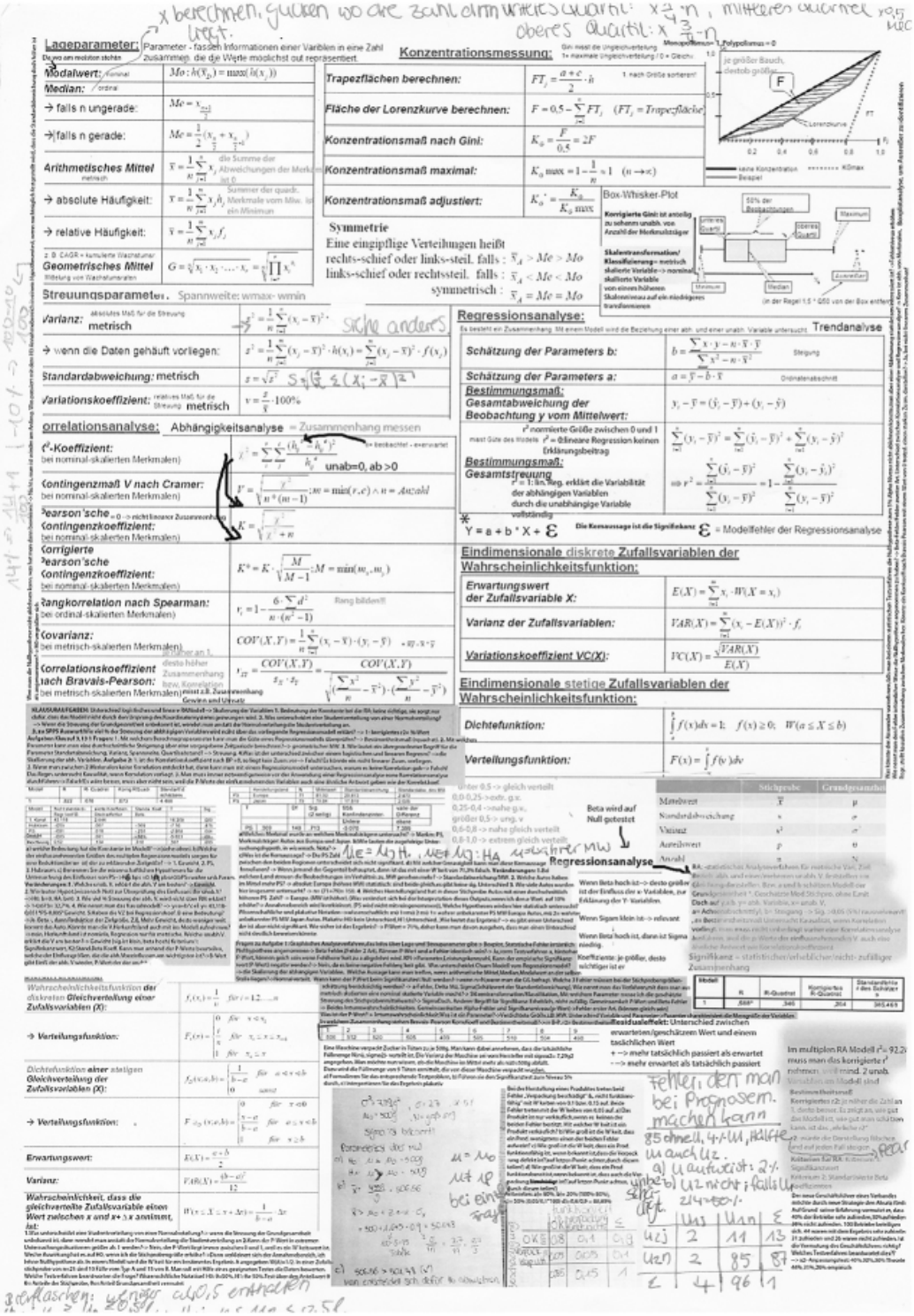

Abbildung 2 Spickzettel

Zusammenfassung des Buchinhaltes

Hier möchte ich versuchen, Ihnen mit einer übersichtlichen Tabelle den Inhalt dieses Buches zu strukturieren, um Ihnen einen Gesamtüberblick der Themen zu geben. Ich hoffe, Sie verstehen dann besser, um was es insgesamt geht.

Kapitel	**Kernaussage**
Beschreibende Statistik	Daten aus einer Stichprobe überschaubar zusammenfassen
1. Grundbegriffe	Einheitliche Sprache schaffen, um Daten analysieren zu können.
2. Mittelwerte und Streuwerte	Daten aus einer Stichprobe mit Kennzahlen überschaubar zusammenfassen (zum Beispiel arithmetischer Mittelwert, Standardabweichung).
3. Korrelation und Regression	Zusammenhänge analysieren (zum Beispiel zwischen Alter und Einkommen).
4. Indizes und Zeitreihen	Zusammenhänge analysieren, dieses Mal über einen Zeitraum betrachtet (zum Beispiel den Trend der Arbeitslosenzahl).
Schließende Statistik	Ergebnisse einer Stichprobe mit einer vertretbaren Irrtumswahrscheinlichkeit auf die Grundgesamtheit übertragen
5. Wahrscheinlichkeitsrechnung	Mit Wahrscheinlichkeiten umgehen können.
6. Wahrscheinlichkeitsverteilung	Gewisse Prozesse haben schon bekannte Wahrscheinlichkeiten, die wir mit einer Verteilung beschreiben können.

Kapitel	Kernaussage
7. Statistische Schätzverfahren	Die erste Möglichkeit, auf Grundlage eines Stichprobenergebnisses Aussagen über die Grundgesamtheit zu machen, mit einem 95 % Vertrauensintervall.
8. Statistische Testverfahren	Die zweite Möglichkeit, auf Grundlage eines Stichprobenergebnisses Aussagen über die Grundgesamtheit zu machen, mit einer Irrtumswahrscheinlichkeit von 5 %.

Tabelle 2 Zusammenfassung Struktur Buchinhalt

Für wen habe ich das Buch geschrieben?

Die Zielgruppe dieses Buches sind Studenten der Wirtschafts-, Sozial-, Natur- und Ingenieurswissenschaften an Universitäten, Fachhochschulen, Verwaltungs-, Wirtschafts- und Berufsakademien. Ich habe das Buch für Studenten so geschrieben, dass Sie mit wenig Zeitaufwand garantiert die Inhalte jeder Statistikvorlesung kompakt überblicken und in der Lage sind, Ihren Willen und Fleiß vorausgesetzt, erfolgreich Ihre Statistikklausur abzuschließen.

Nötiges Vorwissen.

Eigentlich brauchen Sie keine besonderen Vorkenntnisse. Statistik unterscheidet sich sehr von der übrigen Mathematik, daher reichen grob gesagt die vier Grundrechenarten. Alle Formeln, die in diesem Buch auch vorkommen, werde ich versuchen zu entmystifizieren und in verständlicher Sprache zu beschreiben. Ich erkläre Ihnen zum Beispiel, **warum das wissenschaftliche Summenzeichen Σ für mich einer modernen App gleichzusetzen** ist.

Ziel des Buches

Hauptziel dieses Buches ist Ihre zeitlich kompakte, optimale Vorbereitung zum Bestehen Ihrer Statistikklausur. Daher werde ich in diesem Buch, abweichend von üblichen, wissenschaftlich

orientierten Statistikbüchern, sehr viel mit typischen Verständnishinweisen, Interpretationshilfen und mit einer Menge an Beispielen und Übungsaufgaben mit Musterlösungen arbeiten.

Was bedeutet was.

Fett werden Wörter in diesem Buch ausgezeichnet, die beim ersten Blick auf der Seite auffallen sollen, *kursiv* werden Wörter ausgezeichnet, die beim Lesen auffallen sollen.

Im Text kommen selten Fremdwörter oder andere erklärungsbedürftige Wörter vor. Manchmal lässt sich das aber nicht vermeiden. Diese Wörter markiere ich auch **fett** und liste sie hinten im **Glossar**.

Die Symbole in diesem Buch

In diesem Buch werden Symbole verwendet, um Ihre Aufmerksamkeit auf bestimmte Aspekte zu lenken. Hier eine kurze Erläuterung dieser Symbole:

Tipp

TIPP. Dieser Kasten erscheint, wenn ich Ihnen zeigen möchte, wie Sie eine Information oder Problemstellung besser verstehen oder einfach lösen können.

Warnung

Warnung. Dieser Kasten erscheint, wenn ich Ihnen typische Fehler aufzeigen möchte.

Beispiel

Beispiel. Dieser Kasten erscheint, wenn ich auf ein Beispiel verweisen möchte.

Danksagung

Ganz herzlich möchte ich mich bei Herrn Marcel Ferner vom Wiley-VCH Verlag für die durchgängig kompetente und verlässliche

Betreuung bedanken. Weiterhin bedanke ich mich sehr herzlich bei meinem hochgeschätzten früheren Fachkollegen von der Westfälischen Hochschule, Herrn Prof. em. Heinz-Jürgen Pinnekamp, und dem mir freundschaftlich sehr verbundenen Herrn Prof. Ehrenfried Salomon von der Hochschule RheinMain für das sorgfältige Korrekturlesen. Jeder noch im Skript befindliche Fehler wird einzig und allein mein unrühmlicher Verdienst sein, da bin ich mir absolut sicher.

Während des Verfassens dieses Buches ist unsere Tochter Sabeth auf diese Welt gekommen. Ich möchte dieses Buch, mein erstes, meiner Frau Kerstin und meinen drei Kindern Henry, Louis und Sabeth widmen. Mit meiner Familie im Hintergrund fiel mir das Schreiben im häuslichen Arbeitszimmer außergewöhnlich leicht.

Teil I

Beschreibende Statistik