It's a Boy!
An Analysis of Tens of Millions of Birth Records Using R

Susan I. Ranney, Ph.D.[1*]

1. VP Product Development, Revolution Analytics, Inc.
*Contact author: sue@revolutionanalytics.com

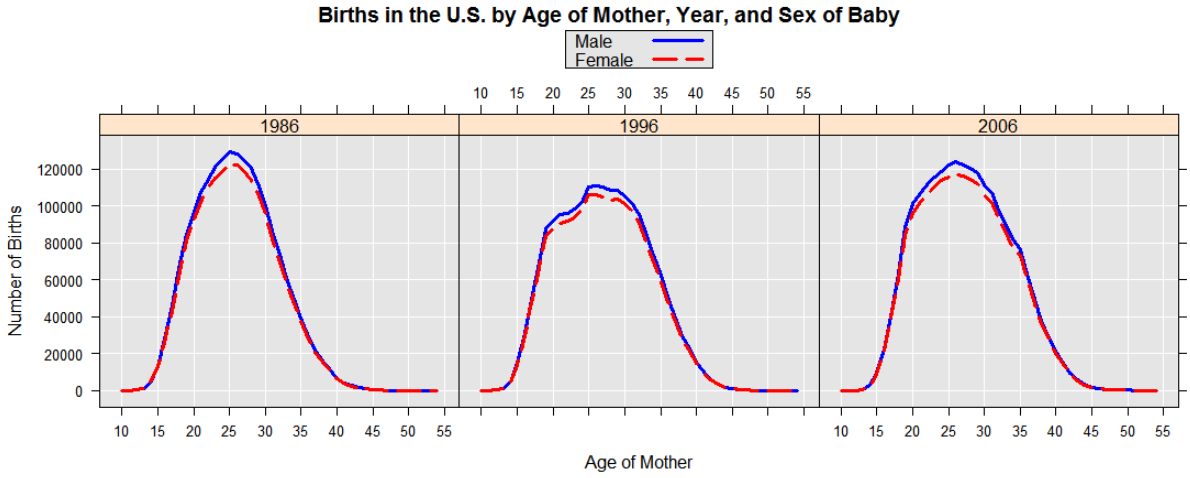
Keywords: Visualization, Data, Birth

The fact that more boys than girls are born each year is well established – across time and across cultures. But there are variations in the degree to which this is true. For example, there is evidence that the sex ratio at birth declines as the age of the mother increases, and babies of a higher birth order are more likely to be girls. Different sex ratios at birth are seen for different racial groups, and a downward trend in the sex ratio in the United States since the early 1970s has been observed. Although these effects are very small at the individual level, the impact can be large on the number of "excess" males born each year. To analyze the role of these factors in the sex ratio at birth, it is appropriate to use data on many individual births over multiple years.

Such data are in fact readily available. Public-use data sets containing information on all births in the United States are available on an annual basis from 1985 to 2008. But, as Joseph Adler points out in *R in a Nutshell*, "The natality files are gigantic; they're approximately 3.1 GM uncompressed. That's a little larger than *R* can easily process." An additional challenge to using these data files is that the format and contents of the data sets often change from year to year.

Using the **RevoScaleR** package, these hurdles are overcome and the power and flexibility of the *R* language can be applied to the analysis birth records. Relevant variables from each year are read from the fixed format data files into **RevoScaleR**'s .xdf file format using *R* functions. Variables are then recoded in *R* where necessary in order to create a set of variables common across years. The data are combined into a single, multi-year .xdf file containing tens of millions of birth records with information such as the sex of the baby, the birth order, the age and race of the mother, and the year of birth.

Detailed tabular data can be quickly extracted from the .xdf file and easily visualized using **lattice** graphics, as shown in the plot below. Trends in births, and more specifically the sex ratio at birth, are examined across time and demographic characteristics. Finally, logistic regressions are computed on the full .xdf file examining the conditioned effects of factors such age of mother, birth order, and race on the probability of having a boy baby.



References

Adler, Joseph (2010). *R in a Nutshell*.

CDC (1985-2008). Birth Data Files, http://www.cdc.gov/nchs/data_access/Vitalstatsonline.htm.

Matthews TJ, Hamilton BE (2005). Trend analysis of the sex ratio at birth in the United States, http://www.cdc.gov/nchs/data/nvsr/nvsr53/nvsr53_20.pdf.