

Informationen für KundInnen der statistischen Beratung am IMSV

31. Mai 2024

Dieses Informationsblatt regelt die Zusammenarbeit zwischen dem Institut für mathematische Statistik und Versicherungslehre (IMSV) und den BeratungskundInnen. Das Ziel ist eine Standardisierung der Vorgehensweise im Rahmen der statistischen Beratung sowie die Festlegung eines Qualitätsstandards für die Zusammenarbeit. Zusätzlich findet sich ein Abschnitt mit Hinweisen, wie die Daten zur Übermittlung ans IMSV erfasst werden sollten.

1 Projektannahme, Erstberatungsgespräch

Zu einem an das Institut herangetragenen Projekt findet zunächst ein Erstberatungsgespräch am IMSV statt.

Im Rahmen des Erstberatungsgesprächs stellt der Kunde/die Kundin das Projekt vor, und es wird über eine künftige Zusammenarbeit im Projekt entschieden. Im Falle einer Zusammenarbeit werden ein Analyseplan sowie ein Zeitplan für die künftige Zusammenarbeit und ein Kostenrahmen vereinbart. Das IMSV behält sich vor, eine Zusammenarbeit abzulehnen.

Am Erstberatungsgespräch nehmen in der Regel zwei BeraterInnen des IMSV teil, wobei mindestens ein erfahrendes Teammitglied anwesend sein sollte. Handelt es sich beim jeweiligen Projekt um eine Abschlussarbeit (z. B. Dissertation oder Masterarbeit), so zeigt die Erfahrung, dass es sehr wertvoll ist, wenn neben dem/der Studierenden auch die Betreuungsperson in die statistische Beratung miteinbezogen wird. (Die Betreuungsperson muss zumindest informiert sein, dass statistische Beratung in Anspruch genommen wird.)

Für NeukundInnen ist die Erstberatung in der Regel kostenlos, sofern sie weniger als eine Stunde dauert und im Rahmen der angebotenen Sprechstunden-Termine stattfindet; diese Termine können direkt online gebucht werden. Umfangreichere Projekte und solche, welche nicht im Rahmen der Sprechstunden-Termine besprochen werden können, sind kostenpflichtig. Die Terminvereinbarung für kostenpflichtige Projekte erfolgt nach Absprache zwischen BeraterInnen und KundenInnen. Beratungstermine, die nicht wahrgenommen werden können, müssen von den KundInnen rechtzeitig abgesagt werden; andernfalls behält sich das IMSV eine Verrechnung vor.

Der Zeitplan für die Projektabwicklung und Auswertung hängt vom Umfang des Projekts und von der personellen Kapazität innerhalb des IMSV ab. Projekte, bei denen in Zusammenarbeit mit dem IMSV die Versuchsplanung erstellt wird, haben Priorität. Das IMSV kann keine Express-Auswertungen garantieren; generell muss damit gerechnet werden, dass Projekte mehr als 2 Wochen Bearbeitungszeit benötigen. Projekte werden nicht übernommen, wenn Auswertung und/oder Bericht in so kurzer Zeit fertig sein sollen, dass im Rahmen der laufenden Aufgaben am IMSV eine gewissenhafte Einarbeitung in das Projekt nicht möglich ist, wodurch eine fachlich adäquate Auswertung in Frage gestellt ist.

2 Datenübernahme / Datenhandling

Die Datenerfassung erfolgt in der Regel durch die BeratungskundInnen. Sollen Daten am IMSV ausgewertet werden, müssen sie in geeigneter elektronischer Form (vgl. Abschnitt 6 über die Struktur der Daten) übermittelt werden. Daten, die nicht den Richtlinien entsprechen, können vom IMSV zurückgewiesen werden.

Nach der Übernahme der Daten werden von den BeraterInnen die relevanten Daten auf ihre Plausibilität hin überprüft (Data-Checking). Die BeratungskundInnen tragen die Verantwortung für die Richtigkeit der Daten. Sollten sich während der Auswertung von der Seite der KundInnen noch Änderungen oder Erweiterungen an den Daten ergeben, so muss mit Verzögerungen und Mehrkosten gerechnet werden.

3 Auswertung

Die Fragestellungen sollen vor der statistischen Analyse schriftlich festgelegt werden. Die Auswertung umfasst die zur Beantwortung der mit den BeratungskundInnen vereinbarten Fragestellungen notwendigen statistischen Analysen. Es werden also *nicht sämtliche mit den Daten durchführbaren Analysen* vorgenommen, sondern nur die relevanten. Die Ergebnisse werden in der vereinbarten Form übermittelt. Ergeben sich aus der statistischen Analyse neue Fragestellungen und wünschen die BeratungskundInnen weiterführende Auswertungen, so muss wie im Erstgespräch die Finanzierung und Projektdauer neu besprochen werden. Das IMSV behält sich vor, ein Projekt vorzeitig zu beenden.

4 Bericht, Publikation

Über geplante Artikel, Vorträge, Poster usw. soll bereits im Rahmen des Erstgespräches gesprochen werden. Wenn in einer solchen Publikation das Institut für mathematische Statistik und Versicherungslehre oder eines seiner Mitglieder erwähnt wird, muss das Endmanuskript durch den/die BeraterIn gegengelesen und genehmigt werden. Nach der Publikation ist dem IMSV die publizierte Fassung zur Verfügung zu stellen. Auch eine allfällige Co-Autorschaft soll frühzeitig geregelt werden; eine solche ersetzt aber grundsätzlich nicht die Bezahlung der Arbeiten des IMSV.

5 Weiterverwendung der Daten

Das IMSV ist daran interessiert, geeignete Daten aus Beratungsprojekten in der Lehre und für statistisch-methodische Publikationen zu verwenden. Dies geschieht nur nach Einholung des Einverständnisses der entsprechenden KundInnen.

6 Datenerfassung zur Übermittlung ans IMSV

Datenschutz

Die übermittelten Daten dürfen keine vom Datenschutzgesetz geschützten Informationen enthalten. Insbesondere dürfen die PatientInnen in medizinischen Studien nicht durch Namen oder andere Angaben identifizierbar sein.

Form der Daten

Die gängigen statistischen Auswertungsprogramme (u. a. auch das am IMSV primär verwendete Programmpaket R) setzen voraus, dass die zu verarbeitenden Rohdaten in einer rechteckigen Datenstruktur angeordnet sind.

Darunter versteht man, dass in einer Tabelle die erhobenen Daten für jeden Fall (d. h. für die Beobachtungseinheiten wie z. B. Personen) in genau der gleichen Abfolge und Zahl von Merkmalen aufgelistet sind. Die zu einem Fall gehörigen Angaben werden jeweils in einer Zeile zusammengefasst. Die Messungen eines bestimmten Merkmals werden in der gleichen Spalte untereinander für jeden Fall eingetragen. Eine so strukturierte Datei enthält also je Fall (=Beobachtungseinheit) eine Zeile und je Merkmal (=Variable) eine Spalte.

Die ersten Felder jeder Zeile sind üblicherweise solchen Variablen zugeordnet, mit denen sich die jeweilige Beobachtungseinheit identifizieren lässt. Häufig ist dies eine einzige Variable, etwa eine Identifikationsnummer für die Personen in einer Stichprobe. Daran schliessen sich die Felder an, in denen die Werte weiterer Merkmale erfasst werden.

Wir empfehlen, die Daten als Dateien aus Tabellenkalkulationsprogrammen (z. B. Excel) oder als Textdateien mit Trennzeichen (z. B. Tabulator) zu übermitteln. Nach Absprache können die Daten auch mit Datenfiles von SAS, SPSS o. ä. übermittelt werden, falls Ihnen ein solches Programm zur Verfügung steht.

Messwiederholungen

Werden einzelne Merkmale für jede Beobachtungseinheit zu verschiedenen Zeitpunkten wiederholt erhoben, wie etwa bei Messung des Pulses unmittelbar vor einer Anstrengung und nach der Anstrengung, so muss für dieses Merkmal für jeden Messzeitpunkt ein *eigenes* Feld und somit eine gesonderte Variable zugeordnet werden:

PatNr	Alter	Puls1	Puls2
1	34	67	164
2	53	74	159
3	41	86	180
⋮	⋮	⋮	⋮

Werden fast alle Merkmale pro Beobachtungseinheit mehr als einmal erhoben, so kann es je nach geplanter statistischer Auswertung sinnvoller sein, diese Erhebungen zeilenweise zu erfassen. Dann werden zwei verschiedene Identifikations-Codes vergeben: einer für jede Beobachtungseinheit (z. B. die PatNr) und einer pro Erhebung (z. B. Zeitpunkt, die Nummer der Untersuchung):

PatNr	Zeitpunkt	Alter	Puls
1	1	34	67
1	2	34	164
2	1	53	74
⋮	⋮	⋮	⋮

Weitere Hinweise zur Dateneingabe

Insbesondere bei der Eingabe mittels Tabellenkalkulation sind folgende Hinweise zu beachten:

- Die Variablennamen müssen in der ersten Zeile stehen. Abgesehen von dieser ersten Zeile sollte die Tabelle nur Datenwerte enthalten (keine Zwischenresultate, Grafiken, Formeln usw.).
- Variablennamen sollen möglichst kurz sein, müssen mit einem Buchstaben beginnen und sollen keine Umlaute, keine Wortzwischenräume und keine Sonderzeichen (z. B. ! % # -) ausser dem Tiefstrich (-) enthalten.
- Die erste Variable soll die eindeutige Fall-Kennung enthalten (z. B. IDNR, PatNr, UPN, SERNO).
- Felder mit numerischen Variablen dürfen nur Ziffern, das Vorzeichen + oder – sowie entweder Dezimalpunkt oder Dezimalkomma enthalten.
- Falls Variablenwerte nicht nur Ziffern, sondern auch alphanumerische Zeichen enthalten (also Zeichenketten wie z. B. T1a oder C02), so sollten Umlaute und Sonderzeichen vermieden werden, um eine problemlose Umwandlung der Dateien zu gewährleisten.
- Klartext (z. B. Freitext-Antworten auf offene Fragen) ist nicht direkt auswertbar und muss deshalb sinnvoll codiert werden.
- Die Stufen kategorialer Variablen können mit alphanumerischen Bezeichnungen (z. B. Wörtern) oder mit numerischen Codes angegeben werden. Wenn verschiedene Variablen die gleichen möglichen Antwortkategorien haben, sollen die entsprechenden Codes für alle Variablen gleich sein, z. B. 0=nein, 1=ja. Geordnete Merkmale sollen aufsteigend (ordnungserhaltend) codiert werden, z. B. 0=nie, 1=selten, 2=häufig, 3=immer. Bitte (besonders bei alphanumerischen Bezeichnungen) darauf achten, dass die Codes immer gleich eingegeben werden (keine zusätzlichen Leerzeichen o. ä.; bei der Eingabe in einer Tabellenkalkulation können Funktionen zur Datenüberprüfung hilfreich sein).
- Fehlende Werte (missing values) sind durch leere Zellen darzustellen oder durch einen speziellen Code zu kennzeichnen (z. B. NA, MV, etc.). Codes, welche auch als reale Werte auftreten könnten (z. B. 99 in einer Variable für das Alter in Jahren) sind zu vermeiden.
- Zur Datenmatrix ist in den meisten Fällen ein Dokument nötig, das Bedeutung und Codierung der Variablen erklärt.
- Beim Import in ein Statistikprogramm gehen sämtliche Formatierungen (Farben, Schriftarten usw.) verloren. Wesentliche Informationen dürfen deshalb nie nur durch unterschiedliche Formatierungen dargestellt werden, sondern müssen immer aus dem unformatierten Inhalt der Zellen erkennbar sein.