STIFTERVERBAND

Leitlinien für die Arbeit mit Mikrodaten am FDZ Wissenschaftsstatistik

Bitte beachten Sie: Am Gastwissenschaftlerarbeitsplatz ist ein Internetzugang nicht zulässig. Sollten Sie für die Auswertung ado-Files oder R-Packages benötigen, müssen sie dem FDZ mindestens 2 Arbeitstage vor dem Besuch bereitgestellt werden (bspw. durch einen zip-Ordner oder Download).

Der von Ihnen gewünschte Output durchläuft vor der Herausgabe eine interne Prüfung, wobei datenschutzrechtliche Aspekte berücksichtigt werden. Der Zeitraum von der Erstellung des Outputs bis zur Freigabe kann bis zu zwei Wochen betragen, bedingt durch den Aufwand und die Kapazitäten im Team.
Die folgenden Informationen sollen Sie bei der Arbeit am FDZ Wissenschaftsstatistik unterstützen und die Outputkontrolle erleichtern. Die Einhaltung der Regeln trägt dazu bei, Ihren gewünschten Output zeitnah bereitstellen zu können und die Kosten für die Prüfung möglichst gering zu halten. Ein Abweichen von diesen Vorgaben kann im äußersten Fall zu einer Blockierung Ihres Outputs führen.

KRITERIEN FÜR DIE ERSTELLUNG DER SYNTAX

Eine verständliche Syntax ist grundlegend, um Ihren Output nachzuvollziehen und die Outputkontrolle durchzuführen. Eine fehlerfreie und verständliche Syntax entspricht weiterhin den Kriterien guter wissenschaftlicher Praxis, da Ihre Arbeitsergebnisse nachvollziehbar und jederzeit reproduzierbar sind. Werden die Kriterien zur Erstellung der Syntax nicht eingehalten, kann der erzeugte Output nicht geprüft oder freigegeben werden. Nachgereichte Erklärungen per Mail oder Telefon sind nicht zulässig.

Variablenbeschreibung

Bitte erstellen Sie eine Variablenliste, die den Namen, eine Beschreibung und eine Auflistung der Ausprägungen aller genutzten Variablen enthält (dies betrifft insbesondere Variablen aus externen Daten sowie im Datensatz neu generierte Variablen). Bei stetigen Variablen geben Sie bitte den minimalen und den maximalen Wert an.

Kommentierung

Alle Schritte der Datenaufbereitung und -auswertung sollten ausführlich kommentiert werden. Dazu zählen u. a. die Erzeugung neuer Variablen, das Verknüpfen von Datensätzen sowie das Erstellen statistischer Ergebnisse und Grafiken. Befehle, die das Anzeigen von Ausprägungen einer Variablen bzw. von Berechnungen ermöglichen (z.B. tabulate, describe, display, tabstat, codebook), müssen sorgfältig kommentiert werden. Dies gilt besonders für den Inhalt der dargestellten Variablen und für die Ausprägungen, die dargestellt werden.

Kennzeichnung von freizugebenden Ergebnissen

Die Syntax sollte insgesamt klar strukturiert sein und einzelne Auswertungsschritte nachvollziehbar darstellen. Bitte kennzeichnen Sie Ergebnisse, die freigegeben werden sollen (bspw. Grafiken), und solche Ergebnisse, die zur Durchführung der Outputkontrolle dienen (bspw. zugrundeliegende Fallzahlen). Die zugrundeliegenden Fallzahlen müssen bspw. bei multivariaten Analysen, Grafiken und Streuungsmaßen angegeben werden. Grundsätzlich sollten Sie immer nur solche Ergebnisse für die Outputkontrolle vorlegen, die Sie unbedingt für Ihre Arbeit benötigen.

Ausgabeformate

Tabellarische Auswertungen bzw. Arbeitsergebnisse sollten in einem Excel-File ausgegeben werden, grafische Auswertungen in den Formaten .pdf, .jpeg, .png (die zugrunedeliegenden Fallzahlen ebenfalls in einem Excel-File).

Reproduzierbarkeit des Outputs

Zu jedem Arbeitsergebnis muss die entsprechende Syntax (do-file, r-Skript) vorliegen. Output, der im Rahmen der Outputkontrolle geprüft werden soll, muss durch die Syntax reproduzierbar sein.

STIFTERVERBAND

KRITERIEN FÜR DIE ZULASSUNG VON OUTPUT

Hier finden Sie grundlegende Regeln für die Zulässigkeit von Arbeitsergebnissen. Eine Missachtung dieser Regeln führt zu Sperrungen in den Inhalten. Das Erstellen von wissenschaftlichen Papern mit inkludierten Ergebnissen ist grundsätzlich nicht zulässig – bitte erstellen Sie keine Textdokumente mit Ergebnissen aus Ihren Analysen.

Anonymität

Der Output darf keine Unternehmens-/Hochschulnamen oder sonstige sensible Informationen enthalten. Darunter fallen auch erhebungsspezifische Identifikationsmerkmale wie die ID, Ident_sv, ident_vvc. Ein Abweichen von dieser Regel führt zu einer Sperrung der Inhalte.

Einzelwerte

Einzelwerte wie bspw. Minima, Maxima und Residuen werden bis auf wenige Ausnahmefälle gesperrt. Es dürfen nur Aggregate veröffentlicht werden. Ausnahmen sind bspw. Minima und Maxima von erzeugten Hilfsmerkmalen (z.B. Konzentrationsmaße) oder Dummy-Variablen.

Fallzahlen

Deskriptive Statistiken werden gesperrt, wenn n < 20; Fallzahlen werden gesperrt, wenn n < 5.
Regressionsanalysen werden gesperrt, wenn n < 100 oder die Fallzahl zwischen Analysen um n +/- 1 abweicht.
Auch die Freigabe von Quantilen orientiert sich an den o.g. Fallzahlen, d.h. jedes Quantil gilt als eigene Zelle. Ein Ergebnis wird gesperrt, wenn zu einem Quantils-Abschnitt weniger als 20 Fälle beitragen. Daraus ergeben sich für die Freigabe von Quantilen bspw. die folgenden Mindestfallzahlen: N=40 für 50 %-Quantile, N=400 für 5%- bzw. 95%-Quantile oder N=2000 für 1%- bzw. 99 %-Quantile.

Dominanz

Ein Wert wird gesperrt, wenn der Beitrag des größten Einzelwerts 90% am Gesamtwert übersteigt. In einer Zelle dürfen weiterhin nicht mehr als 90% der Einheiten einer Zeile/Spalte liegen. Ein Wert wird freigegeben, wenn bei der Ausgabe von Summen die Fallzahlen sowie der größte Einzelwert ausgegeben werden und in den Ergebnissen keine Dominanzfälle auftreten.

Abbildungen

Grafiken/Abbildungen werden gesperrt, wenn zugrundeliegende Fallzahlen oder Werte gesperrt werden müssen. Abbildungen müssen Überschriften enthalten, die den inhaltlichen Bezug deutlich machen, X- und Y-Achse sind inhaltlich zu beschriften. Speichern Sie Abbildungen bitte in einem konventionellen Bildformat ab (Bsp: png, jpg). Ein Abweichen von dieser Regel führt dazu, dass die jeweilige Abbildung nicht freigegeben wird.