Memo

To: Colleagues
From: Roger Bakeman
Date: May 20, 2020
Re: How to Create Box-and-Whisker Plots with Excel

Box-and-whisker plots (Tukey, 1977) have two uses. First, in the preliminary stages of data analysis, they can give us a clear picture of how variable values are distributed, alerting us to potential problems—only a bubble plot, indicating each data point with a circle on the y-axis, would be superior. Second, they are often superior to conventional methods when providing descriptive information in research reports. The main impediment to their use is the lack of an easy-to-use and widely available computer program to produce them. Here I explain how to produce publishable box-and-whisker plots in Excel.

As an example, consider two variables—mother and child mean length of utterance (MLU) in words and number of different words (NDW) collected at two time point (Years 1 and 2). Conventionally, descriptive information is presented in a table (we have all done this).

Table 1. *Descriptive Statistics*

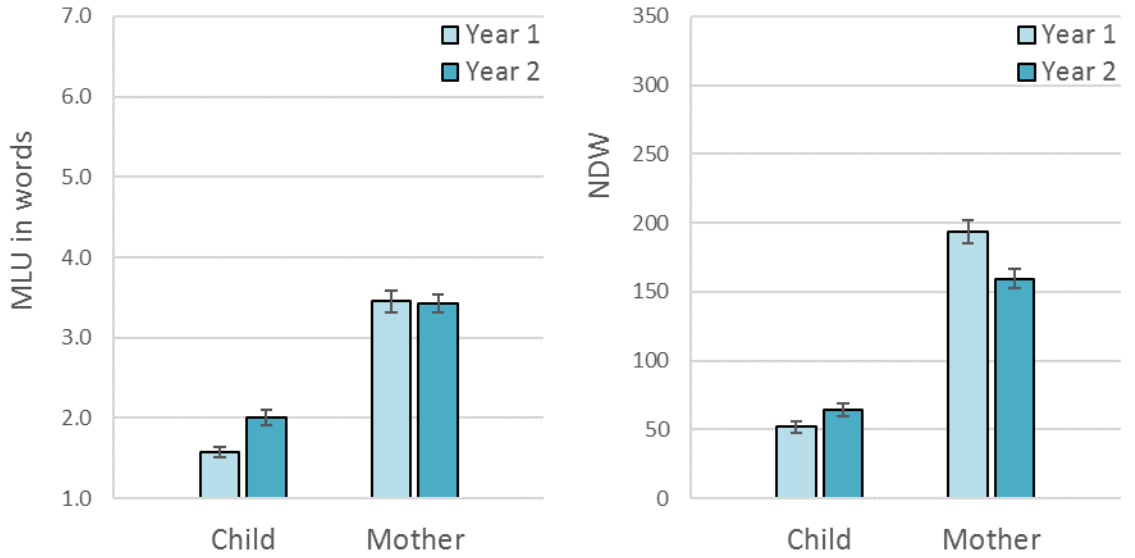
	Year 1			Year 2		
Variable	*M*	(*SD*)	range	*M*	(*SD*)	range
Child						
MLU	1.57	(0.35)	1.00–2.48	2.01	(0.49)	1.11–3.22
NDW	52	(24)	2–107	64	(25)	9–140
Mother						
MLU	5.68	(0.74)	2.04–5.68	3.42	(0.64)	2.00–6.01
NDW	336	(46)	79–336	160	(38)	65–247

Note. *N* = 119 (from the Dallas Language Project; thank you M. O. Caughy and M. T. Owen).

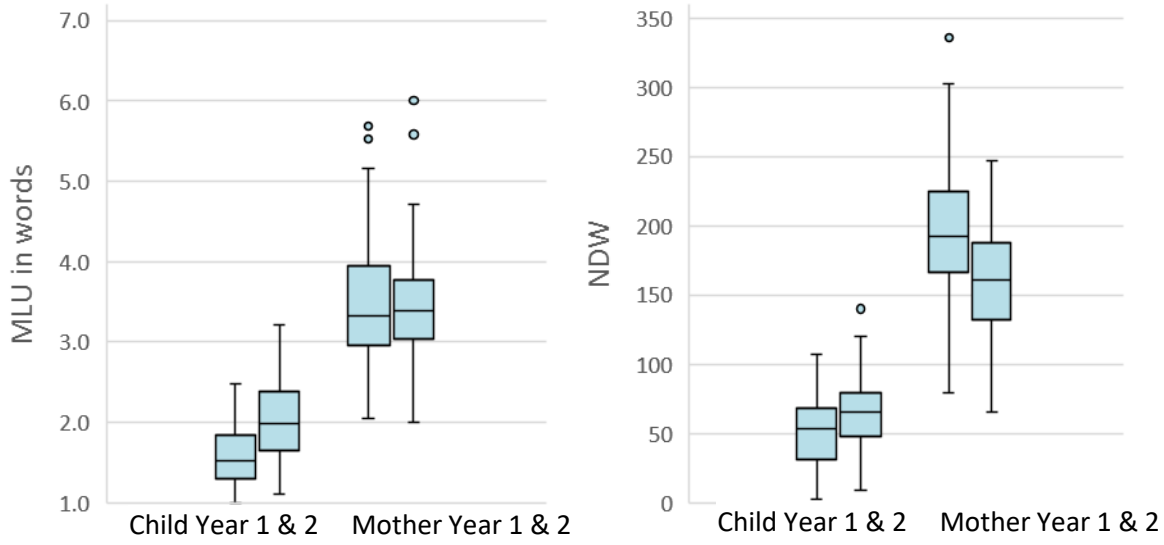
The *SD* is conventional in such table but, I think, not very helpful—although if unduly large or discrepant from other variables with similar ranges, it can suggest problems. Providing the range helps, but often is not included in table like this one.

An alternative to the *SD* is the 95% confidence interval (CI), which I have used occasionally—but I'm not sure it is much better. It extends an equal distance below and above the mean, and so provides no information about whether a distribution is skewed or not. It is often used in bar plots instead of the standard error of the mean (SEM), but it has the same problem there as well. The only merit of the 95% CI for bar plots is that it is about twice as large as the SEM.

Here are conventional bar plots for our example data:

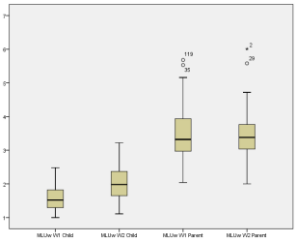


But now consider what box-and-whisker plots looks like for the same data:



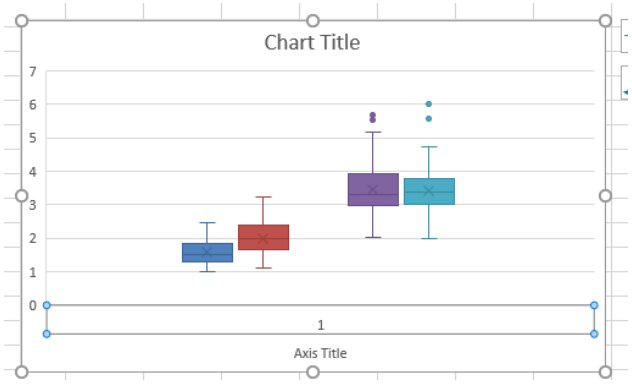
The bar plots—here error bars are 95% CIs—extending as they do to the x-axis, can give a mistaken impression of range. The box includes scores from the 25th to the 75th percentile. The whiskers indicate the lowest and highest scores that are not extreme. Extreme scores, defined as any 1.5 times the interquartile range below the 25th or above the 75th percentile, are indicated with circles.

In sum, unlike standard bar plots, box-and-whisker plots indicate the range, whether the distribution is skewed and how clumped it is, and whether some high scores, low sores, or both are extreme. They are especially helpful when variables are being compared as here, mother to child and Year 1 to 2.

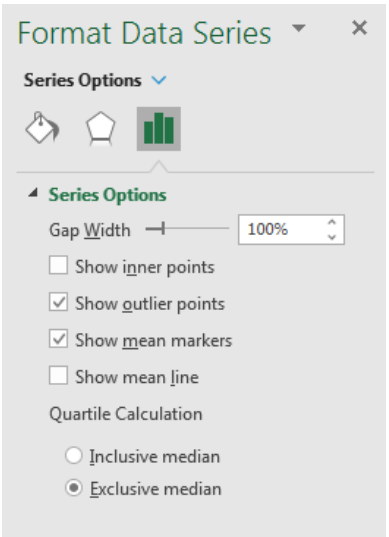


Standard programs do not produce box-and-whisker plots easily. SPSS is fine for a preliminary look but the plots aren't publication quality (at the right is an example). There may be programs that work—I haven't used SigmaPlot since the mid-60s—but I haven't found one yet. Here I used Excel, which can produce good-looking plots but can be a bit annoying and quirky. The box-and-whisker procedure lacks many of the capabilities of other Excel plot types. Here is my experience.

To begin, select a range of values for your variables. They can be organized in columns with the name of the variable in the first row. Then select Recommended Charts on the Insert Ribbon, then the All Charts tab, and then Box & Whisker. For the MLU plot above, I selected five columns; the third was blank go give the spacing between the child and mother plots. (Limitation #1: No way to control spacing between boxes as there is with bars.)

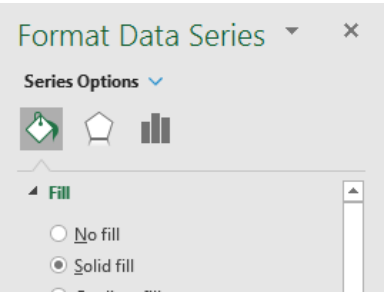


You would expect your variable names (if in the first row) to be displayed—as they are with other Excel plots—but they aren't. Therefore select Add Chart Element on the Design ribbon, then Axis Titles > Primary Horizontal. To the right is what this should look like. Note the 1 above the Axis Title text box. Select it (so the box around it shows) and delete it. (I suspect it was intended for x-axis labeling but was never implemented in the code.) Now you can edit the horizontal axis text box to say whatever you want. You can also add a Primary Vertical axis text box as well, as I did for the examples on the previous page. As with any text box, you can change the font, font size, etc. You can also edit the Chart Tile—or delete this text box as I did for my examples.



Limitation #2: Each box is a separate series. This means each has to be formatted separately—unlike a standard bar graph which allows a series—i.e., all bars in it—to be formatted together. Worse, the defaults aren't my preferences. If you select a box, right click, and select Format Data Series, you will see the dialog box on the right. All you want checked is Show outliers points. Uncheck any other check boxes (leave Exclusive median—more on that in a later memo).

Now select the paint-can icon.



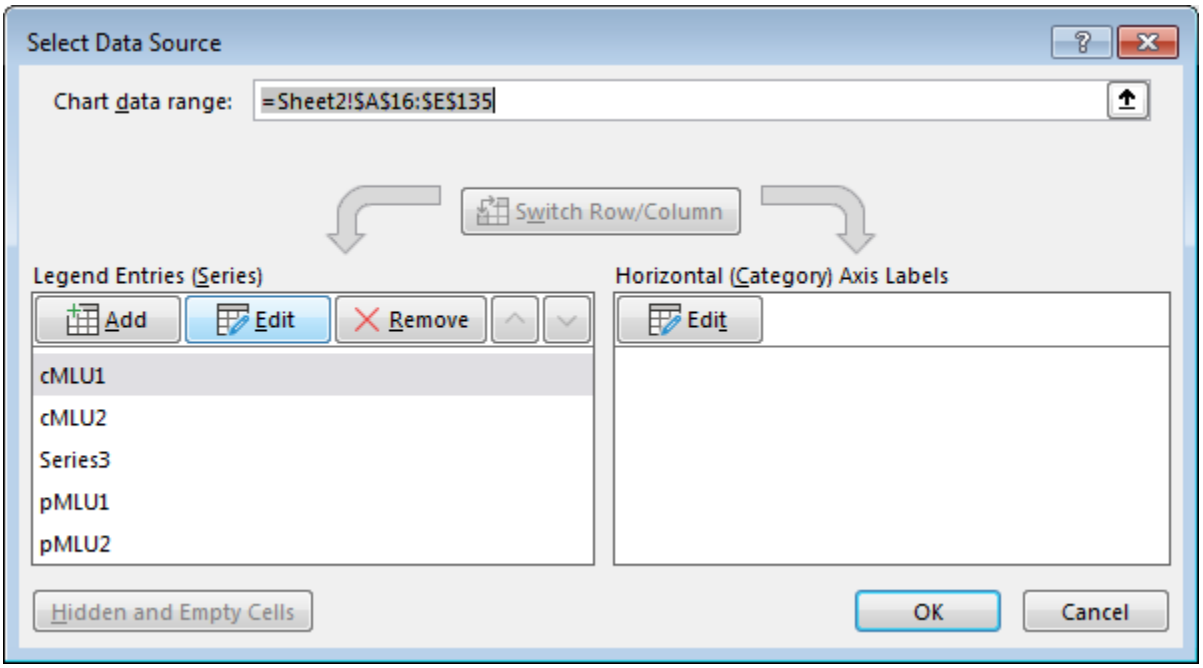
Change Fill to Solid fill and select whatever color you want inside the box. Then change Border to Solid line and select whatever color you want for the line around the box and the whiskers.

Unfortunately, you will need to repeat for each box.

Limitation #3: You can change the Minimum and Maximum values for the y-axis (select the axis, right-click, and select Format Axis), but you can't define a major and minor value as with other Excel plots. All you can change are the tick marks and the format for the numbers.

Limitation #4: You can't control the space between the y-axis and the first box-and-whisker. This space becomes larger, the more box-and-whisker plots you place on the x-axis.

Hint #1: One you have created a box-and-whisker figure that you like, it may be easier to copy and paste that figure and then edit it, instead of selecting a new data range and beginning with the Insert ribbon. Select the figure, right click, select Select Data, and then edit the Legend Entries in the Select Data Source dialog box—see example below. You can Remove an entry (the data for one of the box-and-whisker plots) and add new ones—but only in the last position.
Limitation #5: The ability to shift positions of entries—the up and down buttons—is disabled here, as is the ability to edit (or even specify) horizontal axis labels.



It's possible that the Excel box-and-whisker procedure has more capabilities than I know. If you find something I have missed, be sure to let me know. My hope is that this memo gives you enough information to create box-and-whisker plots with Excel without becoming too frustrated at first.