Destatis
B24 9.07.2024

TPP: Planung und Ziehung der 5%-Stichprobe für das Tax-Payer-Panel 2020

Aus dem Tax-Payer-Panel der Jahre 2001 - 2020 wurde eine 5%-Stichprobe für die Wissenschaft gezogen. Stichprobengrundlage bilden die Datensätze des Tax-Payer-Panels, die x (x=zwanzig, neunzehn, achtzehn, siebzehn oder sechzehn, fünfzehn oder vierzehn, dreizehn oder zwölf, elf oder zehn, neun oder acht, sieben oder sechs, fünf) Jahre besetzt sind. Die Stichprobengrundlage besteht also aus neun verschiedenen, disjunkten Mengen: Grundgesamtheit1 (x=zwanzig), Grundgesamtheit2 (x=neunzehn oder achtzehn), Grundgesamtheit3 (x=siebzehn oder sechzehn), Grundgesamtheit4 (x=fünfzehn oder vierzehn), Grundgesamtheit5 (x=dreizehn oder zwölf), Grundgesamtheit6 (x=elf oder zehn), Grundgesamtheit7 (x=neun oder acht), Grundgesamtheit8 (x= sieben oder sechs), Grundgesamtheit9 (x=fünf).
Aus jeder erwähnten Menge wurde eine 5%-Stichprobe gezogen, damit jede Menge separat für Analysen und Auswertungen verwendet werden kann. Bei der Planung und Ziehung der Stichprobe, die als geschichtete Stichprobe mit disproportionaler Aufteilung angelegt ist, wurde bezüglich der Schichtung und des Aufteilungsverfahren wie folgt vorgegangen, d.h. das beschriebene Verfahren wurde auf jede Menge angewandt.

Schichtung: Um geographische Zugehörigkeit und grundlegende Charakteristiken von Elementen der Grundgesamtheit zu berücksichtigen, wurden folgende Schichtungsvariablen vorgesehen:

(1) Bundesland
(2) Grund- oder Splittingtabelle
(3) Überwiegende Einkunftsart (Gewinneinkünfte, Nichtselbstständige Arbeit, Übrige)

In allen drei Fällen wurde der häufigste Wert über die besetzten Jahren genommen (Modus). Wenn der häufigste Wert nicht eindeutig war, wurde der neuere Wert genommen.

Da erwartet wird, dass Forscher Auswertungen erstellen wollen, bei denen durch Kreuzkombinationen von (1), (2) und (3) definierte Teilgesamtheiten betrachtet werden, soll die Ergebnisqualität auf dieser Ebene kontrolliert werden. Solche Teilgesamtheiten werden im Folgenden auch als *Schichtgruppen* bezeichnet. Diese spielen beim Aufteilungsverfahren eine wichtige Rolle (siehe unten).

Ferner soll die Schichtung genutzt werden, um die Stichprobe bezüglich der Zielvariablen „Gesamtbetrag der Einkünfte“ (GdE) zu optimieren. Entsprechend wurde jede Schichtgruppe gemäß

(4) Median des Gesamtbetrages der Einkünfte über die besetzten Jahre von 2001 - 2020

in sechs Größenklassen eingeteilt (unter 0 Euro, 0-30.000 Euro, 30.000-50.000 Euro, 50.000-100.000 Euro, 100.000-150.000, über 150.000 Euro).
Da in der höchsten Einkunftsklasse, insbesondere in kleinen Bundesländern und im Osten, kleine Besetzungszahlen ein Problem waren, wurden hier Bundesländer nach West/Ost zusammengefasst. Damit ergeben sich für die Elemente mit durchschnittlichem Gesamtbetrag der Einkünfte über 150.000 Euro nur 2*2*3=12 Schichten (West/Ost * Grund/Splitting * Einkunftsart). Die Fusion der Schichten stellt methodisch kein Problem dar, da in der höchsten Einkunftsklasse, aus Gründen der Geheimhaltung, ein fester Auswahlsatz von 85% benutzt wurde. Damit spielen diese Teilpopulationen bei der optimierten Stichprobenaufteilung keine Rolle.
Insgesamt ergaben sich durch die Merkmale (1)-(4) und die eben dargelegten Modifikationen für: x = zwanzig: 478,x = neunzehn oder achtzehn: 479, x=siebzehn oder sechzehn: 481, x=fünfzehn

oder vierzehn: 486, x=dreizehn oder zwölf: 488, x=elf oder zehn: 496, x=neun oder acht: 500, x=sieben oder sechs: 487, x=fünf: 485 Schichten.

Bei der Optimierung der Stichprobe bezüglich der Zielvariablen „Gesamtbetrag der Einkünfte" lag es nahe, die Summe der Stichprobenvarianzen für die besetzten Jahre zu minimieren, da in den meisten Fällen die Einzelergebnisse von Interesse sind. Da bei der Schichtung nach dem *Median* des Gesamtbetrags der Einkünfte Elemente mit starken Schwankungen *zwischen den verschiedenen Jahren* nicht abgesondert werden, wurde in jeder Schicht (außer den 12 oben beschriebenen Schichten mit festem Auswahlsatz) eine zusätzliche Schichtung bzgl. eines relativen Variationsmaßes durchgeführt.
Für jedes Element ist dieses definiert als der Variationskoeffizient der GdEs der besetzten Jahre. Eine Ausnahme bildeten die Elemente mit durchschnittlichen Einkünften unter 15.000 Euro. Hier wird mit der festen Zahl 15.000 normiert, um durch kleine (meist durch Verluste in einzelnen Jahren hervorgerufene) Durchschnittseinkommen erzeugte riesige Variationskoeffizienten aufzufangen.
In Abhängigkeit der Besetzungszahl N der ursprünglichen Schicht wurden diese in weitere Schichten zerlegt:

(5) Feinschichtung nach der relativen Variation der GdE zwischen den Jahren:
   (a) N‹51: Keine weitere Zerlegung
   (b) 50‹N‹101: Zerlegung in zwei Teile am 60%-Quantil
   (c) 100‹N‹401: Zusätzliche Zerlegung am 90% Quantil
   (d) 400‹N‹1601: Zusätzliche Zerlegung am 96% Quantil
   (e) 1600‹N‹4001: Zusätzliche Zerlegung am 99% Quantil
   (f) 4000‹N‹16001: Zusätzliche Zerlegung am 99,6% Quantil
   (g) N›16000: Zusätzliche Zerlegung am 99,9% Quantil

Je nach Besetzungszahl wurden also bis zu sieben Unterschichten erzeugt. Durch diese feinere Zerlegung erhöhte sich die Gesamtzahl der Schichten auf insgesamt 16386 (x=zwanzig: 2217, x=neunzehn oder achtzehn: 1808, x=siebzehn oder sechzehn: 1691, x=fünfzehn oder vierzehn: 1709, x=dreizehn oder zwölf: 1752, x=elf oder zehn: 1769, x=neun oder acht: 1930, x=sieben oder sechs: 1837, x=fünf: 1673). Testrechnungen zeigten, dass die Einführung der feineren Schichtung in (5) zu einer deutlichen Verbesserung der Schätzergebnisse für die Zielvariable GdE in den einzelnen Jahren und insbesondere für deren Veränderung zwischen zwei aufeinanderfolgenden Jahren führt.

Die endgültige Zahl der Schichten ist 16386 (x=zwanzig: 2209, x=neunzehn oder achtzehn: 1797, x=siebzehn oder sechzehn: 1678, x=fünfzehn oder vierzehn: 1693, x=dreizehn oder zwölf: 1735, x=elf oder zehn: 1756, x=neun oder acht: 1920, x=sieben oder sechs: 1833, x=fünf: 1665), da Schichten mit Besetzungszahl 2 größeren Schichten angegliedert werden mussten. (Dies ist eine Folge von (4), nicht von (5)!)

Verteilung der Stichprobe: Die Verteilung der Stichprobe auf die Schichten wurde mittels eines Optimierungsverfahrens durchgeführt. Als Verlustfunktion wurde

$$\mathrm{F} = \sum_{SG} GesVar_{SG} \,/\, EinAbs_{SG}^{2(1-q)} \ ,$$

benutzt. Hierbei ist $GesVar_{SG} = Var_{SG,1} + \ldots + Var_{SG,6}$ die Summe der Stichprobenvarianzen der freien Hochrechnungen der totalen Einkünfte für die besetzten Jahre innerhalb einer Schichtgruppe SG. $EinAbs_{SG}$ ist die Summe der Absolutbeträge der Einkünfte innerhalb von SG und dient der Normierung der Varianz.
Optimale Stichprobenverteilungen werden durch Minimierung der Funktion F ermittelt. Es handelt hierbei im Wesentlichen um Power Allocations nach Bankier. Dabei wird die Konstante q zwischen Null und Eins so gewählt, dass ein verträglicher Kompromiss zwischen Bundesergebnissen ($q = 1$

Neyman Allokation) und einheitlichen Ergebnissen in den Schichtgruppen ($q = 0$, näherungsweise Minimierung der Euklidischen Norm der relativen Fehler in den Schichtgruppen) erzielt wird. Es gibt im vorliegenden Fall auch ein recht natürliches Kompromisskriterium, nämlich die Ergebnisse in den Bundesländern für alle Einkunft- und Tabellenarten zusammen, die für externe Forscher vermutlich auch von großer Bedeutung sind. Bei den vorliegenden Daten haben sich Werte zwischen $q = 0{,}3$ und $q = 0{,}5$ als geeignet herausgestellt. Letztlich wurde $q = 0{,}4$ verwendet.

Aus Geheimhaltungsgründen wird der in einer Schicht erlaubte maximale Auswahlsatz auf 85% beschränkt. Um allzu große Schwankungen der Hochrechnungsgewichte zu vermeiden, wurde ein minimaler Auswahlsatz von 2% (aber mindestens drei Elementen pro Schicht) zu Grunde gelegt. Damit ist der Quotient der Hochrechnungsgewichte zweier Elemente in der Stichprobe maximal (x=zwanzig: 57,01, x=neunzehn oder achtzehn: 55,87, x=siebzehn oder sechzehn: 57,14, x=fünfzehn oder vierzehn: 55,55, x=dreizehn oder zwölf: 57,47, x=elf oder zehn: 57,33, x=neun oder acht: 56,00, x=sieben oder sechs: 53,59, x=fünf: 58,00). (Rundungsbedingt können Ausnahmen auftreten. Dies passiert aber nur sehr selten.) Wie schon zuvor erwähnt, wurde der Auswahlsatz bei der größten Einkommensklasse fest auf 85% gesetzt. (Damit sind diese beim Optimierungsverfahren irrelevant.)

Die Verteilung der Stichprobe wurde durch Minimierung der Funktion F unter den gegebenen Beschränkungen der Auswahlsätze bei vorgegebenem festen Gesamtstichprobenumfang von 5% ermittelt. Die Rechnungen wurden mit dem im Statistischen Bundesamt programmierten SAS Macro OptAlloc durchgeführt.
Erwartungsgemäß führte die Einbeziehung der Ergebnisse auf Schichtgruppenebene in den Optimierungsprozess zu höheren Auswahlsätzen in kleinen Bundesländern als in großen Ländern. Damit wurden aber die gewünschten Annäherungen der Standardfehler auf Länderebene geliefert.

Ziehung der Stichprobe: Die Ziehung wurde mit der SAS Prozedur PROC SURVEYSELECT durchgeführt.

Hinweis: Bei den Grundgesamtheiten 4, 5, 6, 7 und 8 sind die Ausgaben für ein Jahr unbekannt und wurden deshalb imputiert. Für Bundesland, Grund- oder Splittingtabelle und die überwiegende Einkunftsart wurde der häufigste Wert über die besetzten Jahren genommen (Modus). Wenn der häufigste Wert nicht eindeutig war, wurde der neuere Wert genommen. Für den fehlenden Gesamtbetrag der Einkünfte wurde der Median des Gesamtbetrages der Einkünfte über die besetzten Jahre genommen.