Untidy Long Paper Example

Christina Bergmann[1] & The R Unicorn[2]

[1] Ecole Normale Superieure

[2] Institute for Rainbow Studies

Author Note

The authors note that it is quite cold today.

Correspondence concerning this article should be addressed to Christina Bergmann, 29, rue d'Ulm. E-mail: chbergma@gmail.com

Abstract

Enter abstract here (note the indentation, if you start a new paragraph).

Keywords: no keywords

Word count: 666

Untidy Long Paper Example

This sample paper shows how you can combine long texts and code to generate dynamic journal papers. The following text is taken from http://barbarplots.github.io and from http://cogtales.wordpress.com. (Note: Adding http:// to URLs makes them automatically click-able.)

The paper was generated using R (R Core Team, 2016), tidyverse (Wickham, 2016), and papaja (Aust & Barth, 2016) for formatting the paper in APA style. You can create citations for R packages using the citation() command, for example: `citation("papaja")`.

Look, it's a subsection!

Data visualization is a complex topic in the experimental sciences. While there are many ways to display data, many researchers choose to use bar plots. Generally, these plots only depict a group mean and standard error (or deviation). Unfortunately, most data are not as clean as bar plots make them seem, and since bar plots reveal very little about the distribution of the data, this kind of visualization can be misleading (Saxon, 2015; Weissgerber, Garovic, Savic, Winham, & Milic, 2016; Weissgerber, Milic, Winham, & Garovic, 2015). A further issue is that of the bar itself, which implies that the base of the y-axis is meaningful, which is not necessarily the case. The bar can then mislead readers (Saxon, 2015).

More on the barbarplots campaign. For the full text with all figures, visit: https://cogtales.wordpress.com/2016/06/06/congratulations-barbarplots/

Our kickstarter project #barbarplots reached its funding goal and will thus become reality! In the 30-day campaign, 173 backers pledged a total of 3,479 Euro to send #barbarplots t-shirts to editors of major scientific journals. We are very excited and want to thank you for the tremendous support – not only by pledging, but also by spreading the word via email, Facebook, Twitter, and by wearing and carrying tote bags and t-shirts with the following meme around the world.

We also want to thank everyone that joined the discussion on #barbarplots during the 30 days of our campaign. These discussions happened on people's Facebook pages, in the lab, via email, and on Twitter. To not lose all these thoughts widespread along the www, I here try to put together a collection of discussion points – both critical thoughts about the campaign itself as well as musings on data visualization in general. Note that many snippets of answers I borrow from our British t-shirt doctor Rory, who not only in our video, but also in real life proved to be the most eloquent #barbarplots spokesman.

Before going into it, I want to draw your attention to some of the the shoulders this campaign is standing on, for instance Weissgerber et al. (2015), who started the plotting revolution much earlier.

So off we go – I'll start with the maybe most obvious critical remark about our campaign.

But barplots can be a good way to plot!. Sure. That's why we used the following barplot for the cost breakdown of our kickstarter project to demonstrate that count data with a meaningful zero and no distributional properties can be wonderfully represented by a barplot.

This explanation, however, quite readily leads to critique number 2:

But why do you say #barbarplots if you do not actually mean it??. Some people pointed out that our hashtag #barbarplots, our slogan "Friends don't let friends make barplots" and our campaign video with the cats-and-dogs example were catchy but partially misleading.

That's true. But we made a conscious decision in favor of catchiness and simplicity in order to increase the likelihood people would raise their head and smile and share. Researchers take their work seriously and details are important, but they're also humans with an often already overstrained attention span. And they're also smart enough to look beyond the (attention-catching) headline to ponder the more nuanced argument.

We are aware that some people might still be misled by the simplified version of our

message, and could even religiously stop using barplots no matter what. But this is, for sure, not a problem that our slogan is causing.

What are cases in which barplots are not good and why?. In general, we consider barplots not to be the most informative way to represent distributional data. For example, perhaps an effect is driven entirely by a subset of participants in the experiment? Perhaps a null effect arises because some participants have a negative effect and others have a positive effect? Perhaps some items are extremely variable? Perhaps the data are very non-normal and using inferential statistics is inappropriate? And even if your data are perfectly normally distributed, wouldn't it make your paper even more convincing if your visualization method reflected that?

Ok, so what you want to say is that we should not be comparing means?. It is true that the issue of barplots as a data visualisation technique often goes together with the issue of doing statistics by comparing means. However, they are nevertheless separate issues. The focus of the campaign is really about barplots as a visualization technique. We just want to make sure that we think about the choice rather than formulaically applying one way of analysis or visualization.

So... what IS a better plot to represent distributions?. In our campaign, we promote box plots and histograms. Both tell us way more about the distribution of data than barplots, notably about the spread and skewness of the underlying data. And there's more: Scatter plots, violin plots, swarm plots, pirate plots. Below is an example of how the same data would look with different types of plots.

Though this is not the case in the above graphs, the way we plotted in our original meme lead some people to say:

Your histogram has two x axes!! That is very misleading!. It is true that the histogram differs from the other plots in this aspect. Sure, we could have superimposed the distributions so that there was only one x axis (but then the plot would differ from the others in the sense that it would not be a side-by-side but an overlaid representation). We

could have also rotated the histogram 90 degrees to make the y axes match (but then it would differ from the others by being vertically as opposed to horizontally aligned). In short: Yes, there would have been other ways to do it, but it is not clear that there was a clearly better way to do it. Also, we do not want to promote boxplots or histograms as a single alternative to barplots (see above) – all we want is for us to think about our choices.

Still, we want to emphasize that a barplot is often not the most informative alternative for distributional data. That leads me to the last frequent point of criticism:

Isn't a plot supposed to be an abstraction of the data and a means to present them in the simplest and most understandable way possible?. To that, I would first want to state the obvious that simple isn't always good. Like interpreting everything below p = .05 as a victory and above as an epic failure. It's not good, it's not right, and part of its being simple is certainly convention and habitude.

But still, a summary/abstraction/simplification of the raw data cannot be a bad idea to bring your point across? This question gets at the heart of what visualization is for and what scientific publishing is supposed to do. If the visualization is to be as simple as possible to get the most basic point across, maybe just showing means is fine. (To be honest, though, if you're just reporting two means, don't waste space with a graph, just put that in the text.) However, it's conventional for us to go into a lot of obsessive detail in our reporting of methods and statistical analyses – so why not with our graphs too? Showing or summarizing the distribution can be a useful way to put your cards on the table for readers to see that you are not trying to hide or obfuscate any unusual outliers or strange trends.

But of course, if we're giving a 10-min-talk in front of an audience that is used to barplots, walking the audience through the graph would require more time than you have.

Luckily, there are many compromises that can be made – for example, plot a grand mean, along with (less visually prominent) subject means, so that the variability in the data is more apparent. Or a boxplot, which is visually less dense than other alternatives to barplots.

In summary...

Thank you again to all of our backers, tweeters, and reposters. We've really enjoyed working on this campaign, from filming the video to group discussions on and off social media. The campaign is by no means over though, so look for updates as we complete our aim to send out t-shirts to journal editors and continue the discussion. Happy plotting!

Your #barbarplots team,

Page Piccinini,

Christina Bergmann,

Sho Tsuji,

Alexander Martin,

Rory Turnbull,

Adriana Guevara Rukoz,

M. Julia Carbajal

Methods

The data I am using here comes with R and is called Anscombe's Quartet.

Wikipedia description of Anscombe's Quartet

source: https://en.wikipedia.org/wiki/Anscombe's_quartet

Anscombe's quartet comprises four datasets that have nearly identical simple descriptive statistics, yet appear very different when graphed. Each dataset consists of eleven (x,y) points. They were constructed in 1973 by the statistician Francis Anscombe to demonstrate both the importance of graphing data before analyzing it and the effect of outliers on statistical properties. He described the article as being intended to attack the impression among statisticians that "numerical calculations are exact, but graphs are rough." (Anscombe, 1973)

The quartet is still often used to illustrate the importance of looking at a set of data

Table 1

Descriptive statistics of Anscombe's quartet.

key	mean	sd
y1	7.500909	2.031568
y2	7.500909	2.031657
y3	7.500000	2.030424
y4	7.500909	2.030579

Note. This table was created with apa_table()

graphically before starting to analyze according to a particular type of relationship, and the inadequacy of basic statistic properties for describing realistic datasets

Results

Descriptive Statistics

We're now preparing a table by creating a data structure that has exactly the fields we want to display, plus column headers. For advanced users, you can also modify the column headers and other parts of how this table turns out. The second table is showing the same data, but instead of using the APA format, it is generated using kable.

Let's talk a bit about the statistics. So it looks like all means and sds are quite similar. That's part of the point Anscombe (1973) was making.

key	mean	sd
y1	7.500909	2.031568
y2	7.500909	2.031657
y3	7.500000	2.030424
y4	7.500909	2.030578

Data Visualization

First, because this paper is based on a #barbarplots talk I gave last summer, here is of course one of those odious barplots! Look, all datasets look the same! Quelle horreur!

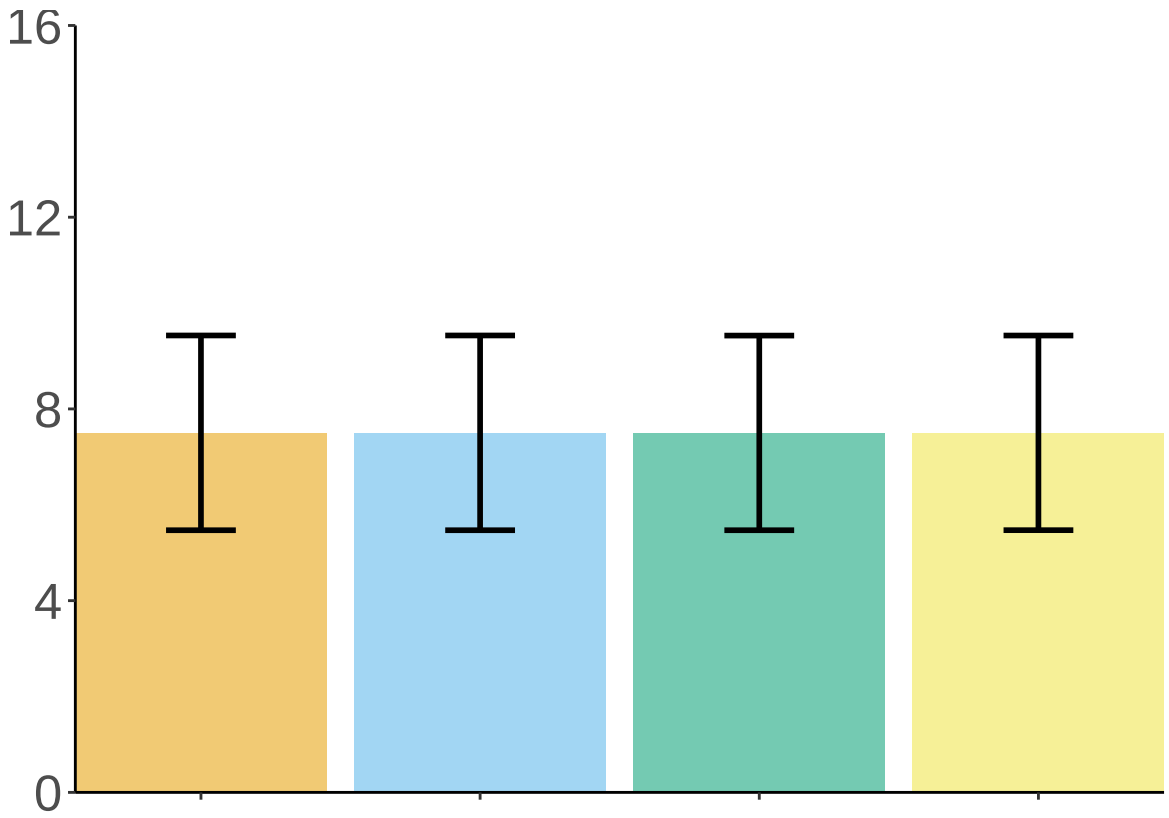

Figure 1. Barplots of one dimension of Anscombe's Quartet

Next, we want to look at some boxplots. Look, some differences between the datasets become visible.

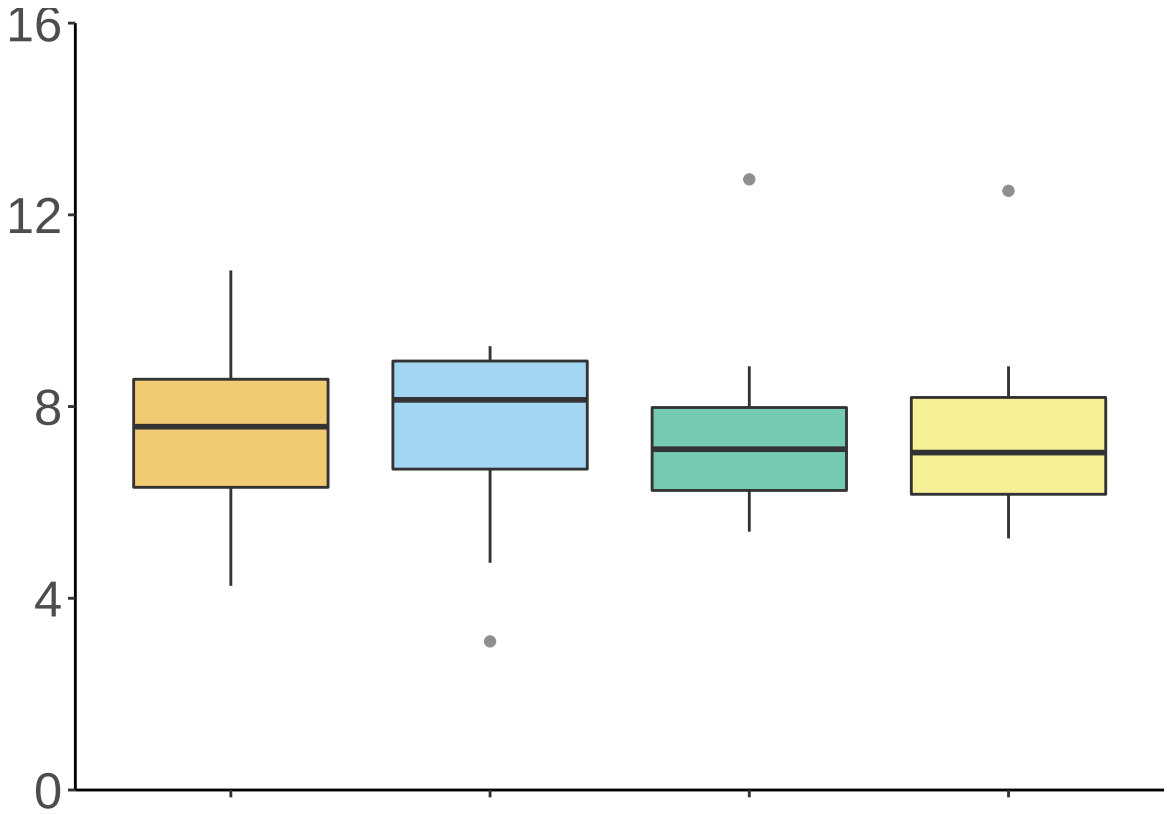

Figure 2. The same data as before in a boxplot.

Plotting both dimensions of the famous Quartet

The first scatter plot (top left) appears to be a simple linear relationship, corresponding to two variables correlated and following the assumption of normality. The second graph (top right) is not distributed normally; while an obvious relationship between the two variables can be observed, it is not linear, and the Pearson correlation coefficient is not relevant (a more general regression and the corresponding coefficient of determination would be more appropriate). In the third graph (bottom left), the distribution is linear, but with a different regression line, which is offset by the one outlier which exerts enough influence to alter the regression line and lower the correlation coefficient from 1 to 0.816 (a robust regression would have been called for). Finally, the fourth graph (bottom right) shows an example when one outlier is enough to produce a high correlation coefficient, even though the relationship between the two variables is not linear.

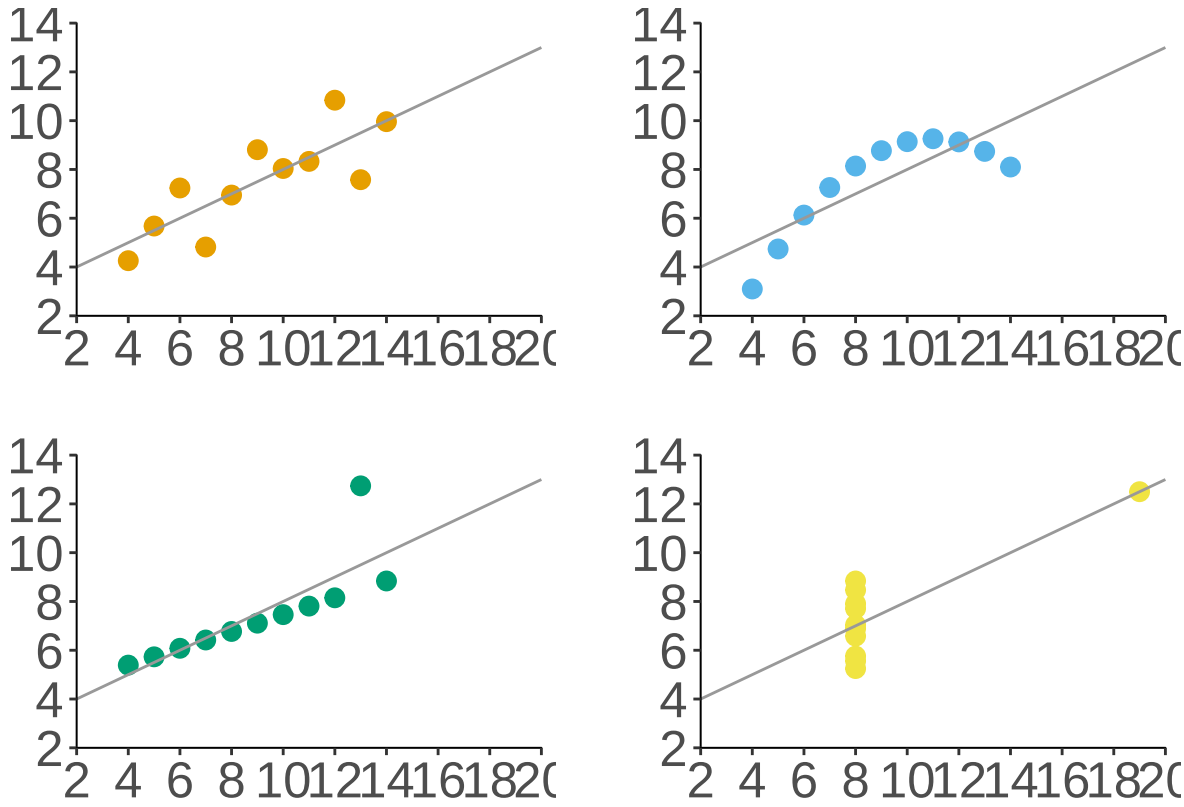

Figure 3. The classical Anscombe Quartet plot. All means and correlations are the same, but the raw data differ quite a bit.

Let's do some stats

```
##
## Call:
## lm(formula = anscombe$x3 ~ anscombe$y3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9869 -1.3733 -0.0266  1.3200  3.2133
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)  -1.0003     2.4362  -0.411  0.69097
## anscombe$y3   1.3334     0.3145   4.239  0.00218 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.019 on 9 degrees of freedom
## Multiple R-squared:  0.6663, Adjusted R-squared:  0.6292
## F-statistic: 17.97 on 1 and 9 DF,  p-value: 0.002176

##
## Call:
## lm(formula = anscombe$x4 ~ anscombe$y4)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.7859 -1.4122 -0.1853  1.4551  3.3329
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -1.0036     2.4349  -0.412  0.68985
## anscombe$y4   1.3337     0.3143   4.243  0.00216 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.018 on 9 degrees of freedom
## Multiple R-squared:  0.6667, Adjusted R-squared:  0.6297
## F-statistic:    18 on 1 and 9 DF,  p-value: 0.002165
```

As suggested by the plots, the linear models are very similar (the correlations even

more so). We can also report results in line, for example the p-value = 0.00. Note that this number is really not correct, because it is rounded to two digits!

References

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, *27*(1), 17–21.

Aust, F., & Barth, M. (2016). *Papaja: Create apa manuscripts with rmarkdown*. Retrieved from https://github.com/crsh/papaja

R Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Saxon, E. (2015). Beyond bar charts. *BMC Biology*, *13*(1), 60.

Weissgerber, T. L., Garovic, V. D., Savic, M., Winham, S. J., & Milic, N. M. (2016). From static to interactive: Transforming data visualization to improve transparency. *PLoS Biology*, *14*(6), e1002484.

Weissgerber, T. L., Milic, N. M., Winham, S. J., & Garovic, V. D. (2015). Beyond bar and line graphs: Time for a new data presentation paradigm. *PLoS Biology*, *13*(4), e1002128.

Wickham, H. (2016). *Tidyverse: Easily install and load 'tidyverse' packages*. Retrieved from https://CRAN.R-project.org/package=tidyverse