Statistical graphics

Making information clear – and beautiful

The obvious way to present information is in a graph. But not all graphs are created equal. A well-designed graph can make clear what an ill-thought-out one conceals. **Jarad Niemi** and **Andrew Gelman** present visualisations of a measles epidemic.

In the toolkit section of the March 2011 issue of *Significance*, Julian Champkin looked at four examples of what we would term "information visualisation", or InfoVis for short. In this issue we will discuss statistical graphics where the importance has shifted from making the graphic beautiful to making it clear. We will look at laboratory confirmed cases for the city of Harare, Zimbabwe, in the face of a measles outbreak that began in autumn 2009 and continued throughout 2010, and consider how best to present the information to the people who most need to make use of it.

Before the outbreak, measles vaccination coverage in Zimbabwe had reached 92%. In early autumn 2009, cases were reported in a number of districts in Zimbabwe. Immediately public health officials responded to these cases by collecting samples for diagnostic testing and vaccinating children in affected villages. By late May 2010, there were 7754 suspected cases, 508 laboratory confirmed cases (61 of Zimbabwe's 62 districts having at least one confirmed case), and 517 deaths. The vast majority of these cases had not been vaccinated previously. From May 24th to June 2nd the World Health Organization (WHO) conducted a mass vaccination campaign for the whole country, vaccinating more than 5 million children. As of December 12th, 2010 there were a total of 13 783 suspected cases, 693 confirmed cases, and 631 deaths. With no confirmed cases in the following months, the outbreak is assumed to have ended. Our data come from the WHO's Zimbabwe epidemiological bulletins.

Default graphs in Excel and R

Figures 1 and 2 respectively provide default plots of the time series of cumulative confirmed measles cases in Excel and in the package familiar to all statisticians, R. The graphs provide the basic outbreak information but not in a complete, concise, or visually appealing way. In the Excel version, the *x*-axis is too busy – the reader must make quite an effort to decode the labels and figure out the timescale involved. It is far from obvious that the epidemic occupied a period of 13 months. The *y*-axis is unlabelled – are its numbers cases or deaths? The default action adds a legend – "Series 1" – that has an uninformative title and is unnecessary when only one series exists. On the positive side,

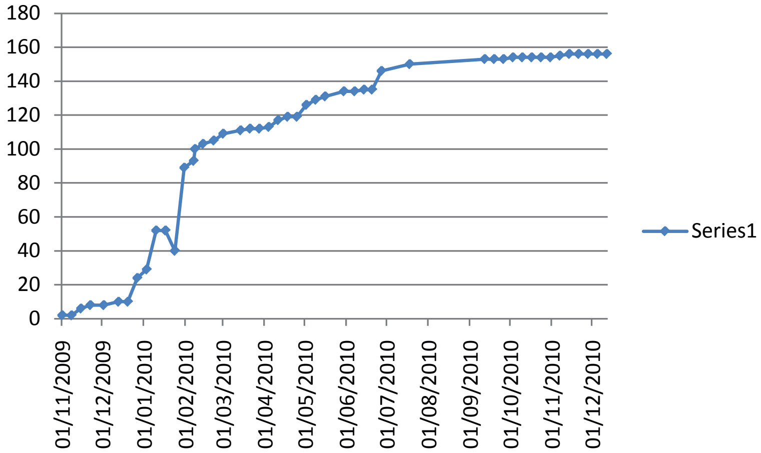

Figure 1. Default creation in Excel of the number of confirmed cases in Harare, Zimbabwe, during the measles outbreak that began in November 2009

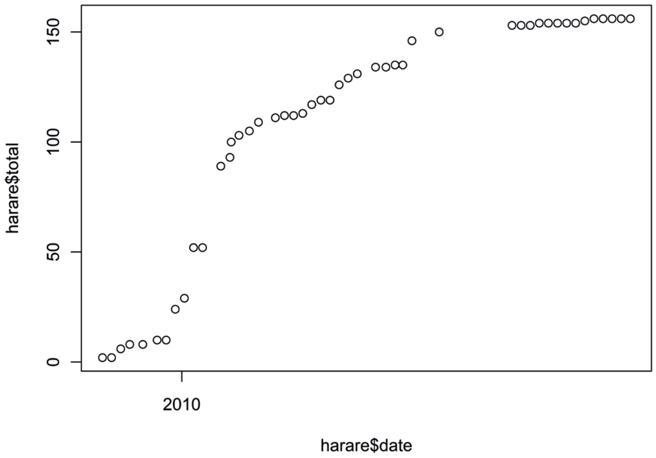

Figure 2. Default creation in R of the number of confirmed cases in Harare, Zimbabwe, during the measles outbreak that began in November 2009

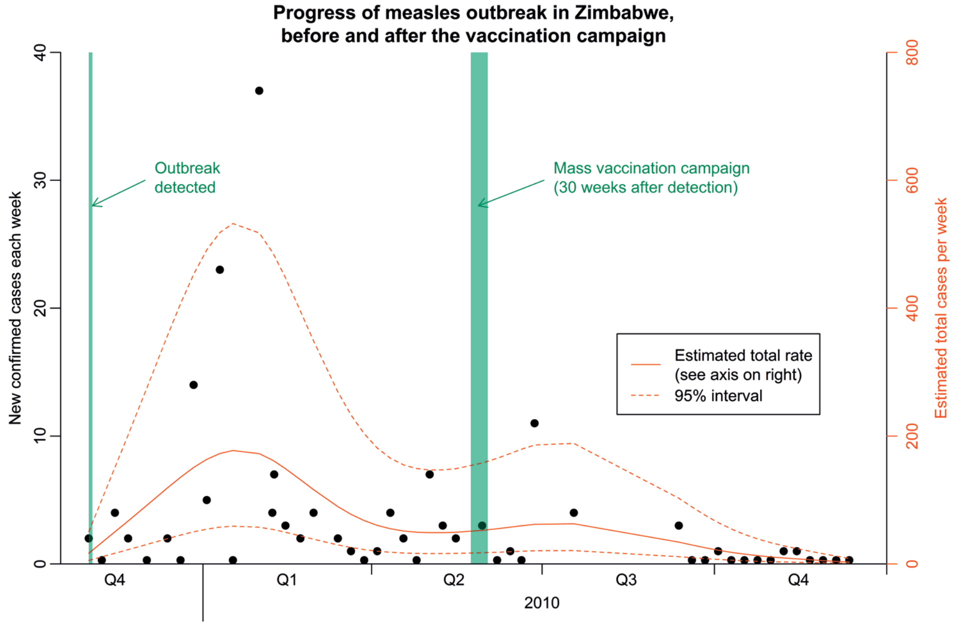

Figure 3. Improved figure displaying the number of confirmed cases (black) and estimated total cases (red) in Harare, Zimbabwe, during the measles outbreak that began in November 2009 and the mass vaccination campaign (green) that was held from May 24th to June 2nd

creating a graphic in Excel requires the user to choose a chart type; our choice was a "marked line plot" and the default blue colour is visually appealing. In contrast, the default graph in R produces a terrible point chart with only one *x*-axis tick demarcating the start of 2010 and only slightly helpful labels. If it was hard to grasp the timescale in Excel, it is impossible in R.

The Excel and R graphs share an even more serious problem, which is that they are labelled and scaled to fill an entire screen. The time series presented here is uncomplicated and could easily fit on a small portion of the page, if the labelling were sized appropriately.

Constructing a more beautiful and informative summary of data and inferences

Our purpose in showing these examples is not to denigrate Excel or R, but rather to point out that using program defaults, while useful for exploratory data analysis, does not produce production-quality figures. When producing quality figures, every decision needs to be made consciously and with intent. Improving figures requires two key decisions:

- Who is your target audience?
- What are you trying to show?

Once these decisions have been made, they will guide your figure creation choices.

In the Harare measles outbreak example, we are interested in showing the timing of the WHO's mass vaccination campaign relative to the outbreak in order to provide public health officials with a retrospective view of the policy decision. Now public health officials are accustomed to visualising outbreaks as the number of infected individuals at a given time, and this statistic is generally more useful to them than the total number of cases; this suggests plotting the number of new confirmed cases rather than the cumulative cases. In addition, we can subdivide the public health officials into those who are interested in this particular outbreak and those who are interested in lessons learned from this outbreak that will inform future policy. The former would likely be interested in the actual dates of this outbreak, while the latter would likely be more interested in the number of weeks since disease discovery. Finally, although understanding the total number of confirmed measles cases is interesting, a more relevant quantity is the estimated total cases and its associated uncertainty.

A few hours of trial and error in R yielded Figure 3, our improved version of the confirmed cases data, along with estimates and uncertainties for total cases, and mass vaccination campaign dates. The figure seamlessly incorporates both confirmed and estimated total cases with uncertainty by adding an additional *y*-axis for the estimated cases. The mass vaccination campaign is indicated with a solid green box, labelled with the time since detection to address the different needs of public health officials, some of whom will be interested in exact dates and others who care about the general time pattern.

Initially, we thought the message of this plot would be that public health officials had vaccinated too late as the main peak of the outbreak occurred much earlier, but after creating the figure we were surprised to see the second peak shortly after the vaccination campaign. Perhaps this is due to the associated public awareness campaign and more people reporting measles cases, but the three data points immediately after the campaign belie that hypothesis. So perhaps a second peak was on its way, but was stymied by the vaccination campaign. In fairness to public health officials, the campaign included all of Zimbabwe, and Harare was one of the earliest outbreak locations.

Most of this figure development was accomplished by trying to put ourselves in the shoes of public health officials and determine what might be important to them, but we used some guiding principles along the way:

- Avoid distracting elements.
- Use informative colour.
- Keep the figure simple.

As an example of avoiding distracting elements, we had initially used a marked line plot, similar to the Excel plot earlier, for the new confirmed cases data – in other words, we joined up the black dots – but the resulting lines had a saw-tooth pattern that dominated the figure without providing meaningful information. Joining the dots served no purpose. We used red, a colour representing emergency or danger, to combine all the figure components related to estimated cases, and green, a colour representing earth and health, to combine the elements associated with the vaccination campaign.

We think Figure 3 is a much improved data summary – it conveys information as well as containing it – but we are under no illusion that it is perfect. In particular, we are not happy with the potentially confusing double *y*-axis and we suspect there is a clearer way to simultaneously display confirmed counts on one scale and estimated totals on the other. At the level of process, the graph was awkward to create. We made countless tweaks to our R script and hard-coded various details such as the positions of the labels, the scaling of the axes, the long tick mark dividing 2009 from 2010, and a slight shifting of the zero points so they would be clear of the *x*-axis (which we wanted to be precisely at zero – contrary to the R default – so make it visually apparent that the estimated rate declines to zero at the end of the time series). Some manual adjustments may always be necessary to prepare a complex presentation-quality graph, but R (as currently configured) is not the ideal environment for this process.

Further graphs for data exploration

Given our curiosity about the timing of the vaccination campaign relative to the outbreak, it is tempting to put the data for multiple cities into Figure 3, but we decided this would unnecessarily complicate the figure.

Instead, the concept of small multiples can be used to compare similar plots, as shown in Figure 4. This figure shows the original Harare data along with that from Bulawayo (the second largest city in Zimbabwe), and Mashonaland (a region in northeastern Zimbabwe). We have chosen to align the plots vertically to emphasise the relative outbreak peaks. If emphasising relative peak heights was of interest, a horizontal series of plots would have been more useful. From this figure it is clear that Harare and Bulawayo had outbreak peaks at approximately the same time, while Mashonaland's peak was later. This suggests the outbreak moved out from cities to individuals in Mashonaland.

In order to create these small multiple plots, we utilised another set of guiding principles:

- Keep the *x*- and *y*-axes on the same scale.
- Eliminate repetitive information.
- Maintain consistency across plots.

The scales for the three axes are repeated and so the figures are immediately comparable for outbreak timing and intensity. For example, Bulawayo had many fewer cases than Harare. For these graphs we decided to remove some of the detail in the axes, as they would unnecessarily complicate the small multiples. In addition to the *x*- and *y*-axes on the small scale, we have also maintained the outbreak estimates and 95% intervals as well as the colour scheme in order to maintain consistency across the plots.

In this way, Figures 3 and 4 work together as a story, and in a way contrary to the usual practice of statisticians. We are generally taught to start with exploratory plots and then move towards presentation-quality graphs. But in this case the visually attractive overview sets the stage for more focused data exploration.

We realise these figures have not been as eye-catching as the InfoVis figures in the March issue of *Significance*, but they serve different purposes. InfoVis essentially gives an outline view, sometimes imprecise, lacking in detail. We hope these figures illustrate how a graphical display of data can be a useful tool in understanding the measles outbreak in Zimbabwe.

We envision a future in which multilayer interactive graphics will become commonplace. The public layer would be an InfoVis figure to catch the reader's attention. The reader could dive a layer deeper to obtain a statistical graphic such as those displayed in this article. Diving one more layer down would produce the data table from which both the statistical graphic and the InfoVis were produced. Each layer would be interactive, allowing the reader to present the data in the way they felt most appropriate for their needs. In this way, we view InfoVis and statistical graphics as complementary tools for understanding data.

Acknowledgements

The authors would like to thank the National Science Foundation, National Security Agency, and Institute of Education Sciences for partial support of this work.

Jarad Niemi is at the Department of Statistics and Applied Probability, University of California, Santa Barbara. Andrew Gelman is at the Department of Statistics and Department of Political Science, Columbia University.

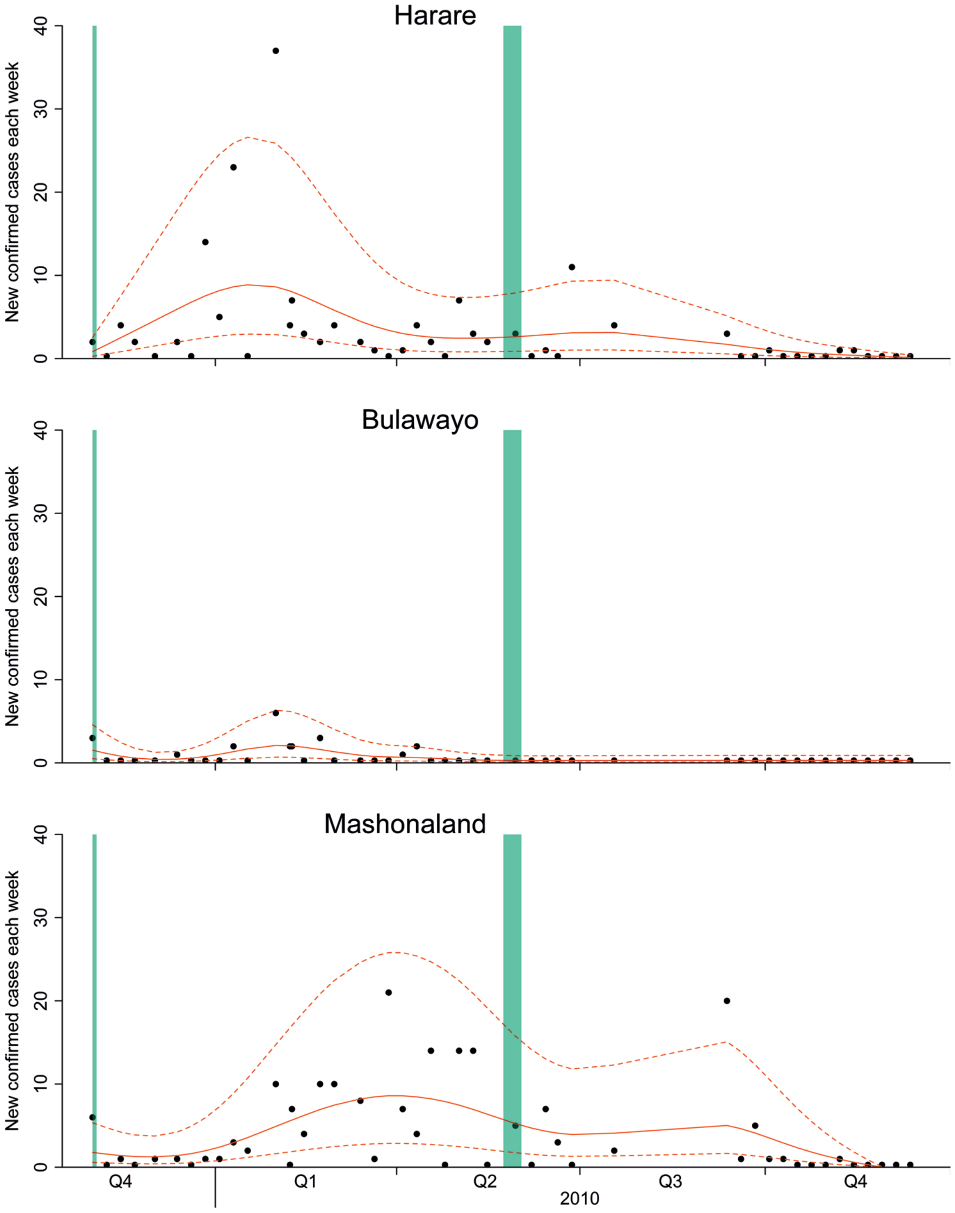

Figure 4. Showing the concept of small multiples in order to compare outbreak profiles in Harare, Bulawayo, and Mashonaland with the number of confirmed cases (black) and estimated total cases (red) during the measles outbreak that began in November 2009 and the mass vaccination campaign (green) that was held from May 24th to June 2nd