CENTER FOR FAMILY & DEMOGRAPHIC RESEARCH

Making Charts & Tables

Fall 2017 Workshop Series

Krista Payne, PhD

- **Order of Presentation**
  - Introduction—Why Charts?
  - Structure of an Excel Chart
  - Different Types of Charts
  - Basic Principles of Design
  - Getting tables from Stata → Excel → Word
  - Charting Interaction Effects

- **Why Charts?**

  "Designing good charts, however, presents more challenges than tabular display as it draws on the talents of both the scientist and the artist. You have to know and understand your data, but you also need a good sense of how the reader will visualize the chart's graphical elements."

  ~Gary Klass

- **The picture superiority effect**
  - Information is better remembered in tests of recall and item recognition when presented as pictures rather than words (e.g. Durso & O'Sullivan, 1983; Gehring, Toglia & Kimble, 1976; Juola, Taylor & Young, 1974; Paivio & Csapo, 1973)

- **The structure of an Excel chart:**
  - **Title**: In academic writing, the title should be used to define the data series, without imposing a data interpretation on the reader. Often, the units of measurement are specified at the end of the title after a colon or in parentheses in a subtitle (e.g. "constant dollars", "% of GDP", or "billions of US dollars").
  - **Legends**: used in charts with more than one data series. Placement is key—they should not be placed on the outside of the chart in a way than reduces the plot area, the amount of space given to represent the data.
    - I will (at times) place the legend to the right of the figure—typically when using vertically stacked bars, b/c I feel they help the reader decipher the information more quickly.
  - **Axis**: The scale of the axis represents the value or magnitude of the main graphical elements of the chart. It is defined by the axis scale. Try to make the scale consistent across figures within your single document.
  - **Axis Titles**: Axis titles should be brief and should not be used at all if the information repeats what is clear from the title and/and axis labels.
  - **Data Labels**: Be sure they are a clear representation of your data. For example, if the dependent variable is hourly wage, I advise using two decimal points. However, if it is yearly income, round to whole dollars.
  - **Gridlines**: If used, they should employ as little ink as possible. You don't want to distract from the main graphical elements in your chart.

- **Source**: Specifying the source of the data is important for proper academic citation, and can also give those "readers in the know" important insights into the reliability and validity of the data. For example, knowing that crime statistics come from the FBI rather than The National Criminal Victimization Survey can be a crucial bit of information.

  Sources:
  http://lilt.ilstu.edu/gmklass/pos138/datadisplay/sections/goodcharts.htm
  http://cas.illinoisstate.edu/jpda/index.shtml
  Companion website for the book: *Just Plain Data Analysis: Finding, Presenting, and Interpreting Social Science Data*, written by: Gary Klass. (New York: Rowman and Littlefield Publishers, 2012). Second Edition. ISBN: 978-1-4422-1508-5

The different types of charts

- **Histograms**: A vertical bar chart that depicts the distribution of a set of data
  - The bars represent the frequency of occurrence by classes of data.
  - Use when you want to see the shape of the data's distribution.
  - Similar to a bar chart, but a histogram groups numbers into ranges.
  - The horizontal axis is continuous like a number line (no gaps between the "columns").
  - Great way to show results of continuous data: weight, height, how much time, etc…
  - A frequency histogram uses vertical columns to show how many times each score occurs.

  Source:
  http://www.mathsisfun.com/data/histograms.html

- **Pie Charts**: Generally used to show percentage or proportional data classified into nominal or ordinal categories.
  - Data that is arranged in one column or row only on a worksheet can be plotted in a pie chart. Pie charts show the size of items in one data series, proportional to the sum of the items. The data points in a pie chart are displayed as a percentage of the whole pie (Excel 2010 help).
  - Pie charts are useful for displaying data that are classified into nominal or ordinal categories.
    - Nominal data are categorized according to descriptive or qualitative information such as county of birth or type of pet owned.
    - Ordinal data are similar but the different categories can also be ranked, for example in a survey people may be asked to say whether they classed something as very poor, poor, fair, good, very good.
  - Pie charts have the following chart subtypes:
    - Pie or pie and bar of pie: display pie charts with user-defined values that are extracted from the main pie chart and combined into a secondary pie chart or into a stacked bar chart. These chart types are useful when you want to make small slices in the main pie chart easier to distinguish.

- Exploded pie: display the contribution of each value to a total while emphasizing individual values. You can change the pie explosion setting for all slices and individual slices, but you cannot move the slices of an exploded pie manually.

Guidelines for using pie charts: Do NOT use pie charts if possible—it is more difficult for the eye to discern the relative size of pie slices than it is to assess relative bar length.
If you MUST use one…

1. ONLY use for data that add up to a meaningful total
2. 3 categories or fewer
3. If there is no other meaningful order, order the slices from maximum to minimum
4. Put "other" in a gray slice
5. Don't use a legend, just label the slices
6. Avoid forcing comparisons across more than one pie chart

Sources:
Excel 2010 help
http://libweb.surrey.ac.uk/library/skills/Number%20Skills%20Leicester/page_53.htm
http://lilt.ilstu.edu/gmklass/pos138/datadisplay/sections/goodcharts.htm
http://www.excelcharts.com/blog/10-x-10-tips-to-improve-your-excel-charts-pie-charts/

- **Column & Bar Charts**: Useful for showing data changes over a period of time or for illustrating comparisons among items. In column charts, categories are typically organized along the horizontal axis and values along the vertical axis.
  - There are chart subtypes:
    - Clustered column: Allows user to compare values across categories. Use when you have categories that represent:
      - Ranges of values (e.g. item counts)
      - Specific scale arrangements (e.g. a Likert scale)
      - Names that are not in any specific order (e.g. item names, geographic names, names of people).
    - Stacked column: show the relationship of individual items to the whole, comparing the contribution of each value to a total across categories.
    - 100% stacked column: compare the percentage that each value contributes to a total across categories.
  - Consider using a bar chart when:
    - The axis labels are long
    - The values show are durations

Source:
Excel 2010 help

- **Line Charts**: Ideal for showing trends over time. Line charts can display continuous data over time, set against a common scale when category data is distributed *evenly* along the horizontal axis, and all value data is distributed *evenly* along the vertical axis. Considered one of the most efficient ways of displaying large amounts of data in ways that provide for meaningful analysis.
  - There are subtypes—"line with markers, staked, 100% stacked"—but I don't find these to be very useful.
  - You can also create them with a double-axis. Good for showing time trends of two different types of variables over the same time point (e.g. marriage rates and state spending on marriage programs). The second axis doesn't have to be another line-chart (e.g. age at first marriage and proportion unmarried births).

  Guidelines for Line Charts:
  1. Almost ALWAYS, time is displayed on the X-axis from left to right
  2. Display as much data with as little ink as possible
  3. Be sure your audience can easily distinguish the lines for separate data series
  4. Beware of *scaling effects*—when two variables with numbers of different magnitudes are graphed on the same chart, the variable with the large scale will generally appear to have a greater degree of variation; the smaller-scale variable will appear relatively "flat" even though the percentage change is the same.
  5. If displaying monetary data over-time, use inflation-adjusted data
  6. Don't use a legend—directly label the series
  7. Spacing between markers should be proportional—if you can't get them to "act right" use a scatter plot instead
  8. If you can't easily see the pattern of each series you may have too many

  Sources:
  Excel 201 help
  http://lilt.ilstu.edu/gmklass/pos138/datadisplay/sections/goodcharts.htm
  http://office.microsoft.com/en-us/excel-help/creating-xy-scatter-and-line-charts-HA001054840.aspx
  http://www.excelcharts.com/blog/10-x-10-tips-to-improve-your-excel-or-not-charts-line-charts/

- **Scatter Plots**: Commonly used to show the relationship between two variables e.g. correlation (from Excel help).
  - A scatter chart has two value axes, showing one set of numeric data along the horizontal axis (x-axis) and another along the vertical axis (y-axis). It combines these values into single data points and displays them in irregular intervals, or clusters.

- Consider using a scatter chart when:
  - You want to change the scale of the horizontal axis
  - You want to make the axis a logarithmic scale
  - *Values for horizontal axis are NOT evenly spaced*
  - There are many data points on the horizontal axis
  - You want to compare many data points without regard to time—the more data you include, the better the comparisons you can make. You want to show similarities between large sets of data instead of differences between data points.

  Guidelines for Scatterplots:
  1. Use two interval-level variables
  2. Fully define the variables with the axis titles
  3. If there is an implied causal relationship between the variables, place the independent variable (the one that causes the other) on the X-axis and the dependent variable (the one that may be caused by the other) on the Y-axis
  4. Use empty circles as markers to let the reader see the overlapping points

  Sources:
  Excel 2010 help
  http://lilt.ilstu.edu/gmklass/pos138/datadisplay/sections/goodcharts.htm
  http://www.excelcharts.com/blog/10-x-10-tips-to-improve-your-excel-or-not-charts-the-scatter-plot/

- **Area Charts**: Show percentage or proportional data classified into nominal or ordinal categories over time. Area charts emphasize the magnitude of change over time, and can be used to draw attention to the total value across a trend.
  - Use these in moderation. Fall victim to the same visualization problems as pie- and doughnut-charts…pies and doughnuts are *also* **area charts**! We avoid them "whenever possible because visual perception in humans can only compare areas as rough estimates" (Few, 2012; p275).
  - The primary issue with area charts is our brains tend to read them as line charts—disregarding the "area" aspect of the chart. The example here isn't as problematic as a regular ole area chart because it is a 100% stacked chart.

Charts to be Cautious of…

1. Doughnut charts
2. Area Charts
3. Radar Charts
4. Circle Charts
5. Unit Charts
6. Funnel Charts
7. Waterfall Charts

- **Basic Principles of Chart Design**
  - Two problems arise in charting that are less common when data are displayed in tables.
    1. Poor choices, or deliberately deceptive, choices in graphic design can provide a distorted picture of numbers and relationships they represent.
    2. A more common problem is that charts are often designed in ways that hide what the data might tell us, or that distract the reader from quickly discerning the meaning of the evidence presented in the chart.

  - Each of these problems is illustrated in the two classic texts on data presentation:
    1. Darrell Huff's *How to Lie with Statistics* (1994)
    2. Edward Tufte's *The Visual Display of Quantitative Information* (1983)
       - I found these helpful when convincing UGs the importance of taking a sociology course.
       - Tufte's unique contribution to art of chart design was to stress the virtue of efficient data presentation.

1. Simplify

- Sort data in a meaningful way
- Minimize ink-to-data ratio
  - The most basic and fundamental rule for Tufte… minimize or eliminate anything from a chart that does not help the viewer understand what the numbers mean.

  "*Just as the purpose of any statistic is to simplify, to represent in one number a larger set of numbers, the purpose of a chart is to simplify numerical comparisons: to represent in several numerical comparisons in a single graphic.* "

- Remove unneeded chart elements
  - Gridlines
  - Chart borders
  - Axis titles—if you have a good title, might not need
  - Legends—can you directly label your chart? If you use a legend, placement is key
  - Markers & data labels—you don't need for ALL data points. Use the last, lowest, highest, or other relevant data points, only.
  - Decimal points (in axis & data labels)
  - Trend lines
  - NO 3D charts!!!

2. **Color vs. Black & White**
   - When in doubt → black & white—use different line types and pattern fill, but be careful with the pattern fill. It can make data labels hard to see.
   - Color can help tell a story
   - Color = branding (e.g. CFDR, NSFMR, BGSU)
     - Use a cohesive and consistent color palette
     - Be mindful of how audience will view
       - Excel vs. Word vs. PDF
       - Color vs. B&W print copy

3. **Do NOT Use Distorted Charts**
   - Do NOT misrepresent your data!
   - Use appropriate and consistent axis and scales

4. **Present Related Charts Simultaneously**
   - One-after-another or side-by-side if possible
   - Emphasizes importance of appropriate axis and scales

5. **Know your Audience**
   - Academic vs. lay folks
   - Undergraduate students vs. graduate students
   - Graduate students vs. professors
   - PAA presentation vs. job talk

6. **TMC = TMI**
   - Too many charts (TMC) is as bad as too much information (TMI) → do NOT overload your audience!

7. **Ask yourself—do you need a chart?**