Mikrodaten-Tools

Berechnung des Stichprobenfehlers im Mikrozensus mit SPSS Complex Samples

Bernhard Schimpl-Neimanns

Mai 2005

URL: http://www.gesis.org/Dauerbeobachtung/GML/Service/Mikrodaten-Tools/Varianz/VarMZ_CS.pdf

ZUMA
Quadrat B2,1
Postfach 12 21 55
68072 Mannheim
Telefon: 0621-1246-263
Telefax: 0621-1246-100
E-Mail: schimpl-neimanns@zuma-mannheim.de

Einleitung

Bei der Berechnung des Stichprobenfehlers für Ergebnisse des Mikrozensus ist zu beachten, dass die Daten nicht durch eine uneingeschränkte Zufallsauswahl zustande gekommen sind, sondern eine geschichtete Klumpenstichprobe bilden. Aus diesem Grund werden bei Verwendung der in den Statistikpaketen verfügbaren Standardverfahren, die von der Annahme einer uneingeschränkten Zufallsstichprobe ausgehen, die Stichprobenfehler nicht korrekt geschätzt.

Die Scientific Use Files des Mikrozensus enthalten ab dem Erhebungszeitpunkt 1996 Informationen über die Schichtung und Klumpung, die näherungsweise für die Berechnung des Stichprobenfehlers verwendet werden können. Für Auswertungen mit den in der empirischen Sozial- und Wirtschaftsforschung gängigsten Statistikpaketen SAS, SPSS und Stata stehen hierfür Programme für verschiedene Schätzer zur Verfügung (Schimpl-Neimanns/Rendtel 2001).[1] Seit der Version 12.0 bietet SPSS mit dem Zusatzmodul Complex Samples unter anderem Möglichkeiten, die Schichtung, Klumpung und Ziehungen mit bis zu drei Stufen sowie unterschiedliche Auswahlwahrscheinlichkeiten zu berücksichtigten. Damit können die in Rendtel/Schimpl-Neimanns (2001) beschriebenen Berechnungen komfortabler durchgeführt werden.

Im Folgenden wird anhand einfacher Beispiele und mit Daten des Mikrozensus-Scientific Use Files 2000 gezeigt, wie SPSS Complex Samples für Schätzungen des Stichprobenfehlers von Gesamtwerten (Totals) und Verhältniswerten (Ratios) eingesetzt werden kann. Hierbei werden nur Designgewichte, nicht aber die Anpassung der Mikrozensusergebnisse an die laufende Bevölkerungsfortschreibung berücksichtigt. Für diese so genannte gebundene Hochrechnung stehen allerdings keine Standardverfahren bereit. Wie dennoch SPSS Complex Samples bei Verwendung von Teilen des Programms VarMZ_A.SPS (Schimpl-Neimanns/Rendtel 2001) genutzt werden kann, den Stichprobenfehler für Gesamtwerte mit gebundener Hochrechnung zu schätzen, wird abschließend gezeigt. Zur Erläuterung der Annahmen, unter denen diese Schätzungen stehen, wird auf die oben genannten Aufsätze verwiesen. Dieser Bericht baut darauf auf und beschränkt sich auf die praktische Umsetzung mit SPSS Complex Samples. Im Anhang wird ergänzend zur SPSS-Syntax die menügesteuerte Vorgehensweise dargestellt.

1 Gesamtwerte (Totals)

Die folgenden Schätzungen konzentrieren sich als Beispiel auf die Zahl von Erwerbslosen (Y). Bei der Abgrenzung von Erwerbslosen und Erwerbstätigen werden die Definitionen der ILO zugrunde gelegt (Rengers 2004; Schmidt 2000). Die interessierende Subpopulation (Z) besteht aus den 15- bis 74-jährigen Erwerbspersonen (Erwerbstätige und Erwerbslose) am Ort der Hauptwohnung. Neben

[1] Siehe auch http://www.gesis.org/Dauerbeobachtung/GML/Service/Mikrodaten-Tools/Varianz/varianz_tools.htm

diesen Variablen werden im ersten Schritt der Variablenkonstruktion die Schichten und das Designgewicht (W) gebildet. Bei der Schichtungsvariablen werden der Einfachheit halber lediglich das Bundesland und die Gebäudeschicht berücksichtigt.[2]

```
title 'MZ-Stichprobenfehler mit SPSS Complex Samples: Totals'.

*  dieses Programm: CS_T.SPS .

get file 'mz2000.sav'
    /keep ef1 ef3 ef4 ef30 ef504 ef505 ef712.
weight off.
missing values all().

*  VARIABLENAUSWAHL .
*  EF1      Land der Bundesrepublik.
*  EF3      Auswahlbezirksnummer (Systemfrei).
*  EF4      Haushaltsnummer (Systemfreie Nr. des Haushalts im Auswahlbezirk).
*  EF30     Alter.
*  EF504    Erwerbstyp.
*  EF505    Bevölkerung am Hauptwohnsitz.
*  EF712    Gebäudegrößenklasse (Gebäudeschicht).

subtitle '1. Schritt: Variablenkonstruktion'.

*  SCHICHTUNG: Bundesland und Gebäudeschicht kombiniert.
*                 ohne Gemeindegrößenklasse (ef708) als Proxyvariable .
*                 für Regionalschichtung unterhalb Bundesland.
compute schicht=ef1*10+ef712.
variable label schicht 'Bundesland & Gebäudeschicht kombiniert (ef1, ef712)'.
formats schicht (f3).

*  DESIGNGEWICHT w: Kehrwert Auswahlwahrscheinlichkeit 70-Prozent-Substichprobe.
compute w=100/0.7.
variable label w 'Designgewicht 100/0,7'.

*  SUBPOPULATION z: Erwerbspersonen, Bevölkerung am Hauptwohnsitz, 15- bis 74-Jährige.
compute z=0.
if (ef504>=1 & ef504<=2 &          /* Erwerbspersonen (ILO-Def.)
   ef505>=1 & ef505<=2 &          /* Bevölkerung am Hauptwohnsitz
   ef30>=15 & ef30<=74) z=1.     /* 15-74-Jährige
variable label z "Subpopulation".
value labels z 1 'ILO-Erwerbspersonen'
               0 'Sonst'.
formats z (f1).

*  Y-Variable y: ILO-Erwerbslose .
recode ef504 (2=1) (else=0) into y.
variable label y "Y-Variable".
value labels y 1 'ILO-Erwerbslose'
               0 'Sonst'.
formats y (f1).
execute.
```

Nach diesen Vorarbeiten wird im zweiten Schritt mit der SPSS-Anweisung *CSPLAN ANALYSIS* der Stichprobenplan definiert, der den Schätzungen zugrunde gelegt werden soll. Zunächst wird das

[2] Bei zusätzlicher Verwendung der Gemeindegrößenklasse als Proxyvariable für die regionale Schichtung des Mikrozensus unterhalb der Ebene der Bundesländer kann es aufgrund dieser Differenzierung vorkommen, dass einige Schichten nur aus einem Auswahlbezirk bestehen. In diesen Fällen kann keine Varianz berechnet werden. Damit diese Auswahlbezirke nicht aus den Berechnungen ausgeschlossen werden, müssen sie mit anderen in möglichst ähnlichen Schichten zusammengefasst werden.

Designgewicht (W) für die Hochrechnung auf die Population bestimmt, das den Kehrwert der Ziehungswahrscheinlichkeiten der Stichprobeneinheiten beschreibt. Vereinfachend wird ein geschichtetes zweistufiges Auswahlverfahren mit einer uneingeschränkten Zufallsauswahl auf jeder Stufe angenommen. Von den Schichtungsmerkmalen der ersten Auswahlstufe werden, wie im ersten Schritt beschrieben, nur das Bundesland und die Gebäudeschicht verwendet. Die Klumpenidentifikation ist mit der Variablen EF3 Auswahlbezirksnummer (primary sampling unit; PSU) gegeben. Die Ziehungswahrscheinlichkeit der PSUs in der ersten Auswahlstufe beträgt ein Prozent. In Bezug auf die Auswahl der Haushalte in der zweiten Stufe wird von einem konstanten Auswahlsatz von 70 Prozent ausgegangen. Auf diese, in der Datei "STSI_SI.csaplan" abgelegten Definitionen kann für verschiedene Schätzungen zurückgegriffen werden.

```
subtitle '2. Schritt: Definition des Stichprobenplans'.

*  ANNAHMEN (Näherungen) .
*    Zweistufige geschichtete Zufallsstichprobe.
*    Schichtung: Bundesland (EF1) & Gebäudeschicht (EF712) -> SCHICHT.
*       1. Stufe: Primäreinheiten (PSU): Auswahlbezirke (EF3).
*       2. Stufe: Sekundäreinheiten: Haushalte (EF4).
*    Uneingeschränkte Zufallsstichprobe für jede Auswahlstufe.

CSPLAN ANALYSIS
   /PLAN FILE='STSI_SI.csaplan'
   /PLANVARS ANALYSISWEIGHT=w
   /PRINT PLAN
   /DESIGN STAGELABEL= 'PSU'  STRATA= schicht  CLUSTER= ef3
      /ESTIMATOR TYPE=EQUAL_WOR
      /INCLPROB VALUE=0.01
   /DESIGN STAGELABEL= 'HAUSHALT' CLUSTER= ef4
      /ESTIMATOR TYPE=EQUAL_WOR
      /INCLPROB VALUE=0.70.
```

```
subtitle '3. Schritt: Schätzung für Gesamtwerte (Totals)'.
CSDESCRIPTIVES
  /PLAN FILE = 'STSI_SI.csaplan'
  /SUMMARY VARIABLES =y
  /SUBPOP TABLE = z  DISPLAY=LAYERED
  /SUM
  /STATISTICS SE CV DEFFSQRT CIN (95)
  /MISSING SCOPE = ANALYSIS CLASSMISSING = EXCLUDE.
```

Schließlich wird im dritten Schritt mit der Anweisung *CSDESCRIPTIVES* die Schätzung der Zahl der Erwerbslosen (Y) angefordert. (Alternativ kann dies auch mit *CSTABULATE* erreicht werden.) Mit der Option */SUBPOP TABLE* werden die Ergebnisse für die Gruppen der Subpopulationen (Z) differenziert. Zusätzlich zum Gesamtwert (*/SUM*) werden mit */STATISTICS* die folgenden Kennziffern angefordert: Standardfehler (*SE*), Variationskoeffizient bzw. relativer Standardfehler (*CV*), Designeffektfaktor des Standardfehlers (*DEFFSQRT*) und ein 95 %-Konfidenzintervall zum Gesamtwert (*CIN*).

Unter den obigen Annahmen erhält man für die Subpopulation der Erwerbspersonen die unten stehenden Ergebnisse. Aufgrund der Stichprobengröße des Mikrozensus-Scientific Use Files beträgt der relative Standardfehler (CV) nur 0,8 Prozent (21,8 / 2.629,1). Wie der Designeffekt zeigt, ist der Standardfehler bei Berücksichtigung des Stichprobendesigns um 15 Prozent größer als bei Annahme einer uneingeschränkten Zufallsstichprobe. Bei Verwendung von Standardverfahren, die auf dieser Annahme einer uneingeschränkten Zufallsstichprobe basieren, würde das Konfidenzintervall entsprechend zu klein ausfallen.

Ergebnisse	
Erwerbslose (in 1.000)	2.629,1 4
Standardfehler (in 1.000)	21,8 4
Variationskoeffizient (in %)	0,83
Designeffekt	1,15
95 %-Konfidenzintervall (in 1.000)	
Untergrenze	2.586,3 4
Obergrenze	2.671,9 4

2 Verhältniswerte (Ratios)

In Anlehnung an das obige Beispiel wird bei der Schätzung der Erwerbslosenquote der Anteil der Erwerbslosen an den Erwerbspersonen berechnet. Das bereits definierte Y-Merkmal "Erwerbslose" bildet den Zähler und das Z-Merkmal "Erwerbspersonen" den Nenner des Verhältnisses. Die Variablenkonstruktion unterscheidet sich somit nicht von den Definitionen im vorigen Abschnitt. Sie wird lediglich um die Variable Bundesgebiet (X) ergänzt, um bei der Analyse die Erwerbslosenquoten zusätzlich für das westliche und östliche Bundesgebiet getrennt auszuwerten. Der in der Datei "STSI_SI.csaplan" bereits definierte Stichprobenplan kann übernommen werden.

```
title 'MZ-Stichprobenfehler mit SPSS Complex Samples: Ratios'.

*  dieses Programm: CS_R.SPS .

subtitle '1. Schritt: Variablenkonstruktion'.
[ (… ) siehe Abschnitt 1 ]

*   SUBPOPULATION X : Bundesgebiet (West/Ost).
recode ef1 (1 thru 10=1) (11 thru 16=2) into X.
variable label X 'Bundesland'.
value labels X 1 'Westliches Bundesgebiet'
                2 'Östliches Bundesgebiet (einschl. Berlin) '.
formats X (f1).
subtitle '2. Schritt: Definition des Stichprobenplans'.
[ (… ) siehe Abschnitt 1 ]
```

Für Verhältniswerte sind im Unterschied zur Schätzung von Gesamtwerten in der Anweisung *CSDESCRIPTIVES* lediglich die Zähler- und Nennermerkmale sowie die neue Subpopulation (X) zu bestimmen. Ansonsten werden alle anderen verwendeten Optionen (siehe oben) übernommen.

```
subtitle '3. Schritt: Schätzung für Verhältniswerte (Ratios)'.
* Zielgröße:   Anteil der Erwerbslosen an den Erwerbspersonen (ILO-Definition).
*    Zähler:    Erwerbslose (y).
*    Nenner:   Erwerbspersonen (z).
*    Subpop:   Bundesland (West/Ost) (X).

CSDESCRIPTIVES
  /PLAN FILE = 'STSI_SI.csaplan'
  /RATIO NUMERATOR = y  DENOMINATOR = z
  /STATISTICS SE CV DEFFSQRT CIN (95)
  /SUBPOP TABLE = X   DISPLAY=LAYERED
  /MISSING SCOPE = ANALYSIS CLASSMISSING = EXCLUDE.
```

		Bundesgebiet	
Ergebnisse	Insgesamt	West	Ost
Erwerbslosenquote (in %)	7,8	5,4	15,9
Standardfehler (in %)	0,06	0,06	0,18
Variationskoeffizient (in %)	0,80	1,1	1,1
Designeffekt	1,13	1,12	1,14
95 %-Konfidenzintervall (in %)			
Untergrenze	7,6	5,3	15,6
Obergrenze	7,9	5,5	16,3

Wie zu erwarten, unterscheiden sich die Erwerbslosenquoten zwischen West- und Ostdeutschland erheblich. Die durch den Variationskoeffizienten wiedergegebene Schätzgenauigkeit ist insgesamt mit 0,8 Prozent sehr hoch und unterscheidet sich nicht nach den Bundesgebieten. Aufgrund der Kovariation des Zähler- und Nennermerkmals liegt der Designeffekt (insgesamt: 1,13) etwas unter dem Wert für das Total (1,15; siehe oben).

3 Gesamtwerte mit gebundener Hochrechnung

Nach der im ersten Abschnitt vorgenommenen Schätzung gibt es im Mai 2000 2.629,1 tausend Erwerbslose. In den Veröffentlichungen des Statistischen Bundesamtes (2001: 30) werden dagegen 3.127 tausend „sofort verfügbare Erwerbslose“ berichtet. Beide Ergebnisse basieren auf der Umsetzung der ILO-Definition. Der Grund für die gegenüber der amtlichen Statistik um rund 16 Prozent niedrigere Schätzung besteht darin, dass in den obigen Auswertungen lediglich das Designgewicht verwendet wurde. Dagegen passen die statistischen Ämter die Mikrozensusergebnisse an Eckzahlen der laufenden Bevölkerungsfortschreibung an. Zusätzlich werden bei der Gewichtung Stichprobenausfälle von rund 2,5 Prozent ausgeglichen. Damit die auf Basis der Mikrozensus-Scientific Use Files geschätzten Werte (näherungsweise) mit den Ergebnissen der amtlichen Statistik übereinstimmen, müssen die entsprechenden Hochrechnungsfaktoren verwendet werden. Für die Varianzschätzung bei der so genannten gebundenen Hochrechnung stehen allerdings in SPSS keine Standardverfahren zur Verfügung.

```
title 'MZ-Stichprobenfehler mit SPSS Complex Samples: Totals bei gebundener Hochrechnung'.

*  dieses Programm: CS_A.SPS .

subtitle '1. Schritt: Variablenkonstruktion'.
[ (… ) zur Konstruktion von Y, Z, W, und X siehe Abschnitte 1 und 2 ]

*  Kopie der in VarMZ_A.SPS benötigten Variablen: .
compute y=y*z.              /* Y-Variable enthält Definition der Subpopulation z
compute psu=ef3.            /* Auswahlbezirk
compute hhnr=ef4.           /* Haushaltsnummer
compute soll_ist=ef750.  /* Personen-Hochrechnungsfaktor

* ---------- ab hier: Zeilen 83-147 aus SPSS-Programm "VarMZ_A.SPS" – Quelle: .
*  http://www.gesis.org/Dauerbeobachtung/GML/Service/ Mikrodaten-Tools/varianz/varmz_a.sps.
*  Kommentare zu Modifikationen in eckigen Klammern [ ].
* y_w:  mit Randanpassung gewichtete Beobachtung   [Zeile 83] .
* compute y_w=y*soll_ist.
* (...) .
* compute schicht = ef1*100 + ef708*10 + ef712. [ Zeile 111 ersetzen durch: ... ]
compute schicht=ef1*10+ef712.    /* [ in diesem Beispiel ohne Gemeindegrößenklasse (EF708) ]
* (...) .
save outfile 'mz_pers.sav'   /* [ Zeile 119; Zeile 120 ergänzen um: X, w, z, ef1, ef3, ef4 ]
     /keep schicht psu hhnr gruppe y y_w soll_ist  X w z ef1 ef3 ef4
     /compressed .
* (...) .
compute u=Soll_Ist * (y-B_dach).   /* [Zeile 147: Konstruktion Hilfsvariable u ]
* ---------- Ende Auszug aus SPSS-Programm "VarMZ_A.SPS" ----------.

subtitle '2. Schritt: Definition des Stichprobenplans'.
[ (… ) siehe Abschnitt 1 ]

subtitle 'Schritt 3.1: Schätzung des Totals bei gebundener Hochrechung (y_w)'.
temporary.
select if (z=1).                    /* Subpopulation: Erwerbspersonen
compute g=soll_ist*100/0.7.    /* gebundene Hochrechung (soll_ist = ef750) und Designgewichtung
weight by g.
crosstabs /y by X.

subtitle 'Schritt 3.2: Schätzung des Standardfehlers zum Total (y_w)'.
*   Beachte: Schätzung des Totals der Hilfsvariablen u ist irrelevant .
*              Variationskoeffizient und Konfidenzintervall müssen getrennt berechnet werden.
CSDESCRIPTIVES
  /PLAN FILE = 'STSI_SI.csaplan'
  /SUMMARY VARIABLES =u
  /SUBPOP TABLE = X  DISPLAY=LAYERED
  /SUM
  /STATISTICS SE
  /MISSING SCOPE = ANALYSIS CLASSMISSING = EXCLUDE.
```

Die gebundene Hochrechung der Gesamtwerte ist durch Verwendung der im Scientific Use File vorhandenen Hochrechnungsgewichte einfach zu erreichen (siehe Schritt 3.1). Für die Berechnung des Stichprobenfehlers der angepassten Gesamtwerte kann man ersatzweise das SPSS-Programm VarMZ_A.SPS verwenden. Hierfür benötigt man nur die Programmteile bis zur Berechnung der Hilfsvariablen *u*. Damit die Daten anschließend mit SPSS Complex Samples weiterverarbeitet werden können, sind nur geringfügige Änderungen notwendig.

		Bundesgebiet	
Ergebnisse	Insgesamt	West	Ost
Erwerbslose (in 1.000)	3.122,6	1.711,9	1.410,7
Standardfehler (in 1.000)	25,1	18,9	16,6
Variationskoeffizient (in %)	0,80	1,10	1,18
95 %-Konfidenzintervall (in 1.000)			
Untergrenze	3.073,4	1.674,2	1.378,2
Obergrenze	3.171,9	1.749,6	1.443,2

Durch die gebundene Hochrechnung kann die vom Statistischen Bundesamt veröffentlichte Zahl der Erwerbslosen praktisch repliziert werden; d. h., das 95 %-Konfidenzintervall deckt diesen Wert (3.127) ab. Für den Gesamtwert, der nur mit dem Designgewicht geschätzt wird (siehe Abschnitt 1), beträgt der Variationskoeffizient 0,83 Prozent. Durch die zusätzliche Berücksichtigung der gebundenen Hochrechnung in diesem Verfahren reduziert sich der Variationskoeffizient hier geringfügig auf 0,80 Prozent insgesamt.[3] Bei anderen Merkmalen, die enger mit den Anpassungsmerkmalen zusammenhängen, kann eine stärkere Verringerung des Stichprobenfehlers erreicht werden.

Zusammenfassend lässt sich in Bezug auf die gebundene Hochrechung festhalten, dass der Aufwand für diese nicht standardgemäß implementierten Schätzungen infolge der notwendigen Verwendung von Programmteilen aus VarMZ_A.SPS höher als bei den anderen Beispielen ist. Ist diese Hürde aber überwunden, können die anschließenden Schätzungen auf einfache Weise für Subgruppen differenziert werden.

Literatur

Rendtel, Ulrich, und *Bernhard Schimpl-Neimanns,* 2001: Die Berechnung der Varianz von Populationsschätzern im Scientific Use File des Mikrozensus ab 1996. ZUMA-Nachrichten 48: 85-116.

Rengers, Martina, 2004: Das international vereinbarte Labour-Force-Konzept. Wirtschaft und Statistik 12: 1369-1383.

Schimpl-Neimanns, Bernhard, und *Ulrich Rendtel,* 2001: SAS-, SPSS- und STATA-Programme zur Berechnung der Varianz von Populationsschätzern im Mikrozensus ab 1996. ZUMA-Methodenbericht 2001/04.

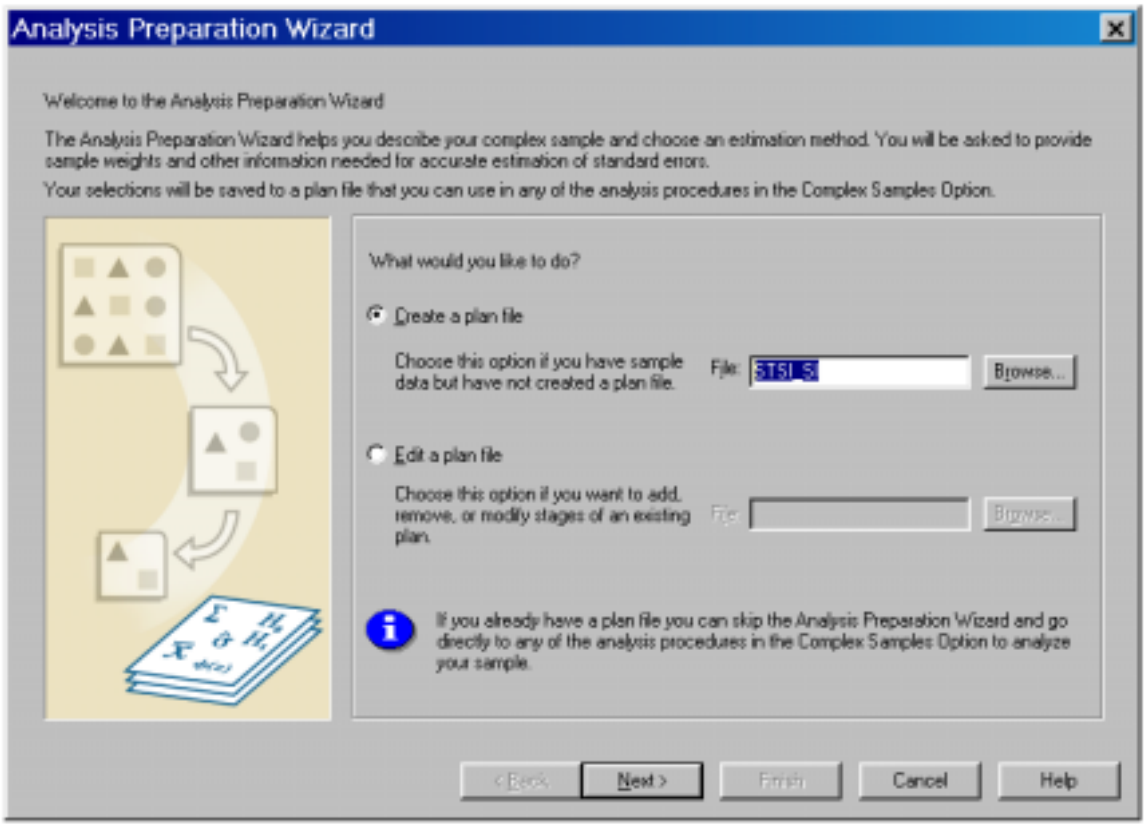
Schmidt, Simone, 2000: Erwerbstätigkeit im Mikrozensus. Konzepte, Definition, Umsetzung. ZUMA-Arbeitsbericht 2000/01.

Statistisches Bundesamt, 2001: Fachserie 1 Bevölkerung und Erwerbstätigkeit. Reihe 4.1.2 Beruf, Ausbildung und Arbeitsbedingungen der Erwerbstätigen 2000 (Ergebnisse des Mikrozensus). Stuttgart: Metzler-Poeschel.

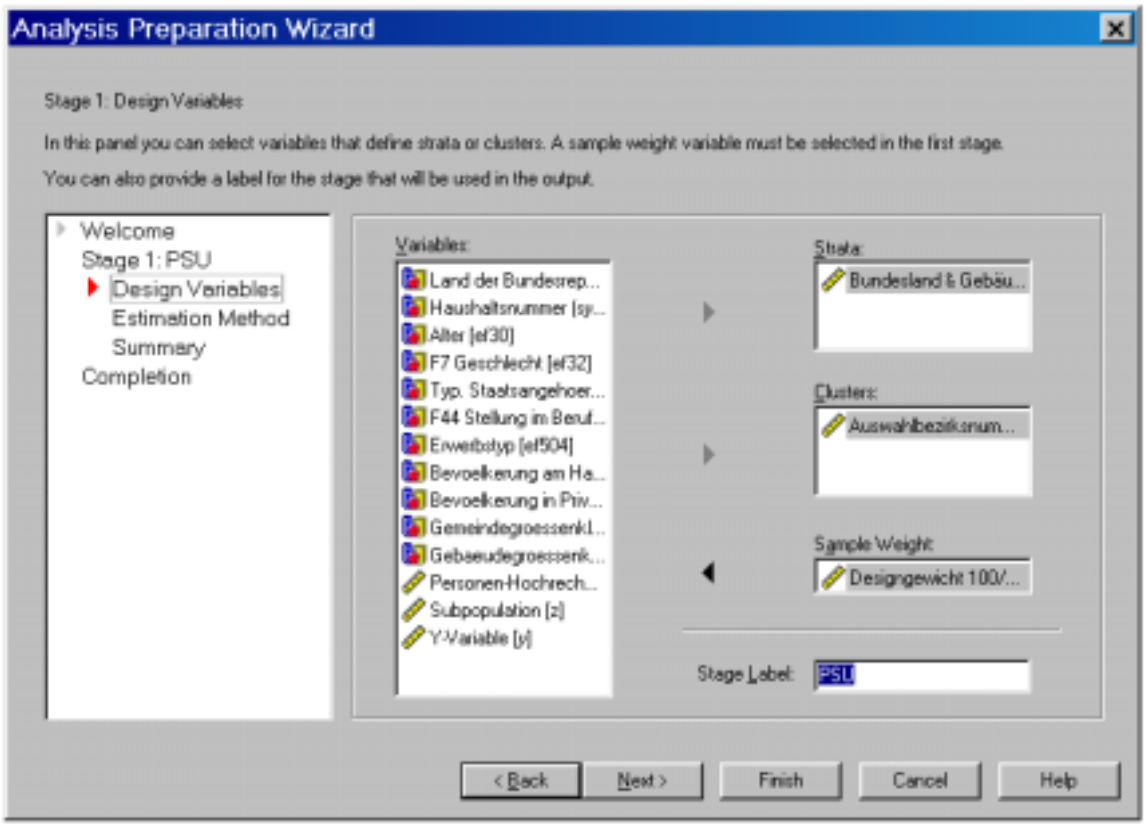
[3] Der Variationskoeffizient und das Konfidenzintervall müssen getrennt berechnet werden, da sich die Werte aus CSDESCRIPTIVES auf den Gesamtwert der Hilfsvariablen *u* beziehen.

Anhang: SPSS Complex Samples - Menüfenster

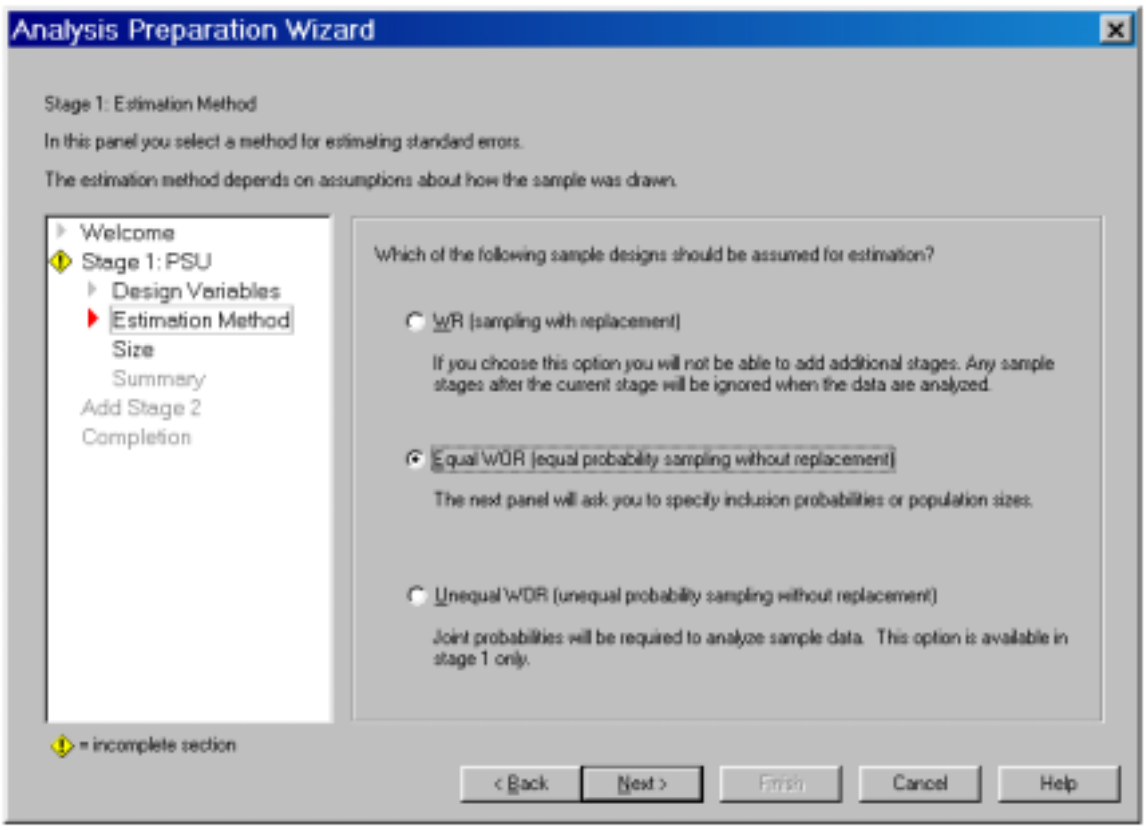
1.1 Definition des Stichprobenplans



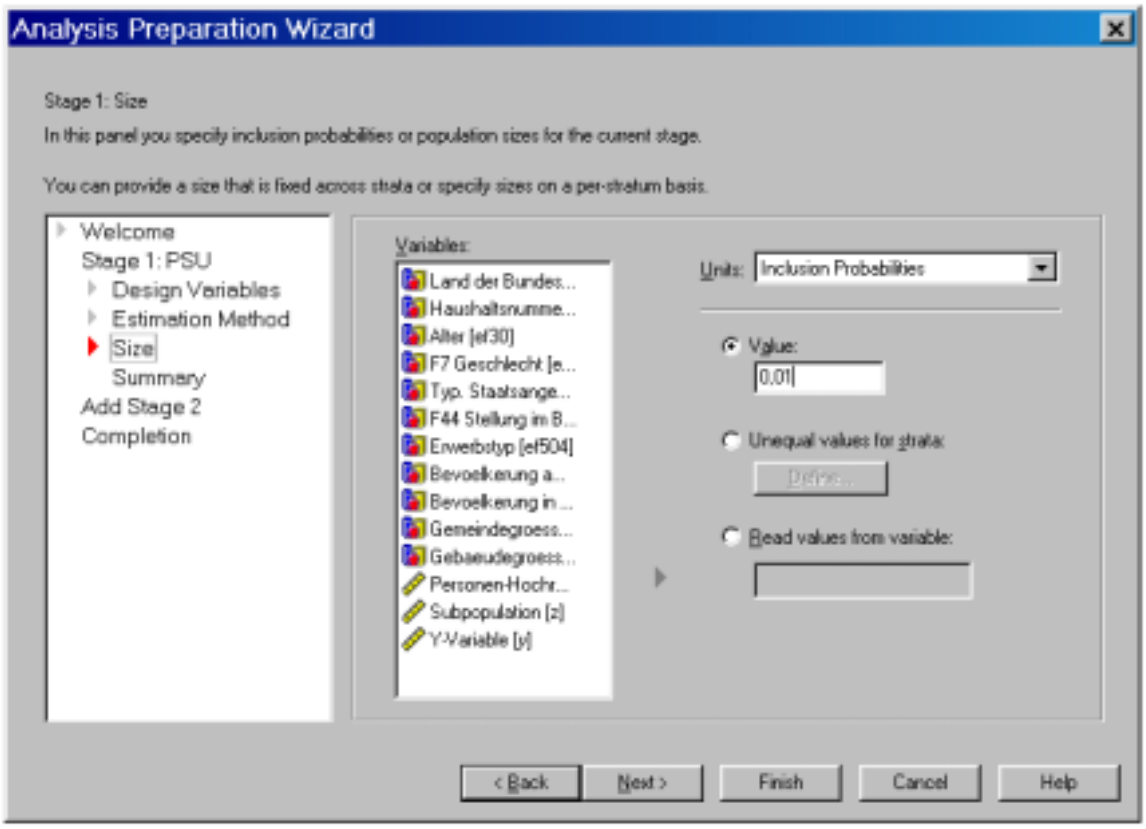
1.2 Definition des Stichprobenplans



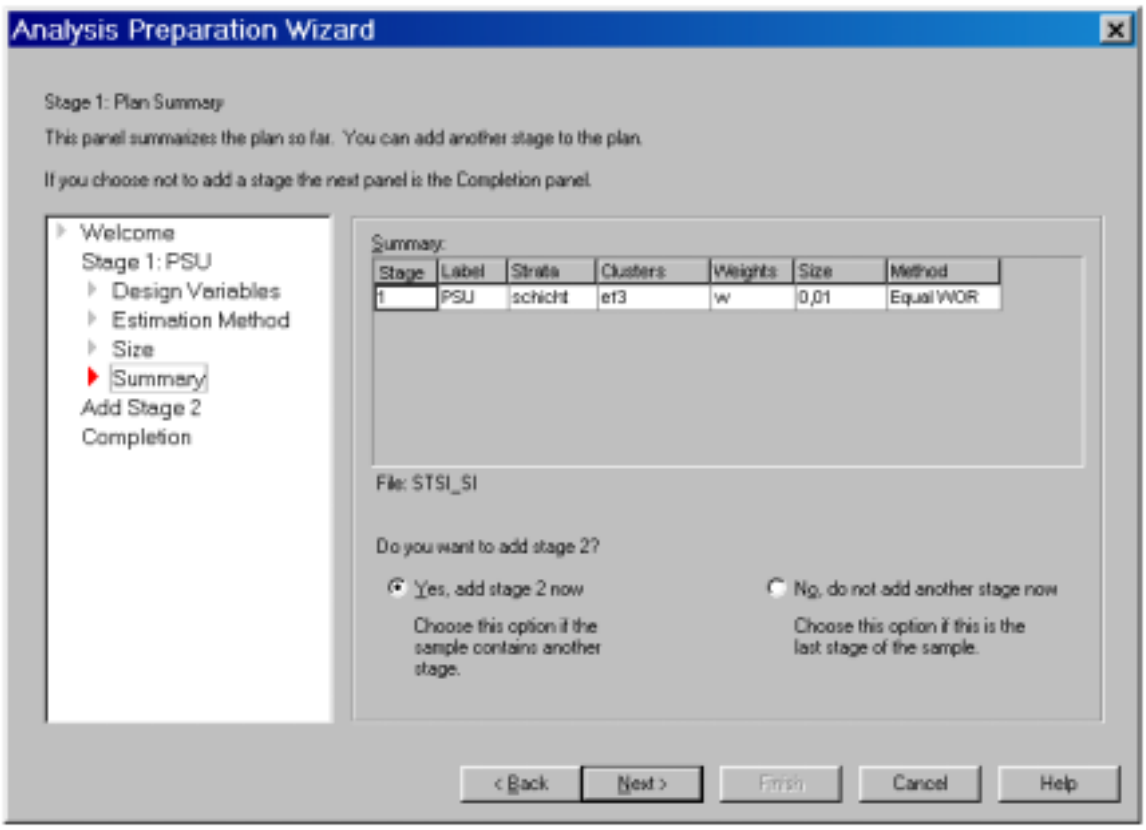
1.3 Definition des Stichprobenplans



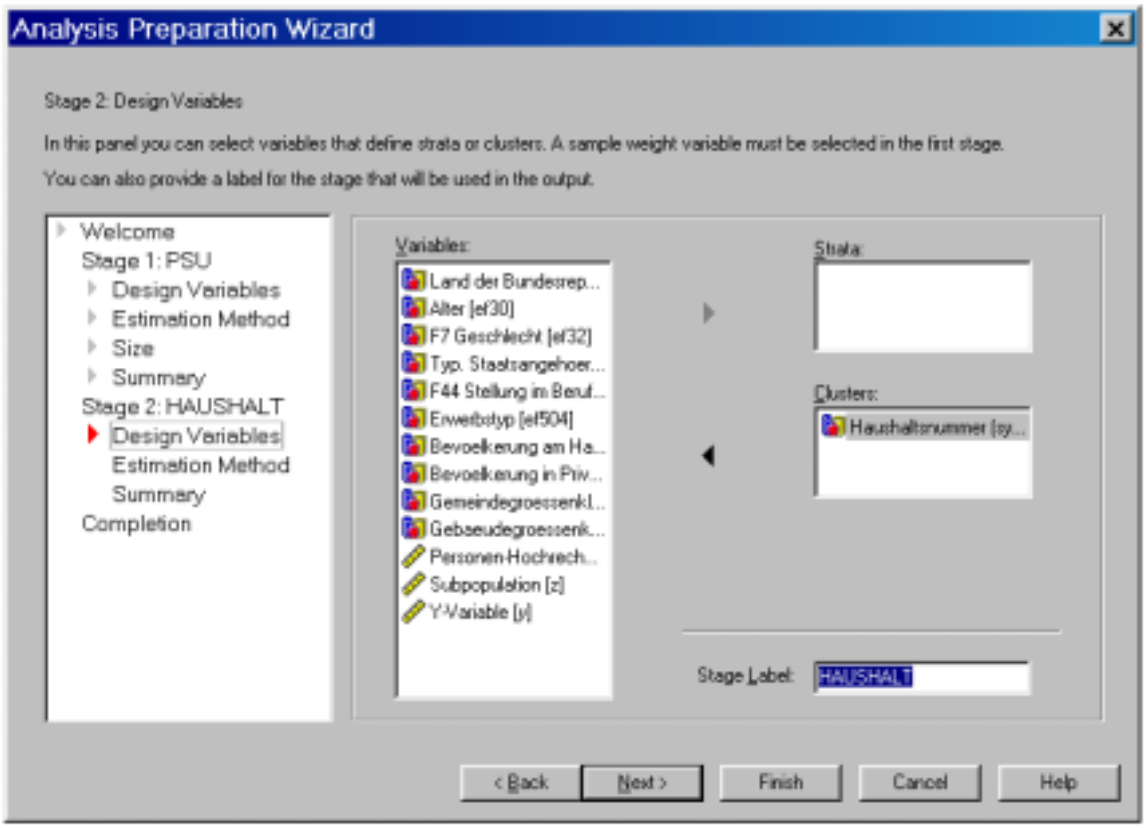
1.4 Definition des Stichprobenplans



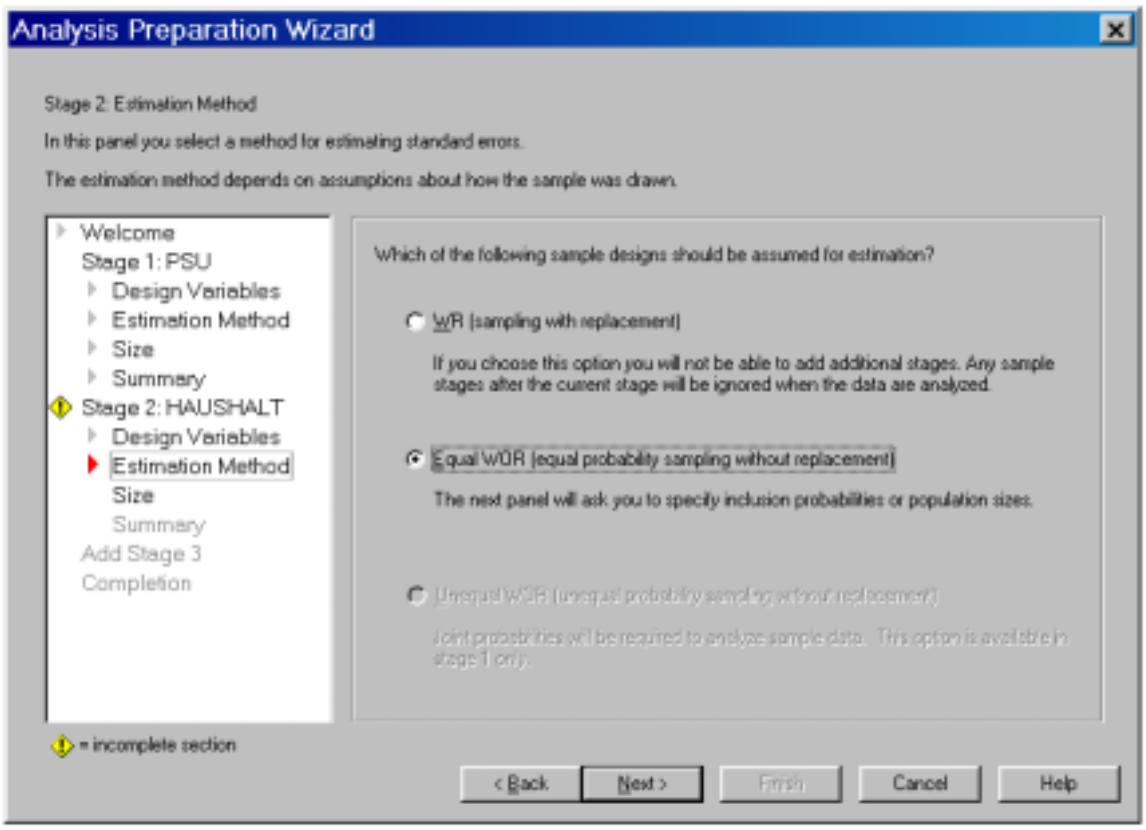
1.5 Definition des Stichprobenplans



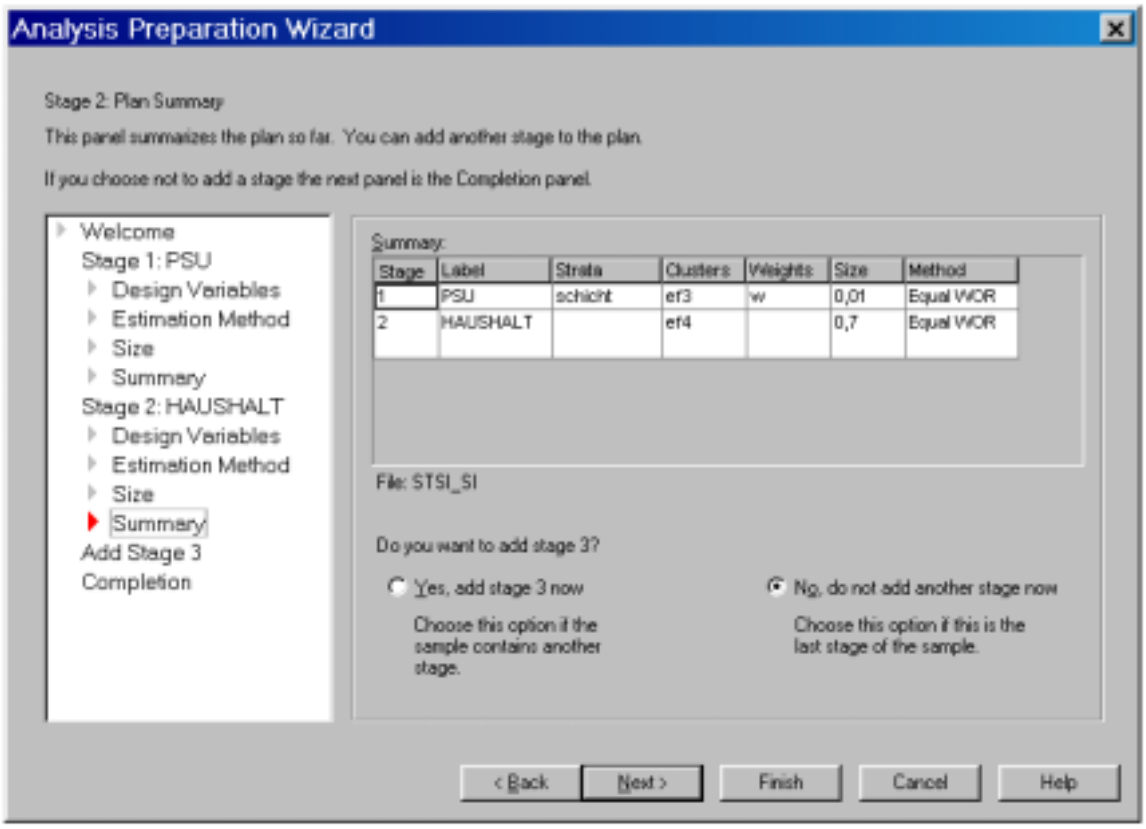
1.6 Definition des Stichprobenplans



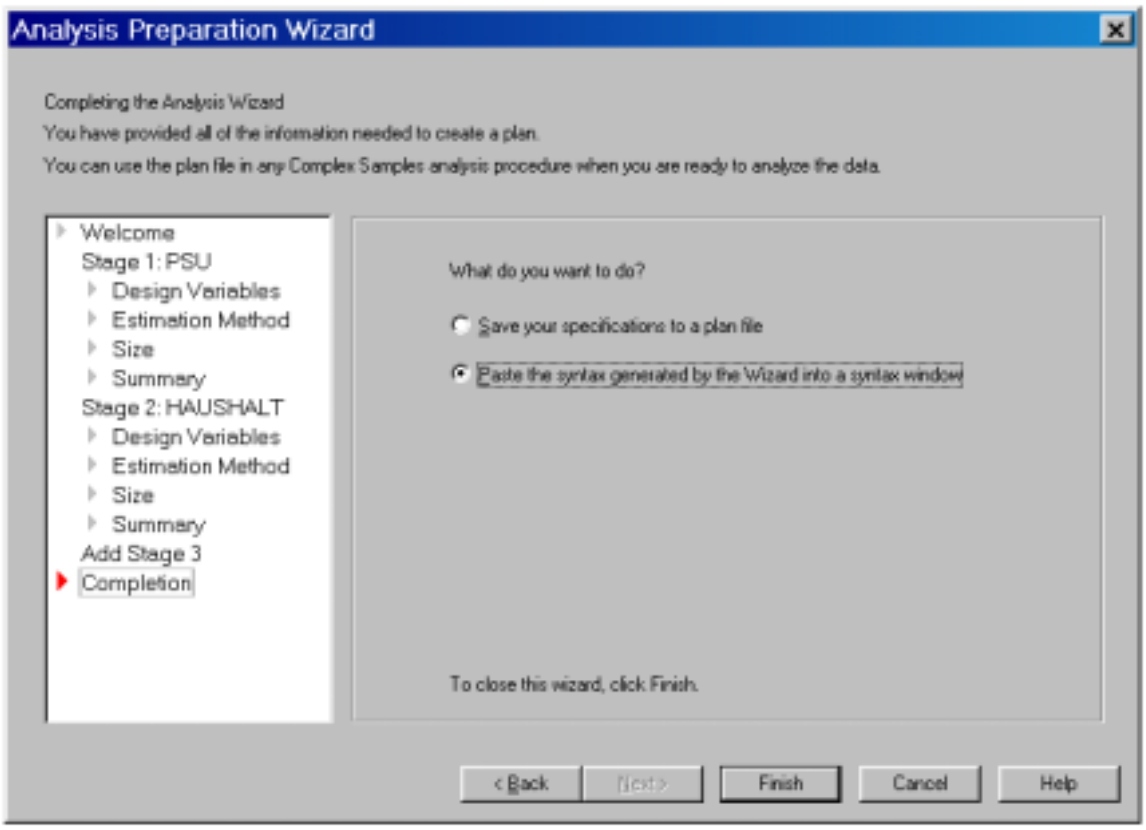
1.7 Definition des Stichprobenplans



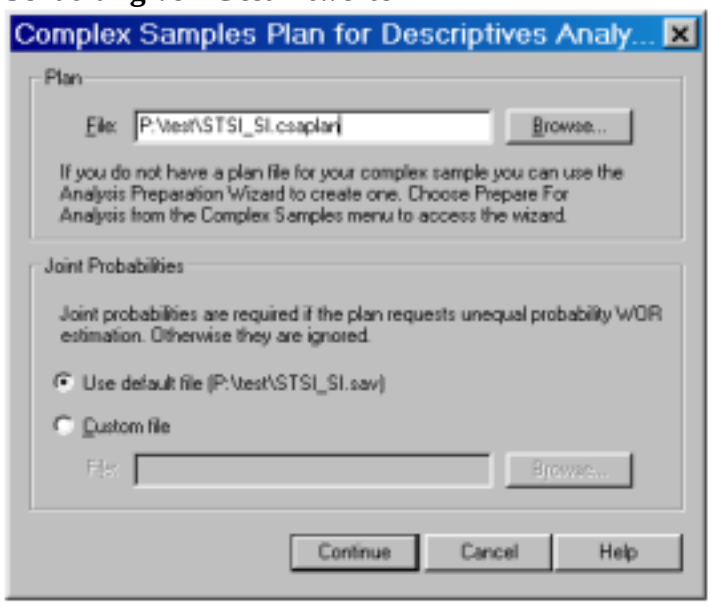
1.8 Definition des Stichprobenplans



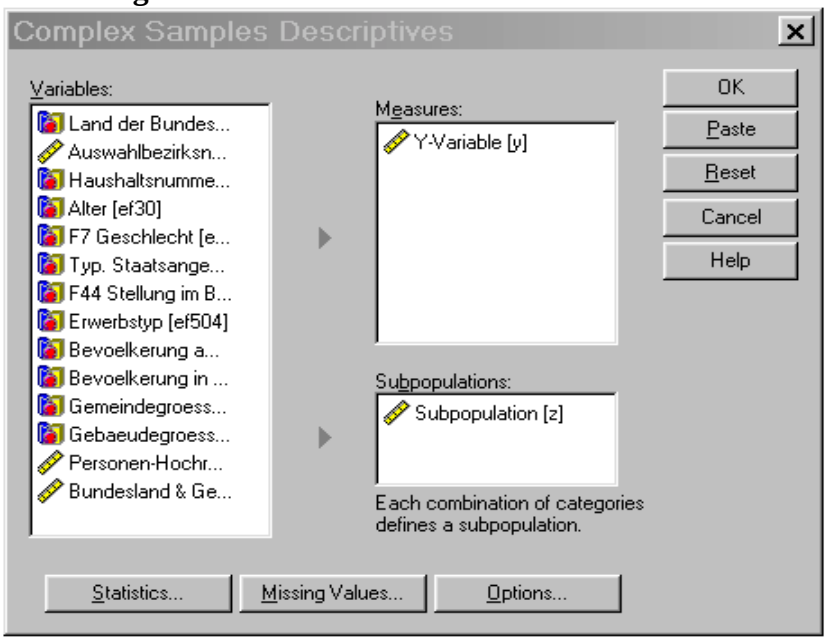
1.9 Definition des Stichprobenplans

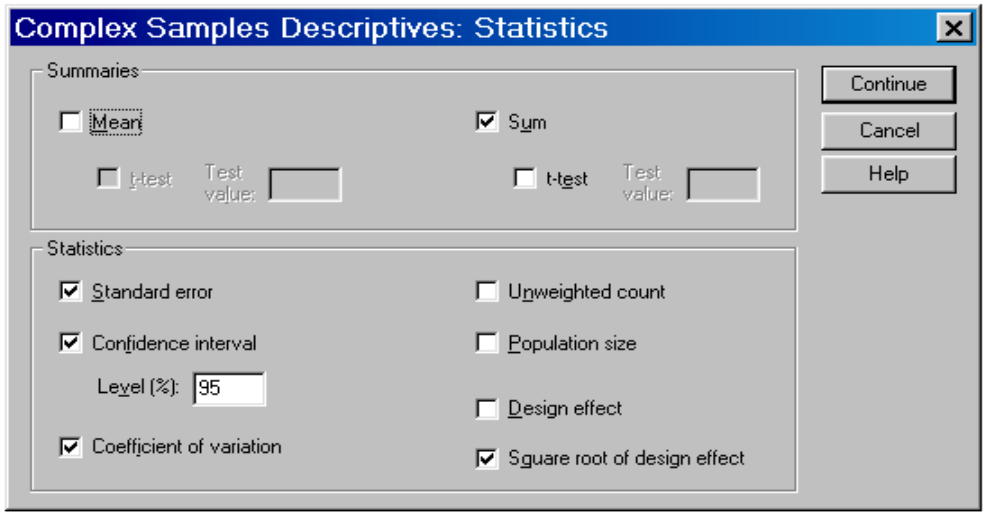


2.1 Schätzung von Gesamtwerten

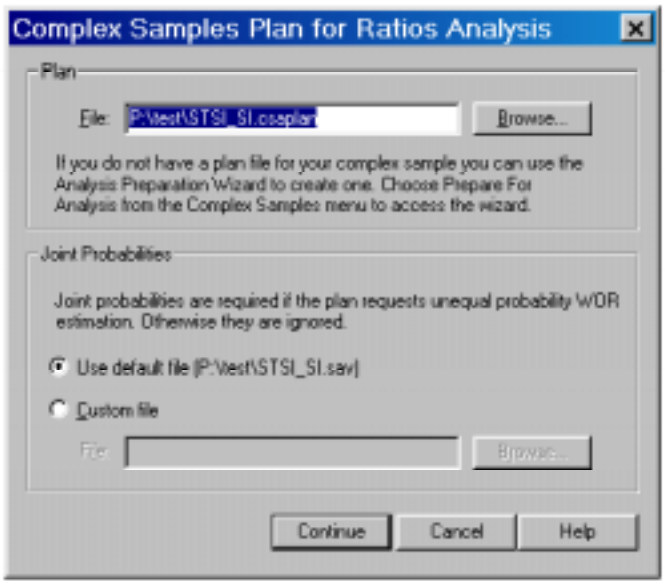


2.2 Schätzung von Gesamtwerten





3.1 Schätzung von Verhältniswerten



3.2 Schätzung von Verhältniswerten