Show the dots in plots

We encourage our authors to display data points in graphs, and to deposit the data in repositories.

'Let the data speak for themselves', the saying goes. Yet in literal terms, this rarely applies to collections of raw numbers, which can often be difficult to interpret. In fact, different presentations of the same dataset can suggest different interpretations. The type of graph, its dimensions and layout, colour palettes and gradients, the data intervals displayed in the axes, specific data comparisons, and above all, the presence or absence of individual data points, error bars and information on statistical significance, can strongly affect how the graphed dataset is interpreted.

An often misused type of visual representation is the bar graph. Bar graphs display data according to categories. However, they are also commonly used to present small samples of continuous data, especially in biomedical fields. There are reasons for this: because of their shape and area, bars are easy to see at a glance; therefore, they are effective when comparing data and visualizing trends; and they make it easy to see the relative position of the data along the axes. However, bar graphs can be misleading. For example, using bars to replace and summarize non-independent sets of data obscures any

patterns across the datasets. Also, different data distributions can generate the same exact bar graph; for example, bimodal and normal distributions, as well as distributions with outliers and unequal number of data points can lead to the same mean and variability. Moreover, providing only statistical parameters (such as mean ± s.d. or mean ± s.e.m., and number of samples) can suggest that the data underlying any particular bar are normally distributed and contains no outliers, when this may not be the case. Graphing error bars with the s.e.m. (which indicates the precision of the mean) is commonly done because they are shorter than error bars representing the s.d. (which instead quantifies variability). We discourage this practice.

All these issues can be avoided by displaying every data point. This journal strongly suggests that the individual data points (in addition to error bars and other statistical information) be graphed, in particular for relatively small samples and for bar graphs, and when statistical significance is claimed. However, when the number of samples is large (typically

when more than 100), scatter plots become crowded, hindering visualization and interpretation of the data. In such cases, box-and-whisker plots are preferable.

As with all other Nature Research journals, we also recommend to authors that the data in the figures and any supplementary datasets be deposited in a public repository (such as figshare; https://figshare.com). The advantages are many: data deposition provides easy access to colleagues who wish to further analyse or make use of the data, increases reporting transparency, encourages the eventual reproducibility of the findings, ensures data preservation, increases the overall usability of datasets, especially when they are large, and enables convenient citation to the data (with a doi). As per the data availability policies of the Nature Research journals, all Articles in this journal include a data availability statement, specifying the availability of the minimal dataset that would be necessary to interpret, replicate and build on the methods or findings reported in the Article (for guidance, see http://go.nature.com/2bf4vqn).

Data presentation should not be an afterthought; the visuals affect how the story is told and perceived. Display items in papers should highlight the relevant data and make their interpretation easy. Considerations such as which comparisons within the dataset are most relevant to the story, how they can be made clear to the reader, and how to display data scatter without hindering interpretation often need careful thought. Bar graphs can make comparisons easier to see at a glance, even for continuous variables when categorized (Fig. 1); yet the individual data points should be displayed. As the grouped bar graph overlaid with a dot plot in Fig. 1 illustrates, the data points themselves show the scatter in the data and allow for additional comparisons (for example, between same-month data points across activities within a time interval). Colours can also be used to best effect (for example, because in Fig. 1 differences in the means for flights of stairs climbed are not meaningful, the relevant bars appear white). Figure captions should be clear, provide all the necessary statistical information, and guide the reader through the story. In fact, an engaging narrative is structured, showcases the protagonists, and provides relevant context. In bar graphs, the data points are the context. Show them. ❑
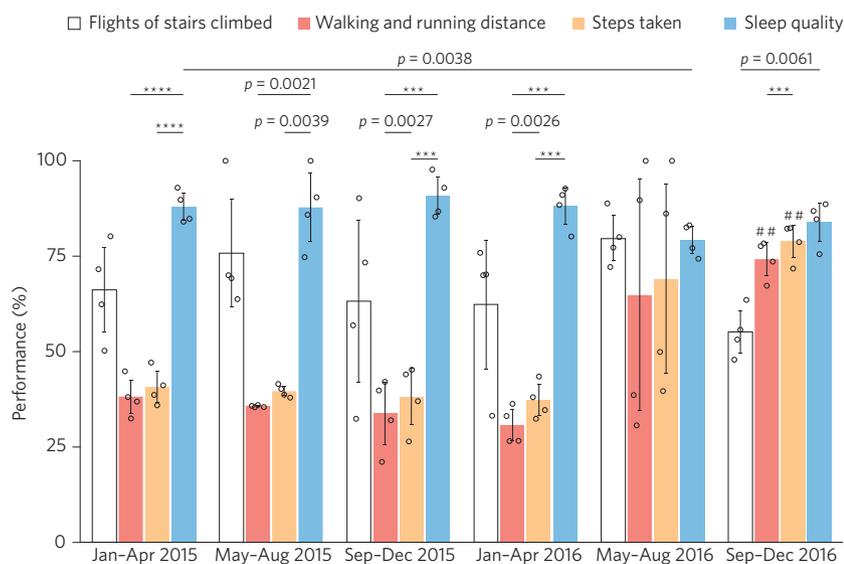
Figure 1 | An individual's monthly activity and sleep quality between January 2015 and December 2016 (data points), ordered from January to December, categorized according to four-month intervals, and normalized by the respective maxima within the two years. Error bars, mean ± s.d. All differences between means with $p < 0.01$ are indicated (within the same category and across categories). ##, $p < 0.01$, for Sep–Dec 2016 and any time interval before May 2016; ***, $p < 0.001$; ****, $p < 0.0001$; two-tailed paired t-tests. Data (available at doi:10.6084/m9.figshare.4928888) courtesy of Pep Pàmies (this journal's Chief Editor), and collected via the iOS Health app.

1