IPUMS Training and Development: Extraction and Analysis

IPUMS USA Exercise 1 - Excel

OBJECTIVE: Gain an understanding of how an IPUMS USA dataset is structured and how it can be leveraged to explore your research interests. As an example, this exercise will use IPUMS USA to explore basic employment differences between men and women workers in 1940.

1/28/2019

IPUMS USA — Training and Development

Research Questions

What proportion of men and women are heads of households? What proportion of men and women are employed? What does the age distribution of employed men and women look like? Is there a difference in average age of employment between men and women?

Objectives

- Create and download an IPUMS data extract.
- Decompress data file and read data into Excel.
- Analyze the data using Excel formulas and built-in features.
- Validate data analysis work using answer key.

IPUMS Variables

- RELATE: Relationship to head of household
- MARST: Marital Status
- EMPSTAT: Employment Status
- SEX: Sex
- AGE: Age

Review Answer Key (page 11)

Will Excel Suit your Needs?

- Excel's strengths lie in being able to visualize data easily. Excel is also good at bringing several spreadsheets together to make coherent data.

- However, Excel is not a large-data processing program like SAS, Stata, and SPSS. It has a limited number of rows and columns, which can make it difficult to fit the large datasets available through IPUMS. In order to be able to use Excel, the user must ensure that the data is small enough for Excel to process. If Excel is not able to read your dataset, try breaking down the dataset into individual samples or considering whether all the data you selected is necessary. If neither of these is possible, Excel may not be the right tool. Excel may also not be able to complete statistically complicated analyses.

Common Mistakes to Avoid

- Forgetting to use appropriate order of operations to ensure correct computation of data.
- Accidentally selecting the wrong data ranges.

Register

• • •

Step 1: Download the Data

Registering with IPUMS

Go to http://usa.ipums.org, click on IPUMS Registration and Login and Apply for access. On login screen, enter email address and password and submit it!

Download the Data

- Go back to homepage and go to Select Data.
- Click the Select Samples box, check the box for the 1940 1% sample, then select “Submit sample selections.”
- Using the drop down menu or search feature, select the following variables:

MARST: Marital status

SEX: Sex

RELATE: Relationship to head

AGE: Age

EMPSTAT: Employment Status

- Click the green VIEW CART button under your data cart. Review variable selection. Click the green Create Data Extract button.
- Because Excel is not able to handle large datasets, we will filter our data so that our extract contains only what is necessary to our research interests.
- Click “Select Cases.” Check the RELATE, AGE, and MARST boxes. Click submit.
- Under the Relate box, highlight (holding control while clicking) only codes 01 and 02, Head/householder and Spouse. We don’t want any other family members in this dataset.
- Under the Age box, select ages 18-65, since we are interested in working-age adults.
- Under MARST, only check 1 and 2 for those who are married.
- The extract system does not automatically create output in CSV format. To add this capability, click "Change" next to Data Format and then select "Comma delimited (.csv)." Click submit.
- Review the ‘Extract Request Summary’, describe your extract and click Submit Extract. You will get an email when the data is available to download.
- Follow the Download and Revise Extracts link on the homepage, or the link in the email.

• • •

Step 2: Decompress the Data

• • •

Step 3: Open the Data

Getting the data into your statistics software

The following instructions are for Excel. If you would like to use a different stats package, see: http://cps.ipums.org/cps/extract_instructions.shtml

- Go to http://usa.ipums.org and click on Download or Revise Extracts.
- Right-click on the ""CSV"" link next to extract you created.
- Choose “Save Target As...” (or “Save Link As...”).
- Save the file in a folder you can find easily.
- Open the folder where you saved your file by navigating through Windows Explorer.
- Right click on the “.csv” file.
- Use your decompression software to extract here.
- Open the .csv file. It should automatically open in Excel.
- Go to File, Save As. Underneath the File Name option, change the “Save As Type” to Excel Workbook (*.xlsx). The full functionality of Excel is now available for use on this dataset.
- Under the View tab, select Freeze Panes, Freeze Top Row to be able to view the headers of your data even when you scroll down past the first page.

• • •

Part I: Frequencies

Analyze the Sample – Part I Frequencies

Get basic frequencies. There are two easy-to-use formulas to count frequencies in Excel: Countif and Frequency. Frequency is a good tool for quick basic frequencies over one column (variable). If it's necessary to check multiple rows or if more than one condition needs to be met for the frequency (ex: male and head of household), countif or countifs is a better tool.

A) Based on the selections we made to the extract, what population of people are we studying?

B) Find the number of males and females in this dataset. Remember the codes for each variable can be found on the IPUMS website.

> *Frequency formula*: =FREQUENCY(range, bins)
> (select the data)

(The bin values in the frequency formula must be entered as an array. Make sure to enter your bins in a separate column and have the formula call those cells for the bins. Then highlight the cells next to the bins, hit F2, and then CTRL+Shift+Enter to fill the bins.)

C) How many men vs. women were heads of household?

> *Counts formula* =COUNTIFS (range, condition, range, condition...)

(COUNTIF/COUNTIFS is a formula that allows the user to specify a condition under which Excel "counts" the cell. In this case, the condition is that sex is male (or female) and relate is 1. Countif is used when there is one condition, whereas countifs is used when there are multiple conditions).

• • •

Part II: Basic Calculat-ions

Basic Calculations in Excel

Excel Operators: + - * /
(Select the data)

A) What percentage of households had a male head of household?

B) What percentage of women are heads of household?

C) What is the ratio of employed men to employed women?
   a. How many employed men are there?
   b. How many employed women are there?
   c. What is the ratio?

• • •

Part III: Visualizing the Data

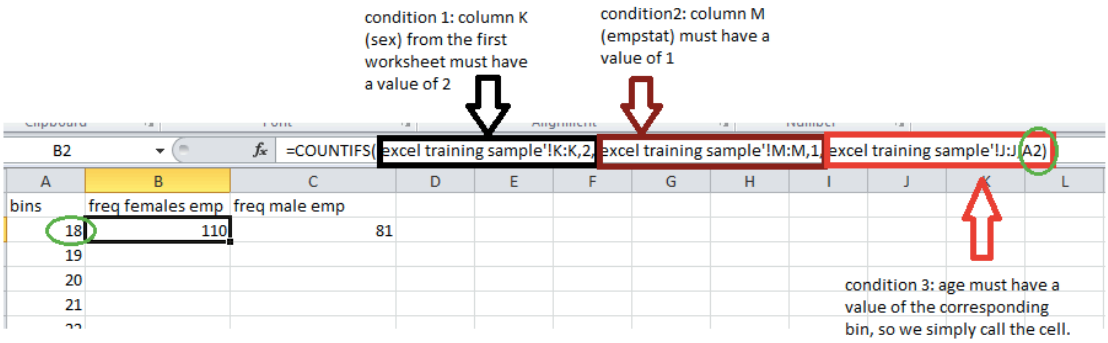
Part III-Visualizing Data

Question: What does the age distribution of working men look like in comparison to working women?

A standout feature of Excel is its ability to easily visualize data in chart or graph form. In order for us to visualize the research question, we must create a table where we can compare the number of men vs. women who are employed at each age.

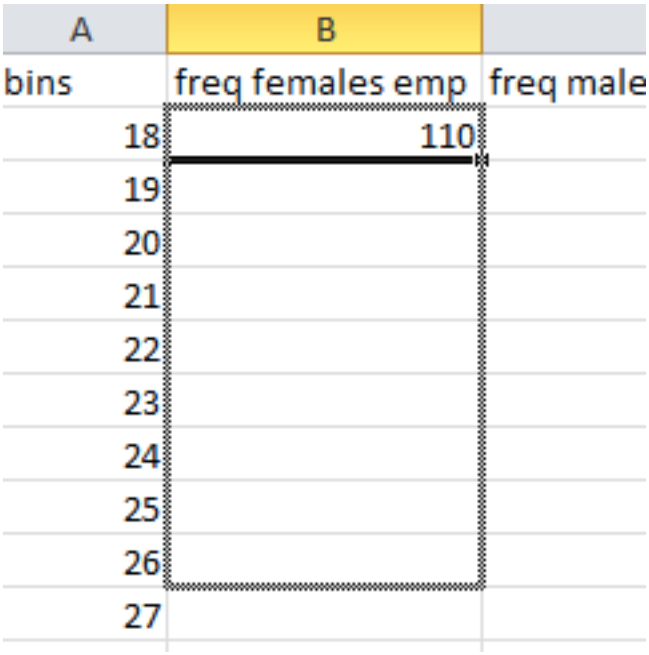
A) Create a frequency chart with age bins.

Create a new tab (worksheet) in the same workbook. In this tab, create the bins for the frequency (create a column of cells with values between 18-65). Use countifs to start the frequency chart for men and women.

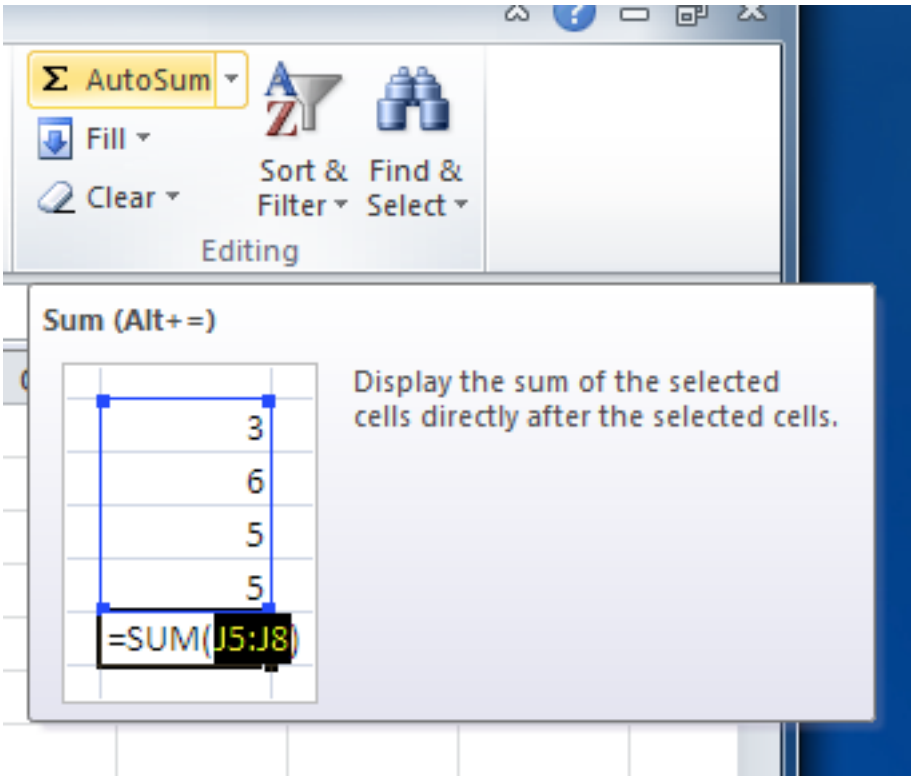


Part III: Visualizing the Data

B) Fill the chart using the drag and fill function (select the cell, drag the dot on the bottom right corner down). Excel will automatically link each cell with the corresponding bin number. In the example below, the filled formula in B3 will use cell A3 (age 19) instead of A2.



C) Check to make sure your frequencies are correct. Use the autosum function (highlight the cell under the last frequency, and click autosum) to make sure all employed men and women are accounted for.



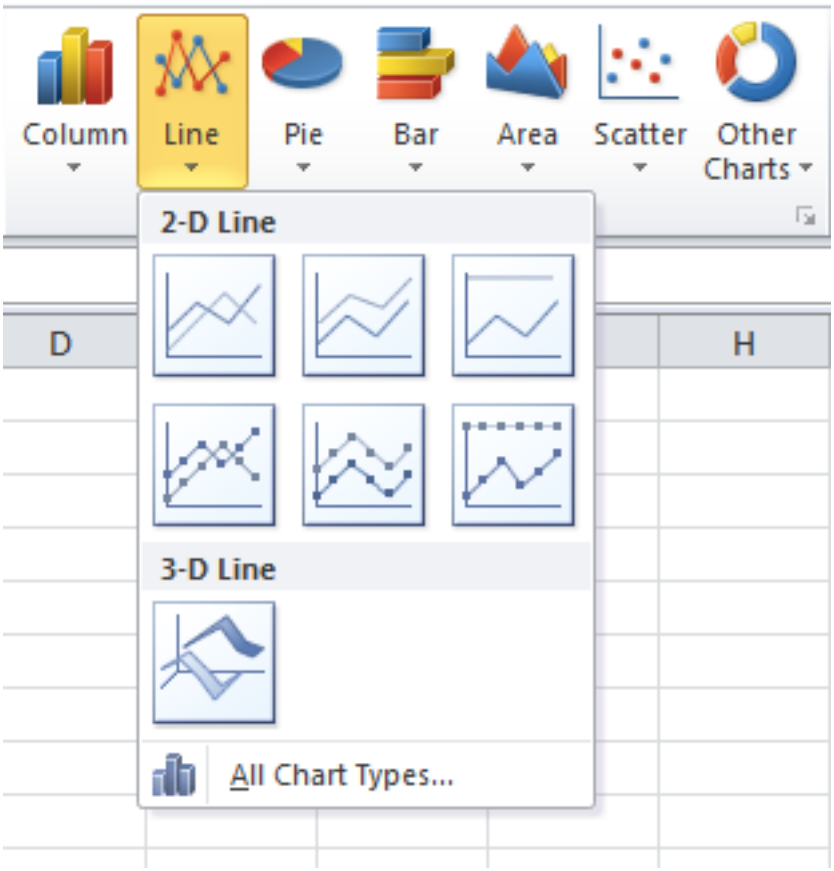
Part III: Visualizing the Data

D) Create a graph of men's and women's ages with x-axis "age" and y-axis "frequency."

a. Under the insert tab, click on line, then select the first option.

b. If the graph does not look correct, right click on it and click on Select Data. Use this menu to ensure that the correct data was auto-selected, and change the selection if necessary. Common errors that excel makes is switching the X and Y axes, and using the bin columns as a third dependent variable rather than the x-axis values.

c. Adding a trendline is possible in this graph by right clicking on a line in the graph and clicking "add trendline." Note the trendline can be edited to reflect a non-linear shape if necessary.



E) Look at the differences between the line for male workers and the line for female workers. Is it possible from this graph to hypothesize that men tend to work at older ages than women? Is there a better way to graph the data to see if this may be true?

• • •

Part IV: Analyzing the Data

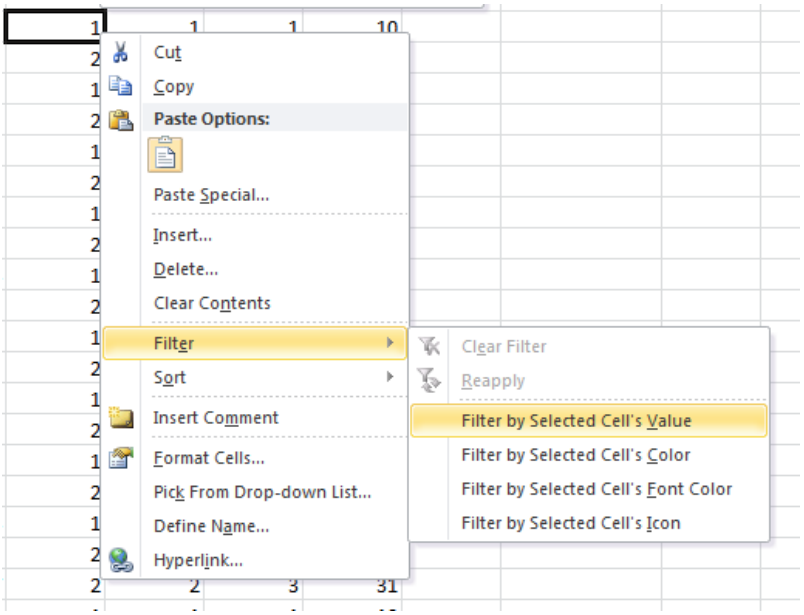
Part IV- Data Analysis

Question: Is the mean age of married working men different than that of women?

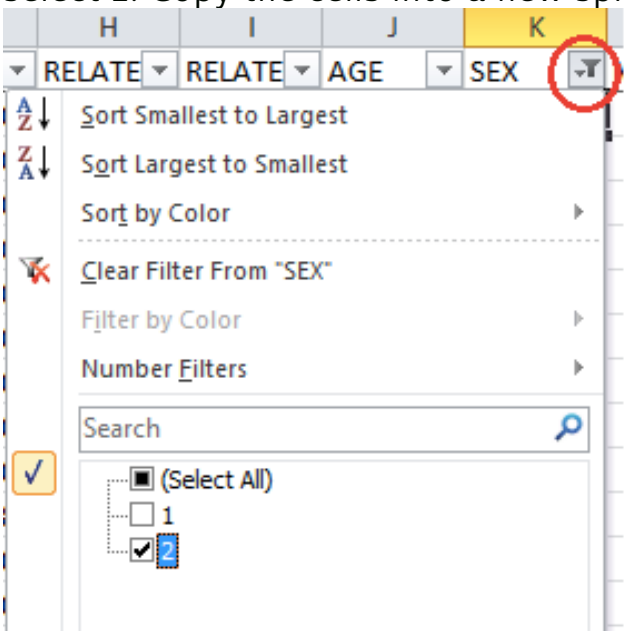
Excel has both formulas and menus that allow for data analysis. Because we have used a few formulas already, we will be using the menu system.

A) Prepare the data for analysis.

To prepare the data for analysis, find the sex column in the original data sheet, click on a cell that has the value of 1 to select it, right click on the cell, find the Filter option, and select "Filter by selected cell's value." Select the entire remaining dataset and copy and paste the cells into a new spreadsheet.



Next to the Sex column title, click on the filter icon, deselect 1 and select 2. Copy the cells into a new spreadsheet.



Part IV: Analyzing the Data

B) Enable Analysis Add-ins

To check whether this is enabled, click on Data and check for “Data Analysis." If it is there, continue to C).

If it’s not there click on the File Tab, click Options, and then click the Add-Ins category. In the Manage box, select Add-ins and then click Go. In the Add-ins available box, check the Analysis ToolPak box, and then click OK.

C) Do an F test to compare variances. What are the results?

Select Data, then Data Analysis. In the data analysis box, select F-Test Two Sample Variances. Click OK. Use the variable ranges to input the ranges for males and females, adjust the alpha if necessary, and then click OK.

D) Do a T test to compare means, assuming a null hypothesis of no differencebetween ages. What are the results?

Navigate back to the tab in which you pasted in the sex-sorted age data. Navigate back to the Data Analysis box, and this time, select T test. There are three types of t-tests that Excel can do, but using the results of our F-test, we can confidently select Two-sample using Unequal Variances. Adjust the ranges of the two datasets if necessary, enter 0 under Hypothesized Mean Difference, and click OK. Notice that you may hypothesize a null hypothesis of something other than zero.

E) Based on the results of the T-test, why would/wouldn't there be a difference in employment ages between genders?

Answers

ANSWERS

Analyze the Sample – Part I Frequencies

Get a basic frequency of the SEX variable.

A) Based on the selections we made to the extract, what population of people are we studying? **Married adults ages 65 and under in 1940.**

B) Find the number of males and females in this dataset? **Males: 250760, females: 261825**

> =FREQUENCY(K:K,P16:P17)
> where K:K is the sex column, and P16:P17 are the bins.

C) How many men vs. women were heads of household? **Male heads of household: 250760, female heads of household: 6246**

> Males =COUNTIFS(H:H,1,K:K,1)
> Females =COUNTIFS(H:H,1,K:K,2)
> Where column H is Relate, and column K is Sex.

Basic Calculations in Excel

A) What percentage of households had a male head of household? **97.57%**

100% * (male heads/ (male head + female heads))
Formula in cell should look something like this:

> =100*Q2/(Q2+T2)

B) What percentage of women are heads of household? **2.39%**

100% * (female heads/total # females)

C) What is the ratio of employed men to employed women?

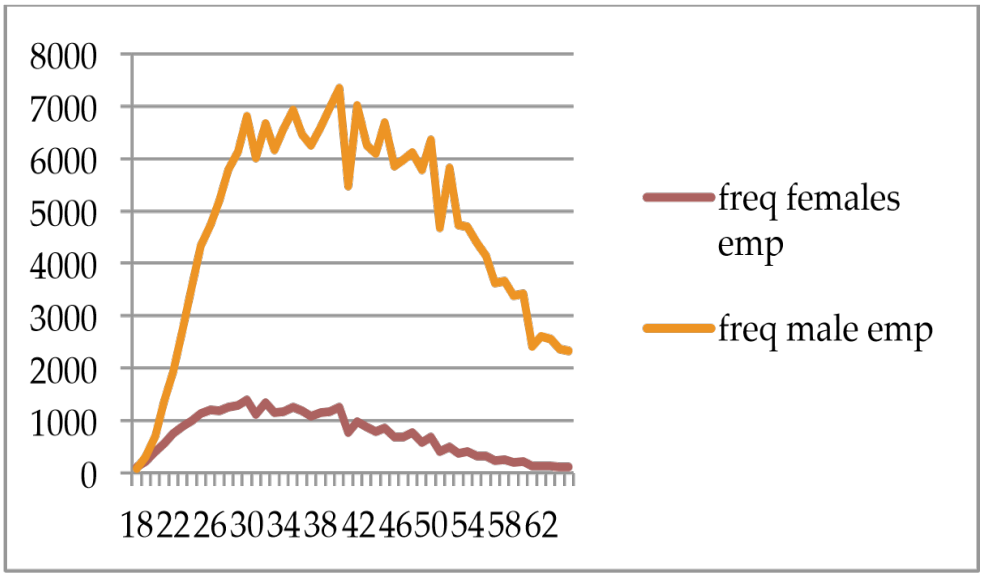
a. How many employed men are there? **226164**
   Use countifs
b. How many employed women are there? **34833**
   Use countifs
c. What is the ratio? **1 female: 6.749 males**
   Use division operator

Answers

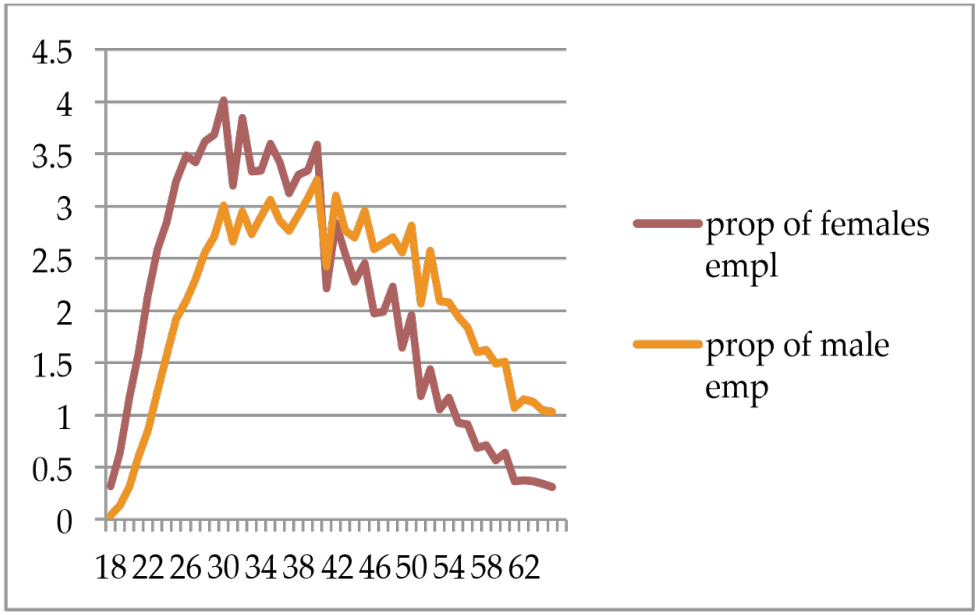
Part III-Visualize the Data

Question: What does the age distribution of working men look like in comparison to working women?

E) Look at the differences between the line for male workers and the line for female workers. Is it possible from this graph to hypothesize that men tend to work at older ages than women? Is there a better way to graph the data to see if this may be true?



The first graph showed us that more men than women are employed during their adult years. A better way to visualize this might be to compare the proportion of workers at each age. To do this, divide the number of women working at each age by the total number of women working and multiply by 100. Do the same for men. Your resulting graph should look like the following graph, which makes it easier to see that more men work at older ages than women; however the level of significance cannot be determined.



Answers

Part IV- Data Analysis

C) Do an F test to compare variances. What are the results?

F-Test Two-Sample for Variances

	Variable 1	*Variable 2*
Mean	41.6822	36.69701
Variance	125.3321	106.7101
Observations	226164	34833
df	226163	34832
F	1.174511	
P(F<=f)one---tail	6.54E---84	
F Critical one-tail	1.013518	

F>f critical one-tail UNEQUAL VARIANCE

Results show unequal variance.

D) Do a T-test to compare means, assuming a null hypothesis of no difference between ages. What are the results?

t-Test: Two-Sample Assuming Unequal Variances

	Variable 1	*Variable 2*
Mean	41.6822	36.69701
Variance	125.3321	106.7101
Observations	226164	34833
Hypothesized Mean Difference	0	
df	48330	
t-Stat	82.88356	
P(T<=t) one-tail	0	
t-Critical one-tail	1.644885	
P(T<=t) two-tail	0	
t-Critical two-tail	1.960013	

If the t-statistic is <t-critical or >t-critical we may reject the null hypothesis. Since 82.9 is greater than both the one tailed and two tailed critical values, and the p value is less than .05, we may reject the null hypothesis.

E) The difference found between men and women working ages may be attributable to a cultural norm of women leaving the workforce when they marry or have children. It may also be a generational difference, with younger women being more likely to work than older women due to cultural changes that happen normally over time.