FORSCHUNGSDATENZENTRUM
der Bundesagentur für Arbeit im Institut für
Arbeitsmarkt- und Berufsforschung

Richtlinien zur Gestaltung von Auswertungsprogrammen und Analyseergebnissen

Inhaltsverzeichnis

Einleitung **3**

1 Warum sollten Nutzende die Richtlinien lesen? **4**

2 Datennutzung in der FDZ-Infrastruktur **6**
2.1 Ablauf von der Analyse bis zur Freigabe der Ergebnisse 6
2.2 Vertragliche Verpflichtungen und Sanktionen 7

3 Richtlinien zur Gestaltung der Auswertungsprogramme **8**
3.1 Die Master-Datei 8
3.2 Auswertungsprogramme 9
3.3 Datensparsamkeit 10

4 Richtlinien für datenschutzkonforme Ergebnisse **12**
4.1 Anzahl der Beobachtungen allgemein 12
4.2 Häufigkeitsauszählungen 13
4.3 Statistische Kennzahlen 14
4.4 Percentiles / quantiles 16
4.5 Gewichtung 17
4.6 Grafiken 17
4.7 LaTeX-Output 18
4.8 Aggregierte Daten 18
4.8.1 Import aggregierter Daten in das Projekt 18
4.8.2 Dokumentation von Aggregationsschritten innerhalb eines Projekts . . . 18
4.8.3 Export/Transfer aggregierter Datensätze aus dem Projekt 19
4.9 Regressionoutput 21
4.10 Ereignisdatenanalyse 21
4.11 Sequenzmusteranalyse 22

Einleitung

Das Forschungsdatenzentrum (FDZ) der Bundesagentur für Arbeit (BA) im Institut für Arbeitsmarkt- und Berufsforschung (IAB) ermöglicht externen Forschenden den Zugang zu Mikrodaten für die nicht-kommerzielle Forschung im Bereich der Arbeitsmarkt- und Berufsforschung. Bei den vom FDZ bereitgestellten Daten handelt es sich um Sozialdaten, die den Datenschutzbestimmungen des Sozialgesetzbuches X (SGB X [1]) bzw. den Vorgaben der statistischen Geheimhaltung unterliegen.

Neben der Erstellung und Dokumentation von Datenprodukten für die Arbeitsmarkt- und Berufsforschung besteht die Hauptaufgabe des FDZ darin, Datennutzende bei der Sicherstellung der Geheimhaltung von Informationen über statistische Einheiten wie Personen, Betriebe oder Unternehmen zu unterstützen.

[1] Zehntes Buch Sozialgesetzbuch - Sozialverwaltungsverfahren und Sozialdatenschutz

1 Warum sollten Nutzende die Richtlinien lesen?

Um den Datenschutz zu gewährleisten muss in der Praxis nach Auswertungen per kontrollierter Datenfernverarbeitung vor Herausgabe von Ergebnisdateien deren Deanonymisierungspotenzial geprüft werden. Da diese Arbeit aufgrund der Vielzahl an Nutzenden und der Fülle an erzeugten Ergebnissen sehr viele Ressourcen bindet, müssen die Prozesse standardisiert werden. Nur so kann das FDZ schnell alle Anfragen abarbeiten und hat noch genügend freie Kapazitäten für Aktualisierung und Weiterentwicklung des Datenangebots. Hierbei ist das FDZ auf die Unterstützung der Nutzenden angewiesen.

Was können Sie zu einem reibungslosen und schnellen Prozess beitragen?

- Lesen Sie folgenden Dokumente **bevor** Sie mit den Analysen in der FDZ-Infrastruktur beginnen:
  - Die 'Anleitungen für Datenfernverarbeitung und Gastaufenthalte am FDZ der BA im IAB' (im Folgenden als „Anleitungen" bezeichnet)
  - Die 'Richtlinien zur Gestaltung von Auswertungsprogrammen und Analyseergebnissen' (im Folgenden als „Richtlinien" bezeichnet)
  - the 'o Stata-Templates für Auswertungsprogramme' (im Folgenden als „FDZ-Templates" bezeichnet)
- Gestalten Sie Ihre Programmdateien nach diesen Dokumenten.
- Überprüfen Sie die Lauffähigkeit Ihrer Programme anhand von Testdaten bzw. Substichproben.
- Überprüfen Sie die Lauffähigkeit der Programme und die datenschutzrechtliche Unbedenklichkeit Ihrer Ergebnisse bevor Sie die Ergebnisse dem FDZ zur Prüfung und Freigabe vorlegen. D.h. in der Regel sollten die Analyseergebnisse so gestaltet sein, dass keine nachträglichen manuellen Löschungen durch das FDZ-Team mehr erforderlich sind.
- Reduzieren Sie die Ergebnisse, die einer Datenschutzprüfung unterzogen werden müssen, auf das Nötigste. Lassen Sie nur Ergebnisse prüfen, die für Publikationen oder Präsentationen vorgesehen sind. Ergebnisse vorläufiger Analysen, Qualitätsprüfungen im Rahmen der Datenaufbereitung etc. sollten dem FDZ nicht zur Prüfung und Freigabe vorgelegt werden.

- Bei Abbrüchen von Programmläufen versuchen Sie zunächst selbst die Fehlerursache zu finden und zu beheben. Überprüfen Sie zunächst, ob Ihre Syntax korrekt ist und die technischen Vorgaben für die FDZ-Infrastruktur (z.B. zur Verwendung von „globals“ und Pfadangaben) richtig umgesetzt sind. Ziehen Sie die FAQ-Liste für JoSuA zu Rate. Erst wenn alle Maßnahmen nicht zum Erfolg führen, kontaktieren Sie das FDZ.

Das FDZ unterstützt Sie hierbei mit Programmiervorlagen und -beispielen FDZ-Templates, die einen einfachen Einstieg sowohl zur Umsetzung der Richtlinien der Syntaxgestaltung als auch der Ausgabe datenschutzkonformer Resultate ermöglichen. Eine ausführliche Beschreibung der Templates finden Sie im Zip -Archiv. Darüber hinaus finden Sie in den folgenden Abschnitten Verweise auf entsprechende Programmierbeispiele in den FDZ-Templates.

Das vorliegende Dokument dient der Erklärung und Vertiefung der Anforderungen zur Syntaxerstellung und zur Ausgabe von datenschutzkonformen Analyseergebnissen, die anhand von Beispielen in den FDZ-Templates dargestellt sind. Sie erfahren hier, wie der Analyseprozess in der IT-Infrastruktur des FDZ unter Einhaltung der Vorgaben idealerweise abläuft. Zudem werden aus datenschutzrechtlicher Sicht als kritisch eingestufte Ergebnisse erläutert sowie Strategien zur Vermeidung solcher Ergebnisse aufgezeigt.

Welche Vorteile haben Sie durch die Einhaltung der Richtlinien?

- Sie erhalten Ihre Ergebnisse schneller.
- Ihre Ergebnisse enthalten wenige oder keine Löschungen und zusätzliche Sperrungen, um das Rückrechnen von gelöschten Werten zu verhindern.
- Ihre Ergebnisse basieren auf ausreichend hohen Fallzahlen und haben daher eine höhere Aussagekraft.
- Durch die Einhaltung der Richtlinien zur Syntaxgestaltung erfüllen Ihre Programme die Anforderungen der „guten wissenschaftlichen Praxis“ und sind damit gut nachvollziehbar. Ihre Ergebnisse können zudem jederzeit reproduziert werden
- Durch eine effiziente Prüfung der Ergebnisse bleiben dem FDZ mehr Ressourcen für andere Arbeiten, wie der Aktualisierung bestehender und Erschließung neuer Datenprodukte oder der Verbesserung der Dienstleistungsqualität.

> **!** Wichtig
>
> Bei Fragen kontaktieren Sie bitte das Forschungsdatenzentrum (iab.fdz@iab.de). Bitte geben Sie immer Ihre FDZ-Projektnummer (**fdzXXXX**) und ggf. die **Job-ID aus JoSuA** an.

2 Datennutzung in der FDZ-Infrastruktur

2.1 Ablauf von der Analyse bis zur Freigabe der Ergebnisse

Die Datennutzung sollte in der Regel in Stata erfolgen. Sind vereinzelte Auswertungen mit Stata nicht durchführbar, besteht die Möglichkeit, dafür mit R oder Octave zu arbeiten. Kontaktieren Sie hierfür bitte im Vorfeld das FDZ.

Die Nutzung der Daten des FDZ erfolgt mittels kontrollierter Datenfernverarbeitung[1] oder im Gastaufenthalt bzw. Remote Desktop Zugang. Für alle Zugangswege setzt das FDZ die Browser-basierte Software JoSuA (Job Submission Application) ein. Nähere Informationen zu JoSuA und den unterschiedlichen Modi finden Sie in den Anleitungen zur Nutzung von Datenfernverarbeitung und Gastaufenthalt bzw. Remote Desktop Zugang. Bitte nutzen Sie für die Datenaufbereitung , die Vorbereitung von Gastaufenthalten und zum Testen der Lauffähigkeit der Auswertungsprogramme den Internal Use Modus in JoSuA. Die für Sie sichtbaren Ergebnisse werden hier einer automatischen Datenschutzprüfung unterzogen und zensiert. Die Ergebnisse können in der JoSuA-Webschnittstelle betrachtet, aber nicht heruntergeladen werden[2]. Die im Internal Use Modus in Ihrem Nutzerverzeichnis gespeicherten Datensätze stehen Ihnen anschließend auch im Presentation/Publication Modus zur Verfügung. Bitte beachten Sie, dass die im Remote Desktop Zugang gespeicherten Datensätze über JoSuA nochmals generiert werden müssen.

Die Programme der ausgewählten Ergebnisse, die Sie im Rahmen einer Präsentation oder Publikation veröffentlichen wollen, laden Sie bitte im Presentation/Publication Modus hoch. Stellen Sie bitte zuvor im Internal Use Modus sicher, dass Ihre Do-Files lauffähig und die den Ergebnissen zugrundeliegenden Fallzahlen[3] der einzelnen Auswertungen ausreichend sind (weitere Informationen zu der erforderlichen Anzahl an Beobachtungen finden Sie im weiteren Verlauf dieses Dokuments). Die im Presentation/Publication Modus hochgeladenen Programme und Ergebnisdateien werden einer Datenschutzprüfung durch das FDZ-Team unterzogen.

[1] Bei Abschluss des Projektes können die Datenaufbereitungsprogramme über den Presentation/Publication Modus zum Nachweis für die Begutachtung beispielsweise durch Zeitschriften hochgeladen werden. Vermerken Sie dies bitte entsprechend im Kommentarfeld.

[2] Die im Internal Use Modus generierten Ergebnisse dürfen nicht abgeschrieben, fotografiert, ausgedruckt oder kopiert werden. Es ist außerdem untersagt, Dritten Einsicht zu gewähren. Jede Verwendung dieser Ergebnisse außer zur Weiterentwicklung Ihrer Programme stellt eine Vertragsverletzung dar.

[3] Wir behalten uns vor, sämtliche Ergebnisdateien innerhalb von JoSuA-Jobs, die offensichtlich nicht zuerst im Internal Use Modus auf Lauffähigkeit und Einhaltung des Datenschutzes getestet wurden, nicht freizugeben.

Hierbei werden in der Regel **keine** weiteren manuellen Löschungen und Zensierungen vorgenommen. In begrenzten Ausnahmefällen übernimmt das FDZ manuelle Löschungen in den Log-Files. Dafür müssen ausreichende Begründungen, warum die Ergebnisse nicht in andere Form ausgegeben werden können, in den Programmdateien hinterlegt werden. Nach erfolgreicher Prüfung durch das FDZ werden Ihnen die Programme und Ergebnisdateien zum Download zur Verfügung gestellt. Sollten noch kritische Ergebnisse enthalten sein, wird das betroffene Log-File oder ggf. der gesamte betroffene Job nicht freigegeben. Sie haben dann die Möglichkeit, die dazugehörige Programmdatei entsprechend zu überarbeiten und nochmals in JoSuA hochzuladen.

2.2 Vertragliche Verpflichtungen und Sanktionen

Trotz der Prüfroutinen im FDZ sind Sie verpflichtet, die übermittelten Ergebnisse auch nochmal selbst zu prüfen. Bei Zweifeln dürfen die Ergebnisse nicht veröffentlicht werden, sondern müssen dem FDZ angezeigt werden. Zudem verpflichten Sie sich im Datennutzungsvertrag auf Geheimhaltung, was unter anderem umfasst, dass keine Rückrechnungen vorgenommen werden, um gelöschte Werte zu ermitteln. Bei Verstößen werden Sie bzw. in schweren Fällen auch Ihr gesamtes Institut bzw. Ihre gesamte Universität für die Dauer von bis zu zwei Jahren von jeder weiteren Datennutzung ausgeschlossen. Des Weiteren können die im Nutzungsvertrag genannten Strafen zur Anwendung kommen. Informationen über Verstöße gegen die o.g. Verpflichtungen werden zudem an andere Forschungsdatenzentren weitergegeben.

3 Richtlinien zur Gestaltung der Auswertungsprogramme

Um die Datenschutzprüfung durchführen zu können, muss es den in einem täglich rotierenden Verfahren wechselnden prüfenden FDZ-Mitarbeitenden möglich sein, sich auch ohne Kenntnisse über den Projektinhalt in den Programmen zurecht zu finden. Aus diesem Grund ist eine ausreichende Kommentierung der Do-Files notwendig. Beispiele zum Aufbau der Do-Files finden Sie in den FDZ templates auf unserer Homepage.

Folgende allgemeine Kriterien müssen bei der Generierung von Auswertungsprogrammen beachtet werden, damit ein Auswertungsprogramm den Richtlinien des FDZ[1] entspricht.

3.1 Die Master-Datei

Die Master-Datei muss folgendermaßen gestaltet sein:

- Alle im Verlauf des Projektes verwendeten Programme sind von einer übergeordneten Master-Datei aufzurufen. Sie muss immer `master.do` heißen. Die Datei `master.do` do in den FDZ-Templates veranschaulicht ihren Aufbau.
- Benutzen Sie zum Zugriff auf Dateien immer die Stata-Globals **$orig, $prog, $data und $log**. Sie stehen sowohl bei der Datenfernverarbeitung mit JoSuA als auch bei Gastaufenthalten und dem Remote Desktop Zugang automatisch zur Verfügung und verweisen auf den korrekten Speicherort. Die Globals dürfen nicht überschrieben werden.
- Nutzen Sie keine cd-Befehle oder relative Pfadangaben.
- Verwenden Sie niemals die Anweisung **set max_memory**.
- Nicht erlaubte Befehle werden durch JoSuA zensiert. In diesen Fällen sehen Sie die Zeile auskommentiert mit *** CENSORED ***.
- Die Master-Datei muss alle im Projektverlauf relevanten Auswertungsprogramme enthalten. Es ist wichtig, dass in der Master-Datei hinter jedem Programmaufruf eine kurze Beschreibung der Inhalte des jeweiligen Unterprogramms erfolgt. Verzichten Sie auf die Option *nostop*.
- Die Auswertungsprogramme sind in der Reihenfolge des Programmablaufs zu nummerieren. Arbeiten mehrere Nutzende zusammen an einem Projekt, so können ihre

[1] Weitere Informationen finden Sie in den Anleitungen

Programme durch Kürzel im Dateinamen unterschieden werden, z.B. `jd02_describe.do`. Dies gilt nicht für `master.do`.
- Alle von Ihnen erzeugten Dateien (Ergebnisdateien, Analysedatensätze, usw.) müssen anhand Ihrer Programme und der Originaldaten replizierbar sein.
- Bereits ausgeführte Programmschritte, die seit der letzten Einsendung über JoSuA nicht verändert wurden, sind im Projektverlauf in `master.do` auszukommentieren. Über JoSuA besteht der Zugriff auf gespeicherte Datensätze in den Verzeichnissen **orig** und **data**. Lediglich Do- und Log-Files aus JoSuA-Jobs werden auf separaten Servern gespeichert.
- Für JoSuA empfehlen wir, im Internal Use Modus die Zeilenlänge mit dem Befehl `set linesize` auf 120 zu setzen, da dies der Breite des Output-Fensters in JoSuA entspricht.

3.2 Auswertungsprogramme

- Ergebnisdateien (auch Grafiken) müssen den gleichen Dateinamen tragen wie das Programm, mit dem diese Dateien generiert werden. Beispielsweise erzeugt `jd02_describe.do` die Ergebnisdatei `jd02_describe.log`. Die zwei in `jd05_graphs.do` erzeugte Grafiken könnten z.B. `jd05_graphs_02_boxplot1.gph` und `jd05_graphs_02_boxplot2.gph` heißen.
- Die Programme müssen für die Datenschutzprüfung in deutscher oder englischer Sprache dokumentiert sein (siehe FDZ templates). Achten Sie dabei darauf:
  - in jedem Do-File immer **zuerst** ein Log-File zu öffnen (vor jeder weiteren Anweisung und vor der inhaltlichen Beschreibung des Do-Files),
  - an den Anfang jedes Do-Files zentrale Informationen (z.B. Autor*in, verwendeter Datensatz) zu schreiben (siehe auch Beispiel in `jd02_describe.do`),
  - Zwischenüberschriften zu verwenden (z. B. “Aufbereitung der Variablen X”, “Aggregation auf Betriebsebene”),
  - die Inhalte und Funktion von Schleifen zu beschreiben und Tabellenbeschreibungen in Schleifen zu programmieren (z. B. mit dem display Befehl),
  - für alle wichtigen Variablen, die neu gebildet wurden, Variablenlabels und ggf. Wertelabels zu vergeben,
  - möglichst sprechende Variablennamen zu verwenden,
  - selbst generierte Variablen direkt vor Auszählungen kurz zu beschreiben und
  - o den Inhalt einer jeden Abbildung und komplexer Tabellen zu beschreiben (siehe Beispiele in den FDZ templates).
- Ergebnistabellen und Grafiken, die im Presentation / Publication Modus ausgegeben werden, müssen sorgfältig dokumentiert werden (siehe auch Beispiele in `jd02_describe.do` und `jd05_graphs.do`):
  - Jede Tabelle/Grafik muss eine Überschrift mit einer fortlaufenden Nummer innerhalb eines Do-Files haben.

  - Jede Tabelle/Grafik benötigt eine Beschreibung mit dem Namen und dem Inhalt der ausgewiesenen Variablen, den relevanten Auswahlkriterien und dem Verwendungszweck der Tabelle/Grafik.
  - Alle anderen Befehle für Ergebnisse, die nur im Internal Use Modus, Remote Desktop Zugang oder im Gastaufenthalt relevant sind, müssen gelöscht oder auskommentiert werden.

- Für die Ausgabe von Regressionsergebnissen empfehlen wir folgendes Vorgehen (siehe Beispiele in `jd03_analyses.do`):

  - Fassen Sie – sofern möglich - mehrere Regressionen in einer Tabelle zusammen (z. B. mit *estimates table*, *estout*, usw.). Lassen Sie sich diese Ergebnisse möglichst nicht zweimal ausgeben! Hierfür können Sie "*quietly*" vor dem Regressionsbefehl verwenden.
  - Wenn Sie nur eine Teilmenge der Regressionskoeffizienten für Ihre Veröffentlichung benötigen, lassen Sie bitte nur diese ausgeben, umfangreichere Ergebnisse können Sie sich zur Prüfung im Internal Use Modus ausgeben lassen.
  - Beschreiben Sie bitte die Besonderheiten, die für die Überprüfung des Datenschutzes relevant sein könnten (vollständig interagierendes Modell, Regressand ist Dummy, usw.). Aus der Kopfzeile der Regressionstabelle muss klar hervorgehen, um welche Ergebnisse es sich handelt (falls es mehrere gibt).
  - o Wir akzeptieren Regressionstabellen, die als TeX-Code ausgegeben werden, aber nur, wenn sie leicht lesbar sind; stellen Sie sicher, dass die Spalten feste Breiten haben.

- Auswertungsprogramme (wie auch eingeschickte Ado-Files) dürfen grundsätzlich keine Daten oder Fallzahlen enthalten. Externe aggregierte Datensätze müssen separat als Stata-Dateien eingeschickt werden. Ausgenommen hiervon sind Klassifikationen und Umrechnungsfaktoren (z. B. Beitragsbemessungsgrenzen, Euro-Umrechnung, BIP-Deflatoren, Verbraucherpreisindizes, Umschlüsselung von Wirtschaftszweigen, Kreisen oder Arbeitsagenturbezirken). Bitte kommentieren Sie, welche Klassifikationen und Umrechnungsfaktoren Sie in Ihrer Programmierung verwenden. Beachten Sie hierbei bitte unsere Anleitungen zum Import aggregierter Datensätze in Abschnitt 2.2.3.

3.3 Datensparsamkeit

- Verwenden Sie den Speicherplatz in Ihrem Verzeichnis **data** sparsam. Wird das Maximum von 30 GB überschritten, können keine weiteren Jobs eingeschickt werden (ausgenommen Löschprogramme, siehe `jd07_cleanup.do`), bis der Speicherbedarf reduziert wurde.
- Speichern Sie nur dann Zwischendatensätze ab, wenn es für die weitere Analyse erforderlich ist. Datensätze sollten mithilfe Ihrer Do-Files jederzeit reproduzierbar sein.

Löschen Sie zwischenzeitlich generierte Datensätze, wenn diese nicht länger benötigt werden (siehe `jd07_cleanup.do`).

- Schränken Sie Ihren Datensatz auf diejenigen Variablen und Beobachtungen ein, die Sie für Ihre Analysen tatsächlich benötigen (siehe `jd01_dataprep.do` und `jd04_mergebhp.do`).
- Zur Entwicklung Ihrer Programme kann es sinnvoll sein, die Lauffähigkeit zunächst an einer Stichprobe zu testen (siehe `jd00_subsample.do`).
- Nutzen Sie den Befehl *compress* mindestens vor dem Speichern von Datensätzen.
- Achten Sie bei der Erstellung von Variablen darauf, sie im sparsamsten Format zu erzeugen (z.B. *generate byte* oder *generate int*).
- Der verfügbare Arbeitsspeicher ist beschränkt auf **30 GB per person and project**.

4 Richtlinien für datenschutzkonforme Ergebnisse

Um die Ergebnisse aus Jobs im Presentation / Publication Modus freizugeben, erfolgt eine datenschutzrechtliche Prüfung am FDZ in einem zweistufigen Verfahren: Zunächst kommt ein eigens dafür entwickeltes Prüfskript zur automatischen Löschung von kritischen Werten zum Einsatz. Dieses Skript kann jedoch nur einfache Ausgabeformate in Stata (z.B. für die Befehle `tabulate oneway`, `summarize`) prüfen und kritische Werte löschen. Anschließend erfolgt immer eine manuelle Prüfung aller Ergebnisse durch die Beschäftigten des FDZ.

Das FDZ nimmt in der Regel **keine manuellen Löschungen** in den Ergebnisdateien vor, sondern prüft lediglich, ob die Anforderungen für die Mindestanzahl erfüllt sind oder zu geringe Fallzahlen durch das automatische Prüfskript des FDZ zensiert wurden. In Ausnahmefällen übernimmt das FDZ manuelle Löschungen, wenn begründet werden kann, dass eine Zusammenfassung von Kategorien einer Variablen, um die Ausgabe von zu kleinen Fallzahlen zu vermeiden, nicht möglich ist. Nachträgliche manuelle Löschung durch die Nutzenden sind nicht möglich.

Sie müssen dafür Sorge tragen, dass Ihre Programme keinen Output mit zu geringen Fallzahlen erzeugen. Beispiele zur Gestaltung der Do-Files sind in den FDZ templates enthalten. Im Folgenden sind die wichtigsten Richtlinien des FDZ zur datenschutzrechtlichen Prüfung der Analyseergebnisse dargestellt. Die folgenden Beispiele sollen Ihnen zum einen veranschaulichen, welche Ergebnisse als datenschutzrechtlich bedenklich eingestuft werden, und zum anderen helfen, die Ausgabe solcher Ergebnisse zu vermeiden.

4.1 Anzahl der Beobachtungen allgemein

- Weisen Sie immer die zugrundeliegenden Fallzahlen bei allen statistischen Kennzahlen oder multivariaten Analysen aus. Bei Fehlen der Fallzahlen werden die betreffenden Ergebnisse nicht freigegeben.
- Sämtliche Ergebnisse müssen aus datenschutzrechtlichen Gründen auf **mindestens 20 Beobachtungen** basieren. Dies gilt für alle statistischen Kennzahlen, Kreuztabellen und auch multivariate Analysen. Die Mindestanforderung von 20 Beobachtungen der betrachteten Einheit gilt sowohl für die Unternehmens-, Betriebs- als auch die Personendaten, die das FDZ anbietet.

- Im Rahmen der Datenschutzprüfung am FDZ werden Log-Files oder ggf. der gesamte Job nicht freigegeben, wenn Werte enthalten sind, die auf weniger als 20 Beobachtungen (Personen und/oder Betrieben und/oder Unternehmen) basieren.
- Weisen Sie bei Auswertungen kombinierter Unternehmens-, Betriebs- und Personendaten in deskriptiven Tabellen sowohl die Anzahl der Beobachtungen für Personen als auch für Betriebe bzw. Unternehmen aus.

4.2 Häufigkeitsauszählungen

Bei Häufigkeitsauszählungen werden Werte < 20 in der Regel nicht freigegeben. In Kreuztabellen muss die Fallzahl jeder einzelnen Zelle >= 20 sein, die Gesamtzahl der Tabelle ist hier nicht ausreichend. Wenn also Auswertungen auf einer zu kleinen Beobachtungszahl basieren, müssen die Programme zu deren Erzeugung von Ihnen so angepasst werden, dass keine zu geringen Fallzahlen mehr ausgegeben werden. Beispielsweise können bei deskriptiven Statistiken Ausprägungen zusammengefasst oder ausgeschlossen werden, damit keine manuellen Löschungen mehr erforderlich sind. **Beispiel 1** zeigt wie mit der Zusammenfassung der Kategorien „100-499 SVB“ und „500-999 SVB“, die Ausgabe von zu kleinen Werten verhindert werden kann. Weitere Hinweise finden sich im FDZ-Template `jd02_describe.do`.

Beispiel 1: Ostdeutschland

Vorher

	Betriebsrat		
Anz. svpfl. Besch.	Ja	Nein	Total
1 1-4 SVB	43	1380	1423
2 5-9 SVB	39	547	586
3 10-99 SVB	594	1322	1916
4 100-499 SVB	573	175	748
5 500-999 SVB	142	16	158
Total	1391	3440	4831

Nachher

	Betriebsrat		
Anz. svpfl. Besch.	Ja	Nein	Total
1 1-4 SVB	43	1380	1423
2 5-9 SVB	39	547	586
3 10-99 SVB	594	1322	1916
4 100-999 SVB	715	175	906
Total	1391	16	4831

Ist eine Zusammenfassung von Kategorien aus wichtigen Gründen nicht möglich, kann in Ausnahmefällen eine manuelle Zensierung der Ergebnisse durch das FDZ-Team erfolgen. Zunächst müssen alle Werte (Fallzahlen, Verteilungsparameter und Regressionskoeffizienten), die auf einer Beobachtungszahl < 20 basieren, gelöscht werden. Um die Möglichkeit des Rückrechnens von gelöschten Werten durch Zwischen- oder Randsummen zu verhindern, müssen ggf. weitere Werte gelöscht oder gerundet werden.

In **Beispiel 2** wird das Vorgehen des FDZ bei der Datenschutzprüfung von deskriptiven Kreuztabellen kurz erläutert. Zunächst wurden hier Werte < 20 aufgrund der zu geringen Fallzahl durch "/" ersetzt (Primärsperrung). Damit keine Rückrechnung des gelöschten Wertes mit Hilfe der Randsumme möglich ist, wurden zudem weitere Werte durch ein ("*") einer Sekundärsperrung unterzogen.

Zudem muss ausgeschlossen werden, dass eine Rückrechnung über mehrere Tabellen hinweg möglich ist. Dies wird in **Beispiel 3** deutlich. In Westdeutschland ist die Fallzahl in den einzelnen Zellen zwar jeweils >=20, da jedoch neben Ost- und Westdeutschland auch die Tabelle für Gesamtdeutschland ausgewiesen ist, wäre eine Rückrechnung der gelöschten Werte von Ostdeutschland (aus Beispiel 2) möglich. Aus diesem Grund werden auch in der Tabelle für Westdeutschland die entsprechenden Zellen zensiert (Sekundärsperrung: "*").

Beispiel 2: Ostdeutschland

Vorher

	Betriebsrat		
Anz. svpfl. Besch.	Ja	Nein	Total
1 1-4 SVB	43	1380	1423
2 5-9 SVB	39	547	586
3 10-99 SVB	594	1322	1916
4 100-499 SVB	573	175	748
5 500-999 SVB	142	16	158
Total	1391	3440	4831

Nachher

	Betriebsrat		
Anz. svpfl. Besch.	Ja	Nein	Total
1 1-4 SVB	43	1380	1423
2 5-9 SVB	3*	54*	586
3 10-99 SVB	594	1322	1916
4 100-499 SVB	573	175	748
5 500-999 SVB	14*	/	158
Total	1391	3440	4831

Aufgrund des Deanonymisierungsrisikos ist es notwendig, Auszählungen nach allen Subgruppen auszuweisen, sobald die Möglichkeit des Rückrechnens besteht. In obenstehendem Beispiel ist es daher notwendig sowohl die Tabelle nach West- als auch nach Ostdeutschland auszuweisen, da durch die Werte für Gesamtdeutschland auf Ostdeutschland rückgerechnet werden kann. Wollen Sie hingegen beispielsweise anstatt eines Ost-West-Vergleichs branchenspezifisch nur das Baugewerbe analysieren, ist keine Ausgabe aller Wirtschaftszweige notwendig, da keine Rückrechnung mit Hilfe der Ergebnisse für die Gesamtwirtschaft möglich ist.

4.3 Statistische Kennzahlen

Statistische Kennzahlen lassen auf den ersten Blick keinen Rückschluss auf die zugrunde gelegten Fallzahlen zu. Das heißt aber nicht, dass die Ausgabe statistischer Kennzahlen, wie z.B. Mittelwerte, nicht problematisch sein kann. Auch hier gilt das Prinzip, dass die angesprochenen Kennzahlen nur dann als sicher eingestuft werden, wenn die Berechnungsgrundlage mindestens 20 Beobachtungen umfasst. Als Spezialfall ist die Ausgabe

statistischer Kennzahlen bei Dummys zu sehen. Bei binär kodierten Merkmalen verteilen sich die Ausprägungen dieser auf lediglich zwei Kategorien. Auch wenn die Gesamtanzahl der Beobachtungen einer Dummy-Variablen größer als 20 ist, kann es sein, dass durch eine schiefe Verteilung nur drei Personen, Betriebe oder Unternehmen auf eine der Ausprägungen fallen. In diesem Fall wird die Auswertung als unsicher angesehen, auch wenn nicht direkt über die Gesamtanzahl der Beobachtungen auf eine geringe Besetzung einer Kategorie geschlossen werden kann. Denn dies kann mit Hilfe des Mittelwerts einfach herausgefunden werden. Um diese Fälle identifizieren und prüfen zu können, ist es erforderlich, dass bei der Ausgabe von Mittelwerten dokumentiert wird, ob es sich um eine Dummy-Variable handelt.

Beispiel 3: Westdeutschland

Vorher

	Betriebsrat		
Anz. svpfl. Besch.	Ja	Nein	Total
1 1-4 SVB	64	2461	2525
2 5-9 SVB	54	847	901
3 10-99 SVB	859	1985	2844
4 100-499 SVB	793	255	1048
5 500-999 SVB	198	22	220
Total	1968	5570	7538

Nachher

	Betriebsrat		
Anz. svpfl. Besch.	Ja	Nein	Total
1 1-4 SVB	64	2461	2525
2 5-9 SVB	5*	84*	901
3 10-99 SVB	859	1985	2844
4 100-499 SVB	793	255	1048
5 500-999 SVB	19*	2*	220
Total	1968	5570	7538

Gesamtdeutschland

Vorher

	Betriebsrat		
Anz. svpfl. Besch.	Ja	Nein	Total
1 1-4 SVB	107	3841	3948
2 5-9 SVB	93	1394	1487
3 10-99 SVB	1453	3307	4760
4 100-499 SVB	1366	430	1796
5 500-999 SVB	340	38	378
Total	3359	9010	12369

Nachher

	Betriebsrat		
Anz. svpfl. Besch.	Ja	Nein	Total
1 1-4 SVB	107	3841	3948
2 5-9 SVB	93	1394	1487
3 10-99 SVB	1453	3307	4760
4 100-499 SVB	1366	430	1796
5 500-999 SVB	340	38	378
Total	3359	9010	12369

Beispiel 4 veranschaulicht die Problematik bei der Ausgabe von Mittelwerten. Bei der Dummy-Variable r61 mussten alle Werte zensiert werden, da aufgrund der schiefen Verteilung nur 12 Betriebe (140*0.0857143) auf eine der beiden Ausprägung fallen.

Beispiel 4

Tabelle 4.9: Vorher

Variable	Obs	Mean	Std. Dev	Min	Max
r60	201	2.373134	0.9192794	1	3
r61	140	0.0857143	0.2809469	0	1
r62a	73	2.219178	2.340742	1	15

Tabelle 4.10: Nachher

Variable	Obs	Mean	Std. Dev	Min	Max
r60	201	2.373134	0.9192794	1	3
r61	140	/	/	/	/
r62a	73	2.219178	2.340742	1	15

Bei Verwendung des `summarize`-Befehls in Stata werden entsprechende Fälle durch das Prüfskript des FDZ automatisch überprüft und ggf. gelöscht. Werden entsprechende Ergebnisse anhand anderer Befehle (z.B. `table`, `tabstat`) erzeugt, kann keine automatische Löschung erfolgen. In diesen Fällen müssen Sie auf die Ausgabe von Ergebnissen mit zu geringen Fallzahlen verzichten, indem Sie den entsprechenden Code aus Ihren Auswertungsprogrammen löschen.

4.4 Percentiles / quantiles

Bei der Ausgabe von Perzentilen muss darauf geachtet werden, **dass mindestens 20 Beobachtungen im jeweiligen Perzentil** enthalten sind. Bei einer detaillierten Ausgabe (1% Perzentile) müssen folglich insgesamt mindestens 2000 Beobachtungen in die Ausgabe mit einfließen, um die Datenschutzvorgaben des FDZ zu gewährleisten. Grundsätzlich gilt, je detailliertere Informationen man erhalten will, umso mehr Beobachtungen müssen für die gesamte Verteilung vorhanden sein:

- Mind. 20 Beobachtungen für die Ausgabe von Mittelwerten (Ausnahme Dummys)
- Mind. 40 Beobachtungen für die Ausgabe von 50%-Perzentilen
- Mind. 80 Beobachtungen für die Ausgabe von 25%- oder 75%-Perzentilen
- Mind. 200 Beobachtungen für die Ausgabe von 10%- oder 90%-Perzentilen
- Mind. 400 Beobachtungen für die Ausgabe von 5%- oder 95%-Perzentilen
- Mind. 2000 Beobachtungen für die Ausgabe von 1%- oder 99%-Perzentilen

Werden Quantilswerte ausgegeben, so ist immer der geringste Abstand zwischen den gewählten Prozentsätzen (auch zur Null und zu 100) ausschlaggebend – werden z.B. die Perzentile 10 – 15

– 30 ausgegeben, muss die Stichprobe mindestens (15-10)/100 * x 20 → x 400 Beobachtungen betragen, um die notwendigen 20 Beobachtungen in jedem Quantil zu gewährleisten.

Bei Verwendung des `summarize`-Befehls in Stata werden entsprechende Fälle durch das Prüfskript des FDZ automatisch überprüft und ggf. gelöscht. Werden entsprechende Ergebnisse anhand anderer Befehle (z.B. `table`, `tabstat`) erzeugt, kann keine automatische Löschung erfolgen. In diesen Fällen müssen Sie auf die Ausgabe von Ergebnissen mit zu geringen Fallzahlen verzichten, indem Sie den entsprechenden Code aus ihren Auswertungsprogrammen löschen oder so anpassen, dass Perzentile mit größeren Abständen ausgegeben werden, für die ausreichend viele Fälle vorliegen.

4.5 Gewichtung

Bei Verwendung jeglicher Hochrechnungs- oder Gewichtungsfaktoren in deskriptiven Analysen sind auch immer analoge ungewichtete Ergebnisse auszuweisen. Die gewichteten und die dazugehörigen ungewichteten Ergebnistabellen müssen dabei **must directly follow** folgen, weil dies die Datenschutzprüfung erleichtert und somit - auch für Sie - beschleunigt. Dies muss auch in Schleifen entsprechend berücksichtigt und kommentiert werden. Fehlt die ungewichtete Ausgabe, kann das entsprechende Log-File bzw. ggf. der gesamte Job nicht freigegeben werden. Weitere Hinweise finden sich im FDZ-Template jd02 `jd02_describe.do`.

Bei gewichteten Regressionen ist die Angabe der ungewichteten Fallzahl notwendig, wenn `iweights` oder `fweights` verwendet werden.

4.6 Grafiken

Zu jeder Grafik ist nachzuweisen, auf wie vielen Beobachtungen die einzelnen dargestellten Werte basieren. Dabei ist es nicht ausreichend die Gesamtfallzahl auszuweisen, sondern es muss die **Beobachtungszahl jedes einzelnen dargestellten Balkens bzw. Punktes** aufgelistet werden. Dies kann über die Angabe der Fallzahlen entweder direkt in den Grafiken oder in Tabellen, die sich unmittelbar vor oder nach der Erstellung der Grafiken befinden, erfolgen. Bei der Darstellung von sehr vielen Datenpunkten in Grafiken ist es nicht erforderlich, für jeden einzelnen Datenpunkt die Fallzahl auszugeben. Hier reicht es, wenn Sie nachweisen, dass die kleinste zugrundeliegende Fallzahl mindestens 20 ist. Ein Beispiel dafür finden Sie im FDZ-Template `jd05_graphs.do`.

Auch für Grafiken gilt die Fallzahlgrenze von mindestens 20 Beobachtungen. Somit sind z.B. Scatterplots auf Individualebene nicht erlaubt, da hinter den einzelnen ausgegebenen Datenpunkten weniger als 20 Beobachtungen stehen. Während bei Histogrammen die Fallzahlen für jeden einzelnen Balken ausgegeben werden müssen, ist bei Kernel Density Plots die Gesamtfallzahl ausreichend. Koeffizienten- bzw. Marginsplots sind unbedenklich, wenn

die Fallzahl in der zugehörigen Regressionstabelle ausreichend ist. Zudem ist die Verwendung der asis-Option beim Speichern der Grafik im gph-Format verpflichtend, um sicherzustellen, dass die Grafik in der derzeitigen Form gespeichert wird und keine nachträgliche Bearbeitung mehr möglich ist. Weitere Hinweise finden sich im FDZ-Template `jd05_graphs.do`.

4.7 LaTeX-Output

Stata-Ergebnisdateien müssen entweder die Endung „.log“ oder „.txt“ tragen. Ergebnisse, die in separate Dateien außerhalb der von Stata angelegten Log-Files geschrieben wurden, können nicht bereitgestellt werden. Sie müssen wieder in das Log-File mit `type Pfadangabe` eingebunden werden. Achten Sie darauf, dass der Output **übersichtlich gestaltet und leicht lesbar** ist (z.B. durch feste Spaltenbreiten). Sollte dies in Ausnahmefällen nicht umsetzbar sein, müssen die Ergebnisse unmittelbar vor dem eingebundenen Text nochmals in lesbarer Form dargestellt werden. Ein Beispiel finden Sie im FDZ-Template `jd03_analyses.do`.

4.8 Aggregierte Daten

4.8.1 Import aggregierter Daten in das Projekt

Externe Kenngrößen auf aggregiertem Niveau (z. B. Arbeitslosenquoten nach Kreisen) dürfen den Daten zugespielt werden, wenn sie den Datenschutzrichtlinien des FDZ entsprechen. Weitere Informationen zu den Voraussetzungen und zum Vorgehen finden Sie in unseren Anleitungen in Abschnitt 2.2.3.

4.8.2 Dokumentation von Aggregationsschritten innerhalb eines Projekts

Wenn Sie die ursprünglichen Mikrodaten für weitere Analysen aggregieren (z. B. auf die Ebene von Branchen, Regionen etc.), stellen Sie bitte sicher, dass dies im Log-File angemessen dokumentiert wird (Überschriften; erläuternde Anmerkungen). Alle Ergebnisdateien, die nach einem solchen Aggregationsschritt erstellt werden (und den aggregierten Datensatz als Hauptdatenquelle verwenden), müssen am Anfang in der Kommentierung folgende Informationen enthalten:

- auf welche Ebene aggregiert wurde
- in welchem Programmschritt (z.B. Do-File) aggregiert wurde (bitte auch in der Master-Datei kennzeichnen)

Fehlen diese Informationen oder sie sind schwer zu finden, können die Ergebnisse nicht geprüft und freigegeben werden.

Wenn Sie diese aggregierten Variablen später ausgeben wollen, beachten Sie bitte, dass hier immer noch die Regeln für die Mindestanzahl von Beobachtungen (siehe Abschnitt 4.1 bis 4.3) gelten. Daher müssen die folgenden Anforderungen erfüllt sein:

- Es muss je aggregierter Variable, die ausgegeben werden soll, eine zusätzliche Variable gebildet werden, welche die Anzahl der entsprechenden gültigen Beobachtungen pro Zelle enthält.
- Ist das gewünschte Aggregat z.B. ein Anteil oder eine Quote muss bei tabellarischen Darstellungen zusätzlich die Anzahl der gültigen Werte in den einzelnen Subgruppen, die zur Erstellung des Aggregats verendet wurden, ausgegeben werden (z. B. Anzahl Männer in Bundesland X und Anzahl Frauen in Bundesland X für die Frauenquote in Bundesland X).

Beispiele zur richtigen Programmierung von Aggregaten finden Sie im FDZ-Template `jd02_describe.do`. Die Anzahl der aggregierten Tabellen sollte im Sinne der Datensparsamkeit auf ein Minimum beschränkt sein [1].

4.8.3 Export/Transfer aggregierter Datensätze aus dem Projekt

Falls Sie einen umfangreicheren aggregierten Datensatz als .dta-Datei erhalten oder in ein anderes Projektverzeichnis transferieren möchten, weil er als Tabelle in einem Log-File nicht richtig dargestellt und geprüft werden kann (oder wenn Sie diesen Datensatz in ein anderes Projektverzeichnis übertragen möchten), sprechen Sie die Vorgehensweise bitte **vorher**, im Idealfall bereits bei der Antragsstellung, mit uns ab. Neben dem Zweck der Aggregation muss geklärt werden, auf welcher Ebene aggregiert wird und wie die enthaltenen Variablen generiert wurden (Summe, Mittelwert, etc.). Erläutern Sie, warum der Export/Transfer notwendig ist.

Die Generierung der aggregierten Daten kann mit den Testdaten vorbereitet werden und sollte während eines Gastaufenthalts erfolgen. Die Programme müssen anschließend nochmals mit JoSuA gestartet werden. Bitte verwenden Sie dafür den Modus Presentation/Publication und vermerken Sie im Kommentarfeld, dass Sie einen generierten Datensatz erhalten bzw. transferieren möchten und den entsprechenden Dateinamen. Speichern Sie diesen Datensatz immer im Unterverzeichnis **data**.

[1] Zu beachten ist hierbei, dass das automatisierte Prüfskript bei bestimmten Stata-Befehlen wie summarize oder tabulate auch aggregierte Ausgaben, beispielsweise auf Kreisebene, trotz einer ausreichenden Anzahl an Personen bzw. Betrieben/Unternehmen innerhalb der Kreise löscht bzw. zensiert. In solchen Fällen würden wir Sie bitten Rücksprache mit dem FDZ zu halten, um eine solche Löschung von Ergebnissen durch eine angepasste Programmierung zu vermeiden.

Da die Prüfung von aggregierten Datensätzen sehr aufwändig ist, müssen folgende Regeln bei der Erstellung berücksichtigt werden:

- Die vollständige Erstellung des aggregierten Datensatzes, von Original-Datensatz bis zum zu exportierenden Datensatz, muss in **einem** Do-File/Log-File erfolgen. Bitte vermeiden Sie unnötige Schritte im Do-File. Ein Beispiel finden Sie im FDZ-Template `jd06_export_aggregate.do`.
- Die Vorgaben zur Dokumentation und Darstellung der Fallzahlen, die im vorherigen Abschnitt beschrieben wurden, gelten ebenso.
- Verwenden Sie aussagekräftige Variablennamen, Variablenlabels und Wertelabels für die zu exportierenden Merkmale.
- Vermeiden Sie sich überlagernde Aggregationen (z.B. Bundesland + Arbeitsmarktregion + Kreis), zumindest wenn die kleinste Aggregationsebene Zellen enthält, die zensiert werden müssen.
- Fügen Sie keine Variablen für weitere Untergruppen im 'wide' Format an. Bleiben Sie im 'long' Format.
- Die Variablen, die die Aggregationsebene angeben, müssen "lvl_1_*, lvl_2_*, ..." heißen, z.B. bei einer Aggregation nach Jahren und Bundesländern sollten die Variablen für die Kategorien mit lvl_1_year und lvl_2_fstate bezeichnet werden.
- Nach jeder aggregierten Ergebnis-Variablen (wenn sie nicht an sich schon eine Zählvariable ist) muss eine Variable, die die entsprechende Anzahl der zugrundeliegenden Beobachtungen enthält, folgen. Diese Variable muss genauso benannt werden, aber das Präfix N_ enthalten.
- Versuchen Sie, den aggregierten Datensatz so klein wie möglich zu halten. Insbesondere sollten Werte, die anhand anderer Informationen im aggregierten Datensatz berechnet werden können (z.B. Summen oder Quoten), erst nach dem Export oder Transfer generiert werden, um den Prüfaufwand zu reduzieren.

Zusätzlich müssen Sie für den Export/Transfer eines aggregierten Datensatzes sicherstellen, dass jeder Zelle mindestens 20 Beobachtungen zugrunde liegen. Sollte dies nicht der Fall sein, haben Sie folgende Optionen:

- **Fassen Sie Kategorien zusammen** (vom FDZ bevorzugtes Vorgehen). Möchten Sie z. B. Aggregate auf Kreisebene erhalten, fassen Sie Kreise, die zu geringe Fallzahlen aufweisen, zusammen.
- Zensieren Sie Zellen, die zu geringe Fallzahlen aufweisen selbst (vom FDZ nicht bevorzugtes Vorgehen). Gehen Sie dabei wie folgt vor:
  - Ist die Fallzahl in einzelnen Zellen des Datensatzes < 20, setzen Sie die Zählvariable und alle anderen Ergebnisse für dies Zelle auf die Missing-Kategorie ".p".
  - Halten Sie die Anzahl der Zellen mit nicht ausreichend großen Fallzahlen so gering wie möglich. Bedenken Sie, dass wir ggf. eine Sekundärsperrung vornehmen müssen (siehe Abschnitt 4.2).

- Ergänzen Sie eine Begründung, warum weitere Aggregationen/Zusammenfassungen für Ihr Projekt nicht vertretbar sind.
- Bitte beachten Sie, dass wir uns das Recht vorbehalten, die Freigabe von Daten für den Export oder Transfer zu verweigern, wenn sie zu gering besetzte Zellen enthalten, die nicht angemessen begründet und dokumentiert sind, oder wenn wir zu dem Schluss kommen, dass der Zeitaufwand für eine angemessene Datenschutzprüfung zu groß wäre.

Für jedes Projekt kann **nur einmal** ein solcher Datensatz weitergegeben werden. Dies gilt auch, wenn der aggregierte Datensatz in einem anderen FDZ-Projektverzeichnis zur Zuspielung bereitgestellt werden soll. Der Export des Datensatzes erfolgt per E-Mail. Transfers in andere Projektverzeichnisse übernehmen die Beschäftigten des FDZ.

4.9 Regressionoutput

Die Gesamtzahl der zur Schätzung verwendeten Beobachtungen muss bei Regressionen mindestens 20 betragen. Ergebnisse multivariater Analysen, die auf weniger als 20 Beobachtungen beruhen, können generell nicht freigegeben werden und sollten daher aus den Programmen entfernt werden.

Mit Hilfe von Regressionen lassen sich unter Umständen allerdings auch indirekt kleine Fallzahlen ermitteln. Deshalb reicht es hier nicht aus, nur die Gesamtanzahl der Beobachtungen zu beachten. Problematisch sind:

- Regressionen, aus deren Koeffizienten sich der unbedingte Mittelwert ergibt, wie beispielsweise bei einer Regression mit lediglich einer Dummy-Variablen oder einer Gruppe von Dummys, z.B. für die Bundesländer
- Regressionen mit zwei oder mehr Variablen in voll interagierten Modellen (alle möglichen Kombinationen der Merkmalsausprägungen).

Hier muss sichergestellt werden, dass die einzelnen Ausprägungen der dichotomen oder kategorialen Variablen mindestens 20 Beobachtungen beinhalten. Dies soll durch eine Auszählung der einzelnen Variablen nach dem Regressionsoutput unter Verwendung von `if e(sample)` gezeigt werden. Jede weitere Variable im Modell lässt keine Rückschlüsse mehr zu. Weitere Hinweise finden Sie im FDZ-Template `jd03_analyses.do`.

4.10 Ereignisdatenanalyse

Bei Anwendung des `sts list`-Befehls im Rahmen der Ereignisdatenanalyse, um sich die Ergebnisse des Kaplan-Meier-Schätzer auszuweisen, ist der Wert zu Beginn der Tabelle (population at risk) für die Datenschutzprüfung relevant. Darüber hinaus muss bei der

Ausweisung von Gruppen in Grafiken der `sts list`-Befehl ebenfalls für jede Gruppe erfolgen.

Bei Weitergabe von Ergebnissen aus Survival-Analysen muss lediglich der **Anfangswert >= 20** sein, nicht jeder einzelne Schritt.

4.11 Sequenzmusteranalyse

Bei der Sequenzmusteranalyse ist sicherzustellen, dass graphische Darstellungen auf einer ausreichend großen Anzahl von Beobachtungen basieren. Hierfür gibt es zwei zulässige Möglichkeiten. Unbedenklich ist es, Sequenzmuster pro Zeiteinheit als Anteile der Zustände anzugeben (z. B. im Monat 1 sind 30 % der Personen in Arbeitslosigkeit und 70 % in Erwerbstätigkeit). Des Weiteren ist es möglich, aggregierte Sequenzmuster für Gruppen darzustellen, wenn jede Gruppe mindestens 3 Beobachtungen beinhaltet. Sequenzmuster für weniger als 3 Individuen, Betriebe oder Unternehmen graphisch darzustellen ist nicht zulässig.