U.S. DEMOGRAPHICS: USING EXCEL TO GET UNDER THE HOOD

Robert Gebeloff

rgebeloff@nytimes.com

One variable that comes up all the time in data reporting is the population. If you cover politics, population has a major impact on representation and voting. If you cover business, regional population shifts are closely related to economic growth. And if you’re a local reporter, becoming familiar with local demographics informs all types of reporting.

In this exercise, we’re going to look under the hood at how population changes occur and measure the cause of population changes in the U.S. states. The U.S. Census Bureau is best-known for the decennial Census, a mammoth survey where the agency attempts to account for every household and every resident in the country. There is also the American Community Survey, in which the agency conducts a detailed survey of a large sample of the population and publishes estimates on topics such as educational attainment, marital status, and home ownership.

Less known, but incredibly important, are the annual updated estimates of the population. These estimates are driven by four measures: Births, Deaths, Internal migration (people moving from place to place within the U.S.), and International Migration (people moving in and out of the United States).

Serious demographers – and some journalists – mine this information regularly for insights into how the nation is changing. For example, when I worked for the Newark Star-Ledger in the 2000s, I used the estimates data to break news that New Jersey was destined to lose a Congressional seat in the 2010, and in ongoing coverage explaining why the state’s population was growing more slowly than other states.

The Web page housing this data is [here](https://www.census.gov/popest/data/national/totals/2014/NST-EST2014-alldata.html) (https://www.census.gov/popest/data/national/totals/2014/NST-EST2014-alldata.html), or you can download the file 2014popestimates.csv from our Github page. While you can double-click on the file and open it, it’s not a best practice; sometimes files contain extraneous information that we don’t want in our spreadsheet, so it’s safer to save the file to our computers, then open the file with Excel in case there are any unexpected issues with the file format.

As you can see, there’s a lot of data in the file. To see more clearly what you’re dealing with, select all of the columns and autofit the column width. (Format-Column-Autofit)

As you can see, when you’re working with datasets, the headers on top of the data are often designed to be compact more than they are designed to be human-understandable. You could probably make out what the headers mean, but to be sure, refer to the code sheet that is provided – this document is often called a “file layout” (2014popestimates\_layout.pdf).

Let’s look at some of the columns included in the dataset. The first few columns are devoted to geography. The “SUMLEV” is described as the geographic summary level – this is a common flag in Census data that signifies whether the entity on that row is a state, county, city etc. This comes in handy when you’re working with datasets that mix geography.

Next comes “REGION” and “DIVISION” – these codes are useful if you want to analyze the data in broader terms, such as the Midwest vs the South, etc. If you want to go in that direction, of course, you need to do more research as to how these codes work. This is where Google is your friend; Google “census regions and divisions” and you’ll find lots of information, including a handy [map](http://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf) (<http://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf>) and a [list](http://www2.census.gov/geo/docs/maps-data/maps/reg_div.txt) that describes the actual codes (<http://www2.census.gov/geo/docs/maps-data/maps/reg_div.txt>).

Then there is “STATE” and “NAME”. Note that the “state” field is described as the state “FIPS” code. You’ll encounter FIPS, which is an acronym for “Federal Information Processing Standard”, a lot as you work with U.S. data about places. FIPS codes are handy for two primary reasons - - they allow you to “join” different data sets together based on a common set of codes, and they help you map the data with mapping software, which relies on the FIPS codes to assign data attributes to the physical shapes displayed on the map.

Finally comes the data. First comes the population estimates, then the components (Births, Deaths, etc). One thing that might seem strange at first – the estimates are mostly annual, except for the first year, 2010, which is described in the code sheets as covering April 1 through July 1 of that year. The reason for this is that the estimates data is a mid-year measure, but in 2010, there was the decennial Census as of April 1. So what the data is showing, in essence, is the “official” population as of the decennial Census, and then the “base” for all of the subsequent July 1 estimates, starting with July 2010.

The data uses terms like “natural increase”. You can guess this has something to do with births and deaths, but it’s never good to guess. If you click on “About Estimates”, there is a “[Terms and Definitions](https://www.census.gov/popest/about/terms.html)” page (<https://www.census.gov/popest/about/terms.html>). This glossary gives you two valuable pieces of information, that in fact, natural increase = Births-Deaths, but also that the natural increase rate provided represents “natural increase during a specified time period as a proportion of an area's population at the midpoint of the time period. Rates are expressed per 1,000 population.”

Ok, now we’ve examined the data layout and the field definitions. Let’s analyze something. Generally when working with datasets, it is always best to preserve the original data in its own worksheet, and do your analysis on subsequent worksheets. In a wide spreadsheet like this, it also makes it easier to see what you’re doing.

So let’s do one quick thing together and then I’ll leave you to explore some questions on your own. Let’s rank the states based on natural increase rate in 2014. First create a new worksheet (right-click on the worksheet tab on the bottom of your screen), and then go back to the original and highlight the columns “STATE”, “NAME, and “RNATURALINC2014” (it’s column ‘BL’, way over to the right). Copy and paste these columns into your new worksheet.

As you’ll notice, the top of the file has summary data for the nation and four regions, then the states, and then Puerto Rico. While this is all interesting data, the standard state rankings include the 50 states plus the District of Columbia. So let’s set aside the others for a moment by cutting and pasting the regions and then Puerto Rico someplace to the right on our sheet. Then delete the empty cells by highlighting them, right-clicking, delete, then ‘move cells up’.

There are two ways we could establish the rankings here. The first way would be to sort the data based on the rate. To do this, we would:

1. Highlight all three columns
2. Click on “Data|Sort”
3. Make sure “my data has headers” is checked
4. Select “RNATURALINC2014” as column, sort on “values”, and sort by “Largest to Smallest”

Sorting is a great way to quickly see your tops and bottoms. But the disadvantage of sorting comes when you want to bring new data into your worksheet -- you need to line up any new variables you want to add to the mix with the existing data, and sorting makes that difficult.

So let’s make a new worksheet, copy and paste the original data again, and use the RANK function instead. Once you have the data pasted into the new sheet, and the regional data and Puerto Rico pasted to the side, type a new header on column D, “RANK”

In cell D2, enter the following formula and hit enter:

=rank(c2,c$2:c$52)

As you type, Excel guides you as to what the RANK function requires, but essentially what we’re saying here is “Take the value in cell C2 and compare it to all the values in cells C2 through cells C52”. We’re also adding a $ symbol before the 2 and 52 in the range. This is called an anchor – this means that even if we copy and paste this formula down our spreadsheet, Excel will fix the range as cells C2 to cells C52.

Note: The default rank order places the highest value as 1. If you’re looking at something like death rates, where high values are bad, you can tell your ranking formula to assign 1 to the lowest value by adding a third parameter to your formula:

=rank(c2,c$2:c$52,1)

Let’s go ahead and copy and paste the formula down the column.

Ok, this is a start. Try exploring this data on your own. Some suggestions you might try:

1. Compare the natural increase rankings with net migration. Are the results similar or different?
2. Which states are more reliant on immigration as a share of population growth – are there states that would be shrinking in population without immigration?
3. Rank the states by total gain/loss due to domestic migration for each year, and then for the 2010-2014 total? Any surprises? Is the overall ranking inconsistent with the ranking for the most recent year?
4. Compared to the 2010 Census population, which states have lost the largest share of population to domestic migration?
5. Pick out one state that is interesting to you and analyze all categories. What are some story angles you might pursue?