Ausarbeitung zum Vortrag

Einführung in die Techniken wissenschaftlicher Umfragen

Martin Spickermann

Empirische Forschungsmethoden in der Softwaretechnik
Blockseminar im Sommersemester 2004

Freie Universität Berlin
Fachbereich Informatik
Arbeitsgemeinschaft Software Engineering

Zusammenfassung

Umfragen sind eine häufig genutzte Methode zur Erforschung von Meinungen und Trends. Diese Ausarbeitung liefert Grundlagen zum Verständnis der Methoden und gibt Tipps für die Erstellung einer erfolgreichen Umfrage.

Abstract

Surveys are frequently used methods for the exploration of opinions and trends. This paper supplies basics for the understanding of the methods and gives tips for the preparation of a successful questionnaire.

Inhaltsverzeichnis

Einführung 4

Die Forschungshypothese 5

Das Umfragedesign 6

Arten von Umfragen 7

Die Zielgruppe 8

Vorbereitung zur Fragenerstellung 9

Der Fragenentwurf 9

Vortest 10

Zuverlässigkeit und Gültigkeit 11

Die Stichprobe 13

Verzerrungen bei der Durchführung 15

Rückläuferquote 16

Vorbereitung der Auswertung 17

Auswertung 17

Kritik an der Vorlage 18

Literaturverzeichnis 20

Grundlage dieser Ausarbeitung ist eine 6-teilige Artikelserie aus den *„Software Engineering Notes“,* einer Publikation der *Special Interest Group on Software Engineering.* Die SIGSoft ist eine Arbeitsgruppe der acm (Association for Computing Machinery), der weltweit ältesten und größten Vereinigung von Wissenschaftlern der Informatik.

Autorinnen der Artikelserie sind:

Barbara A. Kitchenham PhD

Professor of Quantitative Software Engineering

Keele University, Staffordshire, UK

Mitglied der Royal Statistical Society

Dr. Shari Lawrence Pfleeger PhD

President of Systems/Software Inc .

Arlington, Virginia, USA

IEEE Mitglied

Einführung

Umfragen begegnen uns nahezu täglich. Seien es die berühme Sonntagsfrage zu dem möglichen Ausgang, wenn am folgenden Wochenende eine Wahl stattfinden würde, seien es Polls im Internet, bei denen man mit nur einem Klick schnell seine Meinung kundtun kann, oder Telefonanruf von Emnid oder Forsa am frühen Abend.

Die Beliebtheit von Umfragen unter Meinungsforschern liegt vor allem darin begründet, dass es sie im Gegensatz zu Experimenten auf eine vergleichsweise kostengünstige Art einen guten Überblick über Trends, Entwicklungen und Meinungen wiedergeben. Weiterhin ermöglichen Sie die Erhebung und Erforschung von Daten und Zusammenhängen in den unterschiedlichsten Bereichen, darunter auch Bereichen, die mit Hilfe von Experimenten kaum oder gar nicht zu erforschen wä-

ren. Wie würde man mit Hilfe von Experimenten erfahren wollen, welche Meinungen zu aktuellen politischen Entscheidungen und Entwicklungen in der Bevölkerung vorherrscht? Dies ließe sich nur indirekt erkunden, indem man beispielsweise eine experimentelle Wahl stattfinden ließe und aus dem Wahlergebnis die politische Orientierung einer Partei als gewünschte Entwicklungsausrichtung deuten würde.

Auch wenn eine Umfrage auf den ersten Blick trivial anmutet, Umfragen sind mehr als einfach ein paar Fragen zu stellen und sich das Ergebnis anzuschauen. Damit Umfragen und ihre Aussagen anerkannt werden, müssen Regeln beachtet werden. Umfragen bedürfen einer umfangreichen Vorbereitung: Aus einer klaren Hypothese folgt eine sinnvolle Auswahl einer Zielgruppe. Stichproben müssen repräsentativ bestimmt werden, der gesamte Fragebogen muss auf Gültigkeit und Verlässlichkeit untersucht werden. Schließlich ist auch die Auswertung Bestandteil einer Umfrage und unterliegt bestimmten Gesetzmäßigkeiten.

Die Forschungshypothese

Am Anfang einer jeden Umfrage steht eine Frage oder eine Hypothese, die geklärt oder untersucht werden soll. Dies kann die Frage nach einer Meinung zu einer bevorzugten Entwicklungsumgebung sein oder auch die Betrachtung von einem Zusammenhang zwischen Art der Ausbildung und späterer Tätigkeit in einem Unternehmen. In jedem Fall jedoch muss diese Frage klar gestellt und auch messbar sein, denn nach ihr richtet sich der gesamte Prozess der Umfrage. Sie gibt die Sprache der Fragen und die Zielgruppe vor, bestimmt die Art und das Design der Umfrage und entscheidet, ob eine Umfrage überhaupt die passende Forschungsmethode ist.

Ist die Hypothese zu ungenau formuliert, dann wird man eventuell die falschen Personen befragen, einen Wortschatz benutzen, der nicht verstanden wird, oder man erhält ein nicht aussagekräftiges Ergebnis.

Das Umfragedesign

Wenn der zu erforschende Gegenstand klar formuliert ist, folgt daraus meist ein bestimmter Designtypus, der für die Umfrage zu wählen, bzw. der der Umfrage zuzuordnen ist. Man unterscheidet im Einzelnen folgende Umfragedesigns:

Beschreibende Umfragen erklären direkt aus Antworten heraus Zusammenhänge und Tatbestände.

- Die *Momentaufnahme* (cross sectional) ist dabei eine Umfrage, die den Zustand zu einem genau festgelegtem Zeitpunkt erhebt. Es wäre denkbar an einem Montag morgen zu einer bestimmten Uhrzeit an alle Arbeitnehmer einer Firma die Frage zu richten, wie motiviert sie nach dem Wochenende zur Arbeit erschienen sind. Das entscheidende dabei ist, dass der Zeitpunkt feststeht und dass er das Hauptkriterium ist.

- Die *Gruppenumfrage* (cohort) hingegen legt den Schwerpunkt auf die befragte Gruppe. Dies liefert Informationen über einen Sachverhalt oder eine Entwicklung in einem definierten Personenkreis. Damit erkennt man Trends, die von einer Zielgruppe erwartet werden.

- Bei *fallorientierten Umfragen* (case control) steht das Thema im Mittelpunkt. So kann sich das Thema zwar auch an eine bestimmte Gruppe richten - das ist aber nicht der Hauptaspekt. Die Evaluation an unserem Institut ist ein typischer Vertreter dieser Gattung.

Experimentelle Umfragen sind eine Kombination aus Experiment und Umfrage. Das Experiment muss dabei nicht ein Experiment im klassischen Sinne sein, möglich wäre auch, dass eine Fortbildung die Rolle des Experiments einnimmt. So kann es interessant sein, eine Teilnehmergruppe eines Seminars vor und nach dem Seminar zu ihrer Meinung zu befragen. Nimmt man dann noch Antworten einer Vergleichsgruppe hinzu, die nicht an dem Seminar teilgenommen hat, erhält man umfassende Hinweise auf den Nutzen dieser Veranstaltung. Man unterscheidet folgende experimentelle Umfragtypen:

- *Gleichzeitige Kontrollstudien* (concurrent control studies), bei denen die Teilnehmer *zufällig* gruppiert werden, d.h. die Zuordnung der Gruppen, die bei den Experimenten unterschiedliche Aufgaben zu erfüllen haben, findet nach keinem bestimmten Schema statt.

- G*leichzeitige Kontrollstudien*, bei denen die unterschiedlichen Gruppen naturgemäß (*nicht zufällig*) vorhanden sind. Will man unterschiedliche Verhaltensweisen der Geschlechter erkennen, so macht es Sinn, die Geschlechter gleich getrennt zu betrachten.

Vergleichende Kontrollstudien stellen verschiedene Untersuchungen gegenüber. Entweder werden Ergebnisse eines Teilnehmers mit den Antworten einer früheren Befragung dieser Person verglichen (self control studies), oder man zieht als Referenz eine frühere Studie an einer anderen Personengruppe heran (historical control studie).

Das oben angeführte Beispiel mit den Teilnehmern eines Seminars ist also eine Kombination aus einer gleichzeitigen Kontrollstudie, mit nicht zufällig gruppierten Befragten (Teilnehmer und Nichtteilnehmer) und einer vergleichenden Kontrollstudie, denn es werden die Ergebnisse der jeweils gleichen Person vorher und nachher verglichen.

Arten von Umfragen

Die zu untersuchenden Hypothesen können an Komplexität erheblich differieren. Je nachdem, ob die gestellten Fragen einfach oder schwer zu beantworten sind, aber auch abhängig vom Budget der Umfrage, hat man die Möglichkeit, die Befragten zu betreuen und zu unterstützen. Ein Extrem ist hierbei das *Einzelinterview*. Ob nun persönlich oder am Telefon - der Befragte kann jederzeit Verständnisfragen klären, oder der Interviewer kann hilfreiche Tipps zur Beantwortung geben. Bei schwierigen Sachverhalten kann diese Art von Umfrage von Vorteil sein, sie ist allerdings sehr teuer und bringt die Gefahr mit sich, dass der Befragte durch den Interviewer beeinflusst wird.

Das andere Extrem wäre eine Onlinebefragung oder ein per Post versandter Bogen (*unbetreute Umfragen*). Hier ist der Befragte nicht durch eine andere Person beeinflussbar, und Unklarheiten können nicht geklärt werden. Durch die niedrigen Kosten kann man erheblich mehr Personen in eine Erhebung mit einbeziehen.

Zwischenstufen sind denkbar: Man könnte jeweils einen Umfrageleiter einer Gruppe zuteilen, so dass er entweder nur die Fragen einleitet und Beispiele verdeutlich, oder gar Zwischenfragen beantwortet.

Die Zielgruppe

Nicht jeder Mensch kann eine sinnvolle Antwort zu jedem Thema geben. Es ist für den Erfolg einer Umfrage wichtig, die Zielgruppe, also die Personen, die befragt werden sollen, genau einzugrenzen. Ein Beispiel:

Wenn für eine Untersuchung von Interesse ist, ob ein bestimmter beruflicher Ausbildungsgang Voraussetzung für die häufige Benutzung von Mustern in der Softwareentwicklung ist, dann fallen als Zielgruppe Mitarbeiter von Softwarefirmen ins Auge. Doch diese Eingrenzung ist noch nicht ausreichend. So kommt nicht jeder Mitarbeiter in Betracht sondern nur Mitarbeiter, die aktiv in der Entwicklung von Software tätig sind. Auch sollte man vielleicht Mitarbeiter, die gerade aus der Ausbildung gekommen sind, aussparen, da sie noch nicht über den Erfahrungsschatz verfügen, um über ihre Neigung zur Musterbenutzung im Arbeitsalltag berichten zu können. Auch sehr langjährige Mitarbeiter könnten für die Untersuchung uninteressant sein, weil ihr Wissen über Muster häufig nicht mit ihrer Ausbildung sondern mit Erfahrung zu tun hat, schließlich war vor 30 Jahren der Ausbildungsgang eines Informatikers ein völlig anderer und ist nicht mit heutigem Standard vergleichbar.

Hat man eine Zielgruppe geformt, ergibt sich daraus ein Vokabular, das in den Fragen zu verwenden ist, und das für diese Zielgruppe eine bestimmte (vielleicht nicht allgemein gebräuchliche) Bedeutung hat. Gerade Begriffe in der Softwaretechnik wie „Release“ oder „Version“, haben für verschiedene Personen häufig andere Bedeutungen. In diesem Fall könnte ein Amateur unter einer Version ein

Majorrelease (z.B. Windows 2000) verstehen. Ein Marketingmitarbeiter kennt vielleicht noch die Abstufung der Service Packs, ein Programmierer hingegen sieht schon einen Unterschied zwischen der Version von heute zu der von gestern Abend.

Für den weiteren Umgang mit der Zielgruppe und einer eventuell zu bildenden Stichprobe ist es hilfreich die Zielgruppe genau zu erfassen. Wie groß ist sie, wie groß sind relevante Partitionen (Geschlecht, Alter, etc)? Denn daraus folgt, wie groß und von welcher Zusammensetzung eine repräsentative Stichprobe zu sein hat, damit ihr Ergebnis auf die gesamte Gruppe übertragbar ist.

Vorbereitung zur Fragenerstellung

Vor der Erstellung der Fragen empfiehlt es sich, ältere Studien zu ähnlichen Themen zu recherchieren. Dadurch wird vermieden, dass eventuell vorhandene Studien kopiert oder wiederholt werden. Auch können aus vorhandenen Umfragen Schlüsse gezogen werden. Gab es Lücken oder Probleme in der Auswertung, die vermeidbar sind? Beleuchtet die Studie Aspekte, die noch nicht bedacht wurden? Außerdem hat die Übernahme von Teilbereichen gegebenenfalls den Vorteil, dass man sich auf einer schon durchgeführten Prüfung von Gültigkeit und Verlässlichkeit in Teilbereichen abstützen kann.

Eine bedenkenlose Übernahme darf aber nicht erfolgen. Nach Veränderungen oder Abwandlungen muss die gesamte Gültigkeit erneut geprüft werden, da sie im neuen Kontext beeinträchtig sein kann. Genau beleuchtet werden muss, ob die Teilbereiche überhaupt Anwendung finden können. Ist die ältere Studie auf eine andere Gruppe ausgerichtet oder hat sie einen anderen Umfang, kann eine Übernahme problematisch bis unmöglich sein.

Der Fragenentwurf

Um den Befragten die Beantwortung zu vereinfachen oder gar überhaupt erst zu ermöglichen, sind bestimmte Regeln einzuhalten: Jede Frage sollte auf ihre sach-

liche und sprachliche Tauglichkeit in Hinblick auf die Zielgruppe geprüft werden und in einem erkennbaren Zusammenhang mit der zu untersuchenden Hypothese stehen. Die Fragen sollten in einer klaren Sprache gestellt sein, keinen Interpretationsspielraum lassen und einen genauen Zeitraum angeben, damit der Befragte die Antworten leichter geben kann. Um die Ergebnisse nicht zu beeinflussen, sollte man keine negierten Fragen verwenden (diese werden leicht missverstanden) und durch die Einleitung und die Fragestellungen den Probanten auch keine Antworten in den Mund zu legen.

Zu lange Fragebögen bergen das Risiko, dass der Befragte die Motivation verliert. Deshalb sollte man den Umfang des Fragebogens auf eine Bearbeitungszeit von circa 10-15 Minuten begrenzen. Zur schnelleren Beantwortung können standardisierte Antworttypen vorgegeben werden: Ja/Nein oder eine Skalen von sehr gut bis sehr schlecht (oder ähnlich den Fragen angepasst) mit einer empfohlenen Abstufung von 5-7 Schritten.

Zu unterschieden hat man zwei verschiedene Fragetypen, die unterschiedliche Antworten provozieren: *Offene Fragen* sind Fragen, die eine ausführliche, nicht vorgegebene Antwort vom Befragten verlangen. Die Antworten sind unter Umständen aufschlussreicher, aber auch schwerer auszuwerten, da sie ja nicht standardisiert sind. Das Gegenstück bilden *geschlossene Fragen*, die nur mit Ja/Nein oder einer Skala beantwortet werden können.

Vortest

Nach Fertigstellung des Fragenkatalogs muss dieser ausgiebig begutachtet werden. Dies kann man auf zwei Arten erreichen. Einerseits kann man die Fragen von einem unabhängigen, unvoreingenommenen Fachmann des Gebietes durchsehen lassen, andererseits kann man den Fragebogen an einer Focus Group (ausgewählte repräsentativ erscheinende Personen) testen. Letztere Variante hat den Vorteil, dass man gleichzeitig erste Ergebnisse erhält. Diese sind zwar aufgrund der geringen Anzahl an Probanten noch ungenau, geben aber einen Überblick über die zu erwartenden Ergebnisse.

Nach Abschluss dieses Vortests bietet sich unter Beachtung des Feedbacks (besonders in Hinblick auf überflüssige oder ungenaue Fragen) eine Überarbeitung der Fragen an.

Zuverlässigkeit und Gültigkeit

Ob eine Studie auch tatsächlich das widerspiegelt was gewünscht ist, und sie Ergebnisse liefert, die nicht anfechtbar sind, hängt von der Zuverlässigkeit der Antworten (Reproduzierbarkeit) ab.

Als Maß für die *Zuverlässigkeit* gibt es verschiedene Testmethoden: In einem *Wiederholungstest* (test-retest) lässt man den gleichen Probanten zu einem späteren Zeitpunkt erneut die Fragen beantworten. Ist die Korrelation der Antworten dabei bei mindestens 70%, kann man von einer hohen Zuverlässigkeit ausgehen. Problematisch ist hierbei aber, dass der Befragte sich eventuell noch an seine Antworten vom ersten Test erinnert, dass er in der Zwischenzeit auf dem abgefragten Gebiet hinzugelernt hat oder sich über Aspekte des Fragebogens Gedanken gemacht und damit vielleicht auch eine andere Meinung gebildet hat. Dies alles kann den Wiederholungstest beeinflussen.

Um zumindest die Verfälschung durch Erinnerung auszuschalten, kann man den Wiederholungstest auch in einer *geänderten Fassung* (alternate form) ausgeben. Dabei werden die Fragen in der Reihenfolge umsortiert oder umformuliert, was leider zu neuen Problemen führt. So hat eine Studie festgestellt, dass bei den folgenden beiden Fragen auf Frage B erheblich mehr Zustimmung erfolgt, wenn A vor B gefragt wird.

- A: „Sollen US Reporter ohne Zensur aus Russland berichten dürfen?“
- B: „Sollen russische Reporter ohne Zensur aus den USA berichten dürfen?“

Um die *interne Konsistenz* (internal consistency) der Fragen schon während des ersten Durchlaufs überprüfen zu können, gibt es die Möglichkeit im gleichen Test gleiche Fragen in anderer Form zu stellen. Lässt sich belegen, dass die Probanden bei beiden Fragen ähnliche bis gleiche Antworten geben, dann ist die Antwort

zuverlässig. Diese Methode empfiehlt sich aber nur in persönlichen Interviews, denn bearbeitet der Befragte die Fragen selber und merkt, dass ihm die gleichen Fragen wiederholt gestellt werden, führt das zu Irritation und Frustration, da er seine Glaubwürdigkeit angezweifelt sieht oder die Fundiertheit der Umfrage anzweifelt.

Es gibt auch Verfahren, die die Zuverlässigkeit mit Hilfe *statistischer Methoden* überprüfen, ohne dass Fragen doppelt gestellt werden müssen. Auf diese Methoden wird hier allerdings nicht eingegangen. Siehe dazu El Emam / SPICE.

Die *Gültigkeit* einer Umfrage setzt voraus, dass man belegt, warum die Antworten auch das widerspiegeln, was erforscht wird, d.h. dass der gewonnene Schluss auch wirklich zulässig ist. Eine *scheinbare Gültigkeit* (face validity) erreicht man mit Durchsicht der Fragen durch eine beliebige Person, die an der Bogenerstellung nicht beteilig war. Da diese Person nicht geschult ist und aber eine eingefärbte Meinung hat, ist dies Verfahren nicht sinnvoll. Besser ist eine *inhaltliche Gültigkeit* (content validity), die durch Prüfung mehrerer Fachleute gewonnen wird. Aber auch hier gilt, dass dies Verfahren statistisch nicht als sicher einzustufen ist.

Sollten auf dem Gebiet schon ältere, ähnliche Studien existieren, kann man die Fragen und die gewonnenen Antworten vergleichen. Diesen Vorgang nennt man *Kriteriumsvergleichsgültigkeit* (criterion validity). Dies bietet aber nur dann eine Grundlage, wenn an der Gültigkeit der Vergleichsstudie kein Zweifel besteht.

Die beste Variante ist der Beweis der *konstruktiven Gültigkeit* (construct validity). Hierbei erklärt man schlüssig, dass bei bestimmten Voraussetzungen auf die gestellten Fragen nur auf eine bestimmte Art und Weise geantwortet werden kann, was sich wiederum in einem Ergebnis widerspiegelt, dass die Voraussetzungen (und damit die Hypothese oder Frage) bestätig. Ein solcher Beweis ist naturgemäß äußert schwierig zu führen, aber aufgrund seiner Konstruktion aus statistischer Sicht aussagefähig.

Pfleeger und Kitchenham behaupten in ihrer Artikelserie, dass die Mehrzahl aller umfragebasierten Studien in der Softwaretechnik nicht hinreichend in ihrer Gültigkeit belegt seien.

Die Stichprobe

Wie bereits angedeutet, benutzt man für eine Umfrage in den meisten Fällen nicht die gesamte Zielgruppe, sondern man befragt nur eine wohl definierte Stichprobe. Die Gründe hierfür liegen vor allem in den Kosten und dem Aufwand, den eine Umfrage erzeugt, würde man eine große Zielgruppe befragen. Außerdem ist es bei sehr großen Gruppen (vgl. Sonntagsfrage an alle Wahlberechtigten) unmöglich eine solche Studie überhaupt durchzuführen.

Die *Stichprobe* (sample), die man auswählt, muss repräsentativ sein, damit sich im Anschluss an die Studie das Ergebnis von der Probe auf die gesamte Gruppe übertragen lässt. Repräsentativ bedeutet, dass die Zusammensetzung der Teilgruppe und die Zusammensetzung der Zielgruppe in allen für die Untersuchung relevanten Faktoren gleichen muss.

Weiterhin muss auch die Größe der Stichprobe in einem bestimmten Verhältnis zur gesamten Gruppe stehen. Man darf die Gruppe nicht zu klein wählen: Das Ergebnis wäre nicht übertragbar. Zu groß sollte sie aus Kostengründen auch nicht gewählt werden. Als Anhaltswert zur Ermittlung einer geeigneten Größe dient folgende Formel:

$$\text{sample Size} = \left[\frac{(z_a - z_b) * \sigma}{\mu_1 - \mu_2} \right]^2$$

wobei

$\mu_1 - \mu_2$ = erwartete Größendifferenz (aus pilot - test)

σ = Standartabweichung (aus pilot - test)

z_a = ist das obere Ende in der Standartnormalverteilung zu α. α ist Wahrscheinlichkeit eines Type I Fehlers.

z_b = ist das obere Ende in der Standartnormalverteilung zu β. β ist Wahrscheinlichkeit eines Type II Fehlers.

Bsp. $z_a = 1{,}96$ wenn $\alpha = 0{,}05$.

Bsp. $z_b = -0{,}84$ wenn $\beta = 0{,}20$.

Eine Stichprobe ist statistisch betrachtet nur dann auf die Allgemeinheit übertragbar, wenn sie zufällig entstanden, also nicht nach einem System ausgewählt wurde. Üblich sind folgende Verfahren, um zufällige Stichproben aus größeren Gruppen zu erhalten:

Der *einfache Zufall* (simple random) wird angewandt, wenn man eine *sortierte* Liste der Gesamtgruppe vorliegen hat. In diesem Fall wählt man *n* Teilnehmer an zufälliger Position der Liste aus.

Das Gegenstück wäre der *systematische Zufall* (systematic random), man verfügt dabei über eine *unsortierte* Liste und wählt jede *n*-te Person aus.

Unterscheidet sich die Gruppe stark und möchte man bestimmte Unterschiede im Verhalten herausfinden, wendet man den *geschichteten Zufall* (stratified random) an. Dieses Verfahren erlaubt, die Vorsortierung der Gruppe in Teilgruppen (z.B. nach Geschlecht oder Ausbildung). Aus diesem Teilgruppen ermittelt man dann die eigentliche Stichprobe, indem anteilig zu der Größe der Gruppen zufällig Personen ausgewählt werden.

Eine Sonderform stellt der *gebündelte Zufall* (cluster-based sampling) dar. Statt zufällig mehrere Personen auszulosen, fällt der Zufall immer gebündelt auf kleine Gruppen. Dies können alle Mitarbeiter einer Firma sein, wenn man die Firma zufällig bestimmt hat. Diese Methode hat den Vorteil, dass man Korrelationen zwischen zusammengehörigen Personen (z.B. denen aus einer Firma) gegenüber denen andere Bündel erforschen kann.

Doch bisweilen ist es nicht möglich zufällig eine Stichprobe auszuwählen, vielleicht weil man nicht genug Teilnehmer in der zu untersuchenden Gruppe kennt, was bei verschwiegenen Kreisen wie z.B. Hackern denkbar ist, oder es gibt gar nicht genug Teilnehmer, die die Zeit haben zu.

Im letzten Fall, der „bequemen Auswahl“ (*convenience sampling)*, wählt man die Teilnehmer aus, indem man jeden teilnehmen lässt, der teilnehmen will und der den Anforderungen entspricht. Plant man eine Umfrage unter Vorstandsvorsitzenden von börsennotierten Aktiengesellschaften, kann man sich möglicherweise nicht darauf verlassen, dass die ausgewählten Personen Zeit und Interesse haben und antworten. Vielmehr schreibt man mehr mögliche Teilnehmer an, als die Stichprobengröße verlangen würde, und erfasst jede Antwort, die zurückkommt. Umfragen in Internet entsprechen diesem Typus.

Kennt man nicht genug Teilnehmer einer Gruppe, interviewt man die, die bekannt sind, und bittet sie, weitere Probanten vorzuschlagen. Passenderweise nennt man diese Methode *Schneeballauswahl* (snowball sampling).

Als Gegenstück zum geschichteten Zufall bei den zufallsbasierenden Methoden kann man bei der *Quotenauswahl* (quota sampling) ebenfalls vorsortieren und dann die Teilgruppen nicht-zufallsbasierend auswählen.

Eine Sonderform stellt die *Focus Group* dar, die aus einem ausgewählten Personen besteht, die repräsentativ erscheint. Benutzt wird eine Focus Group gerne für Vortests, bei denen ein Eindruck über die Wirkung der Umfrage und dem zu erwartendes Ergebnis gewonnen werden soll.

Zu betonen ist besonders, dass alle vier genannten, nicht-zufallsbasierenden Verfahren eine Daseinsberechtigung für ihren speziellen Einsatzbereich haben, aber nur bedingt übertragbar sind. Insbesondere Umfragen im Internet leiden unter mangelnder Genauigkeit, da weder die Gruppe genau definiert ist, noch ist sichergestellt, dass die Gruppe repräsentativ ist. Bei Onlinebefragungen vor Wahlen erhält die FPD zum Beispiel oft erstaunlich hohe Vorhersagen, die später nie eintreten. Worauf könnte dieses zurückzuführen sein? Neben den Verfälschungen durch absichtliche Mehrfachabstimmung oder Beeinflussung des Ausgangs durch Verlinkung der Umfrage auf einschlägigen Webseiten liegt in diesem speziellen Fall die Vermutung nahe, dass Internetbenutzer nicht repräsentativ zum durchschnittlichen Bundesbürger sind. Internetbenutzer sind beispielsweise jünger und auch gebildeter als der Bundesdurchschnitt [10].

Verzerrungen bei der Durchführung

Verschiedene Faktoren können das Ergebnis einer Umfrage bei Interviews (supervised) beeinflussen. So ist das Verhalten des Interviewers während der Befragung bedeutend: Er kann durch die Art der Fragestellung den Befragten zu unrichtigen Antworten verleiten. Außerdem kann Sympathie oder Antipathie zwischen Interviewer und Befragten das Ergebnis färben. Ebenfalls kompliziert ist es, wenn der Befragte nicht so antworten kann, wie er gerne möchte, weil die vorge-

gebenen Antworten nicht seine Sicht widerspiegeln. Vorteilhaft ist es deshalb, immer auch Antwortmöglichkeiten wie „neutral“ oder „weiß nicht“ anzubieten, um in solchen Fällen unrichtige Antworten zu vermeiden.

Rückläuferquote

Neben der genauen Kenntnis der Zielgruppengröße und der daraus resultierenden Stichprobe ist es wichtig die Menge der Antworten und damit auch die der Personen zu kennen, die nicht geantwortet haben. Denn unter Umständen kann eine zu geringe Rücklaufquote – insbesondere eine einseitige Rücklaufquote – dazu führen, dass die Antworten nicht mehr repräsentativ sind, obwohl die gesamte Stichprobe es noch gewesen wäre. Dies wäre der Fall, wenn zu einem Thema anteilig mehr Amateure als Profis antworten, wenngleich die Stichprobe eigentlich ein ausgeglichenes Verhältnis vorgesehen hätte.

Somit muss man in Anschluss an die Umfrage genau erforschen, wie die Gruppe der Antwortenden zusammengesetzt ist, weshalb einige nicht geantwortet haben und ob dies eventuell das Ergebnis verfälscht. Weiterhin sollte man die Rücklaufquote in der Auswertung angeben und dem Leser nachvollziehbar erklären, ob und welche Auswirkung dies auf die Auswertung hat.

Verschiedene Tipps zum Design der Umfrage können die Rücklaufquote positiv beeinflussen, da sie den Befragten motivieren teilzunehmen oder das Antworten erleichtern. Das beginnt bei einer übersichtlichen Gestaltung der Fragebögen in Hinblick auf die Lesbarkeit und die gleichmäßige Benutzung von Gestaltungsmerkmalen. Bei Befragungen im Internet spielt die Navigation oder das benutze Frontend eine große Rolle. Ebenso dient die Lieferung von Hintergrundinformationen der Unterstreichung der Seriosität. Werden Ansprechpartner und Auftraggeber der Studie genannt, schafft dies Vertrauen in den Umgang mit den Daten und steigert somit die Motivation der Befragten. Es ist nicht empfehlenswert die Motivation durch Geschenke oder Preise zu steigern, da dies die Probanden beeinflussen könnte. Besser ist es den Befragten zu verdeutlichen, welchen Zweck die Studie verfolgt und welchen Vorteil er als Befragter aus der Studie ziehen kann.

Vorbereitung der Auswertung

Zur Vereinfachung der Auswertung oder auch zum Ausgleich eventueller Fehler bei der Konzeption der Umfrage kann es sinnvoll sein, vor Auswertung der Fragen, die Ergebnisse vorzusortieren. Eine Vorsortierung z.B. nach Geschlecht vereinfacht die Auswertung bzw. Analyse der Daten, wenn man Unterschiede in der Meinung von Männern und Frauen erkennen möchte.

Weiterhin müssen Vereinbarungen getroffen werden, wie mit Fragebögen umgegangen werden soll, die nicht vollständig ausgefüllt sind. Man hat die Möglichkeit, diese Bögen gar nicht in die Erhebung einfließen zu lassen, was den Vorteil bietet, dass auf alle Fragen die gleiche Anzahl an Antwortet in die Bewertung eingehen, was wiederum die Analyse von Korrelationen vereinfacht. Unter Umständen können aber auf diese Art sehr viele Fragebögen aus der Analyse herausfallen. Mögliche Folge: Die Stichprobengröße wäre wohlmöglich unzureichend. Alternativ erfasst man alle beantworteten Fragen. Dann haben aber die unterschiedlichen Fragen eventuell unterschiedliche Anzahl an Antworten, was eine Analyse von Zusammenhängen zwischen verschiedenen Fragen beeinträchtigen, oder gar unmöglich machen kann.

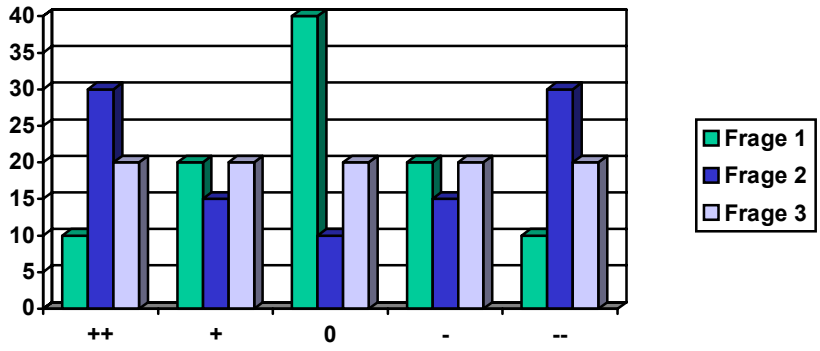
Häufig werden Ja/Nein Antworten oder Skalen in Zahlenwerte (z.B. 0/1 oder Schulnoten) übertragen, damit die rechnergestützte Verarbeitung vereinfacht wird. Dafür gilt es eine einheitliche Kodiervorschrift für die Erfassung zu definieren.

Auswertung

Die eigentliche Auswertung von Umfragen geschieht mit den üblichen Methoden der Statistik (Mittelwerte, Abweichungen, Streuungen), weshalb ich hier nicht weiter auf die Materie eingehe.

Betonen möchte ich aber doch die folgende Fehlerquelle bzw. mögliche Ungenauigkeit, die sich bei der Analyse einstellen kann.

Betrachten wir folgende Grafik:



Abgebildet ist die Verteilung der Antworten auf drei Bewertungsfragen, die jeweils eine Skala mit fünf Stufen (sehr gut, gut, neutral, schlecht, sehr schlecht) als Antwort boten. In allen drei Fällen wäre der Durchschnitt der Antworten neutral. Würde man dies allerdings so als Ergebnis darstellen, dann wäre es irreführend oder zumindest ungenau, da im Fall der Frage 1 zwar tatsächlich eine Häufung bei neutral zu beobachten ist, aber bei Frage 2 hingegen die Antworten sehr viel extremer ausgefallen sind. Während bei Frage 3 sogar alle Antwortmöglichkeiten gleichverteilt sind. In diesem Fall sollte die Streuung oder Standartabweichung auf jeden Fall erwähnt und auch erläutert werden.

Kritik an der Vorlage

Pfleeger und Kitchenham leiten ihre Veröffentlichung mit der Erklärung ein, dass sich die Beobachtungen während ihrer eigenen (nach eigener Aussage unglücklich verlaufenen) Studie ergeben hätten. Sie schreiben, dass sich die Arbeit speziell mit den Aspekten von Umfragen in der Softwaretechnik auseinandersetzt. Jedoch bleibt es aus meiner Sicht völlig unklar, wo denn nun die speziellen Schwierigkeiten dieser Untersuchungsmethode im Bezug auf die Softwaretechnik liegen. Die angegebenen Beispiele sind austauschbar und der Zusammenhang mit der Softwaretechnik eher zufällig.

Weiterhin führen Pfleeger und Kitchenham Beispiele an, wie man Umfragen optimieren könnte und welche Fehler vermieden werden sollten. Diese Tipps sind

sehr schlüssig und einleuchtend, doch leider geben die Autorinnen keinen Hinweis, ob diese Methoden auch wissenschaftlich fundiert sind. Besonders auffällig wird die fehlende Belegung an Aussagen folgender Art:

- Rücklaufquoten erhöht man, indem…
- Softwaretechnikumfragen sind selten hinreichend auf Gültigkeit getestet.

Dennoch haben die Autorinnen einen guten Überblick über die Techniken und Fallen dieser Forschungsmethode geliefert, so dass sich das Material durchaus eignet, um einen schnellen Einstieg in die Materie zu bekommen.

Literaturverzeichnis

[1] Shari Lawrence Pfleeger & Barbara A. Kitchenham: *Principles of Survey Research, Part 1: Turning Lemons into Lemonade*; Software Engineering Notes 26/6, November 2001

[2] Shari Lawrence Pfleeger & Barbara A. Kitchenham: *Principles of Survey Research, Part 2: Designing a Survey*; Software Engineering Notes 27/1, Januar 2002

[3] Shari Lawrence Pfleeger & Barbara A. Kitchenham: *Principles of Survey Research, Part 3: Construct a Survey Instrument*; Software Engineering Notes 27/2, März 2002

[4] Shari Lawrence Pfleeger & Barbara A. Kitchenham: *Principles of Survey Research, Part 4: Questionnaire Evaluation*; Software Engineering Notes 27/3, Mai 2002

[5] Shari Lawrence Pfleeger & Barbara A. Kitchenham: *Principles of Survey Research, Part 5: Populations and Samples*; Software Engineering Notes 27/5, September 2002

[6] Shari Lawrence Pfleeger & Barbara A. Kitchenham: *Principles of Survey Research, Part 6: Data Analysis*; Software Engineering Notes 28/2, März 2003

[7] Shari Lawrence Pfleeger: About me; http://home.earthlink.net/ ~spfleeger/id1.html

[8] Professor Barbara Kitchenham: Component Based Systems Research; http://www.keele.ac.uk/depts/cs/se/e&m/bakcv2.htm

[9] ACM: *Association for Computing Machinery, the world's first educational and scientific computing society*; http://www.acm.org/

[10] Statistisches Bundesamt: *Informationstechnologie in Haushalten und Unternehmen*; http://www.destatis.de/presse/deutsch/pm2003/ p0511024.htm