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:

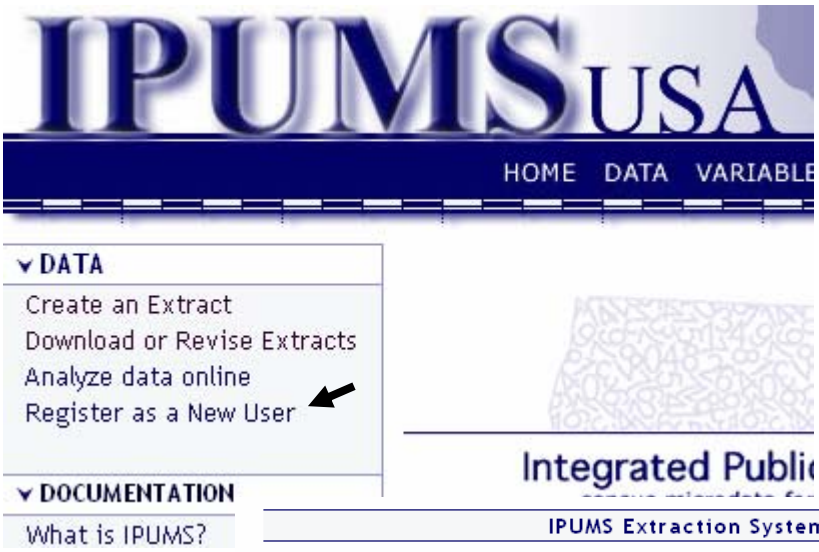
http://www.ipums.umn.edu/

2. Select the *IPUMS-USA* link.

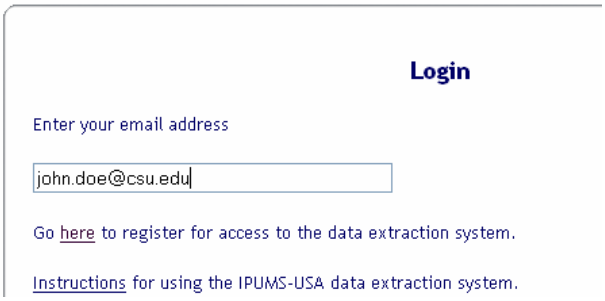
In recent years *IPUMS* has also collected microdata for other countries and this could be a valuable resource for people who want to compare characteristics between the U.S. and other countries.

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

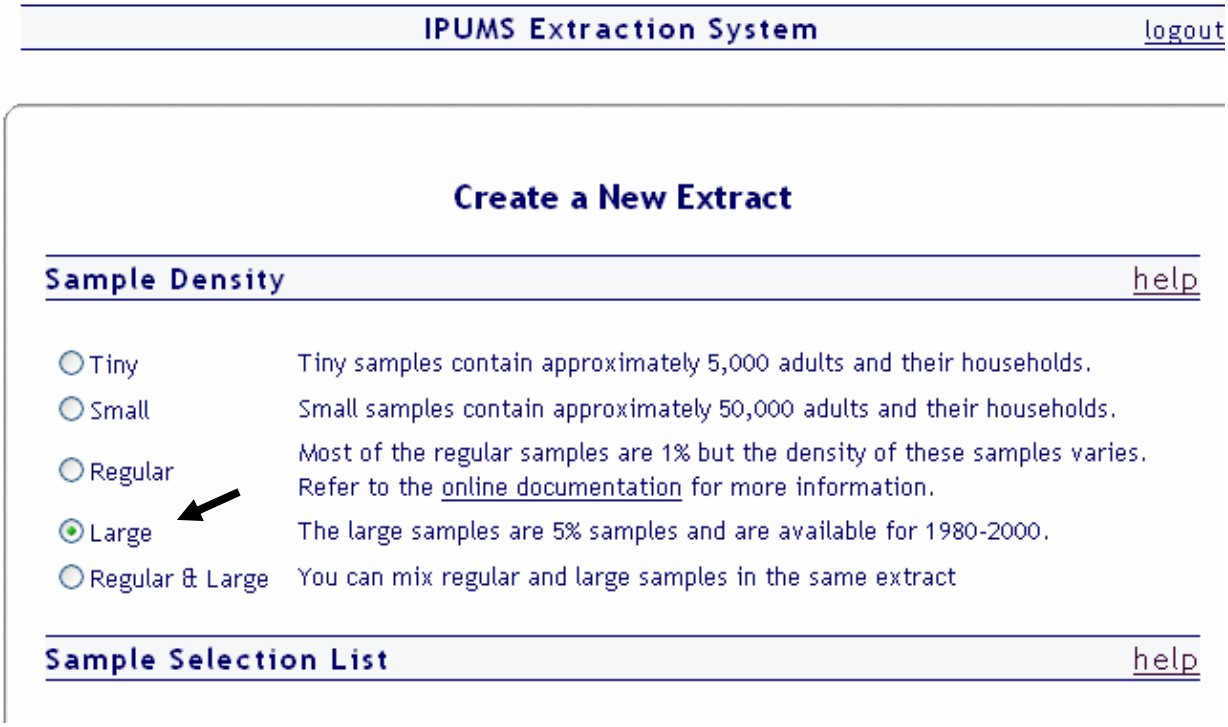
Then select the *Create an Extracs* link.



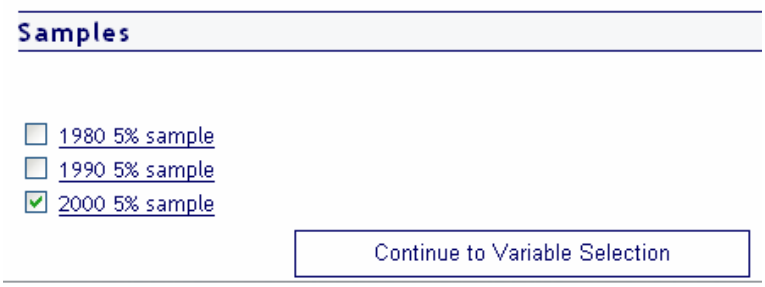
4. When the *Data Extraction System* program starts, you will need to enter your email address as the job name. Then click *Login*.



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.

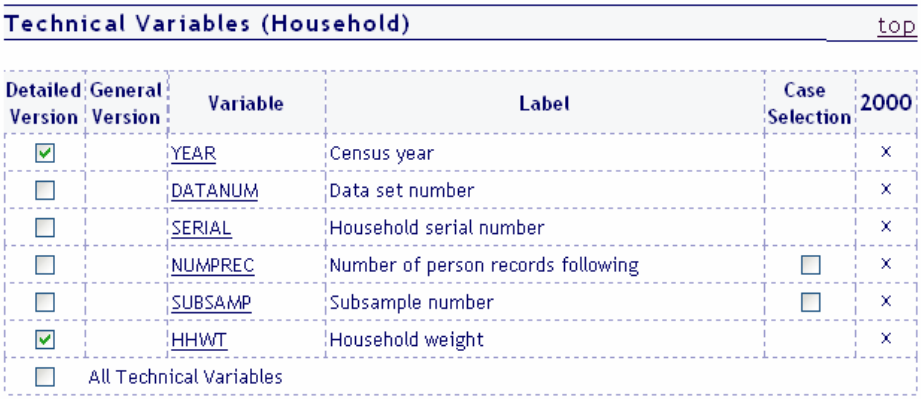


6. On the next *Sample Selection* page be sure to select the *2000 5% State sample* and then click the *Continue to Variable Selection* button.

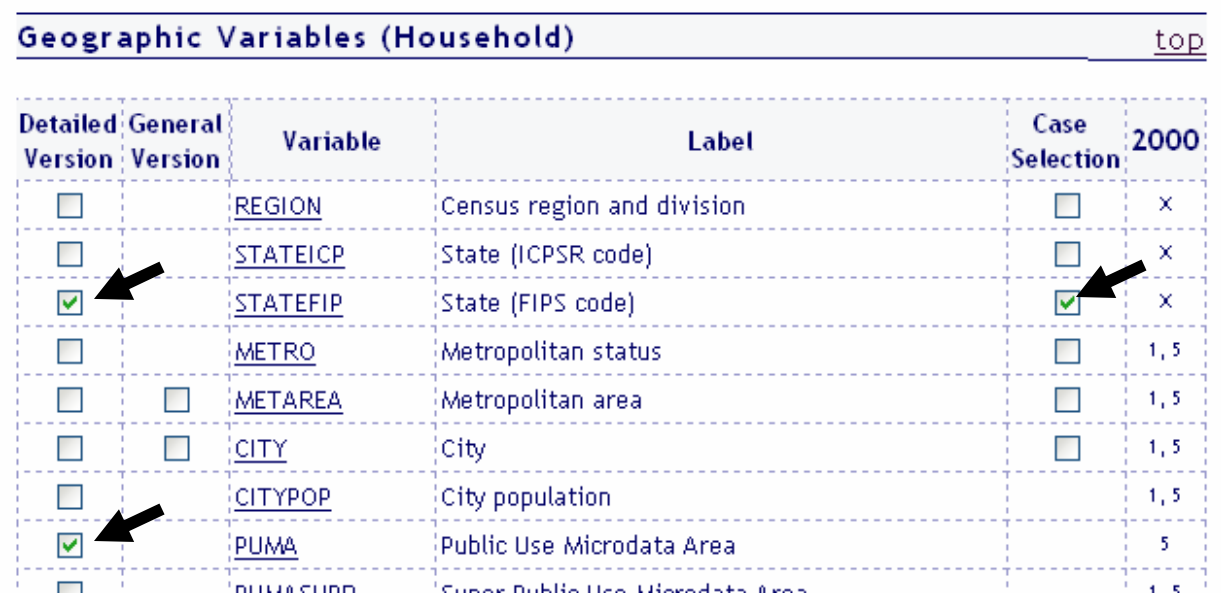


7. On the *Variable Selection* page are listed categories of household and person records. These are links to groups of variables that are listed below. Begin scrolling down this page.

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.



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.

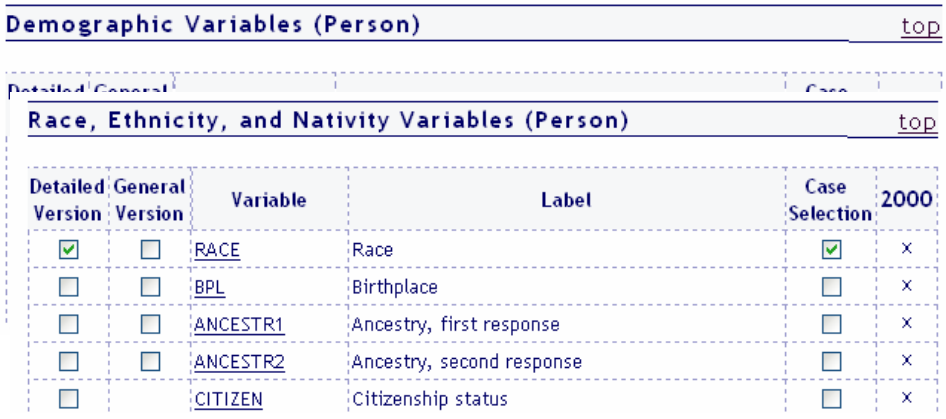


Geographic Variables (Household) top

Detailed Version	General Version	Variable	Label	Case Selection	2000
☐		REGION	Census region and division	☐	x
☐		STATEICP	State (ICPSR code)	☐	x
☑		STATEFIP	State (FIPS code)	☑	x
☐		METRO	Metropolitan status	☐	1, 5
☐	☐	METAREA	Metropolitan area	☐	1, 5
☐	☐	CITY	City	☐	1, 5
☐		CITYPOP	City population		1, 5
☑		PUMA	Public Use Microdata Area		5

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

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



Demographic Variables (Person) top

Race, Ethnicity, and Nativity Variables (Person) top

Detailed Version	General Version	Variable	Label	Case Selection	2000
☑	☐	RACE	Race	☑	x
☐	☐	BPL	Birthplace	☐	x
☐	☐	ANCESTR1	Ancestry, first response	☐	x
☐	☐	ANCESTR2	Ancestry, second response	☐	x
☐		CITIZEN	Citizenship status	☐	x

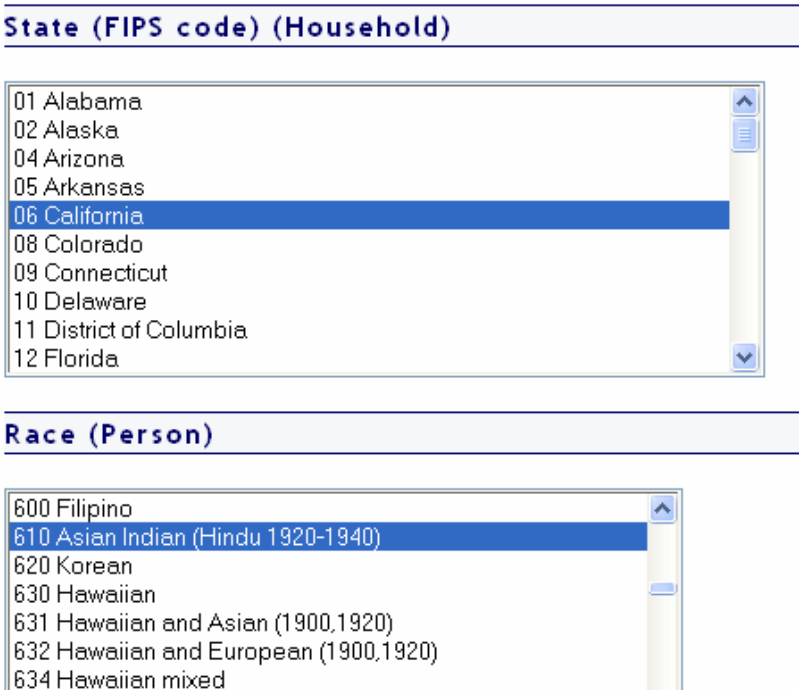
11. Under the *Work Variables (Person) heading* select the *Detailed Version* of the Occupation variable (*OCC*).

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.

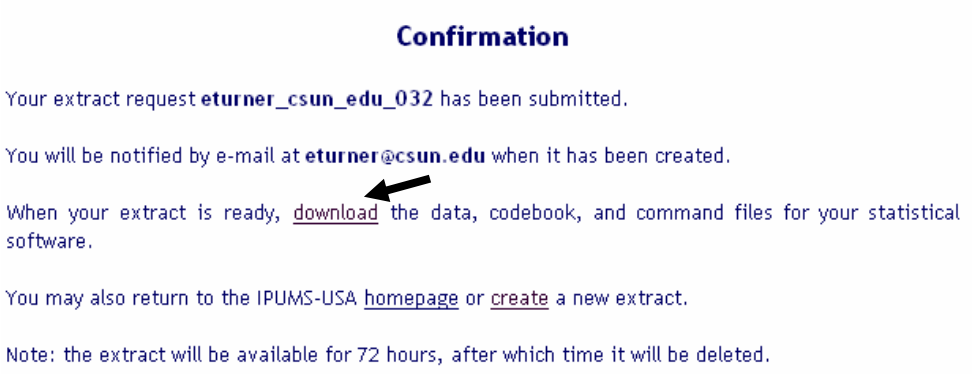
You will get a summary listing of the parameters set for this extract request.

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.

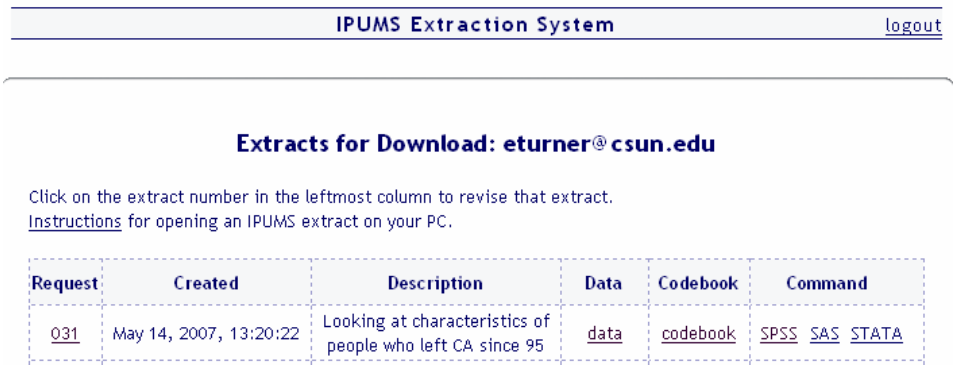
16. You will receive confirmation of your extract request. You can monitor progress if you wish or wait about 15 minutes for an email notice to be sent to you.

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

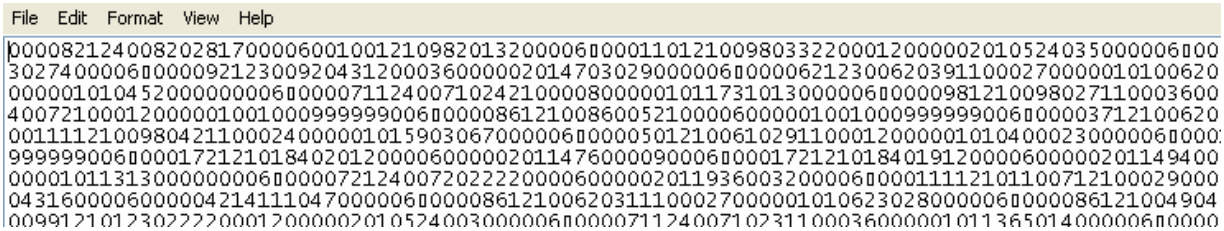


If you want to monitor progress, click the download link shown right. Hit the refresh button from time to time in your browser to update the file listing.

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.



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



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.

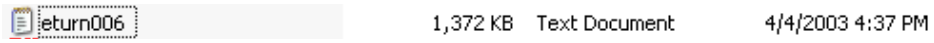
```
data list file ='J:\MyFiles\CAAsianind\eturn006.txt'/
 hhwt      1- 4
 statefip  5- 6
```

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

```
>Error # 31 in column 17.  Text: J:\MyFiles\CAAsianind\eturn006.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.

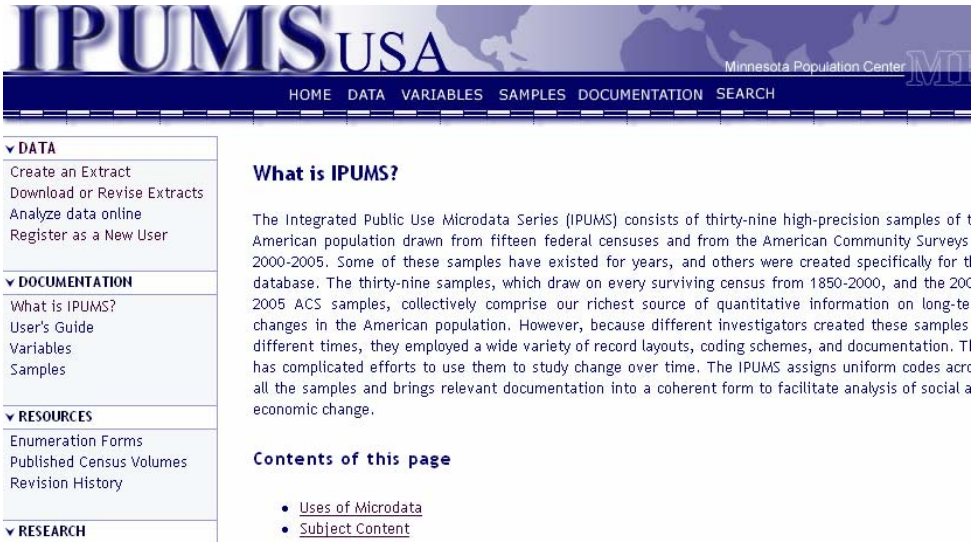
Note that on some machines *Windows* has been set to suppress the file suffix in the listed name as is shown below.



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

B. IPUMS Documentation

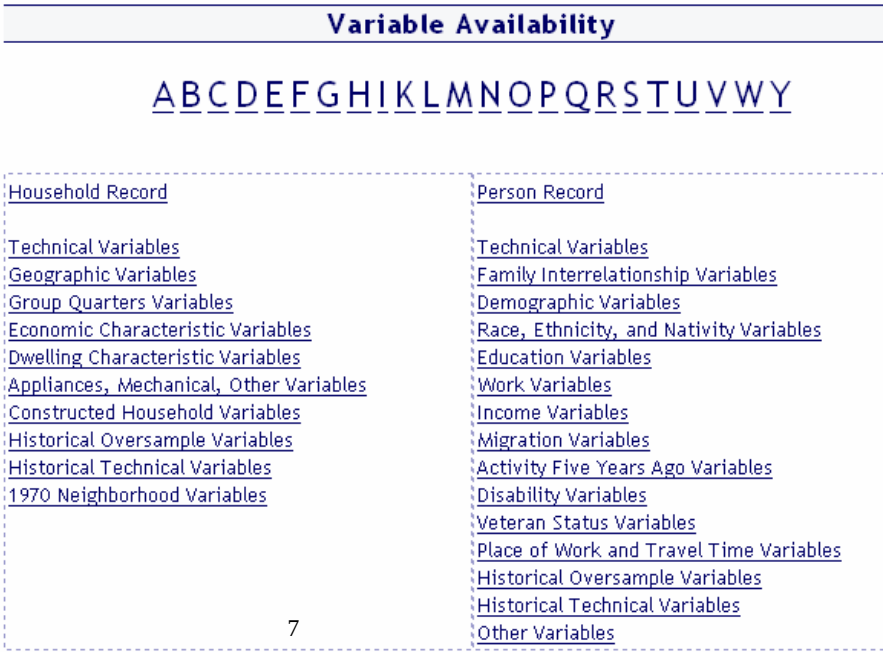
On the main *IPUMS* page is a link to documentation on the data. If you would like to know more about IPUMS and PUMS data select the *What is the IPUMS* link. At right is the *Users Guide* web page.



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.*



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.

○ Category Availability View ◉ 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