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

Empfehlungen zum Aufbau und zur Organisation von Auswertungsprogrammen anhand eines Beispiels in STATA

Beim Arbeiten mit dem IAB-Betriebspanel empfehlen wir grundsätzlich mit einer deskriptiven Analyse zu beginnen. Bitte bedenken Sie, dass Sie eine Sekundäranalyse durchführen und mit Testdaten arbeiten. Beim Erstellen der Auswertungsprogramme bleibt Ihnen daher der Zugriff auf die Originaldaten verschlossen. Damit Sie trotzdem vorab klären können, ob bestimmte Fragestellungen überhaupt mit Hilfe des IAB-Betriebsspanels bearbeitet werden können, stellen wir im Internet verschiedene Arbeitshilfen zur Verfügung. Verschaffen sie sich zunächst anhand des Codebuchs[1] einen Überblick über den Datensatz, die Variablen sowie die Häufigkeitsauszählungen und einige statistische Kennwerte der Befragungsergebnisse. Wenn Sie sich im Vorhinein so mit dem Datensatz vertraut machen, verringern Sie Ihren und unseren Arbeitsaufwand. Dadurch kann die Datenkontrolle schneller durchgeführt werden und Sie müssen auf Ihre Ergebnisse nicht lange warten.

Gehen Sie anschließend sukzessive anhand Ihrer Fragestellung vor. Beginnen Sie mit weiter gefassten Analysen, auch wenn Ihr Auswertungsinteresse tiefer gegliedert ist. Möchten Sie beispielsweise nur Betriebe eines bestimmten Bundeslandes analysieren, ist es trotzdem sinnvoll, zunächst die Zahlen für ganz Deutschland, dann für West- bzw. Ostdeutschland, und erst zum Schluss für das Bundesland zu ermitteln. Auf diese Weise kann Schritt für Schritt die Aussagekraft der Zahlen untersucht werden. Bitte beachten Sie dabei, dass Berechnungen, die auf zu kleinen Fallzahlen basieren nicht zu verwertbaren Ergebnissen führen, da Tabellenwerte < 20 (in Ausnahmefällen <4) von uns aus datenschutzrechtlichen Gründen gelöscht werden. Gleiches gilt für Zahlen, die Hinweise auf die tatsächlichen Verteilungen geben könnten. Kleine Fallzahlen können aber durch gezieltes Zusammenfassen mehrerer Zellen vermieden werden. Beispiele hierfür finden Sie unter Punkt 2. „Aufbereitung der Daten“.
Ein weiterer wichtiger Bestandteil der Syntax ist die Dokumentation. Die Notwendigkeit der Nachvollziehbarkeit von Auswertungsprogrammen zeigt sich vor allem bei Nachauswertungen und der Verwendung mehrerer Auswertungsprogramme. Hierzu gehört beispielsweise, dass Sie neu gebildete Variablen, die Sie für Ihre Auswertungen verwenden, mit Variablenlabels versehen.
Weitere formale Erfordernisse, die Sie bei der Auswertung der Daten beachten sollten, werden ausführlicher unter „Vorgaben des FDZ zur Nutzung von Datenfernverarbeitung und Gastaufenthalten“ http://doku.iab.de/fdz/access/Vorgaben_DAFE.PDF erläutert.

Als Hilfestellung beim Arbeiten mit den Daten des IAB-Betriebspanels wird im Folgenden das Vorgehen bei der Analyse eines fiktiven Auswertungsinteresses besprochen. Die einzelnen Schritte sind jeweils nummeriert, die dazugehörige, vollständige und lauffähige Syntax in STATA finden Sie am Ende des Dokuments.

Die zu beantwortende Frage lautet:

Wie verteilt sich die Summe der Teilzeitbeschäftigten auf Branchen und Betriebsgrößen im Jahr 2001?

Die Anfertigung eines möglichen Auswertungsprogramms kann in drei Hauptschritte untergliedert werden. Der gewünschte Datensatz wird eingelesen und die Auswertung auf die gültigen Querschnittsfälle begrenzt. Anschließend bereitet man die benötigten Variablen neu auf

[1] Das Codebuch wurde eingestellt. Die Fallzahlen zur aktuellen Welle befinden sich auf der FDZ-Homepage unter Arbeitshilfen.

(Rekodierungen) und lässt sich die Anzahl der gewichteten und ungewichteten Fälle ausgegeben. Erst nach diesen Schritten kann die eigentliche Auswertung durchgeführt werden.

1. Einlesen der Daten

Zunächst wird der Datensatz eingelesen. Anschließend wird ein Filter gesetzt bzw. nur die Querschnittfälle im Datensatz behalten um sicherzustellen, dass nur die Betriebe in die Querschnittsgewichtung einbezogen werden, die jeweils zum Stichtag mindestens einen sozialversicherungspflichtig Beschäftigten angestellt hatten. Anstelle des Hochrechnungsfaktors „hr2001q“ (1) besteht alternativ auch die Möglichkeit, das Kriterium „quer2001“ (2) für das Befragungsjahr 2001 heranzuziehen.

Da im Originaldatensatz keine MISSING VALUES definiert sind, müssen Sie die Angaben „weiß nicht“ (= -8) und „keine Angabe“ (= -9) selbst als Antwortausfälle (3) kennzeichnen. Die im Vorfeld von Ihnen definierten MISSING VALUES werden nicht in die statistischen Berechnungen einbezogen.

2. Aufbereitung der Daten

2.1. Brancheneinteilung

In einem weiteren Schritt werden die Variablen so umkodiert, dass sie sich zur Bearbeitung der Forschungsfrage eignen. Die ursprüngliche Einteilung der befragten Betriebe in 41 vorgegebene Branchen, ergibt in einigen Zeilen zu geringe Fallzahlen, die aus Datenschutzgründen gelöscht werden. Wir empfehlen daher durch Gruppieren eine neue Brancheneinteilung anhand der im Codebuch ausgewiesenen Häufigkeitsauszählungen vorzunehmen. In diesem Fall schlagen wir eine Neuner-Branchengliederung vor (4). Wahlweise können Dienstleistungsbetriebe in öffentlicher Hand auch dem öffentlichen Sektor zugeordnet werden (5).

2.2. Betriebsgrößenklasseneinteilung

Auch bei der Betriebsgrößenklassifizierung sollten geringe Fallzahlen vermieden werden. Neben der ursprünglichen Einteilung der Betriebe in zehn Betriebsgrößenklassen besteht beispielsweise auch die Möglichkeit einer Zuordnung zu sechs Betriebsgrößenklassen (6) oder einer Dreierklassifizierung (7).

2.3. Gewichtung der Fälle

Anschließend werden zum Vergleich Kreuztabellen der Variablen „branche“ und „betrgr1“ erstellt. Kreuztabellen ermöglichen es die Dimensionen Branche und Betriebsgröße in einer Tabelle darzustellen und die Anzahl der Betriebe in den einzelnen Zellen zu vergleichen.
Da die Häufigkeitsverhältnisse in der Stichprobe nicht denen in der Grundgesamtheit entsprechen, werden die Fälle gewichtet. Als Gewichtungsvariable wird der Hochrechnungsfaktor „hr2001q“ verwendet. Um nachvollziehen zu können, auf welchen Fallzahlen die gewichteten Werte beruhen, sollten die Kreuztabellen zur Kontrolle mit (8) und ohne Gewichtung (9) ausgegeben werden. Bei umfangreicheren Analysen empfehlen wir, die gewichteten und ungewichteten Befehle in getrennte Programmdateien zu schreiben.
Hochgerechnete Ergebnisse, die auf geringen Fallzahlen basieren, werden nicht ausgewiesen, sondern durch einen Schrägstrich „/“ gekennzeichnet. Die Genauigkeit von Hochrechnungen hängt von der Anzahl der zugrunde liegenden Fälle ab. Die Löschung von geringen Fallzahlen ist in den „Vorgaben des FDZ zur Nutzung von Datenfernverarbeitung und Gastaufenthalten“ beschrieben.

2.4 Aufbereitung der Teilzeit-Rohdaten

Falls Sie später weitere Berechnungen mit den gleichen Daten durchführen möchten, empfiehlt es sich, die Originalwerte der Variable i53tz (Anzahl der Teilzeitbeschäftigten) beizubehalten und mit einem anderen Variablennamen (z.B. teilzeit) weiterzuarbeiten.
Im nächsten Schritt wird, je nach Forschungsinteresse, darüber entschieden mit welcher Erhebungseinheit weitergearbeitet werden soll.
Wer nur Betriebe mit Teilzeitbeschäftigten in seine Berechnungen einbeziehen möchte, setzt zunächst die Summe aller Betriebe ohne Teilzeitbeschäftigte auf -9. Diese werden dann als MISSING VALUE behandelt (10).
Alternativ kann auch die Summer der Teilzeitbeschäftigten in allen Betrieben einer Branche/Betriebsklasse berechnet werden. Zuerst wird die ursprüngliche Variable i53tz rekodiert (z.B. in teilzeit2) und bei Betrieben ohne Teilzeitbeschäftigte, die Summe der Teilzeitbeschäftigten auf 0 gesetzt (11).
In jedem Fall ist darauf zu achten, dass man sich innerhalb einer Auswertung entweder auf Schritt 10 *oder* Schritt 11 festlegen muss.
Entscheidet man sich für die letztere der beiden Alternativen und bezieht auch Betriebe ohne Teilzeitbeschäftigte in die Berechnungen mit ein, kann aus den produzierten Ausgaben auch der Anteil der Betriebe mit Teilzeitbeschäftigten an allen Betrieben der jeweiligen Zelle Größenklasse*Branche bestimmt werden.

3. Auswertung

Nachdem alle Variablen modifiziert wurden, kann nun mit der eigentlichen Auswertung der Daten begonnen werden.
Zunächst wird die Querschnittsgewichtung mit dem Querschnittshochrechnungsfaktor „hr2001q“ wieder aktiviert (12) und anschließend die Verteilung der Summe der Teilzeitbeschäftigten auf Branchen und Betriebsgrößen anhand der Mittelwerte verglichen(13).

Ausgehend von unserer ursprünglichen Fragestellung sind noch weitere Auswertungen denkbar, indem man beispielsweise eine Aufteilung der Daten in Ost- und Westdeutschland (14), nach Bundesländern (15) und/oder für Betriebe mit und ohne Betriebsrat (16) vornimmt. Für diese Analysen müssen die Daten zunächst nach der oder den gewünschten Variablen sortiert und anschließend in mehrere Gruppen aufgeteilt werden.

Beispielsyntax STATA

```
clear
set mem 200m (alternativ Stata 12: set max mem 200m)
set matsize 800
set more off
cap log close

cd ..
use orig\iabbp_2001.dta
log using log\auswertung.log, replace


*(1) Querschnittsgewichtung

keep if hr2001q>0

*(2) Alternative zur Querschnittsgewichtung
```

```
keep if quer2001=="Q"

*(3) Definieren der MISSING VALUES.

mvdecode _all, mv(-9 -8)

*(4) Brancheneinteilung.

gen     branche = 1 if i92b==1 | i92b==2
replace branche = 2 if i92b>=3 & i92b<=18
replace branche = 3 if i92b==19 | i92b==20
replace branche = 4 if i92b>=21 & i92b<=23
replace branche = 5 if i92b==24 | i92b==25
replace branche = 6 if i92b==26 | i92b==27
replace branche = 7 if i92b>=28 & i92b<=32
replace branche = 8 if i92b>=33 & i92b<=38
replace branche = 9 if i92b>=39 & i92b!=.

label variable branche "9er Branchengliederung"
label define branche 1 "Land- und Forstwirtschaft; Bergbau/Energie/Wasser" ///
2 "Verarbeitendes Gewerbe" 3 "Baugewerbe" 4 "Handel und Reparatur" 5 "Verkehr/Nachrichten" ///
6 "Kredit/Versicherungen" 7 "Dienstleistungen überw. F. Unternehmen" ///
8 "sonstige Dienstleistungen" 9 "Öffentlicher Sektor"
label values branche branche

*(5) Zuordnung von Dienstleistungsbetrieben in öffentlicher Hand zum öffentlichen Sektor.

replace branche = 9 if branche >=5 & i80==5

*(6) Betriebsgrößenklasseneinteilung - 6er Klassifizierung.

gen     betrgr1 = 1 if i01ges01 >=1 & i01ges01<=9
replace betrgr1 = 2 if i01ges01 >=10 & i01ges01<=19
replace betrgr1 = 3 if i01ges01 >=20 & i01ges01<=49
replace betrgr1 = 4 if i01ges01 >=50 & i01ges01<=199
replace betrgr1 = 5 if i01ges01 >=200 & i01ges01<=499
replace betrgr1 = 6 if i01ges01 >=500 & i01ges01!=.

label variable betrgr1 "6 Betriebsgrößenklassen"
label define betrgr1 1 "1 bis 9 Beschäftigte" 2 "10 bis 19 Beschäftigte" 3 "20 bis 49 Beschäftigte" ///
4 "50 bis 199 Beschäftigte" 5 "200 bis 499 Beschäftigte" 6 "500 und mehr Beschäftigte"
labe value betrgr1 betrgr1

*(7) Alternative Betriebsgrößenklasseneinteilung - 3er Klassifizierung.

gen betrgr2 = 1 if i01ges01 >=1 & i01ges01<=19
replace betrgr2 = 2 if i01ges01 >=20 & i01ges01<=199
replace betrgr2 = 3 if i01ges01 >=200 & i01ges01!=.

label variable betrgr2 "3 Betriebsgrößenklassen"
label define betrgr2 1 "1 bis 19 Beschäftigte" 2 "20 bis 199 Beschäftigte" ///
3 "200 und mehr Beschäftigte"
label value betrgr2 betrgr2

*(8) Aktivieren der Gewichtung

tab branche betrgr1 [iweight=hr2001q]

*(9) Deaktivieren der Gewichtung.
```

```
tab branche betrgr1

*(10) Aufbereitung der Teilzeit-Rohdaten (Betriebe mit Teilzeitbeschäftigten).

gen teilzeit1 = i53tz
replace teilzeit1 = . if i53a==2

*(11) Alternative zur Aufbereitung der Teilzeit-Rohdaten (alle Betriebe).

gen teilzeit2 = i53tz
replace teilzeit2 = 0 if i53a==2

*(12) Aktivieren der Querschnittsgewichtung.

*(13) Mittelwertvergleich.

table betrgr1 [iweight=hr2001q], contents (n teilzeit1 sum teilzeit1)
table branche [iweight=hr2001q], contents (n teilzeit1 sum teilzeit1)

*(14) Aufteilung der Daten in West und Ostdeutschland.

table betrgr1 [iweight=hr2001q], contents (n teilzeit1 sum teilzeit1) by(wo2001)
table branche [iweight=hr2001q], contents (n teilzeit1 sum teilzeit1) by(wo2001)

*(15) Aufteilung der Daten nach Bundesländern.

bysort wo2001: table betrgr1  [iweight=hr2001q], contents (n teilzeit1 sum teilzeit1) by(bula2001)
bysort wo2001: table branche  [iweight=hr2001q], contents (n teilzeit1 sum teilzeit1) by(bula2001)

*(16) Aufteilung der Daten nach Betrieben mit bzw. ohne Betriebsrat.

bysort wo2001: table betrgr1  [iweight=hr2001q], contents (n teilzeit1 sum teilzeit1) by(i82)
bysort wo2001: table branche  [iweight=hr2001q], contents (n teilzeit1 sum teilzeit1) by(i82)

log close
```

Bitte beachten Sie, dass die Beispielsyntax nur eine Anregung für ein mögliches Auswertungsprogramm darstellt. Selbstverständlich sind für alle hier gezeigten Befehlszeilen und Beispiele auch andere Herangehensweisen denkbar.