The Data (sidebar)

This study depends for its essence on numbers indicating (1) how many people in the United States hold each of the 20 most common first names, (2) how many hold each of the 20 most common surnames, and (3) how many hold each of the 400 combinations of a top-20 first name with a top-20 last name—the "20/20" population.

The most reliable data about populations of names—because it's the most transparent—is that of the U.S. Census Bureau. The Bureau meticulously explains its methods for collecting and editing name data in an online document. However, the U.S. Census data on names comes with a number of drawbacks. (1) The figures are somewhat dated: the numbers of surnames are based on the 2000 census, while those for first names—male and female—are from the 1990 census (ironically, since first names are more subject to changes of fashion and tend to "age" faster). No name data from the 2010 census has been released as of this writing. (2) Although the surname data are in the form of raw numbers, the first-name figures are given only as percentages of the entire (1990) population. And (3), most problematic, the Census Bureau has never provided numbers for first-name/last-name combinations.

For full-name combinations I have had to rely on WhitePages.com, a company whose primary purpose is to provide contact information on individuals, and which only as a by-product provides an online device that answers queries about the numbers of first-name, last-name, and name-combination holders. These data, which I collected in August 2013, are labeled "as of February 2011". The name "WhitePages" might suggest that the data are based solely on landline telephone listings, but the reality is evidently more complex. Presumably not every adult and child is listed in a telephone directory, and yet the sheer numbers of (English) surname-holders in these 2011 data tend to be 12% to 14% higher than the numbers for the same surnames in the 2000 census. (The total U.S. population grew by 9.7% from 2000 to 2010, according to the U.S. Census Bureau.) The site itself reveals very little about the origins of its data. A statement on the site says only "We ingest billions of records every month from a variety of public sources." Elsewhere on the site, under the heading "Our data sources," three entities are named and briefly qualified: (1) Oxford University Press, "Provider of name meanings" (possibly alluding to several reference works by Patrick Hanks and Flavia Hodges); (2) "Social Security Department" [sic], "Provider of birth records" (a careless reference to the U.S. Social Security *Administration*); and (3) "WhitePages.com Searches and Listings" ("Search engine of phone, address, and age information").

My query to "Contact Us" at WhitePages.com, requesting clarification about the data sources, was answered by an individual, Liz Powell, but only to the extent of courteously referring back to the phrases on the website.

In contrast to the numbers that, in general, suggest tallies roughly equivalent to the census figures, the WhitePages data seem to consistently undercount holders of Spanish surnames. The 2011 figures, when compared to those from the 2000 census, show a gain of only 2% for **Rodriguez**, and *decreases* ranging from 0.4% to 5% for the names **Garcia**, **Martinez**, **Hernandez**, **Lopez**, **Gonzalez**, and **Perez**.

In spite of these drawbacks, I have relied on WhitePages.com not only for the data on full-name combinations, but also for that on first and last names separately. I've chosen this approach for the sake of internal consistency in the data. Meanwhile, I have borrowed figures on the ethnicity of surnames from the U.S. Census website, based on the census of 2000.

In October 2010, I carried out a pilot study based on figures for the same names gathered also from WhitePages.com. Almost all the population figures for first names, last names, and full names increased significantly from the earlier data set to the present one (two exceptional *decreases*—for "**David Anderson**" and "**Richard Johnson**"—suggest errors in the 2011 data); nevertheless, the overall patterns revealed in the two studies are remarkably similar and support the same conclusions.

The data given at WhitePages.com support **John** and **Smith** as the most popular first and last name, respectively, in the U.S. The U.S. Census of 2000 agrees on **Smith** as the top surname (28% more popular than second-place **Johnson**), but, as a first name, **John** is surpassed in the census (of 1990—the latest year for which first-name data are available) by **James**, by a narrow margin of 1.4%.

Two other online sources of name data were *not* used for this study. The data at Mongabay.com are the same as those given at the U.S. Census website, repackaged. And the figures at HowManyOfMe.com are labeled "For entertainment purposes only." This site, after describing several sources of inaccuracy, concludes with the request "please don't cite this in any sort of scholarly / semi-scholarly setting. It really isn't accurate enough to be used as a serious source." I appreciate their candor.

In short, my data from WhitePages.com are of unclear origin, and I caution the reader to bear this fact in mind and to recognize the consequently tentative nature of my conclusions.

A rich source of U.S. name data that was not available to me when I made this study is available at <http://www.mynamestats.com/>.

****