IPUMS Data Training Exercise:

An introduction to IPUMS USA

(Exercise 2 for SAS)

Learning goals

- Understand how IPUMS USA dataset is structured
- Create and download an IPUMS data extract
- Decompress the data file and read the data into a statistical package

Summary

In this exercise, you will gain basic familiarity with the IPUMS USA data exploration and extract system to answer the following research questions: What proportion of households in the US has a mortgage? Is the mother's spoken language a consistent determinant of a child's preferred language? How are utility costs changing over time, and are changes in cost different by urban status? You will create a data extract that includes the variables MORTGAGE, VALUEH, LANGUAGE, SEX, AGE, METRO, OWNERSHP, COSTELEC, COSTGAS, ROOMS, UNTSSTR; then you will use the sample code to analyze these data. After completing this exercise, you will have experience navigating the IPUMS USA website and should be able to leverage these data to explore your own research interests.

IPUMS USA: EXERCISE 2 FOR SAS
(UPDATED ON APRIL 28, 2020)

Register for an IPUMS Account

Go to https://usa.ipums.org/usa/ click on Login at the top, and apply for access. On login screen, enter email address and password and submit it!

Make a data extract

- Navigate to the IPUMS USA homepage and click on "Browse Data."

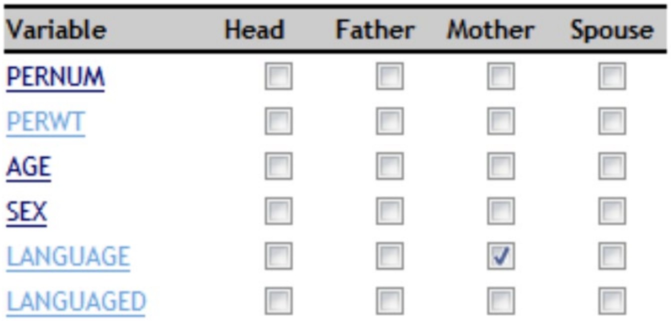
Select Samples – Extract #1: Associations in Household Ownership

- Go to the homepage and click SELECT DATA located at the top of the page.
- On the following webpage, click SELECT SAMPLES.
- Choose the 2010 ACS (1-year) sample by "check marking" the radio box to the left of the sample name.
- Once checked, click SUBMIT SAMPLE SELECTIONS.

Select Variables – Extract #1: Associations in Household Ownership

- Return to the SELECT DATA page. Using the variable table or search feature, find the variables:
  - MORTGAGE: Mortgage status
  - VALUEH: House value
  - LANGUAGE: Language spoken at home
  - SEX: Sex
  - AGE: Age
- Once you have located the variables, click the radio button `Add to cart' on the left side of the page. This selects them to be included in the data extract.
- Once the sample and variables are selected, click VIEW CART -> CREATE DATA EXTRACT
- For this example, we will attach to each person case the language spoken by their mother if she resides in the household.

- To accomplish this, click “ATTACH CHARACTERISTICS” on the EXTRACT REQUEST page. Check the box at the intersection of LANGUAGE and Mother, and SUBMIT.



Variable	Head	Father	Mother	Spouse
PERNUM	□	□	□	□
PERWT	□	□	□	□
AGE	□	□	□	□
SEX	□	□	□	□
LANGUAGE	□	□	☑	□
LANGUAGED	□	□	□	□

- Review and provide a short description for the extract and click SUBMIT EXTRACT.
- You will receive an e-mail when the data is available for download.

Select Samples – Extract #2: Housing Costs

- Go to the homepage and click SELECT DATA located at the top of the page.
- On the following webpage, click SELECT SAMPLES.
- Choose the 2005 through 2010 ACS (1-year) samples by “check marking” the radio box to the left of the sample names.
- Once checked, click SUBMIT SAMPLE SELECTIONS.

Select Variables - Extract #2: Housing Costs

- Return to the SELECT DATA page. Using the variable table or search feature, find the variables:
    - METRO: Mortgage Status
    - OWNERSHP: Ownership of dwelling
    - COSTELEC: Annual electricity cost
    - COSTGAS: Annual gas cost
    - COSTWATR: Annual water cost
    - ROOMS: Number of rooms
    - UNITSSTR: Units in structure
    - CPI99: 1999 Consumer Price Index

- Once you have located the variables, click the radio button `Add to cart' on the left side of the page.

Review and submit your extract

- Click on the "View Cart" button underneath your data cart.
- Review your variable and sample selection to ensure your extract is complete.
    - You may notice a number of additional variables you did not select are in your cart; IPUMS preselects a number of key technical variables, which are automatically included in your data extract.
- Add additional variables or samples if they are missing from your extract, or click the "Create Data Extract" button.
- Review the Extract Request screen that summarizes your extract; add a description of your extract (e.g., "USA Exercise 2: Household Ownership" or "USA Exercise 2: Housing Costs") and click "Submit Extract".
- You will receive an email when your data extract is available to download.

Getting the data into your statistics software

The IPUMS USA extract builder provides raw ASCII data files and the command files necessary for reading the raw data into a stats package. Note that these instructions are for SAS. If you would like instructions for a different stats package, see https://www.ipums.org/exercises.shtml.

Download the data

- Follow the link in the email notifying you that your extract is ready, or by clicking on the "Download and Revise Extracts" link on the left-hand side of the IPUMS USA homepage.
- Right-click on the data link next to the extract you created.
- Choose "Save Target As..." or "Save Link As..."
- Save into your preferred working directory. This tutorial assumes you will save the file into "Documents" (which should pop up as the default location).
- Do the same thing to save the SAS command file.

Decompress the data

- All IPUMS extracts are compressed. There are many applications available for decompressing files. We recommend 7zip for Windows users. Macs can open these types of files without additional software.
- Find the "Documents" folder under the Start menu.
- Double click on the ".dat" file.
- In the window that pops up, press the "Extract" button.
- After the extract has completed, confirm that the Documents folder contains three files that begin with "usa_###".

Read in the data

- Open the "usa_###.sas" file.
- In the syntax editing window, change the first line from "libname IPUMS '. " to "libname IPUMS //Documents...;" using the file directory where you saved your data files.
- After "filename ASCIIDAT", enter the full file location, ending with usa_###.dat;
- Choose "Submit" under the Run file menu.

SAS Code to Review

Code	Purpose
proc freq;	Begins a frequency procedure
proc means;	Begins a means procedure, returns the mean value of a variable
Tables	Required syntax to display frequencies
Where	Selects only specified cases to include in a procedure

Common Mistakes to Avoid

1. Not fully decompressing the data
2. Giving the wrong filepath to indicate the dataset
3. Forget to close a procedure with "run;"
4. Forget to terminate a command with a semicolon ";"

A note on IPUMS USA and sample weighting

Many of the data samples provided by IPUMS USA are based on statistical survey techniques to obtain a nationally representative sample of the population. This means that persons with some characteristics are over-represented in the samples, while others are underrepresented.

To obtain representative statistics, users should always apply IPUMS USA sample weights for the population of interest (persons/households). IPUMS USA provides both person (PERWT) and household—level (HHWT) sampling weights to assist users with applying a consistent sampling weight procedure across data samples. While appropriate use of sampling weights will produce correct point estimates (e.g., means, proportions), many researchers believe that it is also necessary to use additional statistical techniques that account for the complex sample design to produce correct standard errors and statistical tests.

IPUMS USA has provided the variables STRATA and CLUSTER for this purpose. While unnecessary for the following analytic exercises focused on mean and proportional estimates, a further discussion can be found on the IPUMS USA website: ANALYSIS AND VARIANCE ESTIMATION WITH IPUMS USA https://usa.ipums.org/usa/complex_survey_vars/userNotes_variance.shtml

Analyze the Data

Part 1: Frequencies

This part of the exercise uses Extract #1: Associations in Household Ownership.

1. Find the codes page on the website for the MORTGAGE variable and write down the code value, and what category each code represents.

   ____

   ____

   ____

   ____

2. How many people **in the sample** had a mortgage or deed of trust on their home in 2010? What proportion **of the sample** had a mortgage?

   ____

```
proc freq;
     tables mortgage;
run;
```

3. Using weights, what proportion **of the population** had a mortgage in 2010?

   ____

```
proc freq;
     tables mortgage;
     weight perwt;
run;
```

Using household weights (HHWT)

Suppose you were interested not in the number of people with mortgages, but in the number of households that had mortgages. To get this statistic you would need to use the household weight

(HHWT) and select only one person from each household (subset using PERNUM = 1) to represent that household's characteristics.

4. What proportion households **in the sample** had a mortgage? What proportion **of the sample** owned their home?

   ____

```
proc freq;

   where pernum = 1;

   tables mortgage;

run;
```

5. What proportion of households **across the country** had a mortgage in 2010?

   ____

6. What proportion of households **across the country** owned their home? Does the sample over or under-represent households who own their home?

   ____

```
proc freq;

   where pernum = 1;

   tables mortgage;

   weight hhwt;

run;
```

7. What is the average value of:
   a. A home that is mortgaged? _________________________________
   b. A home that is owned? ____________________________________

```
proc means;
     where valueh ^= 9999999 and pernum =1;
     var valueh;
     class mortgage;
     weight hhwt;
run;
```

8. What could explain this difference? *Note: Exclude cases where the house value is missing.*

9. Under the description tab on the website for VALUEH, read the first user note. On the codes page, find the top codes by state for VALUEH, under 2010 ACS/PRCS topcodes by state. How could this complicate your data analysis? Check a histogram of your data to rule out any bias.

```
proc sgplot;
     histogram valueh;
     where valueh^=9999999 and pernum = 1;
run;
```

Part 2: Frequencies

10. What were the three most commonly spoken languages in the US in 2010?

____

```
proc freq order =freq;
   tables language;
   weight perwt;
run;
```

11. Using the code page on the website for LANGUAGE, find the codes for the three most commonly spoken languages.

____

____

12. What percent of individuals who speak English at home:
    a. Has a mother who speaks Spanish at home? ______________________
    b. Has a mother who speaks Chinese at home? ______________________

```
proc freq order =freq;
   where age < 30 and sex = 1;
   tables language;
   weight perwt;
run;
```

13. What percent of men under the age of 30 speak Spanish at home?

____

```
proc freq order =freq;

     where age < 30;

     tables language*sex;

     weight perwt;

run;
```

Part 3: Advanced Exercises

This part of the exercise uses Extract #2: Housing Costs.

14. On the website what are the codes for METRO? What is the code for a single family house, detached in the variable UNITSSTR?

    ____

    ____

    ____

15. What is the proportion of households in the central city who owned their home:

    a. in 2008?

       ____

    b. in 2010?

       ____

```
proc freq;

     where pernum = 1;

     tables metro*ownershp / nocol nopercent;

     by year;

     weight hhwt;

run;
```

Create a graph for annual utility costs by metropolitan status

16. What is the approximate annual cost of *water* for:
    a. A household in the metro area in 2010?

       ____

    b. A household not in the metro area?

       ____

17. What is the approximate annual cost of *electricity* for:
    a. A household in the metro area in 2010?

       ____

    b. A household not in the metro area?

       ____

```
proc format ;
     value in_metro_f 0 = "Not in metro area" 1 = "In metro
     area" ;
run ;

data work.usa_00###_edit1 ;
     set ipums.usa_00### ;
     if metro = 1 then in_metro = 0 ;
     if metro in (2,3,4) then in_metro = 1 ;
     format in_metro in_metro_f. ;
run ;
proc gchart data = work.usa_00###_edit1 ;
     hbar in_metro / discrete type = mean sumvar = costwatr mean
     freq=hhwt ;
     where pernum = 1 and year = 2010 and
     costwatr ^= 0 and costwatr <9990;
run;
```

```
proc gchart data = work.usa_00026_edit1 ;

     hbar in_metro / discrete type = mean sumvar = costelec mean
     freq=hhwt ;

     where pernum = 1 and year = 2010 and

     costelec ^= 0 and costelec <9990;

run;
```

18. In this sample, is there a simple correlation between the number of rooms and the annual cost of electricity?

    ____

    ____

```
proc corr data = work.usa_00026_edit1 ;

     where pernum = 1 and costelec ^= 0 and costelec <9990 and
     rooms>0;

     var costelec rooms;

     weight hhwt;

run;
```

Next, create a graph that will display the average cost of water and gas over time, controlling for the number of rooms and the units in structure. To control for these variables, look at the specific case of a detached, single-family house with 5 rooms.

19. On the website, find the variable description for COSTGAS and note that gas costs are expressed in contemporary dollars. To adjust costs for inflation a price index, CPI99, must be used. Go to the CPI99 variable description page. What year is the index year and how do you apply the inflation adjustment?

    ____

    ____

    ____

20. Has the annual cost of gas for a single family, 5-room home increased since 2005 **in nominal terms**? What about the annual cost of water?

```
proc gchart data = work.usa_00026_edit1 ;

     vbar year / discrete type = mean sumvar = costwatr mean
     freq=hhwt ;

     where pernum = 1 and unitsstr = 03 and rooms = 5

          and costwatr ^= 0 and costwatr <9990;

run;


proc gchart data = work.usa_00026_edit1 ;

     vbar year / discrete type = mean sumvar = costgas mean
     freq=hhwt ;

     where pernum = 1 and unitsstr = 03 and rooms = 5

          and costgas ^= 0 and costgas <9990;

run;
```

21. Has the annual cost of gas for a single family, 5 room home increased since 2005 **in real terms**? *Note: The variable CPI99 assigns an inflation index value according to the year of the observation.*

```
data work.usa_00###_edit2 ;

     set work.usa_00026_edit1 ;

     costgas_real = costgas*cpi99 ;

     label costgas_real = "Cost of gas (1999 dollars)" ;

run ;

proc gchart data = work.usa_00###_edit2 ;
```

```
     vbar year / discrete type = mean sumvar = costgas_real mean
     freq=hhwt ;

     where pernum = 1 and unitsstr = 03 and rooms = 5

          and costgas ^= 0 and costgas <9990;

run;
```

Answers

Part 1: Frequencies

1. Find the codes page on the website for the MORTGAGE variable and write down the code value, and what category each code represents.
   0 N/A; 1 No, owned free and clear; 2 Check mark on manuscript (probably yes); 3 Yes, mortgaged/ deed of trust or similar debt; 4 Yes, contract to purchase

2. How many people in the sample had a mortgage or deed of trust on their home in 2010? What proportion of the sample had a mortgage? 1,523,041 people; 49.75%

3. Using weights, what proportion of the population had a mortgage in 2010?
   47.46%

Using household weights (HHWT)

4. What proportion households in the sample had a mortgage? What proportion of the sample owned their home? *(Hint: don't use the weight quite yet)*
   42.20% of households mortgaged; 23.98% of household owned

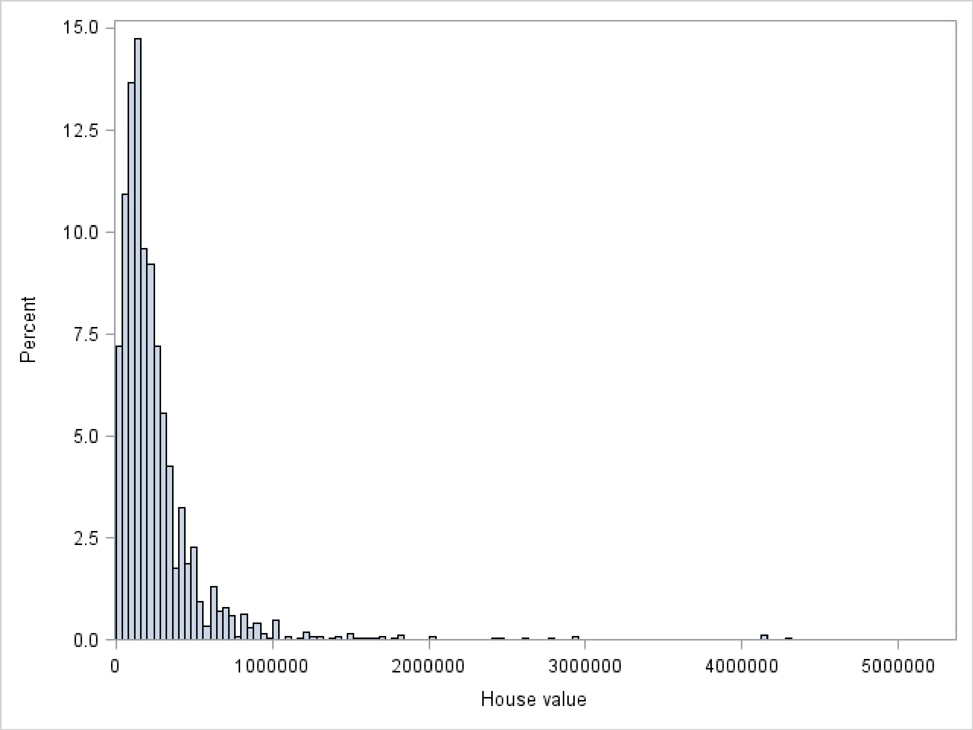
5. What proportion of households had a mortgage across the country in 2010?
   40.53% of households

6. What proportion of households owned their home? Does the sample over or under-represent households who own their home?
   20.07% of households, sample over-represents households that own their own home or have a mortgage.

7. What is the average value of:
   a. A home that is mortgaged? $267,890
   b. A home that is owned? $219,015

8. What could explain this difference?
   Perhaps homes that have already been paid off are older and less expensive, or it takes less time to pay off a home that is worth less.

9. Under the description tab on the website for VALUEH, reader the first user note. On the codes page, find the top codes by state for VALUEH, under 2010 ACS/PRCS topcodes by state. How could this complicate your data analysis? Check a histogram of your data to rule out any bias.
   There doesn't seem to be a significant cluster around the topcodes, so the data sample may not be noticeably biased.



Part 2: Frequencies

10. What were the three most commonly spoken languages in the US in 2010?

    English, Spanish, Chinese

11. Using the code page on the website for LANGUAGE, find the codes for the three most commonly spoken languages. 01 English; 12 Spanish; 43 Chinese

12. What percent of individuals who speak English at home:
    a. Has a mother who speaks Spanish at home? 3.89%
    b. Has a mother who speaks Chinese at home? 0.22%

13. What percent of men under the age of 30 speak Spanish at home? 13.4%

Part 3: Advanced Exercises

14. On the website what are the codes for METRO? What is the code for a single family house, detached in the variable UNITSSTR?
    UNITSSTR: 03 1-family house, detached; METRO: 0 Not identifiable; 1 Not in metro area; 2 Central city; 3 Outside central city; 4 Central city status unknown

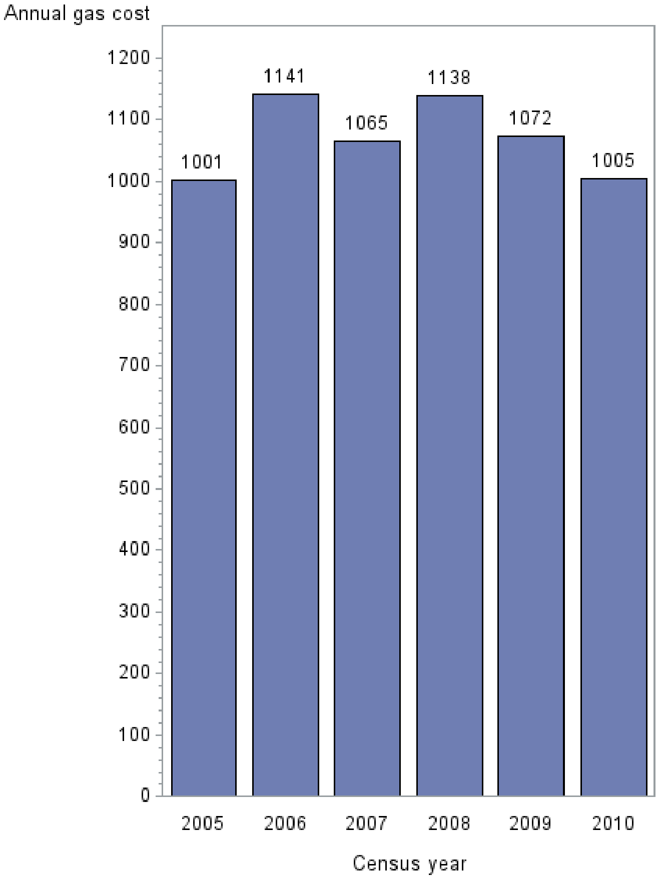
15. What is the proportion of households in the central city who owned their home:
    a. in 2008? 44.51%
    b. in 2010? 42.92%

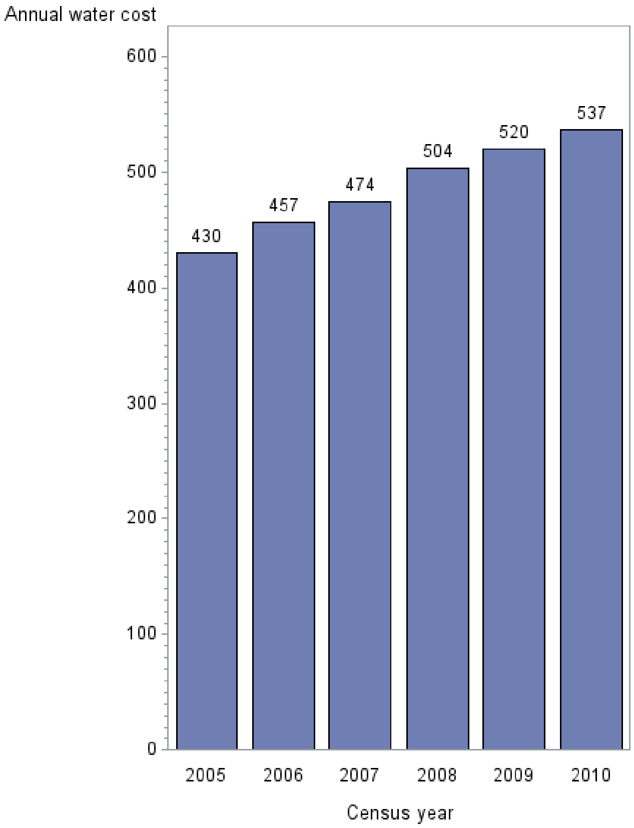
16. What is the approximate annual cost of *water* for:
    a. A household in the metro area in 2010? ~$575
    b. A household not in the metro area? ~$500

17. What is the approximate annual cost of *electricity* for:
    a. A household in the metro area in 2010? ~$1700
    b. A household not in the metro area? ~$1750

18. In this sample, is there a simple correlation between the number of rooms and the annual cost of electricity? There seems to be a weak positive correlation between number of rooms and the cost of electricity. (0.30)

    Next, create a graph that will display the average cost of water and gas over time, controlling for the number of rooms and the units in structure. To control for these variables, look at the specific case of a detached, single-family house with 5 rooms.



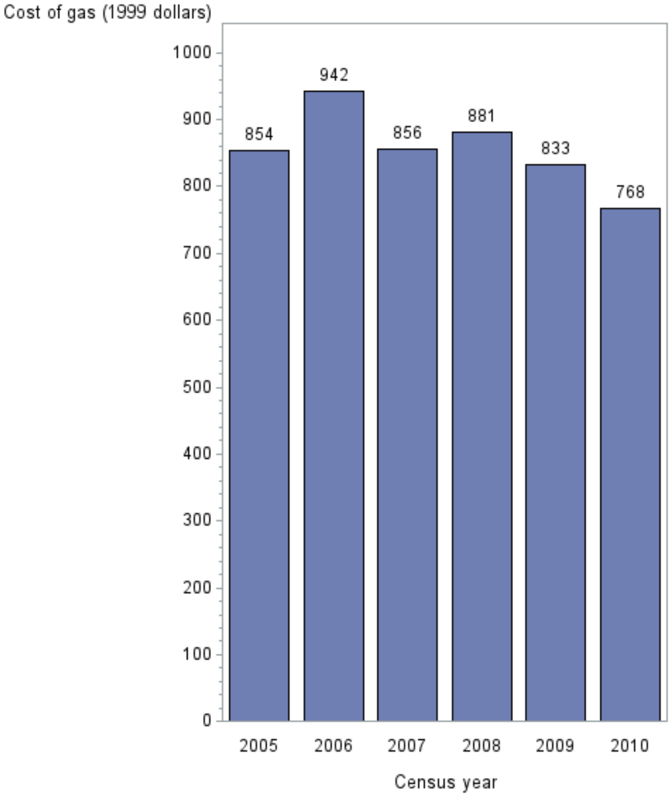


19. On the website, find the variable description for COSTGAS and note that gas costs are expressed in contemporary dollars. To adjust costs for inflation a price index, CPI99, must be used. Go to the CPI99 variable description page. What year is the index year and how do you apply the inflation adjustment? 1999; real costs adjusted for inflation and indexed to the 1999 U.S. dollars are estimated by generating a new variable CPI99 * COSTELEC.

20. Has the annual cost of gas for a single family, 5-room home increased since 2005 in nominal terms? What about the annual cost of water?
 In *nominal terms*, the cost of gas is fluctuated over time, but the cost of water has steadily increased.

21. Has the annual cost of gas for a single family, 5 room home increased since 2005 in real terms?

    <u>In *real terms*, the gas prices fluctuated over time.</u>


Cost of gas (1999 dollars)
Census year