

## Exercise 13. Accessing Census 2000 PUMS Data

**Purpose:** The goal of this exercise is to extract some 2000 PUMS data for Asian Indians for PUMAs within California. You may either download the records for all states or filter the selection to a particular state or attribute. For states, you use the STATEFIP variable to limit your selection. What we would like to determine is if there are any notable changes in occupations between men and women in several selected states. You can determine this by calculating the percent employed in each occupation and noting the major occupational niches.

### A. About PUMS

The *Public-Use Microdata Sample* is a collection of person and household records from the census of Population and Housing. This census file has become quite popular because it allows one to create custom tabulations.

The advantage of custom tabulation is somewhat offset by the limitations in geography. In 1990 and 2000 household records were tabulated by Public-use Microdata Area or PUMA. These units have a minimum of 100,000 persons. Another issue with PUMAs is that they often consist of disconnected areas. Thus, Glendale and San Fernando have been joined to form a PUMA. Apparently the designers tried to aggregate urban places into a PUMA before they would append adjoining rural or suburban space. The last 1990 PUMA in Los Angeles County is particularly poor - consisting of fragments from Signal Hill near Long Beach to Santa Clarita. For mapping purposes, there is a PUMA boundary file you may use to map tabulated variables.

PUMs data are available for a number of decades and these have been organized and integrated together by the Minnesota Population Center for easier access.

PUMs data may also be obtained in raw form from the Bureau of the Census. However, you will have to separate the housing and person records before making any tabulations.

More recent PUMs files consist of a sample of households and the persons in them. Usually these samples consist of a 1% national file and 5% state files, but a few other samples have been created such as a 0.1% national sample and a 3% sample for elderly persons. In 1980, estimates of the total population could be achieved by multiplying all records by a single factor, while in 1990 and 2000 each record must be multiplied by a weighting value.

A PUMs file consists of a housing record followed by person records for that household. The first person is the head of household, followed by the spouse, then children, and then others. The first column in each record identifies the record type with either an "H" or a "P" followed by the relevant data. The records are in text format and contain no delimiting characters. Thus one must be extra careful to properly specify field widths for each variable.

Because the housing and person records are nested together, one can not simply read a PUMs file into a spreadsheet and add the values in a column. The program must recognize each record type as it is input and possibly decide how to link household variables to the persons living there.

One common approach to simplifying the record processing is to link the housing data to each person in the household. This has been done by the *Minnesota Population Center*. The danger here is that you can not add housing variables to get a total since they have been repeated for every person. One can add housing data by selecting data only for persons who are heads of households. Another approach is to subset only household records from the raw PUMs data file.

**1. Log in to the IPUMS web site:**

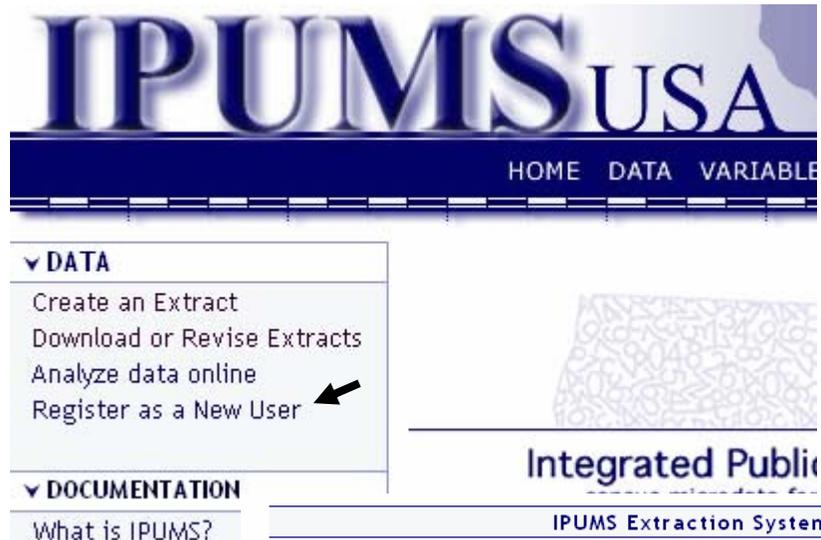
<http://www.ipums.umn.edu/>

**2. Select the *IPUMS-USA* link.**

In recent years *IPUMS* has also collected microdata for other countries and this could be a valuable resource for people who want to compare characteristics between the U.S. and other countries.

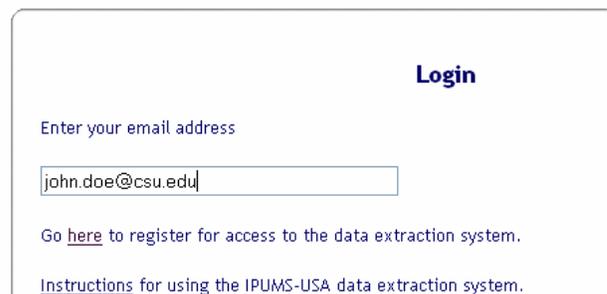


**3. On the PUMS-USA page look for the *Data* links and register as a new user.**



**Then select the *Create an Extracts* link.**

**4. When the *Data Extraction System* program starts, you will need to enter your email address as the job name. Then click *Login*.**



5. On the next page select the *Create New Extract* link. On the following page select the *Large* button and then click the *Continue to Sample Selection* button. We want the 5% sample.

IPUMS Extraction System
[logout](#)

### Create a New Extract

**Sample Density**
[help](#)

Tiny      Tiny samples contain approximately 5,000 adults and their households.

Small      Small samples contain approximately 50,000 adults and their households.

Regular      Most of the regular samples are 1% but the density of these samples varies. Refer to the [online documentation](#) for more information.

Large      The large samples are 5% samples and are available for 1980-2000.

Regular & Large      You can mix regular and large samples in the same extract

**Sample Selection List**
[help](#)

6. On the next *Sample Selection* page be sure to select the *2000 5% State sample* and then click the *Continue to Variable Selection* button.

**Samples**

1980 5% sample

1990 5% sample

2000 5% sample

Continue to Variable Selection

7. On the *Variable Selection* page are listed categories of household and person records. These are links to groups of variables that are listed below. Begin scrolling down this page.

You will note that some items are checked by default and the remainder must be selected as needed by you. Still others can be selected under *Case Selection* to limit the number of extracted records.

Technical Variables (Household) <span style="float: right;"><a href="#">top</a></span>					
Detailed Version	General Version	Variable	Label	Case Selection	2000
<input checked="" type="checkbox"/>		YEAR	Census year		x
<input type="checkbox"/>		<a href="#">DATANUM</a>	Data set number		x
<input type="checkbox"/>		<a href="#">SERIAL</a>	Household serial number		x
<input type="checkbox"/>		<a href="#">NUMPREC</a>	Number of person records following	<input type="checkbox"/>	x
<input type="checkbox"/>		<a href="#">SUBSAMP</a>	Subsample number	<input type="checkbox"/>	x
<input checked="" type="checkbox"/>		<a href="#">HHWT</a>	Household weight		x
<input type="checkbox"/>		All Technical Variables			

8. Under *Geographic Variables (Household)* click the *STATEFIP* and *PUMA* items. Note that each household item will be appended to any person within that household later. To limit the records to a particular state now, you should select the *Case Selection* button to the right of the *STATEFIP* item.

Geographic Variables (Household) <a href="#">top</a>					
Detailed Version	General Version	Variable	Label	Case Selection	2000
<input type="checkbox"/>		<a href="#">REGION</a>	Census region and division	<input type="checkbox"/>	x
<input type="checkbox"/>		<a href="#">STATEICP</a>	State (ICPSR code)	<input type="checkbox"/>	x
<input checked="" type="checkbox"/>		<a href="#">STATEFIP</a>	State (FIPS code)	<input checked="" type="checkbox"/>	x
<input type="checkbox"/>		<a href="#">METRO</a>	Metropolitan status	<input type="checkbox"/>	1, 5
<input type="checkbox"/>	<input type="checkbox"/>	<a href="#">METAREA</a>	Metropolitan area	<input type="checkbox"/>	1, 5
<input type="checkbox"/>	<input type="checkbox"/>	<a href="#">CITY</a>	City	<input type="checkbox"/>	1, 5
<input type="checkbox"/>		<a href="#">CITYPOP</a>	City population		1, 5
<input checked="" type="checkbox"/>		<a href="#">PUMA</a>	Public Use Microdata Area		5
<input type="checkbox"/>		<a href="#">PUMASUPP</a>	Supp. Public Use Microdata Area		1, 5

9. Scroll down to *Demographic Variables (Person)* and select the *RELATE* and the *SEX Detailed* buttons.

Demographic Variables (Person) <a href="#">top</a>					
Race, Ethnicity, and Nativity Variables (Person) <a href="#">top</a>					
Detailed Version	General Version	Variable	Label	Case Selection	2000
<input checked="" type="checkbox"/>	<input type="checkbox"/>	<a href="#">RACE</a>	Race	<input checked="" type="checkbox"/>	x
<input type="checkbox"/>	<input type="checkbox"/>	<a href="#">BPL</a>	Birthplace	<input type="checkbox"/>	x
<input type="checkbox"/>	<input type="checkbox"/>	<a href="#">ANCESTR1</a>	Ancestry, first response	<input type="checkbox"/>	x
<input type="checkbox"/>	<input type="checkbox"/>	<a href="#">ANCESTR2</a>	Ancestry, second response	<input type="checkbox"/>	x
<input type="checkbox"/>		<a href="#">CITIZEN</a>	Citizenship status	<input type="checkbox"/>	x

10. Under the *Race, Ethnicity, ...* category select the *Detailed Version* and the *Case Selection* buttons.

11. Under the *Work Variables (Person)* heading select the *Detailed Version* of the *Occupation* variable (*OCC*).

12. Under the *Income Variables (Person)* heading select the *Detailed Version* button of the *Total personal income (INCTOT)* variable. Later, if you wish, you can compare incomes for Asian Indian men and women and between occupations.

13. Go to the bottom of the *Variables Selection* page and click the *Continue* button.

**14. The Case Selection page will open.** This page allows you to limit the number of records according to one or more conditions. In this case we will limit the search to one state and to only persons of Asian Indian race.

**15. For this exercise we will look at Asian Indians in California. Click on 06 California under the State (FIPS code) window and then scroll down to 610 Asian Indian under the Race (Person) window and select it.**

**Then select the Continue to Extract Request Summary button.**

You will get a summary listing of the parameters set for this extract request.

**State (FIPS code) (Household)**

- 01 Alabama
- 02 Alaska
- 04 Arizona
- 05 Arkansas
- 06 California**
- 08 Colorado
- 09 Connecticut
- 10 Delaware
- 11 District of Columbia
- 12 Florida

**Race (Person)**

- 600 Filipino
- 610 Asian Indian (Hindu 1920-1940)**
- 620 Korean
- 630 Hawaiian
- 631 Hawaiian and Asian (1900,1920)
- 632 Hawaiian and European (1900,1920)
- 634 Hawaiian mixed

**15. If all seems OK, enter a brief description of why you are extracting this information. In this case it is “to examine the differences in occupations among Asian Indian men and women.”**

**Then click the Submit Extract Request button.**

**16. You will receive confirmation of your extract request.** You can monitor progress if you wish or wait about 15 minutes for an email notice to be sent to you.

**17. Your email confirmation will appear similar to that at right.**

**Confirmation**

Your extract request [eturner\\_csun\\_edu\\_032](#) has been submitted.

You will be notified by e-mail at [eturner@csun.edu](mailto:eturner@csun.edu) when it has been created.

When your extract is ready, [download](#) the data, codebook, and command files for your statistical software.

You may also return to the IPUMS-USA [homepage](#) or [create](#) a new extract.

Note: the extract will be available for 72 hours, after which time it will be deleted.

**If you want to monitor progress, click the download link shown right. Hit the refresh button from time to time in your browser to update the file listing.**

**18. Be sure to download the Data, Codebook, desired Command files to a working directory. Note that IPUMS provides commands in SPSS, SAS, and STATA forms. We will use SPSS for this exercise.**

IPUMS Extraction System [logout](#)

---

**Extracts for Download: eturner@csun.edu**

Click on the extract number in the leftmost column to revise that extract.  
[Instructions](#) for opening an IPUMS extract on your PC.

Request	Created	Description	Data	Codebook	Command
<a href="#">031</a>	May 14, 2007, 13:20:22	Looking at characteristics of people who left CA since 95	<a href="#">data</a>	<a href="#">codebook</a>	<a href="#">SPSS</a> <a href="#">SAS</a> <a href="#">STATA</a>

The *data* file will appear similar to that below. Note that without the field descriptions it is useless.

```
File Edit Format View Help
00008212400820281700006001001210982013200006000011012100980332200012000002010524035000006000
30274000060000092123009204312000360000020147030290000060000062123006203911000270000010100620
00000101045200000000600000711240071024210000800000101173101300000600000981210098027110003600
40072100012000001001000999999006000008612100860052100006000001001000999999006000003712100620
00111121009804211000240000010159030670000060000050121006102911000120000010104000230000060000
999999006000017212101840201200006000002011476000090000600001721210184019120000600000201149400
0000101131300000006000007212400720222200006000002011936003200006000011112101100712100029000
04316000060000042141110470000060000086121006203111000270000010106230280000060000086121004904
00991710123027220001200000201052400300000600000711240071023110003600000101136501400000600000
```

**19. Load the sps file into Word or a word processor**  
 The beginning of the SPSS program file appears right.  
 The *Data list file* command provides a description of all the items and their field locations. You will eventually run this file in SPSS to input your data for analysis. However, you first need to make a couple of adjustments.

```
*IPUMS USER: note that you'll need to enter the path to the
*folder containing your data file in the program line below
*these comments.
*
*For instance, if your file is in the 'C:' drive in a
*folder called IPUMS, the "file =" statement below these
*comments would read:
*
*       file = 'C:\IPUMS\eturner_csun_edu_031'
*
*Please alter the following lines accordingly.

data list file = '<insert path to folder containing data file
here>eturner_csun_edu_031.dat' /
year          1-2
hhwt          3-6
statefip      7-8
```

**20. Correctly set the path in the data list file command to the location of your data file and change the .dat to .txt. See below. To get the path, use Windows to locate the data file and then copy the path from the top of the window.**

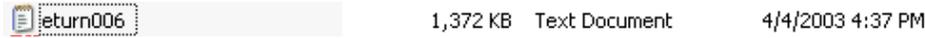
```
data list file = 'J:\MyFiles\CAAsianind\return006.txt' /
hhwt          1- 4
statefip      5- 6
```

Failure to get this correct will result in the following SPSS error message:

```
>Error # 31 in column 17. Text: J:\MyFiles\CAAsianind\return006.dat
>File not found.
>This command not executed.
```

21. You must open your data file (dat file) and resave it as a text file rather than an html file. In other words the file should have a .txt suffix.

Note that on some machines *Windows* has been set to suppress the file suffix in the listed name as is shown below.



22. If you would like to read this data into *SPSS*, proceed to the next exercise.

## B. IPUMS Documentation

On the main *IPUMS* page is a link to documentation on the data. If you would like to know more about *IPUMS* and *PUMS* data select the *What is the IPUMS* link. At right is the *Users Guide* web page.

1. Under the *Contents of this page* links select the *Subject Content* link. The browser will jump to a discussion of the item.

2. Within the *Subject Content* discussion locate the “*Variable Availability*.” link and select it.

You will get a list of Person and Household variables.

3. Under the *Person Record* column select the *Race, Ethnicity, and Nativity Variables*.

Variable Availability	
A B C D E F G H I K L M N O P Q R S T U V W Y	
<a href="#">Household Record</a> <a href="#">Technical Variables</a> <a href="#">Geographic Variables</a> <a href="#">Group Quarters Variables</a> <a href="#">Economic Characteristic Variables</a> <a href="#">Dwelling Characteristic Variables</a> <a href="#">Appliances, Mechanical, Other Variables</a> <a href="#">Constructed Household Variables</a> <a href="#">Historical Oversample Variables</a> <a href="#">Historical Technical Variables</a> <a href="#">1970 Neighborhood Variables</a>	<a href="#">Person Record</a> <a href="#">Technical Variables</a> <a href="#">Family Interrelationship Variables</a> <a href="#">Demographic Variables</a> <a href="#">Race, Ethnicity, and Nativity Variables</a> <a href="#">Education Variables</a> <a href="#">Work Variables</a> <a href="#">Income Variables</a> <a href="#">Migration Variables</a> <a href="#">Activity Five Years Ago Variables</a> <a href="#">Disability Variables</a> <a href="#">Veteran Status Variables</a> <a href="#">Place of Work and Travel Time Variables</a> <a href="#">Historical Oversample Variables</a> <a href="#">Historical Technical Variables</a> <a href="#">Other Variables</a>

A matrix of variables and the decades for which they are available will appear. Note that because of the American Community Survey that values appear by year starting with 2000.

**4. Locate the *BPL* variable (Birthplace) and select it.** The *BPL* web page explains the nature of this variable across the various censuses.

**5. At the top of the page is a link, *Codes and Frequencies*. Select it.**

A code list of places starting with states will appear. These may be important should you not have labels for the code numbers in a program.

The default listing is *Category Availability View*. However, you can change this to *Case-Count View* to see how many records are available for each place as is shown below.

General Codes											
Case counts are for the most commonly requested regular sample for each year.											
<input type="radio"/> Category Availability View <input checked="" type="radio"/> Case-Count View											
	Code	1850	1860	1870	1880	1900	1910	1920	1930	1940	1950
UNITED STATES											
Alabama	001	3263	4698	9659	13436	19977	32537	25839	15337	34695	53331
Alaska	002	1	•	•	16	6188	5425	377	28	74	162
Arizona	004	•	•	18	87	5722	10140	1342	1102	3152	6993
Arkansas	005	682	1512	2874	5173	11270	21836	16514	9578	23506	35175
California	006	47	724	1731	3523	10375	21882	13836	10364	26371	56417
Colorado	008	14	•	69	355	2114	5139	4886	3346	8189	13365
Connecticut	009	4107	4789	4800	5387	6846	10803	9122	5750	13177	18772
Delaware	010	1050	1122	1305	1547	1705	2701	2005	1171	2404	3560