- LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN -

INSTITUT FÜR STATISTIK

Die Entwicklung der Randomized Response Technik bis hin zur Zusammenhangsanalyse

BACHELORARBEIT

ZUR ERLANGUNG DES AKADEMISCHEN GRADES
BACHELOR OF SCIENCE (B.Sc.)

Autor: Nina Scharrer

Betreuer und Gutachter: Prof. Dr. Thomas Augustin

München, den 29. Januar 2014

Abstract

Die Selbstauskunft ist die häufigste und in vielen Fällen auch die einzige Möglichkeit, um Daten von Personen zu gewinnen und diese schließlich mit statistischen Methoden auszuwerten. Gerade aber bei sensiblen Fragen zu illegalen oder unmoralischen Thematiken, bei denen die Wahrheit für die befragte Person peinlich oder gar schädigend sein könnte, ist die Bereitschaft, ehrlich zu antworten, jedoch meist sehr gering. Insbesondere dann, wenn die Anonymität der Befragten nicht hinreichend gewährleistet ist. Bei der direkten Befragung zu solchen sensiblen Themen ist daher eine hohe Anzahl von Antwortverweigerern und Falschantworten zu erwarten. Der tatsächliche Anteil derjeniger Personen, welche das sensible Merkmal tragen, wird folglich unterschätzt. Um dieser Verzerrung bei persönlichen und indiskreten Fragen entgegenzuwirken, führte Warner 1965 ein neues Interview-Verfahren namens Randomized Response Technik ein. Dabei werden die Umfragen mit einem Zufallsexperiment gekoppelt, sodass eine individuelle Zuordnung der Antworten nicht mehr eindeutig möglich ist. Dennoch lassen sich die Häufigkeiten der interessierenden Merkmale unverzerrt schätzen und sich Aussagen auf der Aggregatebene treffen. Warner's Modell wurde schließlich in viele Richtungen weiterentwickelt. Neben dem Unrelated Question Modell von Simmons und dem Forced Question Modell von Boruch, deren Motivation und Ziel darin lag, die Effizienz zu steigern und den Befragten noch mehr Sicherheit zu geben, war insbesondere die Ausweitung der Randomized Response Technik auf nicht dichotome Fragestellungen ein großer Fortschritt. Es konnten nun auch heikle Merkmale mit mehreren Kategorien oder sogar quantitativen Wertebereichen erhoben werden. Durch Betrachtung der Randomized Response Daten als fehlklassifizierte Daten beziehungsweise als Daten mit Messfehler, ist es schließlich möglich, die Schätzer zu korrigieren. Dadurch lassen sich bei Zusammenhangsanalysen zwischen verschiedenen Variablen, die mit Randomized Response erhoben wurden, gute Ergebnisse erzielen. Das Anwendungsfeld der Randomized Response Technik erweitert sich dadurch merklich, da sich auf diese Weise schließlich auch multivariate Hypothesen testen lassen.

II Inhaltsverzeichnis

Abbildungsverzeichnis

1 Randomized Response Theorie

Daten über die menschliche Bevölkerung, über deren Verhalten, sowie deren Eigenschaften basieren zumeist auf Umfragen, in denen die Teilnehmer Auskunft über sich selbst geben. Dies ist häufig der einzige Weg, um an persönliche Merkmale zu gelangen und schließlich statistische Analysen durchführen zu können. Problematisch wird diese Art der Datengewinnung jedoch, wenn die Teilnehmer einer Studie über heikle Merkmale befragt werden, welche sie gerne vor Anderen verheimlichen würden. Darunter fallen illegale oder kriminelle Machenschaften wie Diebstahl oder der Konsum von Drogen. Zudem ist auch bei Handlungen beziehungsweise Eigenschaften, die nicht der sozialen Erwünschtheit entsprechen, wie zum Beispiel die Unterstützung extremer Parteien, die Aussagebereitschaft eher gering. Auch bei Fragen, die den Interviewten peinlich berühren oder zu persönlich erscheinen, besteht eine hohe Tendenz, dass der Befragte sich in seiner Privatsphäre verletzt fühlt. Das kann zur Folge haben, dass der Umfrage-Teilnehmer seine Antwort verweigert oder eine Falschantwort gibt, um möglichen negativen Konsequenzen zu entgehen. Diese beiden Faktoren verzerren offensichtlich die Schätzungen der Häufigkeitsverteilungen der interessierenden Merkmale und führen folglich zu falschen Schlussfolgerungen und Ausssagen über die Bevölkerung. Ziel bei Befragungen zu sensiblen Themen ist es also, die Nonresponse- und Falschaussagenquote weitestgehend zu reduzieren, um eine möglichst genaue Abbildung der Realität zu erhalten.

Diese Überlegung bildete die Grundlage für eine Vorgehensweise der Datenerhebung, die im Jahr 1965 von Stanley L. Warner eingeführt und Randomized Response Technik genannt wurde. Dabei antworten die Teilnehmer einer Umfrage auf eine Frage, die sie zufällig aus zwei oder mehreren vorgegebenen Fragen ausgewählt haben. Dem Interviewer ist dabei nicht bekannt, welche Frage der Teilnehmer beantworten sollte, denn er erhält nur eine Antwort, die auf alle Fragen zutreffen könnte. Er weiß nur, mit welcher Wahrscheinlichkeit die jeweiligen Fragen gestellt wurden. Durch diese Verknüpfung der Umfrage mit einem Zufallsexperiment, wird die Privatsphäre und die Anonymität der einzelnen Teilnehmer gewährleistet und man erwartet folglich eine erhöhte Kooperation. Da die Befragten nun ihre persönliche Situation nicht mehr offenlegen müssen und daher das Preisgeben von eventuellen Peinlichkeiten oder heiklen Eigenschaften verhindert wird, haben sie auch keinen Grund mehr, nicht die Wahrheit zu sagen. Es wird folglich erwartet, dass die Validität einer solchen Randomized Response Umfrage bei heiklen Themen höher ist als bei einer direkten Befragung. Das bedeutet, dass die Daten einer Randomized Response Umfrage eher das beschreiben, was erforscht werden soll und damit die Realität besser abbilden.

Trotz dieses Vorteils wird der Randomized Response Technik in der Praxis häufig mit Skepsis gegenüber gestanden. Ursache hierfür ist zum einen, dass die Randomized Response Anwendung viel Zeit und Kosten fordert. Das Vorgehen muss den einzelnen Befragten erst detailliert erklärt werden, damit diese Vertrauen fassen und sich tatsächlich an die Regeln halten. Dafür werden wiederum gut geschulte Interviewer benötigt. Zudem ist die Randomized Response Technik weniger effizient als die direkte Befragung. Das heißt, durch die Kopplung des Zufallsexperiments an die Befragung erhöht sich die Streuung des erwartungstreuen Schätzers für den Anteil des sensitiven Merkmals in der Bevölkerung. Die Varianz ist dementsprechend größer und der Schätzer somit weniger effizient. Dieser Effizienzverlust kann nur durch eine größere Stichprobe ausgeglichen werden, was wiederum mehr Zeit und Geld kostet. Der ausschlaggebendste Grund für die Zurückhaltung gegenüber der Randomized Response Technik ist aber wohl der geringere Informationsgehalt der erfassten Daten. Durch die Zufallsverschlüsselung ist eine direkte Zuordnung der gegebenen Antworten zu den jeweiligen Fragen, sowie zu den jeweiligen Teilnehmern nicht möglich und es liegen folglich keine Individualdaten vor. Während sich also univariate Parameter wie Anteile oder Wahrscheinlichkeiten noch indirekt schätzen lassen, benötigt die multivariate Statistik, zum Beispiel in Form von Korrelationen, Individualdaten. Damit ist das Anwendungsfeld der Randomized Response Technik stark eingeschränkt. Sudman und Bradburn (1982, S. 81) schrieben dazu:

> „By using this procedure you can estimate the undesirable behavior of a group; and, at the same time, the respondent's anonymity is fully protected. With this method, however, you cannot relate individual characteristics of respondents to individual behavior. That is, standard regression procedures are not possible at an individual level. [...] On the whole, however, much information is lost when randomized response is used."

Jedoch ist die Einschränkung der Randomized Response Daten auf univariate Methoden unberechtigt, wie sich bei weiterer Forschung in diesem Bereich zeigte. Indem Randomized Response Daten behandelt werden, als würden diese die wahren Ausprägungen nur mit einer bestimmten Wahrscheinlichkeit mit einem Fehlerterm „verschmutzen", lassen sich multivariate Methoden anwenden. Sie müssen nur richtig korrigiert werden (vgl. Fox and Tracy, 1984, S.189). Auf diese Weise lassen sich schließlich auch Zusammenhangsanalysen zwischen Merkmalen, die durch Randomized Response erfasst wurden, durchführen. Diese Erweiterung vergrößert das Anwendungsfeld von Randomized Response Daten ungemein und es lassen sich folglich spezifischere Hypothesen testen.

Die vorliegende Arbeit ist wie folgt aufgebaut: Im zweiten Kapitel werden, ausgehend von Warners Original-Modell, verschiedene Randomized Response Modelle für dichotome Fragestellungen vorgestellt und es wird zudem erläutert, inwiefern dabei Zusammenhangsanalysen möglich sind. Das dritte Kapitel beschäftigt sich mit Randomized Response Modellen für kategoriale Daten. Diese beruhen auf dem Prinzip der Fehlklassifikation. Desweiteren wird eine multiattribute Situation modelliert, welche die Anteilsschätzung mehrerer kategorialer Merkmale anhand einer Stichprobe ermöglicht. Abschließend wird in diesem Kapitel dargelgt, wie die Unabhängigkeit zwischen zwei kategorialen Merkmalen, die mittels Randomized Response erfasst wurden, mit dem Chi-Quadrat-Test getestet werden kann. In Kapitel vier werden Randomized Response Daten, die einen quantitativen Wertebereich haben, als Daten mit Messfehler betrachtet, der durch das gekoppelte Zufallsexperiment entstanden ist und durch verschiedene Methoden korrigiert werden kann. Es wird zudem gezeigt, dass auch der Korrelationskoeffizient nach Bravais-Pearson zwischen den beobachteten, fehlerhaften Werten korrigiert und somit als Schätzer für die Korrelation zwischen den wahren, unbekannten Werten zweier Randomized Response Variablen herangezogen werden kann. In Kapitel fünf wird schließlich gezeigt, dass auch Regressionsanalysen durchgeführt werden können, wenn eines der Merkmale mittels Randomized Response erhoben wurde. Dadurch kann geprüft werden, inwiefern eine Randomized Response Variable von bestimmten Kovariablen beeinflusst wird beziehungsweise ob die Randomized Response Variable einen Effekt auf eine betrachtete Zielgröße hat. Danach folgt abschließend ein kapitelübergreifendes Fazit, sowie ein Ausblick.

2 Randomized Response Techniken für dichotome Variablen

2.1 Das Warner Modell

Das ursprüngliche Randomized Response Modell, welches im Jahr 1965 von Stanley L. Warner eingeführt wurde und deshalb im Folgenden als das Warner Modell bezeichnet wird, hat das Ziel, anhand einer Stichprobe den unbekannten Anteil π_A der Bevölkerung mit Merkmal A zu schätzen. Warner betrachtete dabei den Fall, dass die gesamte Population in Personen unterteilt werden kann, die das interessierende, meist heikle Merkmal A tragen, und diejenigen, die das Merkmal A nicht haben und folglich der komplementären Gruppe $\overline{A}$ angehören. Es wird eine einfache Stichprobe vom Umfang n mit Zurücklegen aus der Population gezogen. Den Teilnehmern wird die interessierende Frage nicht direkt gestellt, sondern es wird stattdessen zunächst ein Zufallsexperiment durchgeführt, das entscheidet, welche der beiden Fragen, „Gehören Sie der Gruppe A an?“ oder „Gehören Sie der Gruppe $\overline{A}$ an?“, beantwortet werden soll. Für den Zufallsmechanismus gibt es unzählige Möglichkeiten. Es kann zum Beispiel ein Würfel geworfen werden, wobei bei Augenzahl eins und zwei die erste Frage und bei den Augenzahlen drei bis sechs die zweite Frage wahrheitsgemäß beantwortet werden muss. Eine weitere Alternative wäre ein Beutel mit zwei verschiedenfarbigen Kugeln. Die Farbe gibt dabei an, welche Frage gewählt wird. Dasselbe Vorgehen ist auch mit einem Stapel bestehend aus Karten mit unterschiedlichen Fragen denkbar. Für die Anwendung der Randomized Response Technik muss nur im Vornherein die jeweiligen Häufigkeitsverteilungen, also die Wahrscheinlichkeiten p und $1 - p$, mit denen die sensitive beziehungsweise die komplementäre Frage ausgewählt wird, festgelegt werden. Neben diesen bekannten Größen weiß der Interviewer zudem, ob der Befragte mit „Ja“ oder „Nein“ antwortet. Welche Frage aber letztendlich beantwortet wurde, bleibt geheim. Aufgrund dieses Zufallsmechanismus kann die befragte Person anhand ihrer Antwort nicht mit Sicherheit in eine der beiden Gruppen A und $\overline{A}$ eingeordnet werden. Die Privatsphäre bleibt also geschützt und die Teilnehmer werden deshalb erwartungsgemäß ehrlicher antworten (vgl. Chaudhuri and Mukerjee, 1988, S. 2).

Aufgrund des gekoppelten Zufallsexperiments, lässt sich das Warner Modell anhand eines Baumdiagramms veranschaulichen. Dafür, sowie für die darauf folgende Herleitung des Schätzers $\hat{\pi}_A$, dient die folgende Notation (vgl. Warner, 1965, S. 64, Notation angepasst und erweitert):

$\pi_A =$ wahrer Anteil der Merkmalsträger A in der Bevölkerung

$p =$ Wahrscheinlichkeit, dass die Frage „Gehören Sie der Gruppe A an?“ ausgewählt wurde

$$X_i = \begin{cases} 1, & \text{falls der i-te Befragte der Stichprobe mit „Ja“ anwortet} \\ 0, & \text{falls der i-te Befragte der Stichprobe mit „Nein“ anwortet} \end{cases}$$

$n =$ Stichprobengröße

$n' =$ Gesamtsumme der „Ja“-Antworten in der Stichprobe

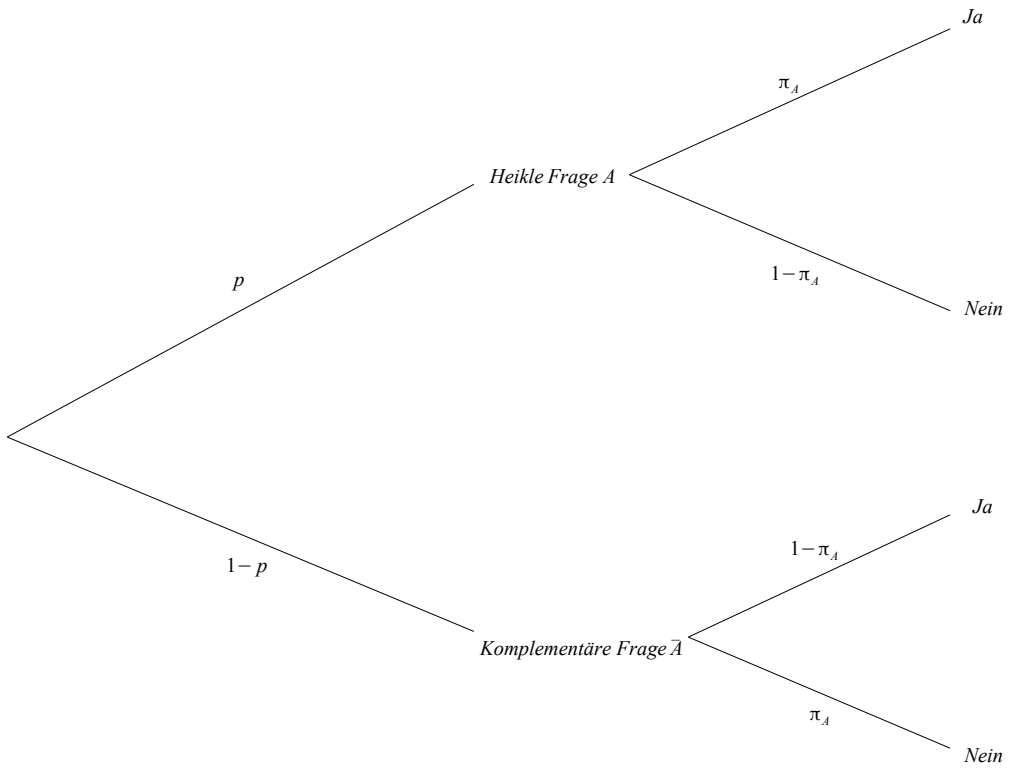

Abbildung 1: Das Warner Modell als Baumdiagramm
(Darstellung orientiert sich an Ostapczuk et al. (2009, S. 270), Notation angepasst)

Anhand des Baumdiagramms wird ersichtlich, dass sich die Wahrscheinlichkeit für eine „Ja“-Antwort zusammensetzt aus der Wahrscheinlichkeit, dass ein Befragter auf die sensitive Frage mit „Ja“ antwortet und der Wahrscheinlichkeit, dass ein Befragter auf die komplementäre Frage mit „Ja“ antwortet. Wie bei Campbell and Joiner (1973, S.229) dargestellt, gilt also:

$$\mathbb{P}(\text{„Ja“}) = \mathbb{P}(\text{„Ja“ bei sensitiver Frage} \quad A) + \mathbb{P}(\text{„Ja“ bei komplementärer Frage} \quad \overline{A})$$

Aufgrund des Multiplikationssatzes für bedingte Dichten lässt sich dies umformen in

$$\begin{aligned}\mathbb{P}(\text{„Ja“}) = {} & \mathbb{P}(\text{Sensitive Frage } A \text{ wurde ausgewählt}) \cdot \mathbb{P}(\text{„Ja“}|\text{ Frage } A) \\ & + \mathbb{P}(\text{Komplementäre Frage } \overline{A} \text{ wurde ausgewählt}) \cdot \mathbb{P}(\text{„Ja“}|\text{ Frage } \overline{A}).\end{aligned}$$

Das wiederum lässt sich wie bei Warner (1965, S.229, Notation angepasst) anhand der zuvor eingeführten Notation schreiben als

$$\mathbb{P}(X_i = 1) = \pi_A \cdot p + (1-\pi_A)\cdot(1-p) \quad \text{bzw.} \quad \mathbb{P}(X_i = 0) = (1-\pi_A)\cdot p + \pi_A \cdot (1-p). \quad (1)$$

Diese beiden Gleichungen geben die Wahrscheinlichkeit an, dass eine Person mit „Ja“ beziehungsweise mit „Nein“ antwortet. Setzt man für die Wahrscheinlichkeit $\mathbb{P}(X_i = 1)$ den Anteil der „Ja“-Antworten in der Stichprobe $\frac{n'}{n}$ ein (vgl. Mangat and Singh, 1990, S.439, Notation angepasst), so ergibt sich aus der ersten Gleichung von (1) die Formel $\frac{n'}{n} = \pi_A \cdot p + (1 - \pi_A) \cdot (1 - p)$ und π_A ist schließlich aufgrund des bekannten p der einzige unbekannte Parameter, den es zu schätzen gilt. Nach dem Auflösen nach π_A erhält man schließlich den erwartungstreuen Schätzer für den Anteil der Merkmalsträger von A in der Bevölkerung:

$$\hat{\pi}_A = \frac{\frac{n'}{n} + (p-1)}{2p-1}, \quad p \neq \frac{1}{2}$$

Da es sich bei $\hat{\pi}_A$ um einen Maximum-Likelihood Schätzer handelt und der gewählte Stichprobenumfang n in der Regel groß ist, kann man $\hat{\pi}_A$ als normalverteilt um π_A mit Varianz

$$Var(\hat{\pi}_A) = \frac{\hat{\pi}_A \cdot (1 - \hat{\pi}_A)}{n} + \frac{p \cdot (1-p)}{n \cdot (2p-1)^2} \quad (2)$$

annehmen (vgl. Mangat and Singh, 1990, S.439, Notation angepasst). Es zeigt sich, dass die Varianz von der zuvor gewählten Auswahlwahrscheinlichkeit p abhängt. Je

nachdem, mit welcher Wahrscheinlichkeit der Teilnehmer also die sensitive Frage auswählt, kann die Streuung des Schätzers $\hat{\pi}_A$ größer oder geringer sein. Um möglichst genaue Aussagen über den Anteil des Merkmals A in der Bevölkerung treffen zu können, ist man offensichtlicherweise an einer geringen Streuung interessiert. Die Wahl des p ist also von großer Bedeutung und wird daher im folgenden Abschnitt, welcher sich auf Warners Publikation (1965, S.66) bezieht, thematisiert.

Ein optimales p lässt sich nicht grundsätzlich festlegen. Stattdessen muss für jede Umfrage spezifisch ein Gleichgewicht aus Privatsphäre und Informationsgewinn gefunden werden. $p = 1$ bedeutet, dass die sensitive Frage mit Wahrscheinlichkeit 1 gestellt wird. Folglich reduziert sich die Randomized Response Technik auf eine einfache direkte Befragung. Werte nahe bei 1 beziehungsweise 0 bedeuten, dass der Befragte mit hoher Wahrscheinlichkeit auf die heikle Frage beziehungsweise die komplementäre Frage antworten muss. Damit lassen sich die Teilnehmer mit sehr geringer Irrtumswahrscheinlichkeit zu einer der beiden Gruppen A oder $\overline{A}$ einordnen. Die Befragten werden jedoch wenig kooperativ sein und der eigentliche Vorteil der Randomized Response Technik geht verloren. Wahrscheinlichkeiten nahe $\frac{1}{2}$ hingegen gewährleisten die Anonymität der Befragten, denn es werden weniger Informationen zur tatsächlichen Gruppenzugehörigkeit in den einzelnen Interviews gewonnen. Neben der erhöhten Teilnehmerbereitschaft ist jedoch auch eine höhere Streuung des Schätzers die Folge. Diese kann nur durch eine Vergrößerung des Stichprobenumfangs n ausgeglichen werden. Die Wahl von p hängt also davon ab, ob der Interviewer seine Priorität auf die Effizienz oder das Vertrauen der Teilnehmer in das Verfahren setzt. Wichtig ist an dieser Stelle nur, dass für p nicht $\frac{1}{2}$ gewählt werden darf. In diesem Fall hängt die Likelihood Funktion $L = [\pi_A \cdot p + (1 - \pi_A) \cdot (1 - p)]^{n'} \cdot [(1 - \pi_A) \cdot p + \pi_A \cdot (1 - p)]^{n-n'}$ (vgl. Warner, 1965, S.64, Notation angepasst) nicht vom Parameter π_A ab und man würde folglich keine Informationen über den zu schätzenden Anteil in der Bevölkerung erhalten.

Vergleicht man nun das Warner Modell mit einer direkten Befragung, so hat es den Vorteil, dass eine geringere Verzerrung durch Falschantworten und Non-Response erwartet werden kann. Das beruht auf dem Anonymitätsgewinn durch die Kopplung an ein Zufallsexperiment. Dies bringt jedoch den Nachteil mit sich, dass sich die Streuung des Schätzers deutlich erhöht. Auf diese Problematik gehen Lensvelt-Mulders et al. in ihrem 2005 veröffentlichten Artikel (S.321) im Detail ein: Die Varianz des Warner-Schätzers besteht offensichtlich aus zwei Komponenten (vgl. Gleichung (2)). Der erste Teil $\dfrac{\hat{\pi}_A \cdot (1 - \hat{\pi}_A)}{n}$ entspricht der gewöhnlichen Stichprobenvarianz einer

direkten Befragung. Der zweite Teil $\frac{p \cdot (1-p)}{n \cdot (2p-1)^2}$ repräsentiert die zusätzliche Varianz, welcher durch die Randomized Response Prozedur hinzukommt. Da die zweite Komponente größer als 0 ist, zeigt sich, dass die Randomized Response Technik einen Effizienzverlust mit sich bringt. Um die Schätzer weiter zu optimieren, setzten sich viele Weiterentwicklungen des Warner Modells zum Ziel, die Varianz zu verringern und damit effizientere Schätzer zu erhalten.

2.2 Das Unrelated Question Modell

Das folgende Kapitel bezieht sich weitestgehend auf die Publikation von Greenberg et al. (1965), in welcher die theoretischen Grundzüge des Unrelated Question Modell dargelegt wurden. Dieses Modell von Walt R. Simmons veränderte Warner's Randomized Response Idee dahingehend, dass der Zufallsmechanismus nun nicht mehr zwischen der sensitiven Frage A und deren Komplement $\overline{A}$ auswählt, sondern stattdessen neben der heiklen Frage eine zusammenhangslose, harmlose Frage Y gestellt wird. Hinter diesem Vorgehen steht die Annahme, dass das Vertrauen der Teilnehmer in die Wahrung der Privatsphäre durch die Technik verstärkt wird, wenn zwei verschiedene Fragen gestellt werden, die nicht miteinander in Verbindung stehen. Bei der harmlosen Frage Y kann dabei zum Beispiel gefragt werden, ob der Teilnehmer in einer bestimmten Region geboren wurde oder ob er heute schon eine Katze gesehen hat. Hierfür gibt es unzählige Möglichkeiten. Das weitere Vorgehen ist analog zum Warner Modell, das heißt, auch hier wird nur die Antwort preisgegeben, wohingegen die beantwortete Frage geheim bleibt. Der Schutz der Privatsphäre und die Anonymität der Befragten wird durch das Unrelated Question Modell nochmals deutlich erhöht, denn der Teilnehmer kann sowohl zu einer der beiden Gruppen A und Y gehören, als auch zu keiner oder beiden. Die Antwortbereitschaft sollte sich daher erhöhen, gerade auch aus dem Grund, dass die Befragten hierbei nicht ausschließlich in Zusammenhang mit dem heiklen Thema gebracht werden und somit bei Unsicherheit und Zweifel immer sagen können, sie hätten die harmlose Frage beantwortet.

Das Problem beim Unrelated Question Modell besteht jedoch darin, dass in der Regel keine Vorinformationen zu der Häufigkeitsverteilung der harmlosen Frage bekannt sind. Der Interviewer weiß also nicht, mit welcher Wahrscheinlichkeit der Befragte erwartungsgemäß „Ja“ antwortet, wenn ihm die harmlose Frage gestellt wurde. Neben dem interessierenden Parameter π_A muss nun also auch noch π_Y, der Anteil der Bevölkerung, der das unabhängige Merkmal Y trägt, geschätzt werden. Dafür

müssen nun zwei unabhängige Stichproben, deren Umfänge n_1 und n_2 nicht unbedingt gleich groß sein müssen, gezogen werden. In beiden Stichproben muss die Umfrage unter den gleichen Bedingungen stattfinden, wobei die sensitive Frage A in Stichprobe eins mit Wahrscheinlichkeit p_1 und in der zweiten Stichprobe mit Wahrscheinlichkeit p_2 gestellt wird. Dabei muss $p_1 \neq p_2$ gelten. Unter der Annahme, dass die Anteile π_A und π_Y in beiden Stichproben jeweils gleich groß sind, lassen sich dann zwei unabhängige Wahrscheinlichkeiten für eine „Ja"-Antwort beobachten. Diese Bedingungen gelten stets, wenn mehrere Stichproben für die Durchführung des Randomized Response Verfahrens gezogen werden müssen. Sie werden somit auch im Folgenden bei anderen Randomized Response Modellen als gültig angenommen.

Auch das Unrelated Question Modell lässt sich als Baumdiagramm darstellen:

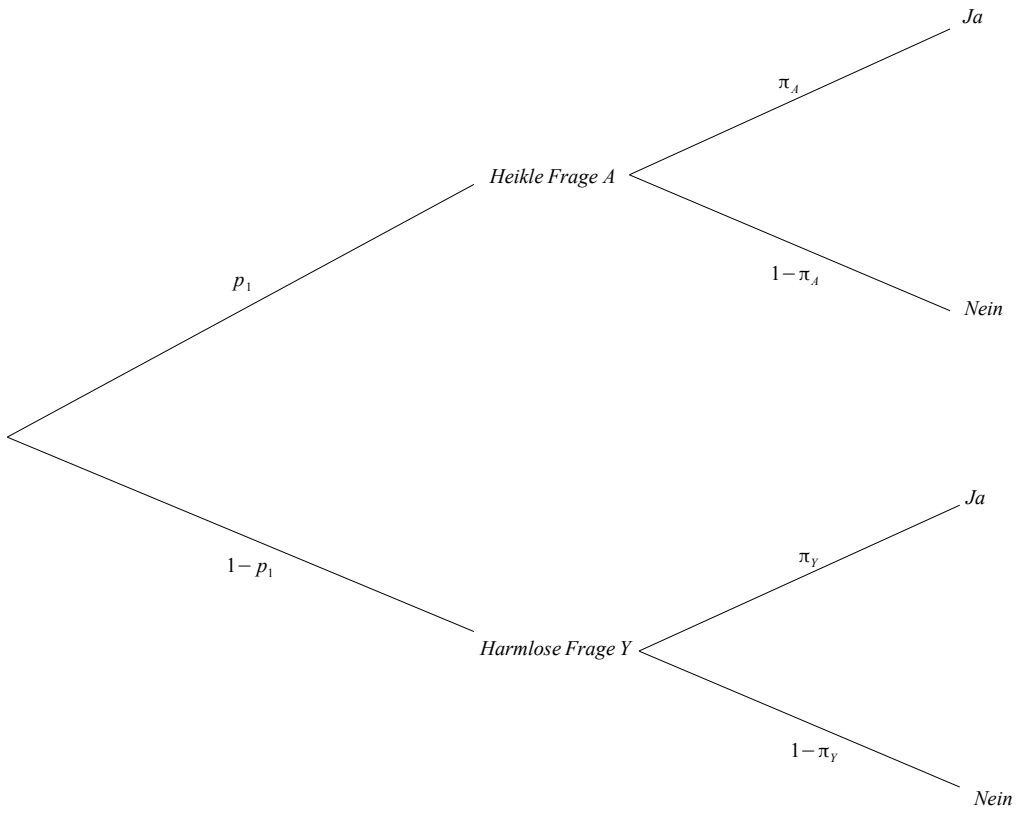

Abbildung 2: Das Unrelated Question Modell als Baumdiagramm
(Darstellung orientiert sich an Ostapczuk et al. (2009, S. 270), Notation angepasst)

Die zum Warner Modell analoge Herleitung ergibt schließlich für die jeweiligen „Ja“-Anteile der beiden Stichproben die folgenden Gleichungen, welche sich auch bei Greenberg et al. (1969, S.524, Notation angepasst) finden:

$$\mathbb{P}_1(X_i = 1) = p_1 \cdot \pi_A + (1 - p_1) \cdot \pi_Y \quad \text{bzw.} \quad \mathbb{P}_2(X_i = 1) = p_2 \cdot \pi_A + (1 - p_2) \cdot \pi_Y$$

Damit ergeben sich zwei Gleichungen mit zwei Unbekannten, welche durch Auflösen und gegenseitiges Einsetzen gelöst werden können. Setzt man außerdem für die Wahrscheinlichkeit einer „Ja“-Antwort den geschätzten Anteil aus den jeweiligen Stichproben ein, also $\frac{n_1'}{n_1}$ und $\frac{n_2'}{n_2}$, so erhält man für die Anteile des sensitiven beziehungsweise des harmlosen Merkmals in der Bevölkerung die folgenden erwartungstreuen Schätzer:

$$\hat{\pi_A} = \frac{\frac{n_1'}{n_1} \cdot (1 - p_2) - \frac{n_2'}{n_2} \cdot (1 - p_1)}{p_1 - p_2},$$
$$\hat{\pi_Y} = \frac{p_2 \cdot \frac{n_1'}{n_1} - p_1 \cdot \frac{n_2'}{n_2}}{p_2 - p_1}$$

Für die Varianz des Schätzers $\hat{\pi_A}$ erhält man

$$Var(\hat{\pi_A}) = \frac{1}{(p_1 - p_2)^2} \cdot [\frac{\frac{n_1'}{n_1}(1 - \frac{n_1'}{n_1})(1 - p_2)^2}{n_1} + \frac{\frac{n_2'}{n_2}(1 - \frac{n_2'}{n_2})(1 - p_1)^2}{n_2}].$$

Damit lässt sich festhalten, dass auch das Unrelated Question Modell von Simmons eine größere Streuung als die direkte Befragung hat. Vergleicht man die Effizienz des Schätzers mit dem des Warner Modells, so stellt man eine Steigerung fest. Es lässt sich bei dieser Technik aber ein noch viel bemerkenswerterer Effizienzgewinn erzielen, wenn der Bevölkerungsparameter π_Y der harmlosen Frage von vornherein bekannt ist (vgl. Edgell et al., 1982, S.90). Grund dafür ist, dass sich das Modell dann auf eine unbekannte Größe, nämlich π_A, reduziert. Das hat zum einen den Vorteil, dass dann nur eine Stichprobe vom Umfang n gezogen werden muss und zum anderen, dass sich die Varianz des Schätzers verkleinert und sich die statistische Effizienz folglich vergrößert. Diese Überlegung bildet die Grundlage für das nächste etablierte Modell.

2.3 Das Forced Response Modell

Das Forced Response Modell von Boruch ist insofern eine Weiterentwickung des Unrelated Question Modells mit bekanntem π_Y, dass es diesen Parameter in den Zufallsmechanismus mit einbaut. Dieser leitet den Befragten wie gewohnt mit einer bestimmten Wahrscheinlichkeit p weiter zur Beantwortung der sensitiven Frage. Der andere Anteil der Teilnehmer wird hingegen aufgefordert, die heikle Frage zu missachten und stattdessen eine vorgegebene Antwort zu nennen. Mit Wahscheinlichkeit Θ ist das ein „Ja“, mit Wahrscheinlichkeit $1 - \Theta$ hingegen ein „Nein“. Diese sind jeweils bekannt. Das zugehörige Baumdiagramm sieht wie folgt aus:

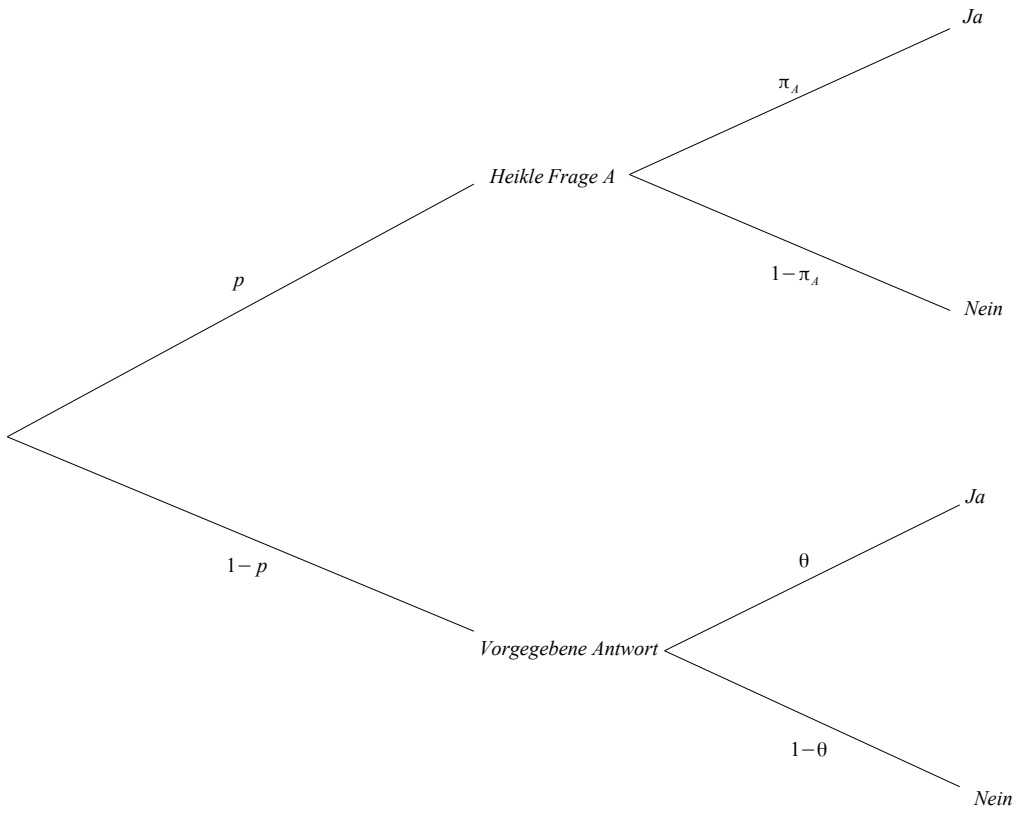

Abbildung 3: Das Forced Response Modell als Baumdiagramm
(Darstellung orientiert sich an Ostapczuk et al. (2009, S. 270), Notation angepasst)

Ein mögliches Zufallsexperiment für diesen Modellaufbau wäre zum Beispiel der Würfelwurf. Bei Augenzahl eins bis vier soll der Befragte wahrheitsgemäß auf die heikle Frage antworten. Bei Augenzahl fünf muss der Teilnehmer mit „Ja“ antworten,

ungeachtet davon, ob er tatsächlich das heikle Merkmal A trägt oder nicht. Bei Augenzahl sechs dementsprechend mit „Nein“.

Um schließlich den Schätzer für den interessierenden Anteil in der Bevölkerung zu erhalten, wird analog zu den vorherigen dichotomen Randomized Response Modellen vorgegangen. Über die bekannten Wahrscheinlichkeiten einer „Ja“- beziehungsweise „Nein“-Antwort erhält man dann, wie bei Lensvelt-Mulders et al. (2005, S.207, Notation angepasst) geschrieben, den Schätzer

$$\hat{\pi_A} = \frac{(\frac{n'}{n} - (1-p)) \cdot \Theta}{p}$$

mit Varianz

$$Var(\hat{\pi_A}) = \frac{1}{p^2} \cdot \frac{\frac{n'}{n}(1-\frac{n'}{n})}{n}.$$

Der Schätzer des Forced Response Modells ist der effizienteste der dargelegten dichotomen Randomized Response Techniken (vgl. Lensvelt-Mulders et al., 2005, S.262). Denoch hat diese Befragungstechnik den gravierenden Nachteil, dass auch Teilnehmer ohne heikles Merkmal A Grund haben, zu verweigern und damit die Schätzung zu verzerren. Denn diese Personen könnten gegebenenfalls befürchten, dass sie bei einer erzwungenen, vorgegebenen „Ja“-Antwort in Verbindung mit dem heiklen Merkmal A gebracht werden, obwohl sie dieses in Wirklichkeit nicht haben. Folglich missachten sie die Vorgaben und beantworten stattdessen die heikle Frage wahrheitsgemäß mit „Nein“. Somit gibt es bei diesem Verfahren eine zusätzliche Quelle für Verzerrungen. Damit lässt sich also festhalten, dass alle bisher vorgestellten Randomized Response Techniken für dichotome Variablen Vor- und Nachteile haben und es nicht eine beste Anwendung gibt. Die Wahl des Schätzverfahrens hängt stattdessen von Hypothese, Thematik, Stichprobe und anderen Faktoren ab.

2.4 Das Cheating Detection Modell

Das folgende Kapitel zum Cheating Detection Modell bezieht sich auf die Publikation von Ostapczuk et al. (2009), sowie der Publikation von Clark and Desharnais (1998).

Während die vorangehenden Modelle auf der Annahme beruhen, dass die Befragten der durch die Randomized Response Technik gewährleisteten Privatsphäre vertrauen und sich folglich an die vorgegebenen Regeln halten, basiert das folgende Modell, welches eine Weiterentwicklung des Forced Response Modells darstellt, auf der Überlegung, dass es durchaus einen Verweigerer-Anteil in der Stichprobe gibt.

Verweigerer können dabei zum einen Menschen sein, die das heikle Merkmal A tragen und das nicht zugeben, obwohl sie laut Zufallsexperiment die Frage wahrheitsgemäß beantworten sollen. Zum Anderen fallen darunter auch Teilnehmer ohne Merkmal A, die keinesfalls in Kontakt mit der heiklen Eigenschaft gebracht werden wollen und sich deshalb gegen die Randomized Response Vorgaben stellen, indem sie nicht, wie der Ausgang des Zufallsexperiment festgelegt hat, die vorgegebene Antwort „Ja“ geben. Verweigerer geben also unabhängig vom Ausgang des Zuallsexperiments die Antwort „Nein“. Das Ziel des Cheating Detection Modells besteht nun darin, den Anteil dieser Verweigerer zu schätzen, um den interessierenden Anteil π_A noch genauer zu spezifizieren. Dafür werden die Teilnehmer in drei disjunkte Gruppen eingeteilt:

- π = Anteil der Personen in der Stichprobe, die das Merkmal A haben und sich an die Vorgaben halten
- β = Anteil der Personen in der Stichprobe, die das Merkmal A nicht haben und sich an die Vorgaben halten
- γ = Anteil der Verweigerer in der Stichprobe

Der vierte Antworttyp, welcher diejenigen Teilnehmer beinhaltet, die unabhängig vom Ausgang des Zufallsexperiments immer mit „Ja“ antworten, ist sehr selten und wird daher von diesem Randomized Response Modell vernachlässigt. Somit lassen sich die drei zuvor dargestellten Anteile zu eins aufsummieren und der Anteil der Verweigerer lässt sich folglich auch als $\gamma = 1 - \pi - \beta$ schreiben.

Das zugehörige Baumdiagramm sieht wie folgt aus:

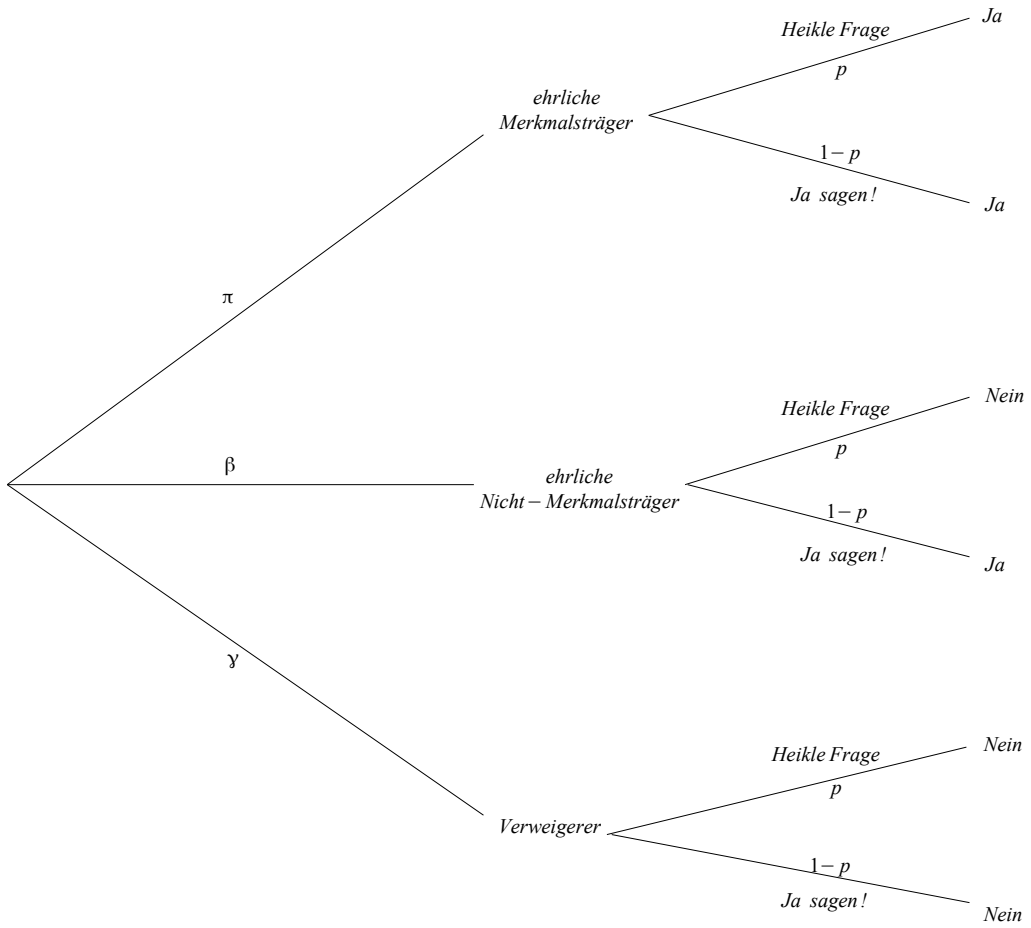

Abbildung 4: Das Cheating Detection Modell als Baumdiagramm
(Quelle: Ostapczuk et al. (2009, S. 270), Notation angepasst)

p gibt also wie zuvor die Wahrscheinlichkeit an, dass ein Teilnehmer die heikle Frage beantworten muss, wohingegen der Befragte mit der Gegenwahrscheinlichkeit $(1-p)$ aufgefordert wird, unabhängig von der heiklen Frage mit „Ja“ zu antworten. Da hier zwei unbekannte Parameter π und β vorliegen, müssen unter den zuvor aufgeführten Bedingungen (siehe S. 9) zwei Stichproben gezogen werden. So lassen sich dann zwei unabhängige Wahrscheinlichkeiten für eine „Ja“-Antwort herleiten:

$$\mathbb{P}_1(X_i = 1) = \pi + \beta \cdot (1 - p_1) \quad \text{und} \quad \mathbb{P}_2(X_i = 1) = \pi + \beta \cdot (1 - p_2)$$

Durch Einsetzten der in den Stichproben beobachteten „Ja“-Anteile, welche durch $\frac{n_1'}{n_1}$ und $\frac{n_2'}{n_2}$ gegeben sind, und anschließendes Auflösen der Gleichungen, ergeben sich schließlich die in Clark and Desharnais (1998, S.163, Notation angepasst) angegebenen Schätzer:

$$\hat{\pi} = \frac{\frac{n_1' \cdot (1-p_2)}{n_1} - \frac{n_2' \cdot (1-p_1)}{n_2}}{p_1 - p_2},$$
$$\hat{\beta} = \frac{(\frac{n_2'}{n_2} - \frac{n_1'}{n_1})}{p_1 - p_2},$$
$$\hat{\gamma} = 1 - (\hat{\pi} + \hat{\beta})$$

Der Anteil $\hat{\gamma}$ gibt also an, welcher Anteil an Verweigerern in der Studie erwartet wird. Unbekannt ist dabei aber deren wahre Kategorie, also ob die Verweigerer in Wahrheit jeweils der Gruppe mit oder ohne Merkmal A angehören. Für die Schätzung des interessierenden Anteils π lässt sich damit das Konfidenzintervall $[\hat{\pi}; \hat{\pi}+\hat{\gamma}]$ aufstellen. Die Untergrenze tritt ein unter der Annahme, dass keiner der Verweigerer das heikle Merkmal A trägt. Die Obergrenze setzt sich hingegen aus dem geschätzten Anteil $\hat{\pi}$ und dem geschätzten Anteil der Verweigerer $\hat{\gamma}$ zusammen, falls man annimmt, dass diese alle Merkmal A haben und es geleugnet haben. Falls die Umfrage-Teilnehmer nicht vollkommen ehrlich waren und sich also nicht alle an die Vorgaben gehalten haben, dann ist das Cheating Detection Modell effizienter als das Original-Modell von Warner (vgl. Clark and Desharnais, 1998, S.165).

2.5 Zusammenhangsanalysen bei dichotomen Randomized Response Variablen

Wenn es um die Analyse von heiklen Merkmalen geht, beschränkt sich die Hypothese in den meisten Fällen nicht nur auf ein einziges sensitives Merkmal, sondern es werden häufig zwei Variablen betrachtet, die möglicherweise in Beziehung zueinander stehen. Ziel ist es dann herauszufinden, ob die Merkmale A_1 und A_2 von einander unabhängig sind und falls das nicht der Fall ist, inwiefern sie miteinander in Verbindung stehen. Die Grundlage für solche Zusammenhangsanalysen besteht darin, dass beide Merkmale durch unabhängige Randomized Response Befragungen erhoben werden (vgl. Drane, 1976, S.566). Die Teilnehmer müssen also im dichotomen Fall für die beiden Merkmale jeweilse einen Randomized Response Durchgang absolvieren. Das Zufallsexperiment wird daher zweimal durchgeführt und leitet den Befragten dann zu zwei unterschiedlichen Fragen weiter, die jeweils zu einem der beiden Merkmale gehören. Somit ist die Unabhängigkeit zwischen den Antworten eines Befragten zu den beiden heiklen Fragestellungen gewährleistet. Sowohl beim Warner Modell, als auch beim Unrelated Question Modell und dem Forced Response Modell gilt,

dass aus der Unabhängigkeit der Antworten die Unabhängigkeit der heiklen Merkmale gefolgert werden kann. Dieser Zusammenhang gilt auch im umgekehrten Fall. Der dazugehörige Beweis findet sich in der Publikation von Drane (1976, S.568-573) und basiert auf der Überprüfung des Zusammenhangs zwischen den beobachteten Antwortwahrscheinlichkeiten zu den beiden Fragen und den wahren, unbekannten Merkmalsanteilen unter Annahme der Unabhängigkeitshypothese, die in Gleichung (3) im späteren Verlauf dieses Kapitels formalisiert wird. Der Zusammenhang basiert auf einer Matrix, wie sie auch in Kapitel 3 eingeführt wird. Wird also die Gültigkeit dieses Beweises angenommen, so bedeutet dies, dass man anstatt der unbekannten, wahren Werte, die gegebenen Randomized Response Antworten auf Unabhängigkeit testen kann, um zu untersuchen, ob zwei heikle Merkmale unabhängig voneinander sind.

Sei dabei λ_{ij} analog zu Drane (1976, S.567) als die Anzahl an Personen definiert, die bei Frage 1 mit $i = 1, 2$ antwortet und bei Frage 2 mit $j = 1, 2$. „1“ steht dabei für ein „Ja“, „2“ entsprechend für ein „Nein“. $\lambda_{i.}$ und $\lambda_{.j}$ sind die dazugehörigen Randhäufigkeiten. All diese Werte lassen sich unter der Annahme, dass die Merkmale unabhängig voneinander sind und zudem die Voraussetzungen einer Zufallsauswahl erfüllt wurden, wie folgt in eine 2×2-Kreuztabelle zusammenfassen:

		A_2		
		„Ja“=1	„Nein“=2	
A_1	„Ja“=1	λ_{11}	λ_{12}	$\lambda_{1.}$
	„Nein“=2	λ_{21}	λ_{22}	$\lambda_{2.}$
		$\lambda_{.1}$	$\lambda_{.2}$	1

λ_{11} gibt also beispielsweise an, wie viele Personen sowohl die erste, als auch die zweite heikle Frage bejaht haben. $\lambda_{1.}$ betrachtet hingegen nur Frage 1 und beinhaltet diejenigen Personen, die darauf mit „Ja“ geantwortet haben. Anhand der gegebenen Häufigkeitverteilung der Antworten lässt sich schließlich mit dem Chi-Quadrat-Unabhängigkeitstests überprüfen, ob die beiden Merkmale A_1 und A_2 voneinander unabhängig sind. In der Publikation von van den Hout and van der Heijden (2002, S.389) wird dargelegt, dass der Standard-Chi-Quadrat-Unabhängigkeitstest auf die Beobachtungen einer Randomized Response Umfrage angewendet und das Ergebnis schließlich auf die wahren Werte übertragen werden darf. Der Test hat dabei das korrekte Signifikanzniveau, doch die Power ist im Vergleich zu einem Test mit direkt beobachtbaren Häufigkeiten reduziert. Die Power beziehungsweise Teststärke eines

statistischen Tests gibt an, mit welcher Wahrscheinlichkeit sich der Test zugunsten der Alternativhypothese entscheidet, wenn diese richtig ist. Eine hohe Teststärke ist daher wünschenswert. Die hier vorliegende Reduktion der Power kann zum Beispiel durch einen größeren Stichprobenumfang ausgeglichen werden, was jedoch einen erhöhten zeitlichen Aufwand mit sich bringt. Es ist damit festzuhalten, dass dieses Vorgehen der Zusammenhangsanalyse bei zwei dichotomen Randomized Response Variablen durchaus gravierende Nachteile mit sich bringt, auch wenn es auf den ersten Blick sehr positiv erscheint, dass die Schätzung anhand der gegebenen Antworten überhaupt möglich ist. Dazu werden auf S. 389 der zugrundeliegenden Publikation weitere Quellen genannt, die sich mit den Hintergründen und der Durchführung des Chi-Quadrat-Tests bei fehlklassifizierten Variablen auseinandersetzen.

Der Chi-Quadrat-Test für die vorliegende Situation ist dabei wie folgt aufgebaut: Die Nullhypothese besagt, dass die gegebenen Antworthäufigkeiten und somit auch die beiden Merkmale A_1 und A_2 von einander unabhängig sind. Die zugehörige Prüfgröße, der Chi-Quadrat-Koeffizient χ^2, basiert auf dem Vergleich der tatsächlich beobachteten Antworthäufigkeiten von A_1 und A_2 mit den Häufigkeiten, die man bei Unabhängigkeit dieser Merkmale erwarten würde. Bei Unabhängigkeit würden sich die absoluten Zellhäufigkeiten in der Tabelle wie folgt aus den Randhäufigkeiten ergeben:

$$\lambda_{ij} = \frac{\lambda_{i.} \cdot \lambda_{.j}}{n} =: \tilde{\lambda_{ij}} \tag{3}$$

Diese, sowie alle folgenden Gleichungen des Kapitels basieren auf der Literatur von Fahrmeir et al. (2010, S.122-125), wobei die Notation angepasst wurde. Die Teststatistik ist schließlich gegeben durch:

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(\lambda_{ij} - \tilde{\lambda_{ij}})^2}{\tilde{\lambda_{ij}}} = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(\lambda_{ij} - \frac{\lambda_{i.} \cdot \lambda_{.j}}{n})^2}{\frac{\lambda_{i.} \cdot \lambda_{.j}}{n}} \tag{4}$$

Für große Abweichungen zwischen den beobachteten und den erwarteten Häufigkeiten, also für große χ^2, wird die Nullhypothese der Unabhängigkeit abgelehnt. Bei einem Signifikanzniveau von α wäre das der Fall, wenn die Teststatistik den Wert $\chi^2_{1,1-\alpha}$ übersteigt.

Liegt keine Unabhängigkeit zwischen den beiden dichotomen Variablen vor, so lässt sich anschließend anhand des Chi-Quadrat-Koeffizienten der Pearsonsche Kontingenzkoeffizient C berechnen. Dieser drückt die Stärke des Zusammenhangs zwischen

A_1 und A_2 aus und berechnet sich durch

$$C = \sqrt{\frac{\chi^2}{\chi^2 + n}}. \tag{5}$$

Da die Obergrenze dieses Zusammenhangsmaßes abhängig von der Anzahl der betrachtetet Dimensionen ist, wird in der Regel der korrigierte Kontingenzkoeffizient

$$C_{korr} = \sqrt{\frac{k}{k-1}} \cdot C = \sqrt{\frac{k}{k-1}} \cdot \sqrt{\frac{\chi^2}{\chi^2 + n}} \tag{6}$$

herangezogen, wobei k das Minimum der möglichen Merkmalsausprägungen der untersuchten Variablen ist. Für den dichotomen Fall gilt $k = 2$ und es ergibt sich schließlich $C_{korr} = \sqrt{2} \cdot \sqrt{\frac{\chi^2}{\chi^2 + n}}$. Der korrigierte Kontingenzkoeffizient kann Werte zwischen 0 und 1 annehmen, wobei ein hoher Wert auf ein hohes Maß an Abhängigkeit zwischen A_1 und A_2 hindeutet. Durch diese Zusammenhangsanalyse lassen sich Hypothesen testen, die auf zwei binären, heiklen Merkmalen basieren. Beispielhaft hierfür wäre die Hypothese „Nichtraucher verzichten in ihrem Leben häufiger auf den Konsum von Alkohol als Raucher.“

3 Randomized Response Daten als fehlklassifizierte Daten

3.1 Fehlklassifikation

Die Thematik der Fehlklassifikation wird in der Publikation von van den Hout and van der Heijden (2002) diskutiert. Sie bildet folglich die Grundlage dieses Kapitels. Auch die verwendeten Formeln wurden aus dieser Quelle (S.270) entnommen.

Das Problem der Fehlklassifikation tritt auf, wenn kategoriale Merkmale in einer Umfrage mit ungenauen oder fehlerhaften Messinstrumenten erhoben werden. Das hat zur Folge, dass die beobachteten Werte nicht notwendigerweise den wahren Ausprägungen der Umfrageteilnehmer entsprechen. Es kann somit zu einer falschen Zuordnung der Teilnehmer zu den einzelnen Kategorien kommen. Die Struktur der Fehlklasssifikation ist dabei unbekannt. Modelliert werden kann die Fehlklassifikation durch sogenannte Fehlklassifikationswahrscheinlichkeiten
$p_{ij} = \mathbb{P}$(beobachtete Kategorie i| tatsächliche Kategorie j), welche die Wahrscheinlichkeit angeben, dass der Teilnehmer, unter der Bedingung, dass er eigentlich Kategorie j angehört, aufgrund der Messung in Kategorie i eingeordnet wird. Eine Fehlklassifikation liegt dann vor, wenn $i \neq j$. Die Wahrscheinlichkeiten p_{ij} können schließlich in eine $k \times k$-Fehlklassifikationsmatrix P zusammengefasst werden, wobei k die Anzahl der Kategorien des interessierenden Merkmals angibt. Damit lässt sich der Zusammenhang zwischen den beobachteten Werten λ_i und den wahren Werten π_i schreiben als:

$$\lambda = P \cdot \pi$$

$$\text{mit } \lambda = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_k \end{pmatrix}, \pi = \begin{pmatrix} \pi_1 \\ \vdots \\ \pi_k \end{pmatrix}$$

λ ist also ein Vektor aus den beobachteten Wahrscheinlichkeiten, mit denen die einzelnen Kategorien $1, \ldots, k$ auftreten, wohingegen π den Vektor aus den wahren Kategorie-Wahrscheinlichkeiten darstellt. P sei eine Übergangsmatrix aus den bedingten Fehlklassifikationswahrscheinlichkeiten p_{ij}, wobei sich die Spalten zu 1 aufsummieren müssen (vgl. van den Hout and van der Heijden, 2002).

Für das binäre Modell von Warner würde sich somit folgende Darstellung ergeben:

$$\begin{pmatrix} \lambda \\ 1-\lambda \end{pmatrix} = \begin{pmatrix} p & 1-p \\ 1-p & p \end{pmatrix} \cdot \begin{pmatrix} \pi \\ 1-\pi \end{pmatrix}$$

3.2 Randomized Response Techniken für kategoriale Variablen

Die Modellierung anhand der Fehlklassifikationsmatrix P dient jedoch nicht vorrangig der Behandlung von dichotomen Fragestellungen. Sie bildet stattdessen die Grundlage für Randomized Response Umfragen, bei denen Merkmale im Fokus stehen, die $t(\geqq 2)$ Kategorien haben. Von diesen t Kategorien ist mindestens eine und maximal $t-1$ Kategorien heikel, also die Bereitschaft, ihre Angehörigkeit zuzugeben, tendenziell gering. Ein Beispiel für ein solches Merkmal ist die Anzahl von begangenen Straftaten oder durchgeführten Abtreibungen. Die Antworten können dabei ganze, positive Zahlen annehmen, einschließlich der 0. Große Zahlen sind dabei eher unwahrscheinlich und kommen nur selten vor. In diesem Fall wären alle Kategorien bis auf die 0, welche für keine Straftat beziehungsweise keine Abtreibung steht, heikel. Ziel ist es nun, die wahren, unbekannten Anteile π_i, wobei $i = 1, \cdots, t$ gilt, der einzelnen Kategorien in der Bevölkerung zu schätzen. Dafür gibt es analog zum dichotomen Fall verschiedene Vorgehensweisen. Einige davon wurden in der Literatur von Chaudhuri and Mukerjee (1988) vorgestellt. Dieses Buch dient damit als Basis für das restliche Kapitel 3.

3.2.1 Das erweiterte Warner Modell

Eine Möglichkeit, die Bevölkerungsanteile eines kategorialen Merkmals, das durch Randomized Response erhoben wurde, zu schätzen, ist eine Weiterentwicklung des Warner Modells. Die Grundidee beinhaltet, dass die Zugehörigkeit zu den einzelnen Kategorien dichotom abgefragt wird. Da sich der Schätzvorgang dieser Technik jedoch mit zunehmender Anzahl an Kategorien stark verkompliziert, wird hier zunächst der Fall von drei disjunkten Kategorien analog zur Quelle von Clark and Desharnais (1967) explizit dargelegt, bevor die Technik dann auf t Kategorien ausgeweitet wird.

Sei also ein heikles Merkmal A gegeben, das $j = 1, 2, 3$ Kategorien hat. Ziel ist es, den wahren Bevölkerungsanteil π_j zu schätzen, der den jeweiligen Gruppen angehört. Da jeder Befragte genau einer der drei Gruppen zugeordnet werden kann, gilt: $\sum_{j=1}^{3} \pi_j = 1$. Ein mögliches Anwendungsbeispiel ist die Verbreitung von unehelichen Schwangerschaften in der Bevölkerung. In diesem Fall würde die Population der Mütter in drei Subgruppen unterteilt werden: Frauen, die während ihrer Schwangerschaft bereits verheiratet sind, Frauen, die während ihrer Schwangerschaft heiraten und Frauen, die bei der Geburt ihres Kindes noch nicht verheiratet sind (vgl. Chaudhuri and Mukerjee, 1988, S.38). Zur Schätzung werden $i = j - 1 = 2$ Stichproben vom Umfang n_1 und n_2 gezogen. In den jeweiligen Stichproben wird schließlich die Gruppenzugehörigkeit durch ein Zufallsexperiment dichotom abgefragt. Dies kann zum Beispiel mit Hilfe eines Kartenstapels geschehen, wobei auf den Karten jeweils eine der Fragen „Gehören Sie Gruppe 1 an?“, „Gehören Sie Gruppe 2 an?“, „Gehören Sie der Gruppe 3 an?“ steht. Die unterschiedlichen Fragen treten dabei mit verschiedenenen, bekannten Wahrscheinlichkeiten auf, die zudem in den beiden Stichproben variieren.

Sei dabei analog zu Abul-Ela et al. (1967)

$$
\begin{aligned}
p_{ij} &= \text{Wahrscheinlichkeit, dass durch das Zufallsexperiment in der i-ten} \\
&\quad\ \text{Stichprobe die Zugehörigkeit zur } j\text{-ten Gruppe abgefragt wird} \\
X_{ir} &= \begin{cases} 1, & \text{falls der r-te Befragte der i-ten Stichprobe mit „Ja“ anwortet} \\ 0, & \text{falls der r-te Befragte der i-ten Stichprobe mit „Nein“ anwortet} \end{cases} \\
n_i' &= \text{Anzahl der „Ja“-Antworten in der i-ten Stichprobe}
\end{aligned}
$$

Das Vorgehen lässt sich mit dem folgenden Baumdiagramm veranschaulichen:

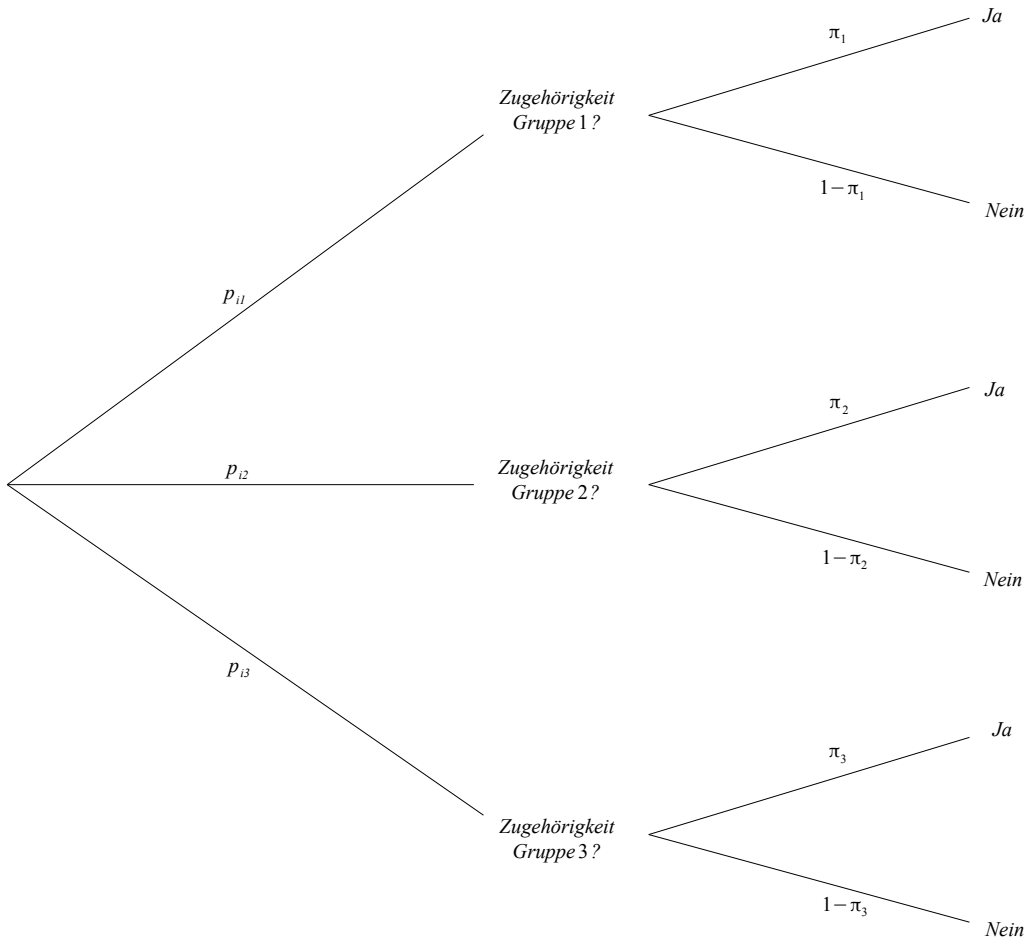

Abbildung 5: Das erweiterte Warner Modell für drei Kategorien als Baumdiagramm (Darstellung orientiert sich an Ostapczuk et al. (2009, S. 270), Notation angepasst)

Für den r-ten Befragten in der 1. beziehungsweise 2. Stichprobe ergeben sich für eine „Ja“-Antwort die Wahrscheinlichkeiten

$$\mathbb{P}_r(X_{1r} = 1) = \sum_{j=1}^{3} p_{1j} \cdot \pi_j = (p_{11} - p_{13}) \cdot \pi_1 + (p_{12} - p_{13}) \cdot \pi_2 + p_{13} = \lambda_1,$$

$$\mathbb{P}_r(X_{2r} = 1) = \sum_{j=1}^{3} p_{2j} \cdot \pi_j = (p_{21} - p_{23}) \cdot \pi_1 + (p_{22} - p_{23}) \cdot \pi_2 + p_{23} = \lambda_2.$$

Diese beiden, sowie auch die nun folgenden Gleichungen stammen aus der Publikation von Abul-Ela et al. (1967, S.992-994), welche diesem Kapitel zugrunde liegt.

Die gemeinsame Likelihood aus „Ja“- und „Nein“-Antworten der beiden Stichproben lautet:

$$L \propto \lambda_1^{n_1'} \cdot (1-\lambda_1)^{n_1-n_1'} \cdot \lambda_2^{n_2'} \cdot (1-\lambda_2)^{n_2-n_2'}$$

Daraus ergibt sich für die Maximum-Likelihood-Schätzer für π_1 und π_2:

$$\hat{\pi_1} = \frac{(\frac{n_1'}{n_1} - p_{13}) \cdot (p_{22} - p_{23}) - (\frac{n_2'}{n_2} - p_{23}) \cdot (p_{12} - p_{13})}{(p_{11} - p_{13}) \cdot (p_{22} - p_{23}) - (p_{12} - p_{13}) \cdot (p_{21} - p_{23})},$$

$$\hat{\pi_2} = -\frac{(\frac{n_1'}{n_1} - p_{13}) \cdot (p_{21} - p_{23}) - (\frac{n_2'}{n_2} - p_{23}) \cdot (p_{11} - p_{13})}{(p_{11} - p_{13}) \cdot (p_{22} - p_{23}) - (p_{12} - p_{13}) \cdot (p_{21} - p_{23})}$$

Außerdem: $\hat{\pi_3} = 1 - \hat{\pi_1} - \hat{\pi_2}$

Dieses Modell kann schließlich auf Merkmale ausgeweitet werden, die $t(\geq 2)$ Kategorien haben. Auch für diese Ausführungen bildet die zuvor genannte Quelle von Abul-Ela et al. die Grundlage. Um die wahren Bevölkerungsanteile $\pi_1, \cdots, \pi_t$ der t disjunkten Gruppen zu schätzen, wobei $0 < \pi_j < 1 \quad (j = 1, \cdots, t)$ und $\sum_{j=1}^{t} \pi_j = 1$ gilt, müssen $s = t - 1$ unabhängige Stichproben der Größe $n_1, \cdots, n_s$ aus der Bevölkerung gezogen werden. In diesen müssen, wie zuvor, Zufallsexperimente durchgeführt werden, wobei jeweils andere Kombinationen der p_{ij} vorliegen müssen. Es gilt auch hier $\sum_{j=1}^{t} p_{ij} = 1$ für $i = 1, \cdots, s$ und zudem $|P| \neq 0$, wobei P eine $s \times s$ Matrix ist, bei der die einzelnen Elemente die Form $p_{ik} - p_{it}$ mit $i, k = 1, 2, \cdots, s$ haben.

Das Baumdiagramm hierzu sieht wie folgt aus:

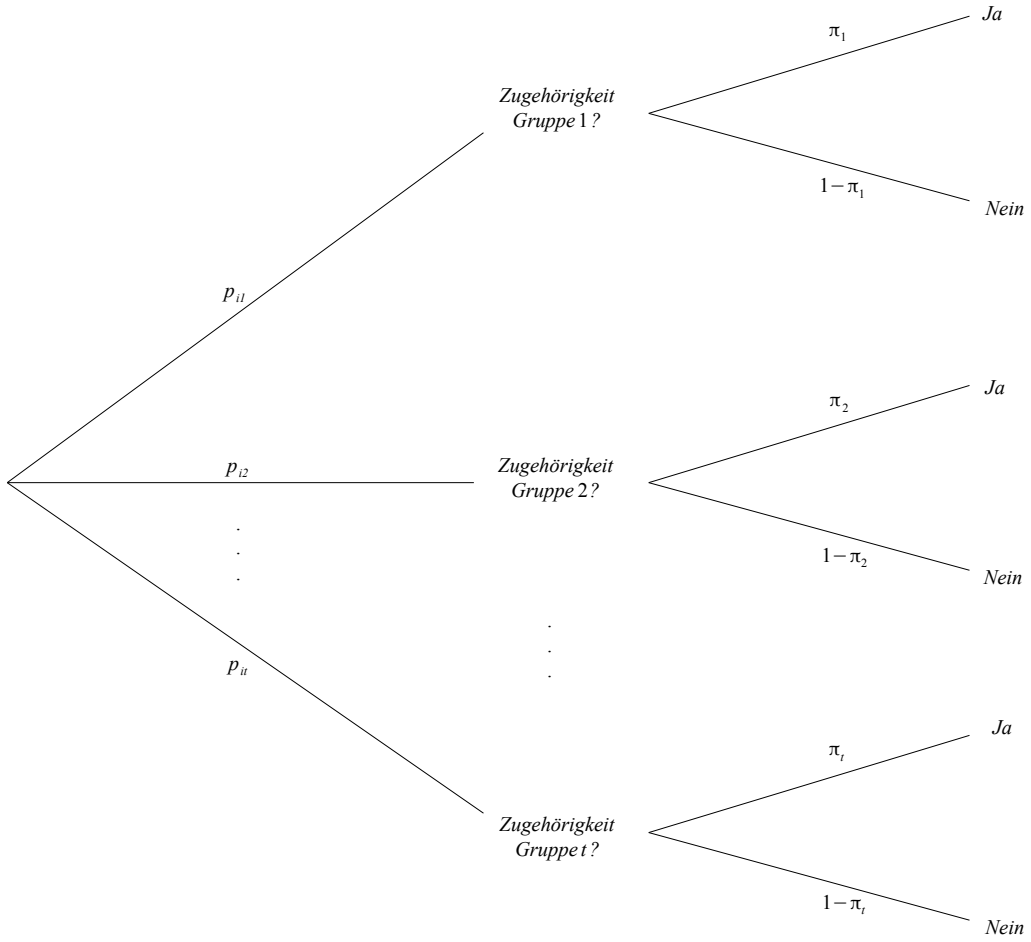

Abbildung 6: Das erweiterte Warner Modell für t Kategorien als Baumdiagramm
(Darstellung orientiert sich an Ostapczuk et al. (2009, S. 270), Notation angepasst)

Analog zum Fall mit drei Kategorien erhält man schließlich für die Wahrscheinlichkeit, dass der r-te Befragte der i-ten Stichprobe „Ja“ antwortet:

$$\mathbb{P}_r(X_{ir} = 1) = p_{i1} \cdot \pi_1 + p_{i2} \cdot \pi_2 + \cdots + p_{it} \cdot \pi_t = \lambda_i$$

Dann ergibt sich aus der gemeinsamen Likelihood für den multinomialen Fall (vgl. Abul-Ela et al., 1967, S. 1006)

$$L \propto \prod_{i=1}^{s} (\lambda_i) \cdot n_i' \cdot (1 - \lambda_i) \cdot n_i - n_i'$$

der gesuchten Maximum-Likelihood-Schätzer der π_i in Matrix-Notation:

$$\hat{\pi} = P^{-1} \cdot C$$

$$\text{mit } \hat{\pi} = \begin{pmatrix} \hat{\pi_1} \\ \hat{\pi_2} \\ \vdots \\ \hat{\pi_s} \end{pmatrix}, P = \begin{pmatrix} p_{11} - p_{1t} & p_{12} - p_{1t} & \cdots & p_{1s} - p_{1t} \\ p_{21} - p_{2t} & p_{22} - P_{2t} & \cdots & p_{2s} - p_{2t} \\ \vdots & \vdots & \cdots & \vdots \\ p_{s1} - p_{st} & p_{s2} - p_{st} & \cdots & p_{ss} - p_{st} \end{pmatrix}, C = \begin{pmatrix} \frac{n_1'}{n_1} - p_{1t} \\ \frac{n_2'}{n_2} - p_{2t} \\ \vdots \\ \frac{n_s'}{n_s} - p_{st} \end{pmatrix}$$

Der geschätzte Anteil der letzten Gruppe lässt sich dann berechnen durch: $\hat{\pi_t} = 1 - \sum_{i=1}^{s} \hat{\pi_i}$ (vgl. Abul-Ela et al., 1967, S.1006). Dieses Modell gründet auf dem zuvor erläuterten Zusammenhang zwischen beobachteten und wahren Werten, der sich durch eine Übergangsmatrix ausdrücken lässt. Der Vektor π der wahren Anteile wird geschätzt, indem die in der Stichprobe beobachteten Werte $\lambda_i = \frac{n_i'}{n_i}$ durch die Matrix P, bestehend aus den bekannten p_{ij}, korrigiert werden. Diese Korrektur kann dadurch begründet werden, dass einige der Personen aufgrund des gekoppelten Zufallsexperiments, in eine falsche Katgorie eingeordnet wurden.

3.2.2 Das erweiterte Unrelated Question Modell

Aber nicht nur das Warner Modell lässt sich auf multinomiale Fragestellungen übertragen, sondern auch das Unrelated Question Modell, wie es bei Chaudhuri and Mukerjee (1988, S.40-42) dargestellt wird. Sei dabei A wie zuvor ein heikles Merkmal mit t Kategorien. Außerdem ist ein harmloses, dichotomes Merkmal Y gegeben, das nicht in Verbindung mit A steht. μ_y sei dabei unbekannt. Um die Anteile in der Bevölkerung zu schätzen, werden t unabhängige Stichproben vom Umfang $n_1, \cdots, n_t$ gezogen. Das gekoppelte Zufallsexperiment ist so aufgebaut, dass bei jedem Teilnehmer der i-ten Stichprobe mit Wahrscheinlichkeit p_{ij} die Zugehörigkeit zur j-ten Gruppe ($j = 1, \cdots, t-1$) oder mit Wahrscheinlichkeit p_{it} die Zugehörigkeit zur Merkmalsgruppe Y abgefragt wird. Dabei gilt wieder $\sum_{j=1}^{t} p_{ij} = 1$. Dies kann zum Beispiel wie im vorherigen Verfahren anhand von Kartenstapeln durchgeführt werden.

Das zugehörige Baumdiagramm sieht dabei wie folgt aus:

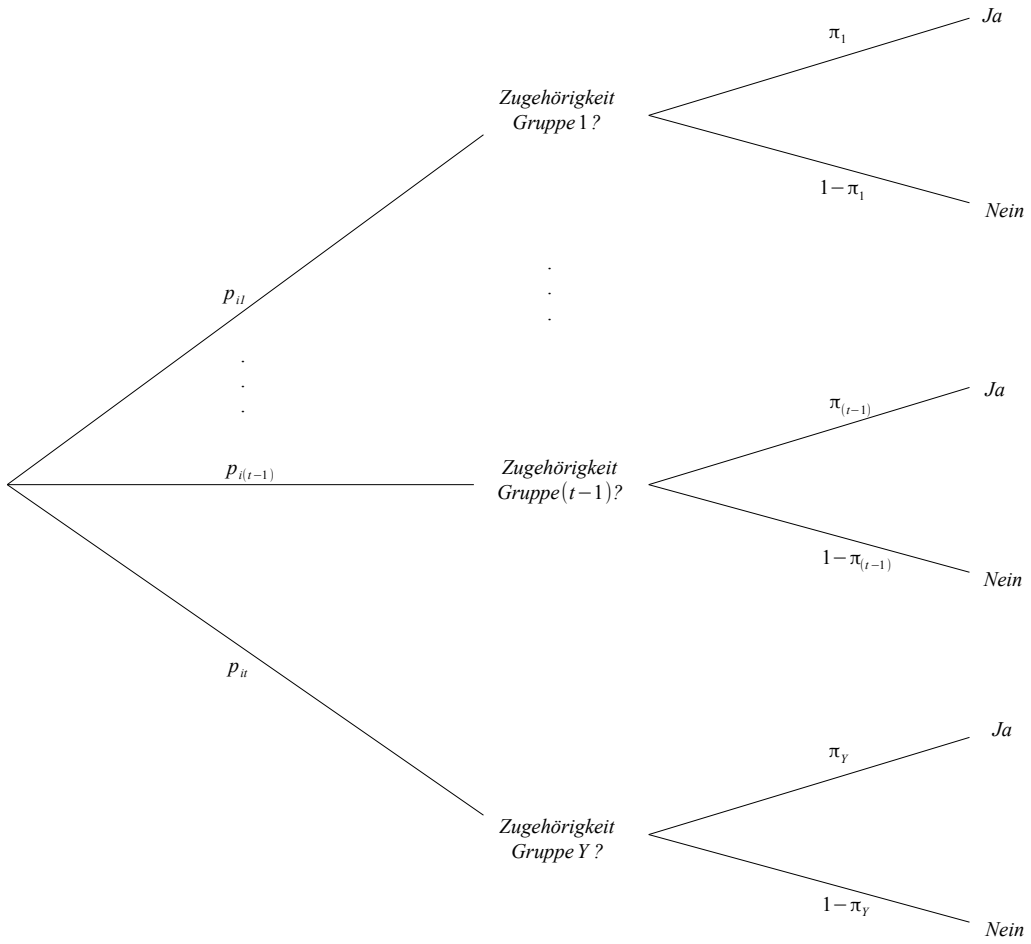

Abbildung 7: Das erweiterte Unrelated Question Modell für t Kategorien als Baumdiagramm
(Darstellung orientiert sich an Ostapczuk et al. (2009, S. 270), Notation angepasst)

Die Wahrscheinlichkeit für eine „Ja“-Antwort beträgt beim r-ten Befragten in der i-ten Stichprobe

$$\mathbb{P}_r(X_{ir} = 1) = \sum_{j=1}^{t-1} p_{ij} \cdot \pi_j + p_{it} \cdot \pi_y = \lambda_i, \quad i = 1, \cdots, t.$$

Man vergleiche dafür Chaudhuri and Mukerjee (1988, S.40). Definiert man schließlich $\lambda = (\lambda_1, \cdots, \lambda_t)'$, $\pi = (\pi_1, \cdots, \pi_{t-1}, \pi_y)'$ und $P = ((p_{ij}))$, so ergibt sich der Zusammenhang $\lambda = P \cdot \pi$ und damit für die geschätzten Anteile:

$$\hat{\pi} = P^{-1} \cdot \lambda$$

$$\text{mit } \hat{\pi} = \begin{pmatrix} \hat{\pi_1} \\ \hat{\pi_2} \\ \vdots \\ \hat{\pi_{t-1}} \\ \hat{\pi_y} \end{pmatrix}, P = \begin{pmatrix} p_{11} & p_{21} & \cdots & p_{t1} \\ p_{12} & p_{22} & \cdots & p_{t2} \\ \vdots & \vdots & \cdots & \vdots \\ p_{1t} & p_{2t} & \cdots & p_{tt} \end{pmatrix}, \lambda = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_t \end{pmatrix}$$

Zudem gilt: $\hat{\pi_t} = 1 - \sum_{j=1}^{t-1} \hat{\pi_j}$ (vgl. Chaudhuri and Mukerjee, 1988, S.41). Auch bei diesem Verfahren werden die beobachteten Anteile nachträglich durch eine Matrix korrigiert, die aus den festgelegten Wahrscheinlichkeiten, mit denen in den einzelnen Stichproben die verschiedenen Kategorien abgefragt wurden, besteht.

3.2.3 Das kategoriale Modell mit Vektor-Antworten

Neben diesen Verfahren für polychotome Fragestellungen, die auf dichotome Antworten basieren, gibt es noch die Möglichkeit von Vektorantworten. Diese wurden in Kapitel 3.3 bei Chaudhuri and Mukerjee (1988, S.42-46) thematisiert. Bei diesem Vorgehen gibt der Befragte bei einem heiklen Merkmal A mit t Kategorien keine „Ja"- beziehungsweise „Nein"-Antwort, sondern nennt seine Gruppenzugehörigkeit anhand einer Zahl zwischen 1 und t. Das Verfahren ist so aufgebaut, dass der Teilnehmer mit Wahrscheinlichkeit p seine wahre Gruppenzugehörigkeit i $(i = 1, 2, \cdots, t)$ Preis gibt oder mit jeweiligen Wahrscheinlicheiten $p_1, p_2, \cdots, p_t$ eine der vorgegebenen Zahlen $1, 2, \cdots, t$ nennt. Ein mögliches Zufallsexperiment kann dabei zum Beispiel eine Urne mit zwei verschieden farbigen Kugeln sein, wobei bei der zweiten Farbe die Kugeln zusätzlich jeweils mit einer der Zahlen 1 bis t versehen sind. Zieht der Teilnehmer eine Kugel der ersten Farbe, so muss er seine wahre Gruppenzugehörigkeit nennen. Zieht er dagegen die andere Farbe, so muss er als Antwort die Nummer angeben, die auf der Kugel geschrieben steht. Die Wahrscheinlichkeiten sind bekannt und es gilt schließlich $p + \sum_{i=1}^{t} p_i = 1$. Für den Anteil der Antwort j in der Stichprobe ergibt sich schließlich die Gleichung

$$\lambda_j = p \cdot \pi_j + p_j = (p + p_j) \cdot \pi_j + p_j(\pi_1 + \cdots + \pi_{j-1} + \pi_{j+1} + \cdots + \pi_t) \quad \text{mit} \quad j = 1, \cdots, t,$$

welche auch bei Chaudhuri and Mukerjee (1988, S.44) zu finden ist. Aufgrund der Gültigkeit von $\sum_{i=1}^{t} \pi_i = 1$ ist die Fehlklassifikationsmatrix dann gegeben durch

$$P = \begin{pmatrix} p+p_1 & p_1 & \cdots & p_1 \\ p_2 & p+p_2 & \cdots & p_2 \\ \vdots & \vdots & \cdots & \vdots \\ p_t & p_t & \cdots & p+p_t \end{pmatrix}.$$

Wie zuvor lassen sich dann die wahren Parameter schätzen durch

$$\hat{\pi} = P^{-1} \cdot \lambda,$$

mit $\hat{\pi} = \begin{pmatrix} \hat{\pi_1} \\ \vdots \\ \hat{\pi_t} \end{pmatrix}$, $\lambda = \begin{pmatrix} \lambda_1 \\ \vdots \\ \lambda_t \end{pmatrix}$.

Für den Gebrauch von Vektor-Antworten gibt es viele verschiedene Verfahren. Vorteil dieses Vorgehens ist, dass nur eine Stichprobe vom Umfang n aus der Population gezogen werden muss und sich dadurch der Aufwand sehr stark verringert.

3.3 Randomized Response Modell für multiattribute Fragestellungen

In vielen Fragestellungen kommen verschiedene kategoriale, heikle Merkmale vor, zwischen denen möglicherweise ein Zusammenhang besteht. Für jedes Merkmal eine unabhängige Stichprobe zu ziehen und die Randomized Response Prozedur durchzuführen, ist sehr aufwendig und kostspielig, gerade wenn es sich um eine große Anzahl von Merkmalen handelt. Das folgende Modell bietet eine Möglichkeit, mit nur einer Stichprobe mehrere kategoriale Merkmale unabhängig voneinander zu erheben. Die hier zu findende Darstellung basiert auf der Literatur von Chaudhuri and Mukerjee (1988, S.47-49).

Es sollen dabei m heikle Merkmale $A_1, \cdots, A_m$ erfasst werden, wobei das j-te Merkmal t_j Kategorien hat. Merkmal A_j lässt sich folglich in die disjunkten Kategorien $A_{j1}, \cdots, A_{jt_j}$ aufteilen. Anhand der Notation $i_j = 1, \cdots, t_j$ mit $j = 1, \cdots, m$ lassen sich alle Kategorien aller betrachteten Merkmale schreiben als $(A_{1i_1}, \cdots, A_{mi_m})$. Die wahren Bevölkerungsanteile der einzelnen Kategorien der jeweiligen Merkmale $A_{1i_1}, \cdots, A_{mi_m}$ sind im unbekannten, lexikografisch geordneten Vektor π zusammengefasst, der aus den Elementen $\pi_{i1}, \cdots, \pi_{im}$ besteht. Um diesen schließlich zu schätzen, wird eine Stichprobe vom Umfang n aus der Bevölkerung gezogen. Jeder der Teilnehmer muss schließlich m Durchläufe absolvieren, sodass jedes Merkmal

durch ein eigenes Randomized Response Verfahren erhoben wird. Bei der Abfrage des $j-$ten Merkmals sei die Notation der Wahrscheinlichkeit, dass ein Teilnehmer, der eigentlich der i_j-ten Kategorie angehört, mit $1, 2, \cdots, t_j$ antwortet, als $p_{1i_j}^{(j)}, p_{2i_j}^{(j)}, \cdots, p_{t_j i_j}^{(j)}$ festgelegt. Der Exponent zeigt also an, welches Merkmal betrachtet wird, der erste Wert im Index gibt an, welche Kategorie bei dem betrachteten Merkmal geantwortet wurde, der zweite Wert im Index hingegen, welche Kategoriezugehörigkeit in Wahrheit vorliegt. Es gilt folglich: $\sum_{u=1}^{t_j} p_{ui_j}^{(j)} = 1$. Die beobachteten Antworten u_j zu den einzelnen Merkmalen $j = 1, \cdots, m$ lassen sich in einem Antwortvektor $(u_1, u_2, \cdots, u_m)$ zusammenfassen, welcher mit Wahrscheinlichkeit $\lambda_{u_1 \cdots u_m}$ in der Stichprobe auftritt. Für alle Merkmale m ergibt sich schließlich $\lambda_{u_1 \cdots u_m} = \sum_{i_1=1}^{t_1} \cdots \sum_{i_m=1}^{t_m} (\prod_{j=1}^{m} p_{u_j i_j}^{(j)})$. Es handelt sich dabei um die Wahrscheinlichkeit, dass bei den m erhobenen Merkmalen der Antworttupel $(u_1, \cdots, u_m)$ genannt wird. Mit λ als lexikografisch geordnetem Vektor der $\lambda_{u_1 \cdots u_m}$ ergibt sich schließlich der Zusammenhang mit π, dem Vektor der wahren Anteile, in Matrixnotation:

$$\lambda = (P_1 \otimes P_2 \otimes \cdots \otimes P_m) \cdot \pi \tag{7}$$

Dabei sei $P_j = ((p_{u_j i_j}^{(j)}))$ mit $j = 1, ..., m$ die invertierbare Designmatrix und $\otimes$ das Kronecker-Produkt (vgl. Chaudhuri and Mukerjee, 1988, S.48). Laut Fahrmeir et al. (2009, S.447) ist das Kronecker-Produkt wie folgt definiert:

Seien A und B Matrizen der Ordnung $n \times p$ und $r \times q$. Dann ist das Kronecker-Produkt von A und B definiert als diejenige Matrix C der Ordnung $nr \times pq$ mit

$$C = A \otimes B = \begin{pmatrix} a_{11} \cdot B & a_{12} \cdot B & \cdots & a_{1p} \cdot B \\ \vdots & \vdots & \cdots & \vdots \\ a_{n1} \cdot B & a_{n2} \cdot B & \cdots & a_{np} \cdot B \end{pmatrix}.$$

Die zugehörigen Rechenregeln für dieses spezielle Produkt zweier Matrizen beliebiger Größe finden sich in der angegebenen Literatur.

Aus Gleichung (7) lässt sich schließlich der unverzerrte Schätzer für den Vektor π herleiten:

$$\hat{\pi} = (P_1^{-1} \otimes \cdots \otimes P_m^{-1}) \cdot \lambda$$

3.4 Zusammenhangsanalysen bei kategorialen Randomized Response Variablen

Bei multiattributen Fragestellungen ist es häufig, wie auch im dichotomen Fall, von Interesse, ob ein Zusammenhang zwischen den betrachteten, kategorialen Merkmalen besteht. Das lässt sich anhand des Chi-Quadrat-Unabhängigeitstests prüfen. Dabei wird für die Nullhypothese angenommen, dass die beiden Merkmale A_1, welches k Kategorien hat und A_2, welches l Kategorien hat, völlig unabhängig voneinander sind. Würde dies gelten, so würden sich die absoluten Häufigkeiten in einer Kreuztabelle aus A_1 und A_2 wie in Gleichung (3) aus den bekannten Randhäufigkeiten ergeben. Die Teststatistik basiert folglich wie in Gleichung (4) auf den Abweichungen zwischen beobachteten und erwarteten Zellhäufigkeiten, nur dass in diesem Fall mehr als zwei Kategorien gegeben sind und sich somit folgende Formel ergibt (vgl. Fahrmeir et al., 2010, S.123):

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(\lambda_{ij} - \tilde{\lambda_{ij}})^2}{\tilde{\lambda_{ij}}} = \sum_{i=1}^{k}\sum_{j=1}^{l} \frac{(\lambda_{ij} - \frac{\lambda_{i.}\cdot\lambda_{.j}}{n})^2}{\frac{\lambda_{i.}\cdot\lambda_{.j}}{n}}$$

Bei einem Signifikanzniveau von α wird die Nullhypothese abgelehnt, falls $\chi^2 > \chi^2_{(k-1)(l-1),1-\alpha}$. Für die Stärke des Zusammenhangs wird, wie im dichotomen Fall, der (korrigierte) Kontingenzkoeffizient herangezogen (vgl. Gleichungen (5) und (6)). Im vorliegenden Fall soll nun der Zusammenhang zwischen m heiklen Merkmalen A_i, die durch ein Randomized Response Verfahren erhoben wurden, getestet werden. Der Aufbau des folgenden Kapitels, sowie die Notation orientieren sich dabei an der Literatur von Chaudhuri and Mukerjee (1988, S.49-51).

Für den Unabhängigkeitstest der m heiklen Merkmale ergibt sich für die Nullhypothese:

$$H_0 : \pi = \pi^{(1)} \otimes \pi^{(2)} \otimes \cdots \otimes \pi^{(m)} \Longleftrightarrow H_0 : \lambda = \lambda^{(1)} \otimes \lambda^{(2)} \otimes \cdots \otimes \lambda^{(m)}$$

Das bedeutet, dass für die m Merkmale Unabhängigkeit angenommen wird, wenn sich der Vektor λ, bestehend aus den beobachtbaren Anteilen aller möglichen Kategorie-Kombinationen, darstellen lässt aus dem Kronecker-Produkt der beobachtbaren Antworten der einzelnen Merkmale. $\lambda^{(1)}$ ist also in diesem Zusammenhang ein Vektor, dessen Komponenten sich aus den Anteilen derjenigen Befragten zusammensetzt, die beim ersten Merkmal die gleiche Kategorie angegeben haben.

Die Teststatistik (vgl. Chaudhuri and Mukerjee, 1988, S.49) ist dann gegeben durch:

$$\chi^2 = n \cdot \sum_{u_1=1}^{t_1} \cdots \sum_{u_m=1}^{t_m} \frac{\hat{\lambda}^2_{u_1 \cdots u_m}}{\prod_{j=1}^{m} \hat{\lambda}^{(j)}_{u_j}} - n = n \cdot \sum_{u_1=1}^{t_1} \cdots \sum_{u_m=1}^{t_m} \frac{(\hat{\lambda}_{u_1 \cdots u_m} - \prod_{j=1}^{m} \hat{\lambda}^{(j)}_{u_j})^2}{\prod_{j=1}^{m} \hat{\lambda}^{(j)}_{u_j}} \quad (8)$$

Dabei darf nicht außer Acht gelassen werden, dass auch hier die fehlklassifizierten Antworten anstatt der unbekannten, wahren Werte auf Unabhängigkeit getestet werden. Der vorliegende Messfehler macht sich schließlich, wie auch schon im dichotomen Fall in Kapitel 2.5, bei der geringeren Power des Tests bemerkbar, was einen deutlichen Nachteil bedeutet.

Zur besseren Übersichtlichkeit wird nun die Gleichung (8) auf einen konkreten Fall angewendet (vgl. Chaudhuri and Mukerjee, 1988, S.50-51). Seien zwei Merkmale A_1 und A_2 gegeben, wobei $t_1 = 3$ und $t_2 = 2$. Dann lautet die Nullhypothese:

$$H_0 : \lambda = \lambda^{(1)} \otimes \lambda^{(2)}$$

Die zugehörige Teststatistik ist dann gegeben durch:

$$\chi^2 = n \cdot \sum_{u_1=1}^{3} \sum_{u_2=1}^{2} \frac{\hat{\lambda}_{u_1 u_2} - \hat{\lambda}^{(1)}_{u_1} \cdot \hat{\lambda}^{(2)}_{u_2}}{\hat{\lambda}^{(1)}_{u_1} \cdot \hat{\lambda}^{(2)}_{u_2}}$$

Eine beispielhafte Hypothese, die zu diesem Modell passt, lautet „Raucher trinken regelmäßig größere Mengen an Alkohol als Nichtraucher.“ A_1 entspricht dann dem Merkmal „Alkoholkonsum“, welches hier zum Beispiel in den drei Kategorien „Im letzten Monat über 1000 ml Alkohol getrunken“, „Im letzten Monat zwischen 250 ml und 1000 ml Alkohol getrunken“ und „Im letzten Monat weniger als 250 ml Alkohol getrunken“ erfasst wird. A_2 ist in diesem Fall dichotom und hat die beiden Ausprägungen „Raucher“ und „Nichtraucher“.

4 Randomized Response Daten als Daten mit Messfehler

4.1 Messfehler

Neben dem systematischen Fehler, welcher unter anderem auf Falschantworten gründet und folglich durch die Randomized Response Technik verringert werden soll, stellt der stochastische Fehler die zweite Art von Messfehlern da. Dieser variiert von Messung zu Messung und hat im einfachen, klassischen Fall die Form: $X = A + U$. A ist dabei eine latente Variable, die nicht direkt messbar ist. Stattdessen ist nur die fehlerbehaftete Variable X beobachtbar. Der zufällige Messfehler U hat den Erwartungswert 0 und ist zudem unkorreliert mit der interessierenden Variable A. Dieser Zusammenhang lässt sich, wie im folgenden Kapitel gezeigt wird, auch auf quantitative Variablen, die mit Randomized Response erhoben wurden, übertragen.

4.2 Randomized Response Techniken für quantitative Variablen

Randomized Response Techniken sind auch bei Erhebungen von quantitativen Daten anwendbar. So kann zum Beispiel das individuelle Einkommen abgefragt werden, welches unzählige Werte annehmen kann. Andere denkbare Fragestellungen sind zum Beispiel die persönlichen Ersparnisse oder das Geld, das für illegale oder unmoralische Zwecke ausgegeben wurde. Überträgt man die zuvor erläuterte Messfehler-Theorie auf die vorliegende Situation, so bedeutet das, dass die wahren Antworten auf die heikle Frage durch die Randomized Response Prozedur mit einem zufälligen Fehler „verschmutzt“ werden, dessen Erwartungswert 0 ist und der mit den wahren, unbekannten Werten nicht korreliert (vgl. Fox and Tracy, 1984, S.192). Die Beobachtungen haben also einen Messfehler, der durch das Zufallsexperiment generiert wurde und in Kauf genommen wird, um die Privatsphäre der Befragten zu schützen. Dieser kann folglich durch das Wissen über die Wahrscheinlichkeitsverteilung im gekoppelten Zufallsexperiment korrigiert werden. Für das zugrundeliegende Modell gibt es analog zum dichotomen und kategorialen Fall verschiedene Möglichkeiten.

4.2.1 Das quantitative Unrelated Question Modell

Das quantitative Unrelated Question Modell ist eine Erweiterung des einfachen, dichotomen Unrelated Question Modells, welches bereits in Kapitel 2.2 dargelegt wurde. Der quantitative Fall wurde detailliert in der Publikation von Greenberg et al. (1971) dargestellt und diskutiert. Diese Quelle dient daher als Grundlage

für den Aufbau und die verwendeten Formeln dieses Kapitels. Die Notation wurde angepasst.

Wie zuvor im dichotomen Fall entscheidet auch beim quantitativen Unrelated Question Modell ein Zufallsexperiment darüber, ob die heikle Frage A oder die harmlose Frage Y beantwortet werden soll. Da es sich bei A um eine Fragestellung handelt, bei der die Antwort metrisch ist, müssen auch die Antworten der Alternativfrage Y den gleichen Wertebereich annehmen, um eine mögliche Zuordnung der individuellen Antworten zu den Fragen zu vermeiden. Die Gesamtverteilung der Antworten besteht folglich aus numerischen Antworten zu beiden Fragen, wobei diese nicht zu unterscheiden und damit zuzuordnen sind. Die Verteilung muss damit statistisch unterteilt werden, um gute Schätzer für die unbekannten Parameter, nämlich Mittelwerte μ_A, μ_Y und Varianzen σ_A^2, σ_Y^2 des heiklen, sowie des harmlosen Merkmals zu erhalten. Es werden dazu zwei unabhängige, sich nicht-überlappende Stichproben vom Umfang n_1 und n_2 erhoben.

Sei dabei analog zu Greenberg et al. (1971, S.244)

p_i = Wahrscheinlichkeit, dass die heikle Frage von einem Teilnehmer in Stichprobe i = 1,2 beantwortet werden muss, $\quad p_1 \neq p_2$

$q_i = 1 - p_i$ = Wahrscheinlichkeit, dass die harmlose Frage von einem Teilnehmer in Stichprobe i = 1,2 beantwortet werden muss

X_{ij} = beobachtbare Antwort eines Befragten j in Stichprobe i, $\quad j = 1, 2, ..., n_i$

$f(x)$ = Wahrscheinlichkeitsfunktion der heiklen Frage, $\quad \mathbb{E}_f(X) = \mu_A$

$g(x)$ = Wahrscheinlichkeitsfunktion der harmlosen Frage, $\quad \mathbb{E}_g(X) = \mu_Y$

$\hat{\mu_A}$ = geschätzter Stichprobenmittelwert der Verteilung des heiklen Merkmals

$\hat{\mu_Y}$ = geschätzter Stichprobenmittelwert der Verteilung des harmlosen Merkmals

X_{ij}, die beobachtbare Antwort eines Individuums j aus Stichprobe i, lässt sich mit Wahrscheinlicheit p_i der heiklen Frage A zuordnen und mit Wahrscheinlichkeit $1-p_i$ der harmlosen Frage Y. Daraus lässt sich die individuelle Wahrscheinlichkeitsfunktion für eine Antwort X schreiben als

$$\psi_i(x_i) = p_i \cdot f(x_i) + (1 - p_i) \cdot g(x_i) \quad \text{mit} \quad i = 1, 2.$$

Aus den erwarteten Antworten

$$\mu_{x_1} = \mathbb{E}(X_1) = p_1 \cdot \mu_A + (1 - p_1) \cdot \mu_Y \quad \text{und} \quad \mu_{x_2} = \mathbb{E}(X_2) = p_2 \cdot \mu_A + (1 - p_2) \cdot \mu_Y$$

in den beiden Stichproben ergeben sich schließlich durch gegenseitiges Einsetzen und Auflösen, sowie durch Ersetzen der μ_{x_i} durch die Anwortmittelwerte der Stichproben $\overline{X}_i$, die geschätzten Mittelwerte der Verteilungen des interessierenden beziehungsweise des harmlosen Merkmals:

$$\begin{aligned} \hat{\mu_A} &= \frac{(1 - p_2) \cdot \overline{X}_1 - (1 - p_1) \cdot \overline{X}_2}{p_1 - p_2}, \\ \hat{\mu_Y} &= \frac{p_2 \cdot \overline{X}_1 - p_1 \cdot \overline{X}_2}{p_2 - p_1} \end{aligned} \tag{9}$$

Die zugehörigen Varianzen finden sich in Greenberg et al. (1971, S. 245).

Wie in der Publikation von Warner (1971) vorgeschlagen, lässt sich das Modell auch wie ein Regressionsmodell schreiben und behandeln. Durch diese Schreibweise wird auch die zuvor erläuterte Übertragung der Messfehler-Problematik auf die vorliegende Situation deutlich. Es sei analog zu Greenberg et al. (1971, S.245, Notation angepasst)

$$\underline{X} = \underline{\underline{P}} \cdot \underline{\beta} + \underline{U}$$

mit

$$\underline{X} = (X_{11}, X_{12}, \cdots, X_{1n_1}, X_{21}, X_{22}, \cdots, X_{2n_2})'$$

$$\underline{\underline{P}} = \begin{pmatrix} p_1 & p_1 & \cdots & p_1 & p_2 & p_2 & \cdots & p_2 \\ q_1 & q_1 & \cdots & q_1 & q_2 & q_2 & \cdots & q_2 \end{pmatrix}'$$

$$\underline{\beta} = \begin{pmatrix} \mu_A \\ \mu_Y \end{pmatrix}$$

$$\underline{U} = (u_{11}, u_{12}, \cdots, u_{1n_1}, u_{21}, u_{22}, \cdots, u_{2n_2})'$$

$\mathbb{E}(\underline{U})$ sei dabei der Nullvektor und die zugehörige Kovarianzmatrix $\underline{\underline{V}}$ sei eine Diagonalmatrix mit $V(X_1)$ auf den ersten n_1 Diagonalfeldern und $V(X_2)$ auf den danach folgenden n_2 Diagonalfeldern.

Mit der Methode der kleinsten Quadrate lässt sich $\underline{\beta}$ dann unverzerrt schätzen:

$$\underline{\hat{\beta}} = \begin{pmatrix} \hat{\mu_A} \\ \hat{\mu_Y} \end{pmatrix} = (\underline{\underline{P}}'\underline{\underline{V}}^{-1}\underline{\underline{P}})^{-1}\underline{\underline{P}}'\underline{\underline{V}}^{-1}\underline{X}$$

Dieser Term entspricht Gleichung (9). Sie geben beide den Schätzer für die Mittelwerte des heiklen Merkmals A, sowie für das harmlose Merkmal Y in der Bevölkerung an.

Vereinfacht wird das Verfahren, wenn die Verteilung des harmlosen Merkmals Y bekannt ist. Deshalb wird in vielen Anwendungsbeispielen Y so gewählt, dass μ_y und σ_Y^2 bekannt sind. Zur Schätzung des Mittelswerts von A benötigt man so nur eine Stichprobe und es ergibt sich bei analogem Vorgehen

$$\hat{\mu_A} = \frac{\overline{X} - (1-p) \cdot \mu_Y}{p}.$$

Man vergleiche dazu Greenberg et al. (1971, S.247).

Das quantitative Unrelated Question Modell bringt jedoch auch gravierende Probleme mit sich. Diese hängen insbesondere mit der harmlosen Frage Y zusammen, deren Auswahl eine große Rolle für dieses Verfahren spielt. Der Mittelwert, sowie die Varianz von Y müssen den entsprechenden Parametern des heiklen Merkmals A sehr stark ähneln, um die individuellen Antworten nicht einer der beiden Fragen zuordnen zu können. Da aber Mittelwert und Varianz von A nicht bekannt sind, ist auch die Wahl eines passenden Y sehr schwierig. Häufig tritt zudem das Problem auf, dass bei einer der beiden Fragen tendenziell eher gerundete Werte angegeben werden, wohingegen bei der anderen Frage unbewusst eher exakte Angaben gemacht werden. Obwohl A und Y den gleichen Wertebereich haben, ist in solchen Fällen ebenfalls eine Zuordnung möglich und der Vorteil der Randomized Response Technik geht verloren (vgl. Eichhorn and Hayre, 1983, S.308).

4.2.2 Das Additive Constants Modell

Das Additive Constants Modell ist eine alternative Vorgehensweise für quantitative Daten, die nicht mit den zuvor herausgearbeiteten Problemen konfrontiert ist. Als Quelle wurde für das folgende Kapitel die Publikation von Himmelfarb and Edgell (1980) herangezogen. Die Notation wurde dabei den vorangehenden Kapiteln angepasst.

Die Zufallsverschlüsselung beim Additive Constants Modell beruht darauf, dass der Teilnehmer zu seiner wahren Antwort auf die heikle Frage eine Konstante addiert, die aus verschiedenen vorgegebenen Werten durch ein Zufallsexperiment ausgewählt wurde. Die Zufallsauswahl kann dabei zum Beispiel anhand eines Kartenstapels durchgeführt werden. Auf den Karten steht dabei jeweils die Aufforderung, zur eigenen, wahren, numerischen Antwort A die vorgegebene Konstante K_i zu addieren, wobei diese auf den verschiedenen Karten variiert. Es gibt c verschiedene Konstanten K_i, wobei jede jeweils mit Wahrscheinlichkeit $p_i \quad (i = 1, \cdots, c)$ auftritt. Für die Konstante können sowohl positive, als auch negative Werte, sowie auch die 0 vorgegeben sein, was bedeutet, dass die Befragten ihren wahren Wert teilweise überschätzt, teilweise unterschätzt und teilweise direkt angeben. Der Interviewer erfährt weder, welche Karte gezogen wurde, sprich welche Konstante hinzugefügt wurde, noch die wahre numerische Antwort. Ihm wird nur die Summe mitgeteilt, das heißt, er erhält Antworten der Form $X = A + K_i$. Auch hier erkennt man, dass die Beobachtungen fehlerbehaftet sind und sich aus den wahren Werten und jeweils einem zufälligen Fehlerterm, der hier der Konstante entspricht, zusammensetzt.

Sei nun g die Wahrscheinlichkeitsfunktion der numerischen Antworten der Befragten und f die unbekannte Wahrscheinlichkeitsfunktion der wahren Merkmalswerte der Befragten. Dann lässt sich der zuvor aufgeführte Zusammenhang darstellen als

$$g(x) = \sum_{i=1}^{c} p_i \cdot f(A + K_i).$$

Daraus ergibt sich

$$\mathbb{E}(X) = \sum_{i=1}^{c} p_i \cdot \mathbb{E}(A \cdot K_i) \overset{\text{unabh.}}{=} \mathbb{E}(A) + \sum_{i=1}^{c} p_i \cdot K_i = \mu_A + \sum_{i=1}^{c} p_i \cdot K_i.$$

Setzt man schließlich für $\mathbb{E}(X)$ den Stichproben-Mittelwert der beobachtbaren, numerischen Antworten $\overline{X}$ ein, so erhält man für den gesuchten Schätzer

$$\hat{\mu_A} = \overline{X} - \sum_{i=1}^{c} p_i \cdot K_i.$$

Diese Gleichung, sowie die zugehörige Varianz des Schätzers findet sich bei Himmelfarb and Edgell (1980, S.526), wobei die Notation angepasst wurde.

In der Anwendung kommen häufig Spezialfälle dieser Technik zum Einsatz. So wird zum Beispiel des öfteren ein Zufallsexperiment durchgeführt, das den Befragten mit

Wahrscheinlichkeit p dazu weiterleitet, die heikle Frage wahrheitsgemäß zu beantworten und mit Wahrscheinlichkeit $(1-p)$ auffordert, zum wahren Wert zunächst eine Konstante zu addieren, bevor die Antwort schließich genannt wird. Hierbei gibt es einige Abwandlungen.

Diese Technik hat nicht nur den Vorteil, dass sie nicht mit den zuvor aufgeführten Problemen umgehen muss, sondern zudem auch effizienter ist. Je kleiner dabei die Konstanten gewählt werden, desto größer ist die Effizienz. Ihr gravierender Nachteil besteht hingegen darin, dass Ausreißer trotz der Addition mit einer Zufallszahl in der Regel noch als solche erkennbar bleiben. Damit ist die Anonymität der Befragten nur teilweise gewährleistet. An diesem Punkt setzen wiederum andere Verfahren, wie das multiplikative Modell für quantitative Daten an. Dieses beruht auf demselben Vorgehen wie das Additive Constants Modell, nur dass hierbei der wahre Wert mit einer ausgewählten Zufallszahl, welche einer bekannten Verteilung folgt, multipliziert wird. Auch die Kombination aus additivem und multiplikativem Modell ist möglich. Dabei erhält der Interviewer Antworten der Form $X = A \cdot S + U$, wobei S und U zwei unabhängige Zufallszahlen sind (vgl. Eichhorn and Hayre, 1983, S.315).

4.3 Zusammenhangsanalysen bei quantitativen Randomized Response Variablen

Das Ziel des folgenden Abschnitts ist es nun, den Zusammenhang zwischen zwei sensitiven Merkmalen, die beide einen numerischen Wertebereich haben und mittels Randomized Response erhoben wurden, zu bestimmen. Als Maß des Zusammenhangs bei zwei quantitativen Variablen X und Y wird in der Regel der Korrelationskoeeffizient nach Bravais-Pearson herangezogen. Die folgende dazugehörige Formel mit angepasster Notation ist unter anderem bei Fahrmeir et al. (2010, S. 136) nachzuschlagen:

$$\rho_{XY} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2 - \sum_{i=1}^{n}(y_i - \overline{y})^2}} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} \tag{10}$$

Im Folgenden soll zunächst die Korrelation zwischen zwei Variablen betrachtet werden, die durch das quantitative Unrelated Question Modell gemessen wurden. Dieser Abschnitt orientiert sich an der Publikation von Fox and Tracy (1984). Danach folgt eine Zusammenhangsanalyse zwischen zwei Merkmalen, erhoben durch das Additive Constants Modell. Dieses basiert auf dem Artikel von Himmelfarb and Edgell (1982).

Seien also nun A_1 und A_2 die beiden latenten Variablen, die durch das quantitative Unrelated Question Modell erhoben wurden, wobei für jeweils ein Merkmal eine Stichprobe gezogen wird. Dabei gilt $p_1 \neq p_2$. Die zugehörigen beobachtbaren Variablen mit Messfehler seien mit X_1 und X_2 notiert. Aus den beobachtbaren Antworten, sowie den Stichprobenmittelwerten lässt sich die Kovarianz der beiden Merkmale X_1 und X_2 berechnen, sowie auch deren jeweilige Standardabweichungen. Damit sind alle Komponenten für den Korrelationskoeffizient nach Bravais Pearson (vgl. Gleichung (10)) zwischen X_1 und X_2 bekannt. Der berechnete Korrelationskoeffizient $\rho_{X_1X_2}$ entspricht jedoch nicht dem Korrelationskoeffizient zwischen A_1 und A_2, da die beobachteten Variablen X_1 und X_2 einen Messfehler enthalten, der durch die Randomized Response Prozedur aufgetreten ist. Um diesem entgegenzuwirken kann der berechnete Korrelationskoeffizient $\rho_{X_1X_2}$ korrigiert werden und man erhält so eine Schätzung für die Korrelation zwischen A_1 und A_2 (vgl. Fox and Tracy, 1984, S.193):

$$\rho_{X_1X_2} = \frac{\sigma_{X_1X_2}}{\sigma_{X_1}\sigma_{X_2}} \overset{\sigma_{X_1X_2}=\sigma_{A_1A_2}}{\Longleftrightarrow} \rho_{X_1X_2} = \frac{\sigma_{A_1A_2}}{\sigma_{X_1}\sigma_{X_2}} \overset{X_1=A_1+U_1, X_2=A_2+U_2}{\Longleftrightarrow}$$

$$\rho_{X_1X_2} = \frac{\sigma_{A_1A_2}}{(\sigma_{A_1}+\sigma_{U_1})(\sigma_{A_2}+\sigma_{U_2})} \Longleftrightarrow \rho_{X_1X_2} = \frac{\sigma_{A_1A_2}}{\sqrt{(\sigma^2_{A_1}+\sigma^2_{U_1})(\sigma^2_{A_2}+\sigma^2_{U_2})}}$$

$$\overset{\sigma_{A_1A_2}=\rho_{A_1A_2}\cdot\sigma_{A_1}\cdot\sigma_{A_2}}{\Longleftrightarrow} \rho_{X_1X_2} = \frac{\rho_{A_1A_2}\cdot\sigma_{A_1}\cdot\sigma_{A_2}}{\sqrt{(\sigma^2_{A_1}+\sigma^2_{u_1})(\sigma^2_{A_2}+\sigma^2_{u_2})}}$$

$$\Longrightarrow \hat{\rho}_{A_1A_2} = \rho_{X_1X_2}\cdot\sqrt{(1+\frac{\sigma^2_{U_1}}{\sigma^2_{A_1}})(1+\frac{\sigma^2_{U_2}}{\sigma^2_{A_2}})}$$

Im ersten Schritt wird die Kovarianz der beobachteten Werte durch die Kovarianz der wahren Werte ersetzt. Warum dies möglich ist, wird im nächsten Abschnitt erklärt, da die Berechnung der Korrelation zwischen zwei Variablen, die durch das Additive Constants Modell erhoben wurden, auf demselben Vorgehen beruht. In den nächsten Schritten folgen mathematische Umwandlungen, die letztendlich zu dem Ergebnis führen, dass die Stichprobenkorrelation mit einem Korrekturfaktor multipliziert werden muss. Dieser setzt sich zusammen aus der Varianz des Messfehlers und den Varianzen der heiklen Merkmale A_1 und A_2. Da diese Parameter nicht bekannt sind, müssen sie wie bei Fox and Tracy (1984, S.193) geschätzt werden.

Seien nun die beiden sensitiven Merkmale A_1 und A_2 mit dem Additive Constants Modell erhoben worden. Neben den jeweiligen Bevölkerungsmittelwerten μ_{A_1} und μ_{A_2}, sowie den Varianzen $\sigma^2_{A_1}$ und $\sigma^2_{A_2}$, ist auch der Korrelationskoeffizient $\rho_{A_1A_2}$ zwischen den beiden latenten Merkmalen unbekannt. Stattdessen sind die beobachteten Variablen X_1 und X_2 durch die Antworten der Befragten gegeben. Diese haben jedoch einen Messfehler und setzen sich wie in Kapitel 5.2 beschrieben, aus den wahren Werten und den zufällig ausgewählten Konstanten zusammen. Die Konstanten sind sowohl voneinander, als auch von den wahren Werten unabhängig. Unter dieser Bedingung der Unabhängigkeit gilt, dass sich die Kovarianz zwischen den wahren Werten einer latenten Variable und einer anderen Variable nicht verändert, wenn zu den wahren Werten jeweils eine Zufallsvariable hinzugefügt wird. Dasselbe gilt auch, wenn beide dieser Variablen in dieser Form transformiert werden (vgl. Himmelfarb and Edgell, 1982, S.282). Auf den vorliegenden Fall übertragen bedeutet dies, dass die Kovarianz zwischen den wahren Werten der beiden interessierenden Merkmale A_1 und A_2 , der Kovarianz der beobachteten Antworten X_1 und X_2 entspricht. Also $\sigma_{A_1A_2} = \sigma_{X_1X_2}$. Die Korrelation zwischen den beobachteten Werten $\rho_{X_1X_2} = \dfrac{\sigma_{X_1X_2}}{\sqrt{\sigma^2_{X_1} \cdot \sigma^2_{X_2}}}$ kann also als ein Schätzer für die Korrelation zwischen den wahren, unbekannten Werten herangezogen werden. Jedoch erhöht sich durch das Randomized Response Verfahren bekanntermaßen die Streuung und so gilt: $\sigma^2_{X_1} > \sigma^2_{A_1}$ und $\sigma^2_{X_2} > \sigma^2_{A_2}$. Die Korrelation zwischen den beobachteten Werten muss also korrigiert werden und so lässt sich die Korrelation zwischen den wahren Werten schätzen durch

$$\hat{\rho}_{A_1A_2} = \frac{\sigma_{X_1X_2}}{\sqrt{\sigma^2_{A_1} \cdot \sigma^2_{A_2}}}. \tag{11}$$

Da die Varianzen der wahren Bevölkerungswerte der beiden Merkmale A_1 und A_2 nicht bekannt sind, müssen sie wie folgt geschätzt werden:

$$\hat{\sigma^2_{A_1}} = \hat{\sigma^2_{X_1}} - \sum_{i=1}^{c} p_i \cdot K_i^2 - (\sum_{i=1}^{c} p_i \cdot K_i)^2$$

$$\hat{\sigma^2_{A_2}} = \hat{\sigma^2_{X_2}} - \sum_{i=1}^{c} p_i \cdot K_i^2 - (\sum_{i=1}^{c} p_i \cdot K_i)^2.$$

$\hat{\sigma^2_X}$ ist dabei die beobachtbare Stichprobenvarianz der Antworten. Die Korrektur steckt also in der Schätzung der unbekannten Streuung der beiden Merkmale. Um jedoch letztendlich eine gute Schätzung für die Korrelation zwischen den wahren

Werten zu erhalten, sollte eine große Stichprobe vorliegen. Gleichung (11) ist schließlich auch gültig, wenn eine der beiden betrachteten Variablen nicht mit Hilfe des Randomized Response Verfahrens erhoben wurde. In dem Fall muss die Varianz dieser Variable nicht korrigiert werden (vgl. Himmelfarb and Edgell, 1982, S.283). Damit wurde gezeigt, dass der Korrelationskoeffizient nach Bravais-Pearson auch bei quantitativen Randomized Response Variablen anwendbar ist. Folglich lässt sich festhalten, dass sich Zusammenhangsanalysen zwischen Variablen, die durch Randomized Response erhoben wurden und die deswegen fehlerbehaftet sind, durch Korrektur der Beobachtungen durchführen lassen, unabhängig davon, welches Skalenniveau vorliegt.

5 Regressionsanalysen bei Randomized Response Variablen

In den vorangehenden Kapiteln wurden Möglichkeiten diskutiert, den Anteil in der Bevölkerung zu schätzen, der ein heikles Merkmal trägt. Außerdem wurden Verfahren aufgezeigt, mit denen Hypothesen geprüft werden können, in deren Mittelpunkt der Zusammenhang zwischen zwei oder mehreren heiklen Merkmalen steht. Häufig ist aber auch von Interesse, von welchen Faktoren dieses heikle Merkmal, das durch ein Randomized Response Verfahren erhoben wurde, beeinflusst wird. Andererseits interessiert man sich zudem häufig dafür, inwiefern diese heikle Randomized Response Variable Auswirkungen auf ein anderes Merkmal hat. Um diesen Fragestellungen nachzugehen eignen sich Regressionsanalysen. Dabei kann die erhobene Randomized Response Variable entweder die zu erklärende Größe sein oder wie bei der zweiten Fragestellung eine Kovariable. Wenn die Randomized Response Variable eine Einflussgröße ist, dann kann wie bei van den Hout and Kooiman (2006) ein lineares Modell geschätzt werden. Die Forschungen in diesem Bereich sind jedoch noch nicht sehr weit vorangeschritten und es gibt nur wenige Studien, welche den theroretischen Regressionansatz mit der fehlklassifizierten Randomized Response Variable an reellen Daten überprüfen.

Anders jedoch in dem Fall, bei dem ein dichotomer Regressand vorliegt, der angibt, ob eine Person das heikle Merkmal trägt oder nicht. Die zugehörigen Wahrscheinlichkeiten hängen dabei von einer oder mehreren direkt messbaren Größen ab. In diesem Fall wird eine logistische Regression herangezogen, um den Zusammenhang zu modellieren. Das Ziel dabei ist, den Effekt der exogenen Variablen auf die tatsächliche Wahrscheinlichkeit, dass das sensitive Merkmal auftritt, zu messen. Als Erster publizierte Maddala (1983, S.54-56) hierzu ein Vorgehen, welches auf die verschiedenen dichotomen Randomized Response Vorgehensweisen anwendbar ist und im Folgenden näher erläutert wird.

Sei dabei die Notation an die vorangehenden Kapitel angepasst und die interessierende dichotome Variable folglich mit A_i beschrieben. Sie nimmt den Wert 1 an, wenn Person i das heikle Merkmal trägt und hat ansonsten den Wert 0. Die analoge, beobachtbare Randomized Response Variable, die angibt, ob eine Person i mit „Ja“ beziehungsweise „Nein“ geantwortet hat, sei mit X_i notiert. Es wird angenommen, dass die wahre Ausprägung des Merkmals A_i bei den einzelnen Befragten, von der individuellen Merkmalskombination der beobachtbaren Kovariablen Z_i abhängt und die verschiedenen Teilnehmer folglich unterschiedliche Wahrscheinlichkeiten haben,

das heikle Merkmal tatsächlich inne zu haben. Für den wahren Anteil der Merkmalsträger, welcher mit π_A notiert ist, wird der folgende Zusammenhang mit den Kovariablen Z_i angenommen, der dem logistischen Regressionsmodell zugrundeliegt:

$$\pi_A = \frac{e^{\beta' Z_i}}{1 + e^{\beta' Z_i}} \tag{12}$$

Aus den Wahrscheinlichkeiten für eine „Ja"- beziehungsweise eine „Nein"-Antwort lässt sich schließlich die Likelihood-Funktion aufstellen. Diese wird im Vergleich zur Likelihood-Funktion bei Variablen ohne Randomized Response, welche wie bei Maddala (1983, S.55) angegeben,

$$L = \prod_{X_i=1} \frac{e^{\beta' Z_i}}{1 + e^{\beta' Z_i}} \cdot \prod_{X_i=0} \frac{1}{1 + e^{\beta' Z_i}}$$

lautet, durch die bekannten Wahrscheinlichkeiten des Zufallsexperiments korrigiert. Somit wird die Fehlklassifikation beim Maximum-Likelihood-Schätzer von β berücksichtigt. Dieser wird wie gewohnt über die Maximierung der Likelihood-Funktion berechnet. Doch kann der Maximum-Likelihood-Schätzer im vorliegenden Fall nicht analytisch berechnet werden, da die Ableitung der Score-Funktion nicht nach den unbekannten Parametern aufgelöst werden kann. Für die Bestimmung der Nullstellen müssen daher, wie bei Fahrmeir et al. (2009, S.473) beschrieben, numerische Verfahren herangezogen werden. Im vorliegenden Fall wird das Newton-Raphson-Verfahren verwendet. Dieses bestimmt, ausgehend von einer Startlösung, iterativ die Nullstellen der Score-Funktion, indem eine Tangente an den Startwert angelegt wird, die daraufhin eine verbesserte Lösung als Nullstelle liefert. Dieses Vorgehen wird so lange fortgesetzt, bis sich die Lösungen nicht mehr ändern. Für eine detaillierte Beschreibung des Newton-Raphson-Verfahren wird an dieser Stelle auf die Literatur von Fahrmeir et al. (2009, S.473) verwiesen.

In der zugrunde liegenden Literatur dieses Kapitels von Maddala (1983) wird das logistische Regressionsmodell für Randomized Response Variablen anhand des Forced Response Modells (vgl. Kapitel 2.3) erläutert. Die nun folgende Notation weicht jedoch davon ab, da sie an die vorangehenden Kapitel der Arbeit angepasst wurde. Beim Forced Response Modell, bei dem die Befragten entweder die heikle Frage beantworten oder ein vorgegebenes „Ja" beziehungsweise „Nein" geben müssen, lauten die Wahrscheinlichkeiten für eine gegebene „Ja"- beziehungsweise „Nein"-Antwort

$$\mathbb{P}(X_i = 1) = p \cdot \pi_A + (1-p) \cdot \Theta \quad \text{und} \quad \mathbb{P}(X_i = 0) = p \cdot (1 - \pi_A) + (1-p) \cdot (1 - \Theta).$$

Anhand dieser Wahrscheinlichkeiten, bei denen p und Θ bekannt sind, sowie Gleichung (12) ergibt sich schließlich die zu maximierende Likelihood

$$L = \prod_{X_i=1} (p \cdot \frac{e^{\beta' Z_i}}{1 + e^{\beta' Z_i}} + (1-p) \cdot \Theta) \cdot \prod_{X_i=0} (p \cdot \frac{1}{1 + e^{\beta' Z_i}} + (1-p) \cdot (1-\Theta)),$$

aus der anschließend der gesuchte Maximum-Likelihood-Schätzer berechnet werden kann.

Diese Erweiterung der Randomized Response Technik, bei der das heikle Merkmal durch Kovariablen Z_i beschrieben wird, kann auch auf das Modell von Warner, sowie das Unrelated Question Modell übertragen werden. Man vergleiche dazu Sheers and Dayton (1988, S.969-971). Voraussetzung für diese Anwendung ist dabei, dass die Kovariablen eine bekannte Verteilung haben. Da es sich bei diesen Variablen aber in der Regel um keine heiklen Merkmale handelt, können sie direkt abgefragt werden. Sheers and Dayton (1988, S.971-973) zeigten, dass durch diese Art der Modellierung sogar eine Effizienzsteigerung gegenüber dem Standard Randomized Response Vorgehen erzielt werden kann, wenn passende erklärende Variablen vorliegen.

Zusammenfassend lässt sich damit festhalten, dass die Regressionsanalyse bei Randomized Response Variablen ein weiterer bemerkenswerter Fortschritt ist, um aus Randomized Response Daten die bestmöglichen Informationen zu gewinnen. Die hier dargelegte Vorgehensweise bildet nur die Basis für bereits durchgeführte beziehungsweise zukünftige Forschungen. Nicht in Betracht wurden in dieser Arbeit Regressionsanalysen gezogen, bei denen die zu erklärende Größe eine kategoriale beziehungsweise metrische Randomized Response Variable ist. Ebenso wurde das Modell der Randomized Response Einflussgröße nicht im Detail betrachtet. Gezeigt wurde aber in diesem Kapitel, dass Regressionsanalysen mit Randomized Response Daten möglich sind und dieses Vorgehen durchaus sinnvoll ist, da effizientere Schätzer resultieren können und zudem Effekte zwischen Merkmalen aufgedeckt werden können. Da in der bestehenden Literatur bereits zahlreiche Ansätze und Durchführungen von Regressionsanalysen mit fehlklassifizierten Daten existieren, sollten diese Modelle auch auf Randomized Response Daten anwendbar sein, da es sich hierbei um fehlklassifizierte Daten handelt, bei denen die bedingten Fehlklassifikationswahrscheinlichkeiten bekannt sind.

6 Fazit

6.1 Zusammenfassung

Zusammenfassend lässt sich festhalten, dass die hier dargestellten Verfahren nur ein kleiner Ausschnitt aus den bestehenden Randomized Response Methoden sind. Aus dem ganzen Repertoire an Verfahren eine Technik auszuwählen, die für alle denkbaren Fragestellungen den bestmöglichen Schätzer liefert, ist natürlich nicht möglich. Lensvelt-Mulders et. al. (2005, S. 323) schrieben dazu:

> „A thorough look at the literature on RRT reveals that 35 years of research have not led to a consensus or a description of best practices. Many statistical improvements have enhanced the method's efficiency and reliability, and numerous varieties of randomized response procedures have been developed."

Diese Vielzahl an Weiterentwicklungen, Verbesserungen und Verallgemeinerungen hat letztendlich dazu geführt, dass aus verschiedensten Randomized Response Daten aussagekräftige Ergebnisse gewonnen werden können. Der Anwendungsbereich hat sich damit in den letzten Jahrzehnten stark vergrößert, sodass die Randomized Response Technik auch zunehmend Anwendung in der Praxis findet. Durch die Möglichkeit, Zusammenhangsanalysen zwischen verschiedenen Variablen, die durch Randomized Response gemessen wurden, durchführen zu können, wird die Technik für multiple Hypothesen aus verschiedensten Bereichen interessant. Denn obwohl keine Individualdaten vorliegen, kann durch entsprechende Korrektur der Beobachtungen Rückschlüsse auf den Zusammenhang der wahren, unbekannten Werte gezogen werden. Dabei ist es irrelevant, ob die Merkmale dichotom, kategorial oder quantitativ sind. Sogar Regressionsanalysen lassen sich mit Variablen durchführen, die mit Randomized Response erhoben wurden. Damit kann zum einen überprüft werden, inwiefern die Randomized Response Variable von anderen Größen beeinflusst wird und zum anderen, ob sie einen Effekt auf andere interessierende Merkmale hat.

Jedoch darf bei all diesen Fortschritten nicht die geringere Effizienz der Schätzer im Vergleich zur direkten Befragung außer Acht gelassen werden. Es muss daher von Studie zu Studie abgewogen werden, worauf die Priorität gesetzt wird: Ist es für die Umfrage wichtiger den Bias aufgrund von Falschantwortern und Verweigerern zu verringern und die Befragten durch Sicherung ihrer Privatsphäre zur Kooperation zu bewegen oder nimmt man stattdessen eine höhere Verzerrung in Kauf und erhält dafür bei geringerem Aufwand effizientere Schätzer. Diese Entscheidung hängt von

vielen Faktoren wie der Thematik oder den verfügbaren Mitteln ab und lässt sich daher nicht allgemein beantworten. Bei verschiedensten Studien, gerade in den Fällen, bei denen sehr heikle Merkmale im Fokus standen, konnten mit der Randomized Response Technik sehr zufriedenstellende Ergebisse erzielt werden, wohingegen bei anderen das Resultat sehr ernüchternd war. Es wird damit deutlich, dass die Randomized Response Technik durchaus einige gravierende Nachteile mit sich bringt. So besteht das Problem der Verweigerer, also derjenigen Personen, die sich nicht an die Instruktionen halten, was letztendlich eine Verzerrung der Schätzer zur Folge hat. Außerdem wurde die Güte der Technik meist nur bei face-to-face Umfragen, sowie bei hochsensiblen Themen getestet. Eine Durchführung des Zufallsexperiments bei online oder Telefoninterviews ist dagegen sehr schwierig. Denoch bleibt die Randomized Response Technik eine der am weitesten verbreiteten und erforschten Alternativen zur direkten Befragung.

6.2 Alternative und Ausblick

Neben der Randomized Response Technik gibt es natürlich auch andere sogenannte Dejeopardizing Techniken, deren Ziel darin besteht, den Anteil der Träger eines kritischen Merkmals zu schätzen und dabei die Anonymität der Befragten zu bewahren. Ein weitverbreitetes, alternatives Vorgehen zur Randomized Response Technik ist dabei die Unmatched Count Technik, die auch häufig als Item Count Technik bezeichnet wird. Detailliert beschrieben wird diese Technik unter anderem bei Coutts and Jann (2011). Ausgangspunkt des Verfahrens ist, dass die Stichprobe in zwei Gruppen unterteilt wird. Während die Kontrollgruppe eine Liste von harmlosen Fragen beantworten muss, wird in der anderen Gruppe zusätzlich die heikle Frage gestellt. Die befragten Personen nennen letztendlich keine direkten Antworten, sondern geben an, wie viele Fragen aus der Liste sie mit „Ja“ beantwortet haben. Unter der Annahme, dass die zweite Gruppe ohne die Antwort auf die zusätzliche, kritische Frage dieselbe Anzahl an „Ja“-Antworten gegeben hätte wie die Kontrollgruppe, lässt sich der Anteil der Leute, welche die sensitive Frage bejaht haben, aufgrund der ungleichen Anzahl an Fragen in den beiden Gruppen schätzen. Wie bei der Randomized Response Technik wird also bei der Unmatched Count Technik die Anonymität der Teilnehmer gewahrt. Zudem hat diese Methode im Vergleich zur Randomized Response Technik den Vorteil, dass kein Zufallsexperiment durchgeführt werden muss. Das erhöht zum einen das Vertrauen der Teilnehmer in die Technik und zum anderen wird der Aufwand deutlich verringert. Einige Studien zeigten, dass die Unmatched Count Technik im Vergleich zur Randomized Respon-

se Technik trotz der einfacheren Durchführung bessere Schätzer hervorbringt (vgl. Coutts and Jann, 2011, S.183-186).

Diese Erkenntnisse machen deutlich, dass in diesem Bereich weitere Forschungsarbeit äußerst sinnvoll wäre. Tests zur optimalen Länge der Fragenlisten sowie der optimalen Wahl der Fragen sollten durchgeführt werden, um die Anwendung zu verbessern. Außerdem sollte der Frage nachgegangen werden, inwiefern sich die Unmatched Count Technik auf nicht-dichotome Fragestellungen ausweiten lässt und inwiefern Zusammenhangsanalysen möglich sind. Nur so ist ein detaillierter Vergleich mit der Randomized Response Technik möglich und es kann geprüft werden, ob die Randomized Response Technik gegenüber existierenden Alternativen bestehen kann. Um das sehr theoretische Randomized Response Verfahren auch in der Praxis noch stärker zu etablieren und konkurrenzfähig gegenüber anderen Erhebungsverfahren zu machen, sind zukünftige Forschungen in diesem Bereich notwendig. Hilfreich und sinnvoll wäre dafür die Implementierung von Softwares, mit deren Hilfe Randomized Response Daten auf verschiedene Weisen korrigiert werden können. Zum jetzigen Zeitpunkt beschränken sich die bestehenden Softwares auf wenige Packages in R, wie zum Beispiel simex, welches Funktionen zur Messfehler-Korrektur von Schätzern bereitstellt. Für Stata entwickelte Ben Jann eine Software, um logistische Regressionen für Randomized Response Daten durchführen zu können. Die Erweiterung von solchen praktischen Anwendungs-Möglichkeiten ist damit ein sinnvolles Ziel für zukünftige Forschungen, um die Vorteile der Randomized Response Technik weiter auszuschöpfen.

7 Literatur

Literatur

Abul-Ela, A. A., B. G. Greenberg, and D. G. Horvitz (1967). A Multi-Proportions Randomized Response Model. *Journal of the American Statistical Association 62*, 990–1008.

Campbell, C. and B. L. Joiner (1973). How to Get the Answer Without Being Sure You've Asked the Question. *Journal of the American Statistical Association 27*, 229–231.

Chaudhuri, A. and R. Mukerjee (1988). *Randomized Response: Theory and Techniques*. Dekker.

Clark, S. J. and R. A. Desharnais (1998). Honest Answers to Embarrassing Questions: Detecting Cheating in the Randomized Response Model. *Psychological Methods 3*, 160–168.

Coutts, E. and B. Jann (2011). Sensitive Questions in Online Surveys: Experimental Results for the Randomized Response Technique (RRT) and the Unmatched Count Technique (UCT). *Sociological Methods and Research 40*, 189–193.

Drane, W. (1976). N the theory of randomized responses to two sensitive questions. *Communications in Statistics: Theory and Methods 5*, 565–574.

Edgell, S. E., S. Himmelfarb, and K. Duchan (1982). Validity of Forced Response in a Randomized Response Model. *Sociological Methods and Research 11*, 89–100.

Eichhorn, B. H. and L. S. Hayre (1983). Scrambled randomized response methods for obtaining sensitive quantitative data. *Journal of Statistical Planning and Inference 7*, 307–316.

Fahrmeir, L., T. Kneib, and S. Lang (2009). *Regression: Modelle, Methoden und Anwendungen* (2 ed.). Springer.

Fahrmeir, L., I. Künstler, I. Pigeot, and G. Tutz (2010). *Statistik: Der Weg zur Datenanalyse* (7 ed.). Springer.

Fox, J. A. and P. E. Tracy (1984). Measuring Associations with Randomized Response. *Social Science Research 13*, 188–197.

Greenberg, B. G., A. A. Abul-Ela, W. R. Simmons, and D. G. Horvitz (1969).

The Unrelated Question Randomized Response Model: Theoretical Framework. *Journal of the American Statistical Association 64*, 520–539.

Greenberg, B. G., R. R. Kuebler, J. R. Abernathy, and D. G. Horvitz (1971). Application of the Randomized Response Technique in Obtaining Quantitative Data. *Journal of the American Statistical Association 66*, 243–250.

Himmelfarb, S. and S. E. Edgell (1980). Additive Constants Model: A Randomized Response Technique for Eliminating Evasiveness to Quantitative Response Questions. *Psychological Bulletin 87*, 525–530.

Himmelfarb, S. and S. E. Edgell (1982). Note On: "The Randomized Response Approach": Addendum to Fox and Tracy. *Evaluation Review 6*, 279–284.

Lensvelt-Mulders, G., J. Hox, and P. van der Heijden (2005). How to Improve the Effciency of Randomized Response Designs. *Quality and Quantity: International Journal of Methodology 39*, 253–265.

Lensvelt-Mulders, G., J. Hox, P. van der Heijden, and C. J. Maas (2005). Meta-Analysis of Randomized Response Research: Thirty-Five Years of Validation. *Sociological Methods and Research 33*, 319–348.

Maddala, G. (1983). *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge University Press.

Mangat, N. S. and R. Singh (1990). An alternative randomized response procedure. *Biometrika 77*, 439–442.

Ostapczuk, M., M. Moshagen, Z. Zhao, and J. Musch (2009). Assessing Sensitive Attributes Using the Randomized Response Technique: Evidence for the Importance of Response Symmetry. *Journal of Educational and Behavioral Statistics 34*, 267–287.

Sheers, J. and C. Dayton (1988). Covariate randomized response models. *Journal of the American Statistical Association 83*, 969–974.

Sudman, S. and N. M. Bradburn (1982). *Asking Questions*. Jossey-Bass.

van den Hout, A. and P. Kooiman (2006). Estimating the linear regression model with categorial covariates subject to randomized response. *Computational Statistics and Data Analysis 50*, 3311–3323.

van den Hout, A. and P. van der Heijden (2002). Randomized Response, Statisti-

cal Disclosure Control and Misclassification: a Review. *International Statistical Review 70*, 269–288.

Warner, S. L. (1965). Randomized Response: A Survey Technique for Eliminating Evasive Answer Bias. *Journal of the American Statistical Association 64*, 520–539.

Warner, S. L. (1971). The Linear Randomized Response Model. *Journal of the American Statistical Association 66*, 884–888.

Erklärung

Hiermit versichere ich, dass ich meine Abschlussarbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel benutzt habe.

Datum: .....

(Unterschrift)