

FEHLENDE WERTE

Seminararbeit zum Zusatzkurs Amtliche Statistik; Prof. Dr. Augustin

Matthias Speidel

Einleitung

„Die Statistik ist die Wissenschaft der verantwortungsvollen Datenanalyse“.

Datenanalyse setzt das Vorhandensein von Daten voraus. Häufig beinhaltet ein Datensatz jedoch Fehlende Werte. Auch das Fehlen der Daten muss verantwortungsvoll berücksichtigt werden.

Diese Arbeit schaut auf die Entstehung und, im Folgeschluss, auf die Vermeidung von Fehlenden Werten. Zur weiterführenden Betrachtung werden drei Klassen von Fehlenden Werten vorgestellt. Als Konzept zur Bestimmung der Klassen wird die FehlendWerte-Differenzmatrix vorgestellt. Je nach Klasse muss die Handhabung mit den Fehlenden Werten differenziert erfolgen. Welche Möglichkeiten es dazu gibt, und deren Grenzen bezüglich der Informationsverwertung zeigt der letzte Teil der Arbeit auf.

Welche Auswirkungen haben fehlerhafte Ergebnisse

Ein Blick in die Presse zeigt immer wieder, wie Statistiken als Entscheidungsgrundlage oder Bestätigung für weitreichende Handlungen dienen. Wenn allerdings sich die Statistiken als unkorrekt erweisen, sollte das nachdenklich machen, denn was ist mit den fehlerhaften Statistiken, deren Fehlerhaftigkeit nicht erkannt wird? Als Beispiel sei hier die vermeintliche ansteigende Geburtenrate, befördert durch die Politik der Familienministerin genannt. Wenige Zeit nach der Bekanntgabe der Daten wurden korrigierte Ergebnisse veröffentlicht, die weiterhin eine sinkende Geburtenrate beziffern.¹ Die besondere Bedeutung von der Richtigkeit einer Statistik verdeutlicht die Tatsache, dass in Unternehmen werden Entscheidungen über Investitionen, Standortverlagerungen und Entlassungen auf Grundlage sowohl von unternehmensinternen Statistiken als auch solche des Statistischen Bundesamtes bzw. privater Wirtschafts-Institute getroffen oder rechtfertigt.

Wo entstehen Fehler

Wie eingehend beschrieben haben Fehler in der Statistik weitreichende –fast immer negative- Auswirkungen. Warum sind die Auswirkungen negativ? Statistiken werden erhoben um die unendlich komplexe Realität für den menschlichen Verstand fassbar zu machen und daraus - bei unterstelltem Wohlwollen – Entscheidungen zum Wohle der Betroffenen abzuleiten. Eine Differenz zur Realität bedeutet somit eine Differenz zur optimalen Lösung. Damit eine Statistik nicht ausschließlich zum Wohle eines einzelnen Entscheidungsträgers oder Interessengruppe gereicht, haben sich Grundsätze basierend auf Neutralität, Objektivität und Wissenschaftlicher Unabhängigkeit etabliert.²

¹ So DIE WELT: http://www.welt.de/die-welt/article1456180/Experte_sieht_Elterngeld_als_Ursache_fuer_steigende_Geburtenrate.html
zurückhaltender DIE ZEIT: <http://www.zeit.de/online/2008/34/geburten-anstieg-analyse>

² u. A. zu finden auf der Seite des Statistischen Landesamtes Berlin: <http://www.statistik-berlin.de/wir/amtliche-statistik/wir1.htm>

Im Ablauf einer statistischen Messung lassen sich zwei grundlegende Arten von Fehlern feststellen: Fehler in der Analyse (Fehlertypen 1; 2; 3; 8) und Fehler in den Daten (Fehlertypen 4; 5; 6; 7).

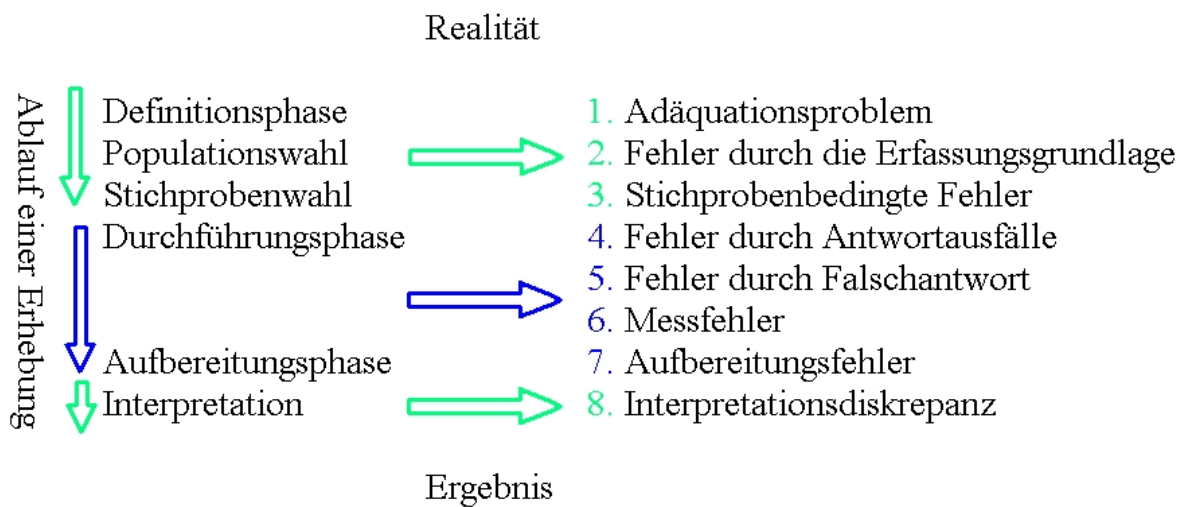


Abbildung 1: Die Phasen einer Erhebung und Ihr Fehlerpotential (frei nach Radermacher)

1. Das Adäquationsproblem: Ein abstraktes Konstrukt, beispielsweise „Armut“, muss auf eine reale Ebene der Empirie gebracht werden. Unweigerlich stellen sich hier die Fragen: „Was ist Armut?“, „Lässt sich Armut in Euro messen?“, „Ist Armut absolut oder relativ?“ etc. Des Weiteren werden valide (wird das Richtige gemessen?) Messinstrumente, die reliable (sind die Messergebnisse zuverlässig?) Ergebnisse liefern, benötigt.³

2. Fehler durch die Erfassungsgrundlage: Wen oder was will ich untersuchen? Oder anders formuliert „Auf wen sollen meine Erkenntnisse übertragen werden können?“. Sollen Erkenntnisse zur regionalen Bevölkerungsstruktur in Land A gewonnen werden, so ist eine Befragung der Bevölkerung des Landes Y unzulänglich. Des Weiteren stellt sich auch hier das Adäquationsproblem. Bei der Untersuchung nach der Armut könnte die Altersarmut interessieren, stellt sich nun die Frage, ab welchem Alter an einer Person „Altersarmut“ gemessen werden kann. Werden Einheiten durch die falsche Bestimmung der Erfassungsgrundlage von der Möglichkeit, gezogen zu werden, ausgeschlossen, so bildet diese Menge die so genannten „Undercoverage“.⁴

3. Stichprobenbedingte Fehler: Diese Fehler treten auf, wenn die Wahl der Stichprobe die interessierende Grundgesamtheit (bezüglich des interessierenden Merkmals) nicht repräsentativ ist. Wobei von „Repräsentativität“ bezüglich eines Merkmals gesprochen wird, wenn die Stichprobe sich diesbezüglich nicht von der Grundgesamtheit unterscheidet. In Haßloch wird in dem Test-Supermarkt von einer „konsum-repräsentativen“ Kundschaft ausgegangen, und so können neue Produkte vor bundesweiter Einführung auf ihren Erfolg bei den Haßlochern getestet werden. Ob nun allerdings das Konsumverhalten für das neue Produkt XY repräsentativ ist, lässt sich erst im Nachhinein bestimmen!⁵

³ Groves (1989) *Survey Errors and Survey Costs* spricht von der interessanten Maßzahl Accuracy := 1 / Variance. Sie wird stellenweise gleichgesetzt mit Validität und Reliabilität. Es sei aber angemerkt, dass ein Messinstrument reliabel sein kann, ohne valide zu sein.

⁴ Analog dazu die „Overcoverage“ durch Personen die fälschlicherweise in die Erfassungsgrundlage gelangen.

⁵ Einen gut zu lesenden Einblick in die Welt der Konsumforschung bietet NZZ Folio 11/06 der Neuen Züricher Zeitung:

Bestimmte Quoten müssen dazu nicht zwangsläufig erreicht werden. Es ist z.B. nicht unter allen Umständen eine Geschlechteraufteilung von 1:1 nötig, um eine repräsentative Stichprobe bezüglich eines anderen Merkmals X zu erreichen. Nimmt man hingegen an, dass Geschlecht und Merkmal X in keinem Zusammenhang stehen, so sollte das Geschlecht einer Binomialverteilung mit $P_i = 0,5$ folgen.

Man beachte aber: es soll der Zusammenhang zwischen dem Geschlecht und dem Merkmal X in der Studie erstmalig untersucht werden, denn wäre er bekannt, so wäre die Studie ohne Nutzen / Informationsgewinn.

4. Fehler durch Antwortausfälle: Für eine Einheit stehen keine Werte zu einer oder mehreren Merkmalsausprägungen zu Verfügung. Diese Art von Fehlern bildet den Schwerpunkt dieser Arbeit. Es soll untersucht werden, warum es Antwortausfälle gibt, wie sie sich vermeiden lassen, und wie mit ihnen umgegangen werden kann.

5. Fehler durch Falschantwort: Eine Einheit hat bewusst oder unbewusst Angaben gemacht, die nicht der Realität entsprechen. Hierbei gibt es nachprüfbare Falschantworten (Gewicht, Größe, Notendurchschnitt etc.) und unprüfbare Falschantworten (Einstellung zum Umweltschutz etc.).

6. Messfehler: Fehler die bei der Messung entstehen können, sei es durch das Instrument, den Befrager, den Befragten oder „zufällig“.

7. Aufbereitungsfehler: Beispielsweise ist die falsche Eingabe eines Wertes vom Fragebogen in den Computer ein Aufbereitungsfehler. Dies kann unbewusst (Vertippen) oder bewusst (bei einer vermeintlichen Korrektur von Werten) geschehen.

8. Interpretationsdiskrepanz: Von den statistischen Ergebnissen wird versucht eine Brücke in die Realität zu schlagen. Durch Unkenntnis, unterschiedliche Auffassungen/Sichtweisen oder bewusste Beachtung/Missachtung von Ergebnissen kann sich bei der Interpretation ein falscher Eindruck von der gemessenen Realität einstellen.

Wie kann man sie vermeiden?

Für die Adäquation ist es von elementarer Bedeutung ein genaues Bild von dem zu untersuchenden Sachverhalt zu haben. Die enge Zusammenarbeit mit einer Person vom Fach ist unabdingbar um Fehler bei der Auffassung des Problems zu vermeiden.

Für das Messinstrument ist die Validität nur schwer sicherzustellen und theoretisch unmöglich zu überprüfen. Allerdings ist eine nicht-Validität folgendermaßen zu überprüfen: Besteht zwischen zwei Sachverhalten in der Theorie ein Zusammenhang, so sollte sich dieser mit dem Messinstrument auch empirisch Nachweisen lassen. Ist dies nicht der Fall, so kann das Instrument nicht valide sein. Eine weitere Möglichkeit bietet der Vergleich von einem sehr validen (aber womöglich auch sehr teurem) Messinstrument und einem vielleicht weniger validen (aber dafür günstigerem) Messinstrument.

Vor eine groß angelegten Untersuchung sollten pre-Tests eine „Bewährungsprobe“ für das Instrument darstellen. Er dient im übertragenen Sinne dazu, die Stellschrauben des Instrumentes zu justieren, und auch zum Beispiel missverständlich gestellte Fragen vorzeitig aufzudecken.

Einige Fehler kann man auch bei der Wahl der Stichprobe machen (falsche Quote in einer Quotenstichprobe, zu kleiner Stichprobenumfang, kein echter Zufallsmechanismus in einer Zufallsstichprobe etc.).

Bei den Stichprobenbedingten Fehler besteht grundsätzlich das Problem, dass man zur Fehlervermeidung vollständige Informationen über die Grundgesamtheit mit dem Interessierenden Merkmal braucht. Aber wäre das gegeben bräuchte man keine Erhebung machen.

Des Weiteren darf das Auswahlverfahren nicht mit dem Untersuchungsmerkmal in Verbindung stehen. Das Beispiel einer telefonischen Befragung ob ein Telefonanschluss vorhanden ist, sollte dies ausreichend verdeutlichen. Wo und wie gefragt wird, schränkt somit unweigerlich die Verallgemeinerung der Ergebnisse ein. Deshalb ist es von größter Bedeutung, dass die Befragung allgemein gehalten wird und wenn eine Zufallsstichprobe ohne systematische Verzerrung benötigt wird, jede Einheit der Erfassungsgrundlage dieselbe Wahrscheinlichkeit hat in die Stichprobe zu gelangen.⁶

Die Gründe für Antwortausfälle sind meist von psychologischer (oder unternehmensstrategischer) Natur, denn das Erheben von Daten bedeutet immer ein Informationsgewinn für den Nutzer der erhobenen Daten. Wenn nun der Informationsgewinn durch die Studie zum Nachteil eines Studienteilnehmers werden kann, sind Antwortverweigerungen quasi vorprogrammiert (z.B. bei der Frage nach der Investitionsstrategie eines Unternehmens). Das Zusichern (und Einhalten!) von Anonymität, sodass keine Rückschlüsse auf den Befragten möglich sind, wird die Auskunftsbereitschaft erhöhen. Bundesstatistiken greifen auch deshalb auf das Prinzip der Auskunftspflicht auf Seite der Befragten bei gleichzeitiger Geheimhaltungspflicht auf Seiten des Statistischen Bundesamtes zurück. Des Weiteren ist die Befürchtung von rechtlichen Konsequenzen durch die Angaben, Anlass sich seiner Antwort zu verweigern. Auf Bundesstatistischer Ebene wird dem durch die „Informelle Einbahnstraße“ - einem Verbot der Rückführung von Informationen, beispielsweise an Steuerämter - begegnet.

Eigentliche Antwortverweigerer greifen allerdings auch auf die Möglichkeit der Falschantwort zurück. Gründe dafür können das Bestreben keine Aufmerksamkeit zu erregen oder die „Soziale Erwünschtheit“ sein. Wo schon ein Nichtantwort einem sozial unerwünschtem „Nein“ gleichkommt, ist das „erzwungene“ „Ja“ nicht weit.⁷ Als Lösung bieten sich schriftliche, anonyme Fragebögen oder indirekte Fragen an.

Ein Beispiel der raffinierte Art, Anonymität unter Ausnutzung des Zufalls zu erlangen⁸, zeigt C. R. Rao (1995) in seinem Buch *Was ist Zufall? Statistik und Wahrheit*: von Interesse ist der Anteil in der Bevölkerung welcher Marihuana konsumiert. Da der Konsum (bzw. der Besitz) nicht nur sozial unerwünscht, sondern auch noch strafbar ist, ist dies eine Frage, die viel Fehlende Werte oder Falsche Werte verursachen wird. Doch nun zum Vorgehen: die Befragten sollen geheim eine Münze werfen, und bei Zahl die Frage beantworten, ob sie Drogen nehmen, bzw. bei Wappen ob Ihre Telefonnummer auf eine gerade Ziffer endet. Der geschätzte Anteil ist nun: $2 \cdot (\text{Anzahl der ‚Ja‘ Antworten} - n/4)$

Die Anzahl der Fragen und die Dauer der Befragung spielen ebenfalls eine Rolle für das Antwortverhalten (Aufmerksamkeit und Motivation lassen nach). Abhängig von der Befragung haben sich Kennzahlen etabliert: Bei einer mündliche Befragung maximal 45 Minuten oder 60 bis 80 Fragen. Ist die Befragung schriftlich 10 bis 20 Minuten oder 40 bis 60 Fragen. Nach Groves empfehlen sich am Telefon zwei kurze (unter Umständen nicht ein mal relevante) Fragen welche die Bereitschaft des Teilnehmer steigern sollen.

Messfehler entstehen wenn das Messinstrument nicht geeignet ist das Merkmal zu bestimmen (Siehe Validität und Reliabilität). Wird wiederholt der selbe Merkmalsträger gemessen, sollte das Messinstrument die selben Ergebnisse mit einer möglichst kleinen Varianz (was eine

⁶ Ein Klassiker, der häufig zur Beschreibung von falscher Korrelation verwendet wird, ist die im Supermarkt gezogene Stichprobe Hühnereier. Er kann hier auch zur Demonstration von nicht-Repräsentativität herangezogen werden; denn, bezüglich natürlicher Länge und Breite, repräsentative Eier sucht man hier vergebens.

⁷ Die Frage „Mögen Sie Kinder?“ ist das Paradebeispiel schlecht hin.

⁸ Grundlegend nach Warner, S. L. (1965). Siehe dazu: <http://www.jstor.org/pss/2283137>

hohen Accuracy entspricht) liefern.⁹ Nicht nur das Messinstrument kann Messfehler erzeugen, auch bei falscher Handhabung durch den Anwender des Instruments entstehen Messfehler. Besonderheiten in denen das Messinstrument nicht verlässliche Ergebnisse liefert, sollten berücksichtigt und analysiert werden.

Um Fehler während der Datenaufbereitung zu reduzieren, empfiehlt es sich, für diese Aufgabe gut geschultes Personal einzusetzen.

Ist die Erhebung abgeschlossen, so liegt die Analyse der Daten meist alleinig in der Hand des Statistikers, weswegen dieser hier größte Sorgfalt walten lassen sollte. Die Wahl der richtigen Modelle und die Überprüfung der Annahmen wie z.B. die Normalverteilungsannahme sollten selbstverständlich sein.

In der Statistik wird häufig von einer Variablen ausgegangen, deren Individuen (Elemente) unabhängig voneinander derselben Verteilung folgen. Wo diese Annahme nicht offensichtlich ist, sollte sie mit einem Kollegen des entsprechenden Faches erörtert werden.¹⁰

Der letzte Schritt einer Erhebung ist der Rückschluss auf die interessierende Population (die Erfassungsgrundlage bzw. Grundgesamtheit). Dass hierbei Fehler gemacht werden, ist nicht auszuschließen, aber der Verdienst der Statistik ist es, dass dieser Fehler quantifiziert werden kann. Man denke dazu beispielsweise an den *Fehler 1. Art* (α -Fehler: die Nullhypothese wird fälschlicherweise verworfen) und den *Fehler 2. Art* (β -Fehler: die Nullhypothese wird fälschlicherweise beibehalten) in der Test-Theorie.

Bei der Interpretation der Ergebnisse ist oftmals Vorsicht geboten. Aufgabe eines Statistikers ist es zu erklären was eine statistische Größe (z.B. Odds-Ratio, Standardabweichung etc) aussagt oder nicht. Die inhaltliche Interpretation warum eine Zahl diesen oder jenen Wert erzielte ist nun Aufgabe eines Substanzwissenschaftlers. Plakativ gesprochen: der Statistiker ist für die Aussage eines Wertes, der Substanzwissenschaftler für die Entstehung des Wertes zuständig.

Sei nach dem Risiko gefragt an Lungenkrebs zu erkranken im Vergleich von Rauchern zu Nichtrauchern. Der Statistiker erklärt dass eine Odds-Ratio von 1,8 bedeutet, dass Raucher im Vergleich zu Nichtrauchern 1,8 mal häufiger an Lungenkrebs erkranken. Der Mediziner nun gibt als mögliche Gründe eine Mutation in den Zellen etc an. Ein Statistiker soll zwar zum einen nicht selbst interpretieren, zum anderen ist er aber auch angehalten, darauf zu achten, dass die Ergebnisse nicht falsch interpretiert werden. Der Fehler aus Korrelationen Kausalschlüsse zu ziehen ist dabei besonders „beliebt“.

Wie auch bei der Vorbereitung der Statistik ist die Kommunikation des Vorgehens und der Ergebnisse wichtig für die richtige Erörterung der Ergebnisse. Ergeben sich weitere Fragen, so kann aus den Erfahrungen und Resultaten der letzten Untersuchung eine neue Untersuchung entstehen. Und der Prozess beginnt von Neuem.

Welche Arten von Fehlenden Werten gibt es?

Rubin (1976) führte drei Begriffe zu Fehlenden Werten ein. Da Name und Fehlendmechanismus in meinen Augen beim „Missing At Random“ doch weit auseinander liegen, möchte ich im Weiteren zusätzlich eigene Begrifflichkeiten verwenden.

Man spricht von einem „Missing Completely At Random“ Fehler (Zufällig Fehlend), wenn das Fehlen des Wertes vollkommen zufällig ist, und mit keiner der abgefragten Variablen in Verbindung steht. Als Beispiel sei eine Bundesweite Studie genannt; die Antworten sollen elektronisch übermittelt werden. Geht man nun davon aus, dass 5% der Antworten nicht zugestellt werden können, so ist das als zufälliger Fehler aufzufassen. Ist der Ausfall

⁹ Die Küchenwaage, die bei jedem Gewicht konstant „500 g“ anzeigt, liefert zwar Ergebnisse mit Varianz = 0 allerdings sind die Ergebnisse nicht valide.

¹⁰ Hat Baum A einen Einfluss auf Baum B, wenn diese 12 Meter auseinander stehen? Was ist bei 6 und 3 Metern? Die Fragen kann qualifiziert nur ein Botaniker beantworten.

allerdings auf die Region Augsburg begrenzt, da dort die Kommunikationsnetze ausgefallen sind, so ist es nun problematisch in der Auswertung Angaben zur Region mit ein zu beziehen. Deutlich vorsichtiger muss vorgegangen werden wenn die Beantwortung einer Frage B mit einer Frage A (die nicht zwangsläufig auch konkret gestellt sein muss) in Zusammenhang steht. Die Frage nach dem Alkoholkonsum wird unter Umständen deutlich seltener beantwortet, wenn vorangehend die Frage nach der beruflichen Position gestellt wurde, unabhängig davon welchen Wert der Alkoholkonsum erreicht. Dieses Fehlen wird „Missing At Random“ genannt (Abhängig Fehlend).

Der problematischste Fall sind Fehlende Werte, bei denen das Fehlen allein von der Höhe der Ausprägung abhängt, denn dieses Fehlen weist eine Struktur auf, die nicht rekonstruiert werden kann (häufig „sensible“ Fragen wie zum Einkommen, oder oben genannter Alkoholkonsum). Dies wird als „Missing Not At Random“ bezeichnet (Nicht Zufällig Fehlend).

Wie kann man die Arten bestimmen?

Ein Vorgehen zum Erkennen von Abhängig Fehlenden Werten ist, wie Göthlich (2005) es in einem anderen Zusammenhang anführt, jede Variable in vollständige und unvollständige Teildatensätze zu trennen. Man begibt sich quasi auf die Suche nach der Variablen, die mit dem Fehlen in Verbindung steht. Dazu werden nun die Unterschiede zwischen dem vollständigen und unvollständigem Teildatensatz betrachtet. Diese werden deutlich, wenn die Verteilungen zwischen den Teildatensätzen verglichen werden. Als Element einer FehlendeWerte-Differenz-Matrix bilden die (signifikanten) Differenzen nun einen Indikator dafür, welche Variable A mit dem Fehlen der Werte in Variable X in Verbindung steht.

Ist das Fehlen dadurch nicht erklärbar, so kann es Zufällig Fehlend oder aber auch Nicht Zufällig Fehlend sein. Denn Unabhängigkeit zwischen den Fehlenden Werten einer Variable und allen anderen Variablen sagt lediglich aus, dass das Fehlen mit keiner ihnen in Verbindung steht, komplett zufällig muss es deswegen noch lange nicht sein (Die Unabhängigkeit ist notwendig aber nicht hinreichend). Wer nun versucht durch viele abgefragte Variablen einen Grund für das Fehlen zu entdecken der erhöht aber auch das Risiko zufällig zu signifikanten Ergebnissen zu kommen.¹¹

Da Typen Zufällig Fehlend und Nicht zufällig Fehlend „Gegensätze“ darstellen, ist ihre Unterscheidung essentiell. Es ist theoretisch nicht zu überprüfen, aber ein möglicher Ansatz unter der Annahme, dass Zufällig Fehlende Antwortausfälle keine besondere Struktur bezüglich der Fehlenden Werte aufweisen, wäre auf Gleichverteilung zu testen, sodass für jede Variable ungefähr das selbe Verhältnis von Fehlenden Werten zu allen Werten gilt. Weist eine Variable deutlich mehr Fehlende Werte auf, so wird das Fehlen nicht Zufällig Fehlend sein.

Eine Variable A ist somit

-vermutlich „Zufällig Fehlend“ wenn sie keine auffällig hohe Fehlende Werte Rate besitzt und mit keiner anderen Variablen im Zusammenhang steht.

-„Abhängig Fehlend“ wenn das Fehlen der Werte mit einer anderen Variablen im Zusammenhang steht.

-„Nicht Zufällig Fehlend“ wenn sie eine auffallend hohe Fehlende Werte Rate besitzt und mit keiner anderen Variablen im Zusammenhang steht

Probleme bei der Einteilung können allerdings entstehen, wenn

¹¹ Im Zusammenhang von multiplen Testproblemen wird nun zum Niveau α/k getestet (Bonferroni-Korrektur)

- das Fehlen eigentlich Zufällig Fehlend ist, aber beispielsweise die Fragestellung eine hohe Fehlende Werte Quote verursacht.
- Nicht Zufällig Fehlende Werte „Normalität“ suggerieren und dadurch als Zufällig Fehlend eingeschätzt werden
- Mehrere der Fehlend-Mechanismen Ursache für das Fehlen sind

Welche Möglichkeit gibt es mit fehlenden Werten umzugehen?

Statistikprogramme wie SPSS beziehen Fehlende Werte in die Analyse nur insoweit mit ein, dass eine Fehlerkennzahl (Anzahl der Vollständigen Daten / Anzahl der Beobachtungen oder Anzahl der Fehlenden Werte) angegeben wird. Eine „Complete Case Analysis“ betrachtet dabei nur Datensätze in denen zu keiner Variablen die Daten fehlen. Bei steigender Anzahl der Variablen steigt allerdings auch die Wahrscheinlichkeit, dass es Fehlende Werte gibt, weshalb bei der Completely Case Analysis immer auf die Fallzahl geachtet werden sollte, gerade bei einer großen Anzahl an Variablen. Die „Available Case Analysis“ betrachtet die Einheiten bei denen vollständige Informationen zu den aktuell interessierenden Variablen vorliegen. Werden Geschlecht, Alter und Gewicht in Beziehung gesetzt, so ist in der Available Case Analysis das Fehlen von Werten zur Größe irrelevant. So wird zwar der Informationsverlust gegen über der Complete Case Analysis reduziert, allerdings erhält man viele verschiedene Teilstichproben (auch mit unterschiedlichen Stichprobenumfängen), was einen Vergleich von Ergebnissen erschwert.

Zu beachten: diese Vorgehen implizieren indirekt, dass in Fehlenden Werte keine Information enthalten ist und aus diesem Grund nicht in die Analyse eingehen. Für Zufällig Fehlende Werte trifft diese Annahme zu, denn man könnte das Fehlen als systemfreie Kürzung des Datensatzes auffassen, was lediglich eine Reduktion des Stichprobenumfangs (und somit größere KI und PI).

Größte Vorsicht ist allerdings bei Abhängig oder Nicht Zufällig Fehlenden Werte geboten! Denn hier ist Information im Fehlen an sich enthalten. Wie die Differenz-matrix gezeigt hat, ist das Fehlen bei Abhängig Fehlenden Werten von einer Variablen der beobachteten Werte (oder noch weiteren, bisher unbekannt, Variablen) und bei den Nicht Zufällig Fehlenden Werten von einer unbekannt Variable oder der Ausprägung der beobachteten Variable abhängig. Gibt es nun einen Zusammenhang zwischen dem Fehlen von Werten in der Variablen X und der Variablen Y, so ist es äußerst problematisch die Fehlenden Werte zu vernachlässigen, wenn man in der Analyse X mit Y in Verbindung stellt. Mit Nicht Zufällig Fehlenden Werte steht man vor einem Problem, das mit dem Ignorieren der Fehlenden Werte nicht zufrieden stellend gelöst ist.

Da allerdings multivariate statistische Verfahren technisch nicht mit Fehlenden Werten arbeiten können, müssen die Lücken im Datensatz geschlossen werden.

Einen Ansatz stellen die Imputationsverfahren (von it. „imputare“ dt. „unterstellen“, „zuschreiben“) dar. Der Grundgedanke dahinter ist, dass Fehlende Werte durch die „Echten“ Werte ersetzt werden. Da diese Werte nicht (oder nur schwer durch wieder zuordnen, und erneuter Befragung¹²) zu ermitteln sind, begibt man sich auf die Suche nach Werten, die möglichst nahe an der Realität sind. Mittelwertschätzungen und Regressionen sind Möglichkeiten, doch sei hier darauf hingewiesen, dass 1,4 Kinder in der Realität nicht auftreten, auch wenn sie das korrekte Ergebnis bei einer Mittelwertimputation sind. Für Kategoriale Merkmale bietet sich nun ein Matching an. Anhand von adäquaten Variablen wird nun eine passende Einheit ermittelt, und deren Ausprägung in der interessierenden Variablen imputiert. Ein Problem der Imputation ist die Unterschätzung der Unsicherheit, so hat sich die Multiple Imputation manifestiert. Ein Fehlender Wert wird hierbei $m > 1$ mal

¹² Das allerdings würde auch die Gefahr der Falschantwort erhöhen.

imputiert und die verschiedenen Ergebnisse analysiert. So werden auch unterschiedliche Verteilungsannahmen in den verschiedenen Variablen berücksichtigt.¹³

Die adäquate Variablen lässt sich anhand einer hohen Korrelation zur Zielvariable finden (unabhängig davon Beachtung des Fehlendmechanismus!); stetige Variablen werden dabei kategorisiert (denn $P(x)=0$), wobei die Einteilung der Klassen nicht zu groß oder zu klein ausfallen sollte um eine gewisse Anzahl an Fällen in den Klassen sicher zu stellen. So kommt man beispielsweise von den nicht realen 1,4 Kindern zu der hoffentlich realistischeren Annahme von 1 oder 2 Kindern für eine Frau.

Welche Auswirkungen, welchen Sinn haben die verschiedenen Methoden?

Was die Imputationsverfahren *nicht* können, ist neue Daten bzw. Informationen liefern. Wer diesen Punkt übersieht, läuft auch Gefahr sein Modell nach der Imputation zu überschätzen. Die Regressionsimputation liefert „neue Werte“ –basierend auf den tatsächlich beobachteten Werten- die nun perfekt auf der Regressionsgeraden liegen. Als Folge davon ist die Gesamtstreuung in $Y_{\text{imputiert}} \leq$ die Gesamtstreuung in $Y_{\text{tatsächlich}}$. Durch das Einbauen eines Störterms kann man versuchen bei obiger Ungleichung Gleichheit herzustellen. Diese Arbeitshilfe ist nun (hoffentlich) natürlicher und näher an der Realität, allerdings immer noch lediglich eine Arbeitshilfe und kein Informationsgewinn! In ein zu schlechtes Bild sollen sie dennoch nicht gerückt werden. Sie helfen die vorhandenen Teilinformationen nicht zu verlieren und mit den Vollständigen Daten eine umfassende Analyse durchzuführen.

Zusammenfassung

Die Notwendigkeit von korrekten statistischen Ergebnissen ist evident. Mit der verantwortungsvollen Datenanalyse ist das korrekte Vorgehen während jeder Stufe einer Statistik untrennbar verbunden. Für dieses Bestreben ist es notwendig sich der Fehlerquellen bewusst zu sein und angemessen darauf zu reagieren. Zur Fehlervermeidung stellen sich zwei Punkte als essentiell heraus: erstens die enge Kooperation zwischen Substanzwissenschaftler und Statistiker ; zweitens die Zusicherung von Anonymität.

Treten nun doch Fehlende Werte auf, so ist ein ‚Zufälliges Fehlen‘ unbedenklich. ‚Abhängig Fehlende‘ Werte stehen im Zusammenhang mit einer weiteren Variablen und sind im Zusammenspiel mit dieser Variablen nur unter Vorbehalt zu interpretieren. Eine Variable, deren Fehlende Werte ‚Nicht Zufällig Fehlend‘ sind, ist allgemein nur mit größter Vorsicht zu genießen. Sind aus rechentechnischen Gründen alle Lücken im Datensatz zu schließen, so ist das Bestreben, Werte nah an der Realität zu finden. Imputationsverfahren sind dabei nicht als Informationsgewinn, sondern als Informationserhaltung anzusehen.

Meiner persönlichen Überzeugung nach hilft die Berücksichtigung des oben genannten, falsche Ergebnisse zu reduzieren und ist somit ein Beitrag zur verantwortungsvollen Datenanalyse.

¹³ Joseph L. Schafer bietet dazu auf seiner Seiten Informationen an: <http://www.stat.psu.edu/~jls/mifaq.html>

Literaturangaben

Fahrmeir, L. et al. (2007): *Statistik*

Göthlich, S. (2005): *Zum Umgang mit fehlenden Daten in großzahligen empirischen Erhebungen*

Groves, R. (1989): *Survey Errors and Survey Costs*

Küchenhoff, H. und Kauermann, G. (2008): *Erkenntnisse aus Stichproben*

Radermacher, W. und Körner, T. (2006): *Fehlende oder fehlerhafte Daten in der amtlichen Statistik. Neue Herausforderungen und Lösungsansätze*

Schnell, R. (1991): *Wer ist das Volk?*

Internetquellen

- http://www.welt.de/die-welt/article1456180/Experte_sieht_Elterngeld_als_Ursache_fuer_steigende_Geburtenrate.html
- <http://www.zeit.de/online/2008/34/geburten-anstieg-analyse>
- <http://www.statistik-berlin.de/wir/amtliche-statistik/wir1.htm>
- <http://www.nzzfolio.ch/www/21b625ad-36bc-48ea-b615-1c30cd0b472d/showarticle/ffc70d1-99f5-4326-912f-dfc7f23cbc48.aspx>
- <http://www.jstor.org/pss/2283137>
- <http://www.stat.psu.edu/~jls/mifaq.html>

08. Juni 2009

Matthias Speidel