The *DEMOGRAPHY-STATISTICS-INFORMATION TECHNOLOGY* Letter
FROM: Griffith Feeney feeney@gfeeney.com
Letter No. 8, 15 November 2014

Dot Plot or Bar Graph?

Dot plots were invented by William S. Cleveland to provide improved visualizations of the kinds of data displayed using bar graphs. The bar graph may be thought of as an adaptation of the histogram for use when the horizontal axis variable is categorical. The dot plot may be thought of as an adaptation of the scatter plot for use when the vertical axis variable is categorical.

Consider the bar graph of unemployment rates in European countries presented in a recent *eurostat* news release, reproduced in Figure 1.

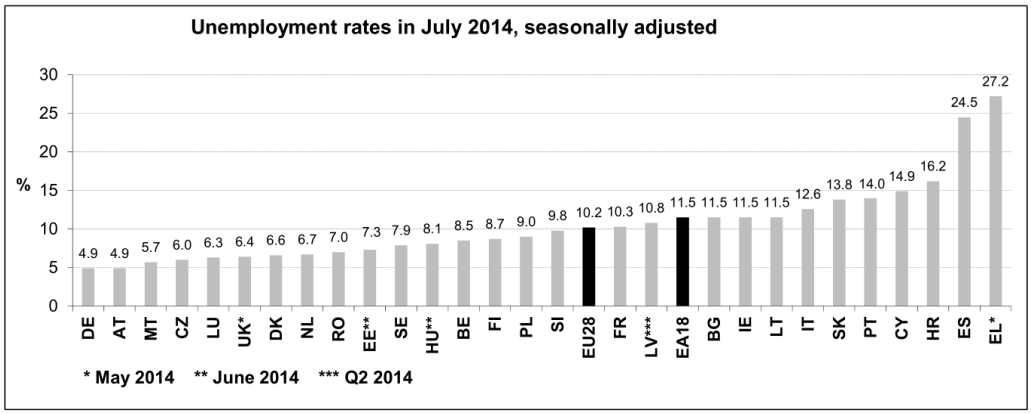

Figure 1 JULY 2014 UNEMPLOYMENT RATES IN EUROPE. Bar graph reproduced from page 1 of *eurostat* news release 129/2014 - 29 August 2014. Definitions of the two letter country codes and of EU28 and EA18 are given in a note on page 2 of the news release. The exceptions indicated by the asterisks (*) are explained implicitly by the "Seasonally adjusted unemployment, totals" table on page 3 of the news release.

The plot is meticulously constructed, but we need the two letter country codes identify individual countries. We may guess UK, FR, IT, but how many of us will know that ES is Spain, EL Greece, and HR Croatia? We can look up the codes on the following page of the news release, but this is clumsy almost to the point of defeating the purpose.

Dot plotting the unemployment rates

Figure 2 shows a dot plot of the same data. Comparison with Figure 1 shows that large, inked-in areas formed by the bars in a bar graph are superfluity. The plot marks on the reference lines in Figure 2 convey the same information with less distraction and create space that can be used to provide more information. The orientation allows country names to be displayed.

A notable feature of both plots is how the length of the unemployment rate scale required to accommodate the two countries with the highest rates limits visual discrimination of the remaining countries. Discrimination can be improved by plotting logarithms of the rates. Cleveland tends to use logarithms base 2 and to

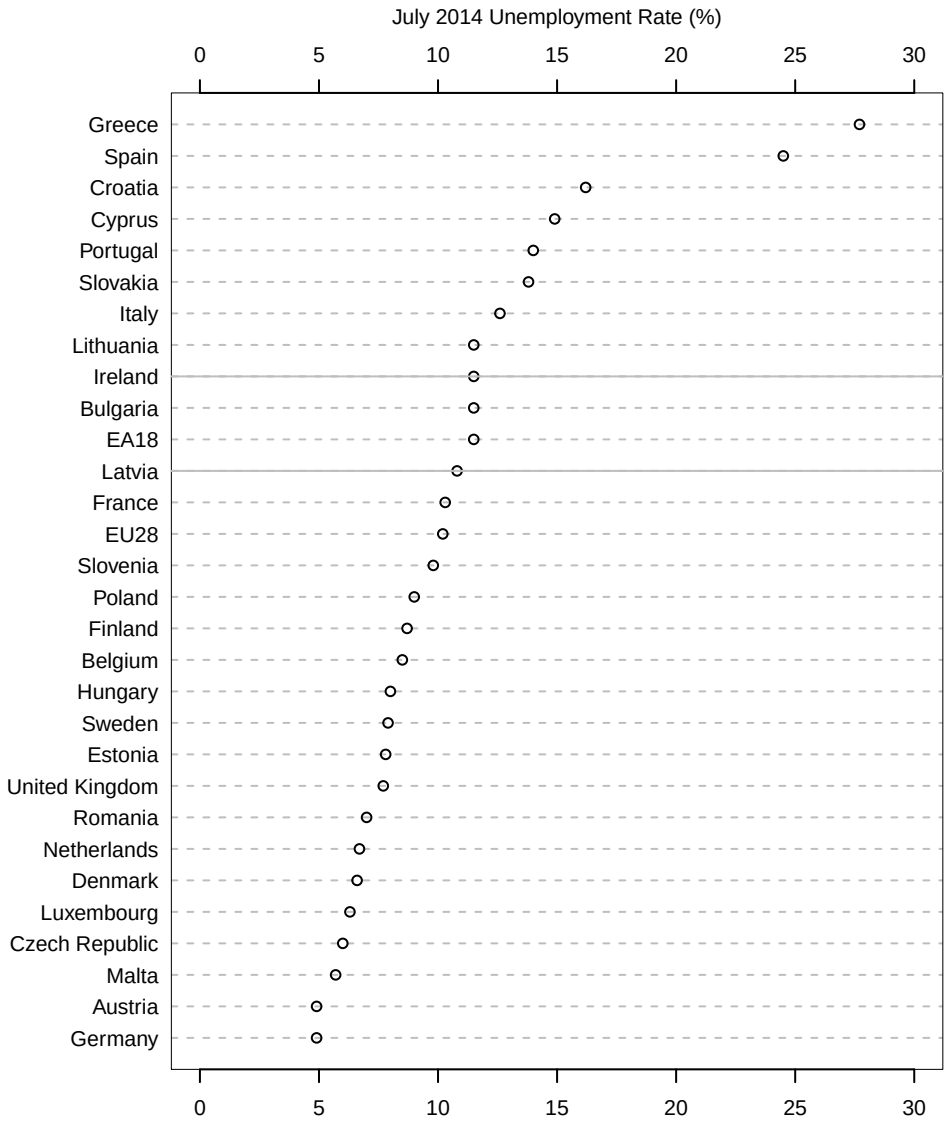

Figure 2 DOT PLOT OF THE UNEMPLOYMENT RATES. The unemployment rates in Figure 1 displayed as a dot plot. The light gray lines highlight the rates for the euro area (EA18) and the European Union (EU28). If no July 2014 rate is available for a country, the most recent available rate for the country, indicated by the asterisk notes, is shown. (* May 2014, ** June 2014, *** Q2 2014. The euro area includes Belgium, Germany, Estonia, Ireland, Greece, Spain, France, Italy, Cyprus, Latvia, Luxembourg, Malta, the Netherlands, Austria, Portugal, Slovenia, Slovakia and Finland. The European Union comprises the 28 countries shown.)

show both value and log value scales. This is useful for data analysis, but may be inappropriate for presentation of official statistics.

Adding youth unemployment rates

Figure 3 adds youth unemployment rates. Note how easily the additional variable is accommodated. Paired bar graphs can display two values for each case, but this is clumsy indeed next to the spare elegance of the dot plot.

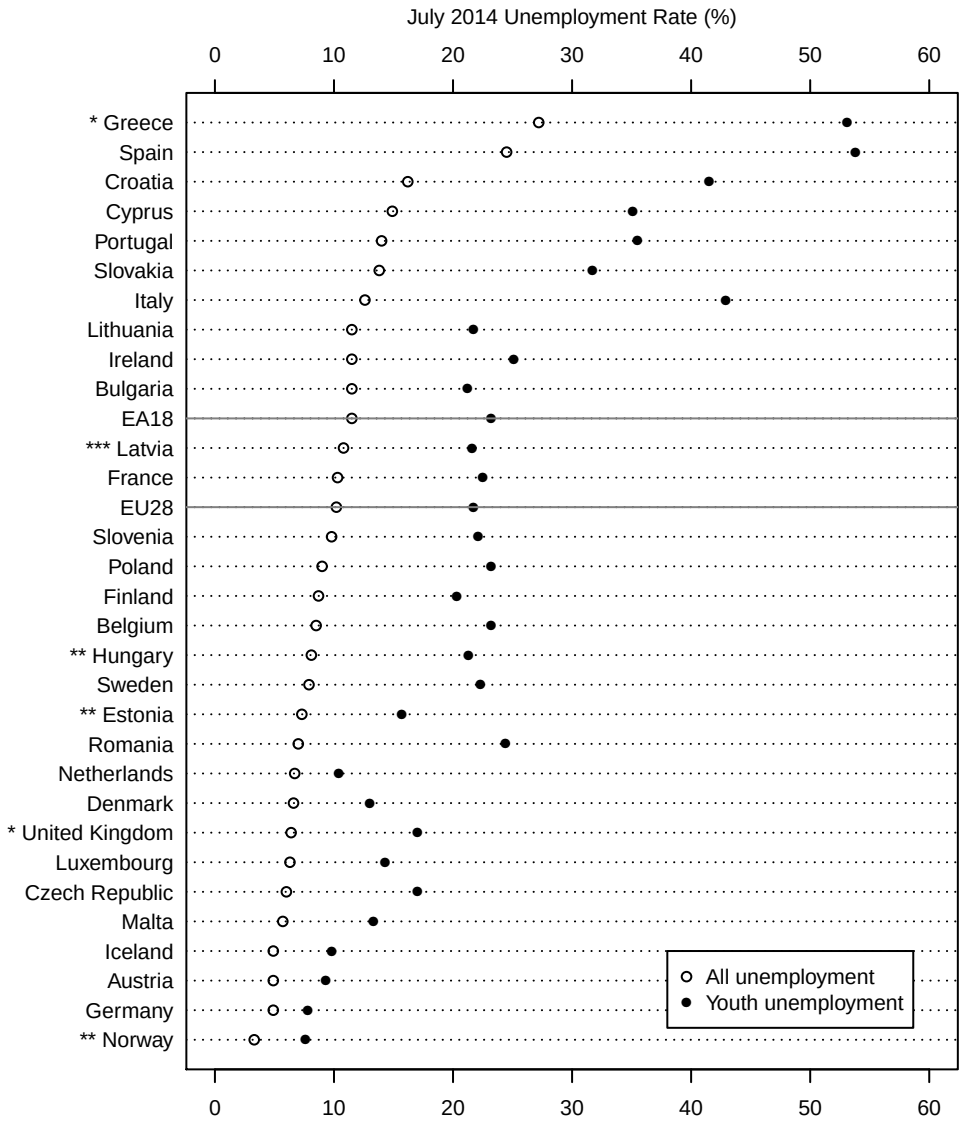

Figure 3 UNEMPLOYMENT AND YOUTH UNEMPLOYMENT IN EUROPE This dotplot adds youth unemployment rates to the overall unemployment rates in Figure 2. Note how easily dot plots accommodate two values for each named entity. Youth unemployment rates are given in the table on the fifth page of the *e*urostat news release, but no visualization is provided.

We see instantly and effortlessly that the youth unemployment rates are uniformly higher. The highest—54% for Spain and 53% for Greece—, are very high indeed. Countries with higher overall unemployment tend to have higher youth unemployment, but the relationship is strongest for the seven highest unemployment countries. Unemployment rates decline slowly as we move from Lithuania (over 10%) to Ireland to Bulgaria and so on down to Romania, but the youth unemployment rates show no evidence of a downward trend.

Total fertility for districts of Sierra Leone

Figure 4 shows total fertility estimates for the 15 districts of Sierra Leone, from

the *Demographic and Health Survey* (DHS) conducted in 2013. The grayscale bars show one and two standard error intervals.

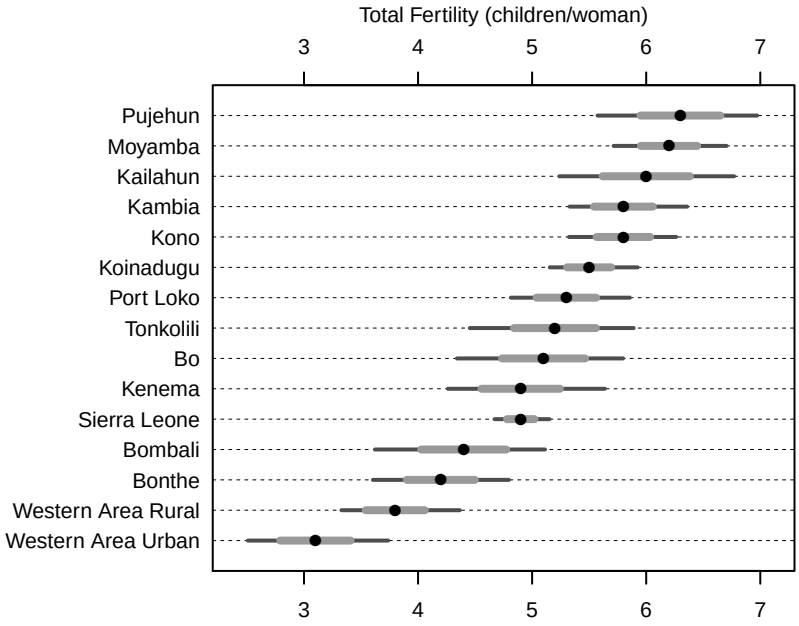

Figure 4 TOTAL FERTILITY FOR DISTRICTS OF SIERRA LEONE The gray bars show ± 1 and ± 2SE (standard error) intervals. Their length in relation to differences in total fertility indicates the limited ability of the survey to estimate total fertility at the district level. This visualization pulls together information from the table presenting the total fertility values and 15 sampling error tables in Annex B. It conveys much more information that the table presented in the report, which suggests a level of precision that does not in fact exist. (Total fertility from Table 5.2, page 65, of *Sierra Leone Demographic and Health Survey 2013* (Statistics Sierra Leone and Ministry of Health and Sanitation, Freetown, Sierra Leone, and ICF International, Rockville, Maryland, USA, July 2014). Standard errors (SEs) and upper and lower limits of 2-SE confidence intervals from pages 331-344 of Appendix B.)

The total fertility estimates are shown in Table 5.2 of the DHS report, but sampling errors are relegated to a 23 page appendix of fine print tables. Variables are not shown in alphabetical order. Finding the information for a particular variable requires a tedious search. Figure 4 thus presents information that cannot be readily extracted from the report—and information that could not easily be displayed on a bar graph.

Table 5.2 of the report shows the estimates to one digit after the decimal place, which might be read as suggesting accuracy to two significant figures. Figure 4 show that this is far from the case. The length of the ± 1SE bars approaches one child per woman and is several times larger than many differences between districts. We cannot reliably conclude, for example, that there is any difference in total fertility between the three districts with the highest total fertility.

Conclusion

Why do we see dot plots so rarely? One would like to be able to say "because they are new", but they were invented over 30 years ago.

Familiarity and habit are doubtless one reason. The Chinese have a saying, "To move mountains and change the course of rivers is easy; but to change a man's habits—this is difficult."

Then again data analysis is not always the main concern. Bar graphs can be used to put large, solid blocks of color on the page, conveying a sense of solidity suited to the glossy, four color world of annual reports.

Having a menu option for dot plots in Excel's "Chart Wizard" would surely promote promote their use. Why doesn't Excel do dot plots? Ignorance of the development team? Managerial veto—perhaps on the ground that "nobody uses them"?

Dot plots can be made with Excel (see below), and it is not very difficult, but it is more difficult than clicking on an existing menu option and let the "Wizard" do the work. Dot plots can be made easily in R (r-project.org), if you use R. But R has a steep learning curve if you have grown up using Excel.

Resources

William S. Cleveland, "Graphical methods for data presentation: Full scale breaks, dot charts, and multibased logging", *The American Statistician* **38**(4):270-280 (1981). Cleveland and Robert McGill, "Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods", *Journal of the American Statistical Association*, **79**(387):531-554 (September 1984) and "Graphical Perception and Graphical Methods for Analyzing Scientific Data", *Science*, New Series, **229**(4716):828-833 (August 30, 1985).

See also Cleveland's books *The Elements of Graphing Data*, 2nd edition, Hobart Press, Summit, New Jersey, USA (1985; 1st edition 1984, Wadsworth Advanced Books and Software, Monterey, California, USA) and *Visualizing Data*, Hobart Press, Summit, New Jersey, USA (1993).

Naomi B. Robbins's *Creating More Effective Graphs* (Wiley-Interscience, Hoboken, New Jersey, USA, 2005) references a macro by Kenneth Kline, available at ftp.wiley.com/public/sci_tech_med/graphs/, visited 13 November 2014, that facilitates producing dot plots in Excel.

R is available at r-project.org. Using R via the RStudio IDE his highly recommended. RStudio is available at rstudio.com. The installation provides an option for installing R and RStudio in a single operation.

DEMOGRAPHY, STATISTICS, and INFORMATION TECHNOLOGY is an every-other-month email letter from Griffith Feeney, Ph.D., an international consultant based in Scarsdale, New York, USA. For more information see griffithfeeneyconsulting.com.

This letter and past letters are available at demographer.com. To SUBSCRIBE, send email to feeney@gfeeney.com with "subscribe" in the subject field. To UNSUBSCRIBE, send email to the same address with "unsubscribe" in the subject field.