Statistical Graphics Considerations

Dawn Koffman
Office of Population Research
Princeton University
May 2017
Statistical Graphics Considerations

Why this topic?

“Most of us use a computer to write, but we would never characterize a Nobel prize winning writer as being highly skilled at using a word processing tool.

Similarly, advanced skills with graphing languages/packages/tools won’t necessarily lead to effective communication of numerical data.

You must understand the principles of effective graphs in addition to the mechanics.”

“... quantitative visualization is a core feature of social-scientific practice from start of finish. All aspects of the research process from the initial exploration of data to the effective presentation of a polished argument can benefit from good graphical habits.

... the dominant trend is toward a world where the visualization of data and results is a routine part of what it means to do social science.”

But ... for some odd reason

“ ... the standards for publishable graphical material vary wildly between and even within articles – far more than the standards for data analysis, prose and argument.

Variation is to be expected, but the absence of consistency in elements as simple as axis labeling, gridlines or legends is striking.”

Why??

maybe this is changing ...

design of statistical graphs was the topic of (at least) 3 articles in academic journals in 2014:

Kieran Healy and James Moody,
Data Visualization in Sociology,
Annu. Rev. Sociol. 2014 40:5.1-5.5.

Nicolas P Rougier, Michael Droettboom, and Philip E. Bourne,
Ten Simple Rules for Better Figures.
PLOS Computational Biology. September 2014.

Jonathan A. Schwabish,
An Economist’s Guide to Visualizing Data,
What is a statistical graph?

“A statistical graph is a visual representation of statistical data.

The data are observations and/or functions of one or more variables.

The visual representation is a picture on a two-dimensional surface using symbols, lines, areas and text to display possible relations between variables.”

A statistical graph allows us to...

- **see the big picture**
  Graphs reveal the big picture: an overview of a data set.
  An overview summarizes the data’s essential characteristics, from which we can discern what’s routine vs. exceptional.

- **easily and rapidly compare values**
  Graphs make it possible to see many values at once and easily and rapidly compare them.

- **see patterns among values**
  Graphs make it easy to patterns formed by sets of values.
  For example, patterns may describe correlations among values, how values are distributed, or how values change over time.

- **compare patterns among sets of values**
  Graphs let us compare patterns found among different sets of values.

From Steven Few, Perceptual Edge: [http://www.perceptualedge.com/blog/?p=1897](http://www.perceptualedge.com/blog/?p=1897)
primary goals of a statistical graph

- explore and understand data by accurately representing it
- allow viewer to easily see comparisons of interest (including trends)
- communicate results in a clear and memorable way

how to do this is somewhat subjective ...

- few hard and fast rules
- many trade-offs
- many guidelines which some may disagree with
- iterative process is often helpful
  ... design and “build” multiple version of “same graph”

purpose of this workshop is to encourage you to consider

- techniques
- guidelines
- tradeoffs

and then to determine what *you* think makes the most sense for your particular case
One more thing ... a disclaimer ...

Let me state clearly ... I intend no criticism of graph authors, either individually or as a group.

Shortcomings show only that we are all human, and that under the pressure of a large, intellectually demanding task like designing and building a statistical graph it is much too easy to do things imperfectly. Additionally, many design considerations involve trade-offs, where there may be, in fact, no “best” solution.

Lastly, I have no doubt that some of the “better graphs” I show will provide “bad” examples for future viewers – I hope only that they will learn from the experience of studying them carefully.

Workshop Organization

Preface  Why this topic?

What is a statistical graph?

I  Introduction

Tables vs graphs  (When)

Audience and setting  (Where)

II  Representing data  accurately

III  Highlighting  comparisons  of interest  (How)

IV  Simplicity and  clarity

V  Summary

VI  Conclusions
Tables vs Graphs

tables: look up individual, precise values

graphs: see overall distribution (shape, pattern) of data
make comparisons
perceive trends
often more useful when working with large sets of data
A graph trying to also serve as a look-up table ...

Percent Annual Increase in National Health Expenditures (NHE) per Capita vs. Increase in Consumer Price Index (CPI), 1980-2012

A much nicer way to show a graph and table ...

Average Monthly Temperatures (°F)

<table>
<thead>
<tr>
<th></th>
<th>Jan</th>
<th>Feb</th>
<th>Mar</th>
<th>Apr</th>
<th>May</th>
<th>Jun</th>
<th>Jul</th>
<th>Aug</th>
<th>Sep</th>
<th>Oct</th>
<th>Nov</th>
<th>Dec</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phoenix</td>
<td>52.1</td>
<td>55.1</td>
<td>59.7</td>
<td>67.7</td>
<td>76.3</td>
<td>84.6</td>
<td>91.2</td>
<td>89.1</td>
<td>83.8</td>
<td>72.2</td>
<td>59.8</td>
<td>52.5</td>
</tr>
<tr>
<td>Raleigh</td>
<td>40.5</td>
<td>42.2</td>
<td>49.2</td>
<td>59.5</td>
<td>67.4</td>
<td>74.4</td>
<td>77.5</td>
<td>76.5</td>
<td>70.6</td>
<td>60.2</td>
<td>50.0</td>
<td>41.2</td>
</tr>
<tr>
<td>Minneapolis</td>
<td>12.2</td>
<td>16.5</td>
<td>28.3</td>
<td>45.1</td>
<td>57.1</td>
<td>66.9</td>
<td>71.9</td>
<td>70.2</td>
<td>60.0</td>
<td>50.0</td>
<td>32.4</td>
<td>18.6</td>
</tr>
</tbody>
</table>

Anscombe’s Quartet

4 data sets that have
nearly identical summary statistics

each has 11 non-missing pairs of values

constructed in 1973 by statistician
Francis Anscombe to demonstrate
importance of graphing data and
effect of outliers

<table>
<thead>
<tr>
<th>SUMMARY STATISTICS</th>
<th>Set 1</th>
<th>Set 2</th>
<th>Set 3</th>
<th>Set 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>mean value of x</td>
<td>9</td>
<td>9</td>
<td>9</td>
<td>9</td>
</tr>
<tr>
<td>mean value of y</td>
<td>7.5</td>
<td>7.5</td>
<td>7.5</td>
<td>7.5</td>
</tr>
<tr>
<td>variance of x</td>
<td>11</td>
<td>11</td>
<td>11</td>
<td>11</td>
</tr>
<tr>
<td>variance of y</td>
<td>4.1</td>
<td>4.1</td>
<td>4.1</td>
<td>4.1</td>
</tr>
<tr>
<td>correlation between x and y</td>
<td>0.816</td>
<td>0.816</td>
<td>0.816</td>
<td>0.816</td>
</tr>
<tr>
<td>linear regression (best fit) line is:</td>
<td>y=0.5x+3</td>
<td>y=0.5x+3</td>
<td>y=0.5x+3</td>
<td>y=0.5x+3</td>
</tr>
</tbody>
</table>

hard to see the forest when looking at the trees
graph allows simple visual examination of effect of outlier on model summary
# US States: percent of population under age 16, 2000-2010

<table>
<thead>
<tr>
<th>State</th>
<th>2000</th>
<th>2001</th>
<th>2002</th>
<th>2003</th>
<th>2004</th>
<th>2005</th>
<th>2006</th>
<th>2007</th>
<th>2008</th>
<th>2009</th>
<th>2010</th>
<th>region</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arizona</td>
<td>23.81</td>
<td>23.72</td>
<td>23.63</td>
<td>23.56</td>
<td>23.43</td>
<td>23.33</td>
<td>23.17</td>
<td>23.09</td>
<td>23.04</td>
<td>22.8</td>
<td>22.62</td>
<td>4</td>
</tr>
<tr>
<td>California</td>
<td>24.39</td>
<td>24.16</td>
<td>23.94</td>
<td>23.75</td>
<td>23.52</td>
<td>23.23</td>
<td>22.91</td>
<td>22.59</td>
<td>22.34</td>
<td>22.09</td>
<td>21.94</td>
<td>4</td>
</tr>
<tr>
<td>Colorado</td>
<td>22.72</td>
<td>22.64</td>
<td>22.49</td>
<td>22.43</td>
<td>22.21</td>
<td>22.17</td>
<td>22.03</td>
<td>21.92</td>
<td>21.8</td>
<td>21.74</td>
<td>21.67</td>
<td>4</td>
</tr>
<tr>
<td>District of Columbia</td>
<td>18.01</td>
<td>17.74</td>
<td>17.98</td>
<td>18.1</td>
<td>17.78</td>
<td>16.75</td>
<td>16.29</td>
<td>15.85</td>
<td>15.34</td>
<td>15.71</td>
<td>14.74</td>
<td>3</td>
</tr>
<tr>
<td>Florida</td>
<td>20.25</td>
<td>20.13</td>
<td>20.02</td>
<td>19.86</td>
<td>19.73</td>
<td>19.58</td>
<td>19.41</td>
<td>19.2</td>
<td>19.02</td>
<td>18.82</td>
<td>18.68</td>
<td>3</td>
</tr>
<tr>
<td>Georgia</td>
<td>23.61</td>
<td>23.58</td>
<td>23.58</td>
<td>23.55</td>
<td>23.42</td>
<td>23.46</td>
<td>23.26</td>
<td>23.24</td>
<td>23.11</td>
<td>22.89</td>
<td>22.79</td>
<td>3</td>
</tr>
<tr>
<td>Idaho</td>
<td>25.1</td>
<td>24.92</td>
<td>24.7</td>
<td>24.5</td>
<td>24.5</td>
<td>24.58</td>
<td>24.5</td>
<td>24.58</td>
<td>24.37</td>
<td>24.51</td>
<td>24.58</td>
<td>4.4</td>
</tr>
<tr>
<td>Indiana</td>
<td>22.95</td>
<td>22.89</td>
<td>22.77</td>
<td>22.66</td>
<td>22.62</td>
<td>22.47</td>
<td>22.36</td>
<td>22.23</td>
<td>22.16</td>
<td>22.03</td>
<td>21.88</td>
<td>2</td>
</tr>
<tr>
<td>Kansas</td>
<td>23.39</td>
<td>23.21</td>
<td>22.99</td>
<td>22.95</td>
<td>22.79</td>
<td>22.62</td>
<td>22.44</td>
<td>22.56</td>
<td>22.51</td>
<td>22.59</td>
<td>22.7</td>
<td>2</td>
</tr>
<tr>
<td>Louisiana</td>
<td>23.97</td>
<td>23.65</td>
<td>23.39</td>
<td>23.16</td>
<td>22.96</td>
<td>22.72</td>
<td>22.01</td>
<td>22.01</td>
<td>22.05</td>
<td>21.95</td>
<td>21.79</td>
<td>3</td>
</tr>
<tr>
<td>Maine</td>
<td>20.83</td>
<td>20.53</td>
<td>20.14</td>
<td>19.82</td>
<td>19.48</td>
<td>19.18</td>
<td>18.88</td>
<td>18.61</td>
<td>18.56</td>
<td>18.05</td>
<td>18.07</td>
<td>1</td>
</tr>
<tr>
<td>Maryland</td>
<td>22.76</td>
<td>22.66</td>
<td>22.48</td>
<td>22.27</td>
<td>22.1</td>
<td>21.83</td>
<td>21.5</td>
<td>21.17</td>
<td>20.91</td>
<td>20.72</td>
<td>20.58</td>
<td>3</td>
</tr>
<tr>
<td>Minnesota</td>
<td>23.04</td>
<td>22.84</td>
<td>22.57</td>
<td>22.36</td>
<td>22.23</td>
<td>22.05</td>
<td>21.79</td>
<td>21.64</td>
<td>21.56</td>
<td>21.45</td>
<td>21.37</td>
<td>2</td>
</tr>
<tr>
<td>Mississippi</td>
<td>24.09</td>
<td>23.83</td>
<td>23.61</td>
<td>23.47</td>
<td>23.33</td>
<td>23.19</td>
<td>22.96</td>
<td>22.88</td>
<td>22.76</td>
<td>22.68</td>
<td>22.32</td>
<td>3</td>
</tr>
<tr>
<td>Montana</td>
<td>22.23</td>
<td>21.83</td>
<td>21.38</td>
<td>20.98</td>
<td>20.65</td>
<td>20.43</td>
<td>20.36</td>
<td>20.31</td>
<td>20.08</td>
<td>20.02</td>
<td>19.68</td>
<td>4</td>
</tr>
<tr>
<td>Nebraska</td>
<td>23.05</td>
<td>22.91</td>
<td>22.86</td>
<td>22.77</td>
<td>22.76</td>
<td>22.71</td>
<td>22.5</td>
<td>22.49</td>
<td>22.38</td>
<td>22.34</td>
<td>22.3</td>
<td>2</td>
</tr>
</tbody>
</table>
audience and purpose ... differences may lead to different design decisions

yourself

to check data correctness
to examine a variable distribution, outlier values, relationships with other variables
to examine model fit

others

to display distributions/relationships of variables that are important to your results ("your story")

**to most accurately and most clearly present your results**

- experts vs novice in subject matter? (use acronyms or abbreviations as axis labels?)

- experts vs novice at interpreting statistical graphs?
consider graph’s setting ...

presentation ... limited time ... CLARITY, CLARITY, CLARITY
... some audience members sitting farther away:
larger font size, higher contrast, brighter colors, axis labeling at top, etc.

class ... limited time

poster session ... more time plus possible interaction with author
(but usually little text) ... graphs need to be able to “stand on their own”

print journal article, book, report ... possibly more time plus full text,
but may be limited by publication constraints such as
graph size, number, color and resolution
... may be able to provide more details in an appendix

web (online article, book, report or blog post) ... possibly more time plus full text
usually less limited by publication constraints
... usually can provide links to further detail
II Representing data accurately

“The representation of numbers, as physically measured on the surface of the graphic itself, should be directly proportional to the quantities represented.”
Edward Tufte

“Visual connections should reflect real connections.”
Hadley Wickham

“Avoid distorting what the data have to say.”
Edward Tufte

Life Expectancy in the Americas, by Country, 2012

Source: 2012 World Population Data Sheet by Pop. Ref. Bureau
Life Expectancy in the Americas, by Country, 2012

Source: 2012 World Population Data Sheet by Pop. Ref. Bureau
Life Expectancy in the Americas, by Country, 2012

Source: 2012 World Population Data Sheet by Pop. Ref. Bureau
North: 67.3%
South: 69.1%
East: 70.1%
West: 75.1%
Central: 82.4%
Does it make sense to use a line graph when showing values for a nominal variable?

Line graphs need interval scales for slopes to be meaningful ...

Amazing website by Jennifer Bryan
http://shinyapps.stat.ubc.ca/r-graph-catalog/
do not change scale part way along an axis:
consider when to use two (dual) scales for the same axis ...

two scales measuring the same data with different labels:
Two scales showing two variables on the same graph? Does the slope of one curve relative to the slope of the other curve make sense? Is the intersection point meaningful?
a possible alternative?
in 2012, is Chinese wage growth higher or lower than US wage growth???
Q: How can ggplot2 be used to make plot with 2 axes, one on left, another on right?

A: "It's not possible in ggplot2 because I believe plots with separate y scales (not y scales that are transformations of each other) are fundamentally flawed.

- They are not invertible: given a point on the plot space, you can not uniquely map it back to a point in the data space.

- They are relatively hard to read correctly compared to other options.

- They are easily manipulated to mislead: there is no unique way to specify the relative scales of the axes, leaving them open to manipulation.

- They are arbitrary: why have only 2 scales, not 3, 4 or ten?

You also might want to read Stephen Few's lengthy discussion on the topic Dual-Scaled Axes in Graphs Are They Ever the Best Solution? “
consider scale when displaying odds ratios ...

Kristin Bietsch, Program in Population Studies PhD candidate, OPR Notestein seminar, November 2014
consider scale when displaying odds ratios ... 

Odds ratio of ever having had an HIV test for women who had watched any TV during the last year

should odds ratios be graphed using a log scale?

somewhat controversial ....
From *American Journal of Epidemiology*, Jun 29, 2011
Letter to the editor by Kenneth Rothman, Lauren Wise and Elizabeth Hatch

“The conventional answer to this question seems to be yes, even to the extent that some have called on the International Committee of Medical Journal Editors to ban graphs of ratio measures that do no employ a log scale. The policy of the American Journal of Epidemiology nearly achieved this ban; ...”

“So can there be a compelling reason for an arithmetic [linear] scale? ... One could argue that rate differences have primacy over ratio measures .... Plotting ratio measures on a logarithmic scale, however, does not scale effects according to rate differences, whereas the arithmetic scale does ....”

References.
Relative measures of association, such as hazard ratio, odds ratio, and risk ratio, are often used to convey comparative information in medicine and public health. Graphical presentation of such ratios is common practice in technical papers. However, there are two crucial features that must be taken into account when presenting ratios in graphical format: (1) the baseline value for a ratio is 1; and (2) ratios are expressed on a logarithmic rather than arithmetic scale.

Szklo and Nieto have nicely summarised these two conditions using three examples of ratios with values of 0.5 and 2.0 (figure). Part A uses a baseline of zero and an arithmetic scale. The visual impression given is that the risk ratio of 2.0 is four times larger than the ratio of 0.5. Part B is correct in using a baseline of 1 but wrong in using an arithmetic scale, which gives the impression that the ratio of 2.0 is twice that of the ratio 0.5. In reality, risk ratios of 2.0 and 0.5 are identical in magnitude but work in opposite directions. Part C shows the correct presentation, using a baseline of 1 and a logarithmic scale.
pop2012 value mapped to radius of bubble ... doubling value results in quadrupling area!

pop2012 value mapped to area of bubble ... Canada = 35, US=314 (about 9 times more)
over-plotting hides data points

Techniques to accurately display data density include:
- adjust point size
- adjust point fill
- adjust point shape
- adjust point transparency
- use data stratification
- use point jittering
adjust point size
adjust point fill

Crude Birth Rate by Crude Death Rate, for US Counties, 2012

Source: CO-EST2012-alldata: US Census Bureau, Population Division
adjust point size and fill
adjust point shape
adjust point transparency
use data stratification

Crude Birth Rate by Crude Death Rate, for US Counties, 2012

Source: CO-EST2012-alldata: US Census Bureau, Population Division
use point jittering – moving overlapping points a bit

– trade-off: sacrifice positional precision for more accurate display of data density

III Highlighting **comparisons** of interest

“At the heart of quantitative reasoning is a single question: *Compared to what?”*

- Edward Tufte

---

Highlight comparisons

1. Determine true quantity (or quantities) of interest, for example ... 
   - magnitude of A and magnitude of B? 
     or 
   - difference between magnitude of A and magnitude of B? 
     or 
   - ratio of magnitude of A to magnitude of B?

2. Make sure the data is easily seen
   - size, contrast, not hidden by other data markers (points, lines, areas), labels, legends, tick marks, or gridlines

3. Show the data, not just summary measures, when possible

4. Involve perceptual tasks high on Cleveland’s list of performing accurate judgements
   - position along a common scale
   - position along identical, non-aligned scales
   - length
   - angle, slope
   - area
   - color

5. Consider proximity, alignment and ordering
determine true quantity of interest

To show the difference between A and B, graph the difference between A and B. You may want to graph A and B on their own too, but don’t stop there.

show imports and exports to and from England

to show balance of trade, imports - exports, graph that also

R graph Catalog by Jennifer Bryan  http://shinyapps.stat.ubc.ca/r-graph-catalog
sometimes surprisingly difficult to calculate the difference between curves:
Wealth inequality has widened along racial, ethnic lines since end of Great Recession.
http://www.pewresearch.org/fact-tank/2014/12/12/racial-wealth-gaps-great-recession/
don’t ask viewers to do extra work

R graph Catalog by Jennifer Bryan http://shinyapps.stat.ubc.ca/r-graph-catalog
Fig 8.4 Task Data Unified

+ Revised ○ Original

- Connect a stereo
- Create a mixed tape
- Program radio stations
- Record on one side of a tape
- Program the CD player
- Set the clock
- Set the timer to record
- Listen to a song on a CD
- Listen to a tape
- Listen to the radio

Time (minutes)

Fig 8.5 Task Data: Improvement

- Connect a stereo
- Create a mixed tape
- Record on one side of a tape
- Program the CD player
- Program radio stations
- Set the clock
- Listen to a song on a CD
- Set the timer to record
- Listen to a tape
- Listen to the radio

Time (minutes)

R graph Catalog by Jennifer Bryan [http://shinyapps.stat.ubc.ca/r-graph-catalog](http://shinyapps.stat.ubc.ca/r-graph-catalog)
Fig 8.3 Task Data
Average Time Spent on Stereo Tasks: Original vs Revision

- Connect a Stereo: Original 18.6, Revised 8.8
- Set the Clock: Original 4.0, Revised 2.3
- Program Radio Stations: Original 11.3, Revised 7.7
- Listen to the Radio: Original 0.9, Revised 1.0
- Listen to a Tape: Original 2.6, Revised 2.2
- Record on One Side of a Tape: Original 9.7, Revised 5.8
- Listen to a Song on a CD: Original 3.8, Revised 2.1
- Program the CD Player: Original 6.4, Revised 2.6
- Create a Mixed Tape: Original 15.9, Revised 9.6
- Set the Timer to Record: Original 3.7, Revised 2.2

Fig 8.6 Task Data: Percent Improvement

- Program the CD player
- Connect a stereo
- Listen to a song on a CD
- Set the clock
- Set the timer to record
- Record on one side of a tape
- Create a mixed tape
- Program radio stations
- Listen to a tape
- Listen to the radio

Percent Improvement

R graph Catalog by Jennifer Bryan  [http://shinyapps.stat.ubc.ca/r-graph-catalog](http://shinyapps.stat.ubc.ca/r-graph-catalog)
show the data - make sure it can be easily seen.

consider:
- size
- contrast
- overplotting
- hidden by tick marks, legends, labels, gridlines, reference lines, text annotations

R graph Catalog by Jennifer Bryan [http://shinyapps.stat.ubc.ca/r-graph-catalog](http://shinyapps.stat.ubc.ca/r-graph-catalog)
show the data, not just summary measures, when possible
show the data - display and compare **distributions** of continuous variables.

**how?**
- box plot
- violin plot
- histogram
- density diagram

**why?**
look for shape of distribution: normal, uniform, bi-modal, skewed, etc.
also look for outliers, data errors, missing data
understand data before modeling
box plot:
box showing median, iqr and contiguous values up to 1.5 times upper and lower quartiles

violin plot:
symmetric shape showing density of data values

Country TFRs by Area, 2012

Country TFRs: Density Distribution, Median and IQR by Area, 2012
histogram: frequency shows number of values within each bin of continuous values

histogram may show proportion of values within each bin, rather than frequency
density curve:
similar to histogram, but is smooth and continuous

violin plot:
symmetric shape showing density of data values
Density curve: similar to histogram, but is smooth and continuous

Violin plot: symmetric shape showing density of data values
1. Position along a common scale

2. Position along identical, nonaligned scales

3. Length

4. Angle-slope

5. Area

6. Color hue and color intensity

Involves perceptual tasks high on William Cleveland’s list of performing accurate judgements.

“Order is based on the theory of visual perception, on experiments in graphical perception, and on informal experimentation.”


via Data + Design, https://infoactive.co/data-design/titlepage01.html
Pie charts may be the most criticized graph form, but are surprisingly common.

They encode values in angles and areas, which are hard for humans to judge.

It is easier to judge position along a common scale, which is why many think dot plots are more effective than pie charts.

angle/area position along a common scale
Fig 2.1 Structured Data Set

Fig 2.2 Structured Data Set: Dot Plot

Mobile OS Share (September 2014)

R graph Catalog by Jennifer Bryan http://shinyapps.stat.ubc.ca/r-graph-catalog
“A table is nearly always better than a dumb pie chart; the only worse design than a pie chart is several of them for then the viewer is asked to compare quantities located in spatial disarray both within and between pies. ... Given their low data-density and failure to order numbers along a visual dimension, pie charts should never be used.”


“Pie charts have severe perceptual problems. Experiments in graphical perception have shown that compared with dot charts, they convey information far less reliably. But if you want to display some data, and perceiving the information is not so important, then a pie chart is fine.”

Pie charts are bad! **Die pie chart, DIE**

Pie charts are bad when you want to accurately compare two numbers.

**But:**
As good as bars for estimating percentage of whole.
Better than bars for comparing compound proportions (A + B vs C + D)

perceptual tasks – using length vs position along a common scale

trends for categories on left and right are easy to see, but trends for categories in middle are hard to judge

stacked bar charts: difficult to decode because they lack a common baseline for judging length

R graph Catalog by Jennifer Bryan [http://shinyapps.stat.ubc.ca/r-graph-catalog](http://shinyapps.stat.ubc.ca/r-graph-catalog)
Humans are fairly good at comparing differences in length, but only when things share a common reference point. (Cleveland, William S. and Robert McGill. “Graphical Perception and Graphical Methods for Analyzing Scientific Data.” Science 229.4716 (1985): 828-833.)
when to use, or not use, stacked bar charts? importance of a common baseline for comparisons.

Solomon Messing Blog:  
https://solomonmessing.wordpress.com/2014/10/11/when-to-use-stacked-barcharts/
Men in 2000

Men in 2014

Changes between 2000 and 2014

stacked bar chart

a line graph alternative when x axis displays time? (aggregation of time periods and vehicle types)

Solomon Messing Blog:
https://solomonmessing.wordpress.com/2014/10/11/when-to-use-stacked-barcharts/
combine bar chart and dot plot??
- hard to read country names
- hard to compare men and women (bars and points are far apart)
- emphasizes data for women

combine bar chart and dot plot??
- hard to read country names
- hard to compare men and women (bars and points are far apart)
- emphasizes data for women

Percentage of Employed Who are Senior Managers, by Gender, 2008

[Bar chart showing the percentage of employed senior managers by gender for various countries, with the emphasis on data for women.]


two bar charts
- hard to read country names
- hard to compare men and women (bars and points are far apart)
  - emphasizes data for women
Percentage of Employed Who are Senior Managers, by Gender, 2008


dot plot
- hard to read country names
- hard to compare men and women
- emphasizes data for women
bar chart
- graph ratio rather than percent for women and percent for men?
- sort using ratio rather than percent for women

color dimensions: hue, chroma, luminance (hcl)

hue: unordered (position along color wheel)

chroma (purity): ordered
~ how much gray is added to pure color

luminance (lightness): ordered
~ how much black or white is added to pure color

Maureen Stone, Choosing Colors for Data Visualization, 2006.
use color dimensions
- to distinguish groups
- to highlight particular data
- to encode quantitative values
only vary color for a reason

varying color and pattern and length

From Stephen Few, Perceptual Edge:
hue is unordered and not perceived quantitatively ... usually a poor choice for indicating magnitude

how to order these colors from smallest to largest??

Two ways to encode quantitative data using color
- sequential scale:
  single hue where color varies from light to dark
  or
  single hue where color varies from pale to pure
- diverging scale: two hues with a neutral color in between, where each hue varies from light to dark or pale to pure

How to order these colors from smallest to largest?

From Stephen Few, Perceptual Edge:
use hue to distinguish groups

use equally spaced hues along color wheel, for example:
use color to highlight

use soft colors to display most information and bright and/or dark colors for emphasis

via Data + Design, https://infoactive.co/data-design/titlepage01.html

use color to encode quantitative information

use a single hue where color varies from light to dark or pale to pure

use color to encode quantitative information

use two hues with a neutral color in between, where each hue varies from light to dark or pale to pure to create a diverging scale

Based on Solomon Messing Blog: https://solomonmessing.wordpress.com/2014/10/11/when-to-use-stacked-barcharts/
proximity - grouped bar charts are difficult because it’s hard to make comparisons between values that aren’t near each other ... **try to put values to be compared near each other**

- hard to compare data for each group (males, both, females) across countries, because other bars get in the way
- non-zero baseline (again)

**Life expectancy at birth, top 10 OECD countries**

Life Expectancy at Birth
Top Ten OECD Countries 2010

Japan
Switzerland
Italy
Australia
Iceland
Spain
Sweden
France
Canada
Israel

Male
Both
Female

Darkhorse Analytics Blog:  http://darkhorseanalytics.com/blog/too-many-bars/
ease comparisons - align things vertically
ease comparisons – use common axes
ordering – don’t sort alphabetically
relationships among subsets
consider having two continuous variables, plus a third categorical variable.
how to compare the relationships both within and between categories?
- distinguish between categories using characteristics such as color, shape, fill

R graph Catalog by Jennifer Bryan: [http://shinyapps.stat.ubc.ca/r-graph-catalog](http://shinyapps.stat.ubc.ca/r-graph-catalog)
a graph showing **small multiples** (sometimes called a *trellis, lattice, grid, panel or facet graph*) shows a series of small graphs displaying the same relationships for different subsets of data.

“Small multiple designs, multivariate and data bountiful, answer directly by visually enforcing comparisons of changes, of the differences among objects, of the scope of alternatives. For a wide range of problems in data presentation, small multiples are the best design solution.”


---

**Fig 6.7 Superposed Data Sets 2**

---

**Fig 6.8 Avoiding Superposed Data**

---

R graph Catalog by Jennifer Bryan: [http://shinyapps.stat.ubc.ca/r-graph-catalog](http://shinyapps.stat.ubc.ca/r-graph-catalog)
small multiples: “a series of graphics, showing the same combination of variables, indexed by changes in another variable.”

- Edward Tufte

Using the same axis scale and keeping proximity, alignment and ordering in mind, allows for easy comparisons both within and between the small multiples.
small multiples

- countries ordered by most recent data point rather than alphabetically

- scale labels on outer edges only, rather than one set per panel

- only used three labels for the 11 years on the plot

- did not overdo the vertical scale either

- extra large scale used for top row. This doesn’t follow principle of small multiples but draws attention to the top left corner.

Points vs Opponent Points for each NBA Team in 2013
sorted by Pearson Correlation

Knicks 0.02  Nets 0.11  Pacers 0.15  Spurs 0.16  Pistons 0.21  Clippers 0.23
Kings 0.29  Bulls 0.30  76ers 0.30  Grizzlies 0.31  Heat 0.31  Hornets 0.31
Thunder 0.32  Lakers 0.33  Jazz 0.33  T-Wolves 0.35  Celtics 0.36  Blazers 0.36
Nuggets 0.37  Cavs 0.40  Rockets 0.40  Warriors 0.41  Pelicans 0.42  Hawks 0.43
San 0.45  Magic 0.48  Mavs 0.53  Wizards 0.55  Bucks 0.58  Raptors 0.60

Average Annual Change in Mean Family Income, 1950-2010, by Quintile and for the Top 5 Percent


classic example:
Cleveland’s use of small multiples allowed easy detection of data entry error

R graph Catalog by Jennifer Bryan:  http://shinyapps.stat.ubc.ca/r-graph-catalog
4 variables shown –
 month number (x) 
 measure (y) 
 state 
 group

**Men**

![Graph showing Men's data]

**Women**

![Graph showing Women's data]

**Children**

![Graph showing Children's data]
use **common scales** to facilitate comparison across small multiples
IV Simplicity and clarity

“Chart junk can turn bores into disasters, but it can never rescue a thin data set.”
- Edward Tufte

“Non-data components of tables and graphs should be displayed just visibly enough to perform their role, but no more so, for excessive salience could cause them to distract attention from the data.”
- Stephen Few

to simplify and increase clarity, consider:

eliminating legend and labeling directly
eliminating colored shapes on lines, but keeping endpoints that are key to the message
eliminating distractions
- background shading
- gridlines
- border

no extra information

also, align labels on y axis ...
show same number of places
to right of decimal point

Percent Annual Increase in National Health Expenditures (NHE) per Capita vs. Increase in Consumer Price Index (CPI), 1980-2012

- when possible, use direct labeling rather than legends
- don’t use too many tick marks or axis labels
- show x axis labels in upright position
- don’t repeat dollar signs or percent signs

Annual Percent Change in National Health Expenditures (NHE) Per Capita and Consumer Price Index (CPI), 1980-2012.


- when possible, use direct labeling rather than legends
- don’t use too many tick marks or axis labels
- show x axis labels in upright position
- don’t repeat dollar signs or percent signs
typically, put outcome variable on vertical (y axis)

but, consider putting outcome variable on x axis if that makes long labels easier to read

via Data + Design, https://infoactive.co/data-design/titlepage01.html
don’t show more decimal places than necessary in axis labels

why show 3 decimal places when only 2 are used?

consideration even more important when displaying small multiples where axis labels are repeated
- use axis labels with units in thousands or millions rather than axis labels having long strings of zeros
- use axis title to indicate units
- is y axis title (country) needed??
axis titles give reader a summary description of variable being displayed
- include unit (percent, dollars, thousands, millions)

via Data + Design, https://infoactive.co/data-design/titlepage01.html
reduce clutter ... 
- remove x axis title “Year”
- one legend (upper left)
- two y axis titles
(left graphs only?)
reduce clutter ...
- remove x axis title “Year”
- one legend (upper left)
- two y axis titles
  (left graphs only?)

- why change color and pattern?

  ↓

don’t vary two visual dimensions, when varying one will do

“The number of ‘information-carrying’ visual dimensions should not exceed the number of dimensions in the data.”
- Edward Tufte
reduce clutter ...
- remove x axis title “Year”
- one legend (upper left)
- two y axis titles
(left graphs only?)

finally ... use common x and y scales to enable comparisons both within and between small multiples
consider whether labels are obscuring data patterns ...

from Carl Sagan's *The Dragon of Eden: Speculations on the Evolution of Human Intelligence*.
but what is uncommon? depends on the viewer

Fig. 2. Sample sizes from the 12 studies (in total, 3866 people participated, of whom 1677 (43%) completed a follow-up survey):  ■ initial and follow-up  □ initial only

Annual Percent Change in National Health Expenditures (NHE) per Capita and Consumer Price Index (CPI), 1980-2012


- consider using a reference line to show an important value across an entire graph if it won’t interfere with displaying the data

- is x axis title (year) needed?
annotate with text

sometimes a few words are helpful

use text to highlight or explain

via Data + Design, https://infoactive.co/data-design/titlepage01.html
annotate with text

why are y axis labels laying on their sides??
“upright “ y axis labels (orientation)

in general, consider modifying tool’s default settings
showing too few data points can sometimes give the wrong impression ("two data points do not define a trend"), especially for measures that vary cyclically

via Data + Design, https://infoactive.co/data-design/titlepage01.html
provide context

National Infant Mortality Rate (IMR) per 1,000 Live Births, 2014

Tanzania, 43.7

US, 6.2
lastly, consistency ...

within and between graphs

be careful when there are multiple authors and multiple tools are used

axis label and axis title fonts, color, axis label orientation, ...
let the data stand out by eliminating decoration

animation created by Darkhorse Analytics shows how communication can be greatly enhanced by eliminating clutter and de-emphasizing supporting elements. Every aspect of a figure should be there on a “need to have it” basis.

https://speakerdeck.com/player/87bb9f00ec1e01308020727faa1f9e72#
V Summary
Summary of Statistical Graphics Considerations

Choice of table vs graph
- focus on the forest or the trees

Audience and Setting

Correctness
- scale considerations
  - for bar graphs, include natural baseline (usually 0), because relative length of bars is basis for comparison
  - for line graphs, need an interval scale, otherwise relative line segment slopes are meaningless
  - graphs with dual scales show meaningless points of intersection and meaningless relative slopes, unless two scales are simply two names for the same value
  - for circles (bubbles), use area not radius to represent magnitude of values
  - when displaying odds ratios, where range is 0 to infinity, use log scale rather than linear scale

- data density and data hiding considerations (overplotting)
  - consider point size, fill, shape, transparency, and data stratification
  - use jittering to sacrifice positional precision for more accurate display of data density
Summary of Statistical Graphics Considerations

Comparisons

- determine **true** quantity of interest
  - difference/ratio between A and B rather than A and B
    - absolute difference or percent difference

- make sure data is easily seen
  - size of data markers and contrast against background
  - not hidden by other data markers (points, lines, areas), labels, legends, tick marks, or gridlines

- show the data, not just summary measures, when possible
  - for continuous data, consider box plots, violin plots, histograms or density plots to compare groups

- involve perceptual tasks high on Cleveland’s list of performing accurate judgments
  - position along a common scale
  - position along identical, non-aligned scales
  - length
  - angle, slope
  - area
  - color

- consider proximity, alignment and ordering
Summary of Statistical Graphics Considerations

Comparisons

- dot plots rely on judgment of position along a common scale
- small multiples rely on judgment of position along identical, non-aligned scales
- bar charts rely on judgment of length
  - stacked bar charts ... difficult without a common baseline
  - grouped bar charts ... difficult when bars are not adjacent (proximity)
- pie charts rely on judgment of angle
- bubble charts rely on judgment of area
- use color for a reason:
  - to indicate groups (via hue)
  - to highlight particular data (via hue or intensity)
  - to encode quantitative data (via intensity)
Summary of Statistical Graphics Considerations

Simplicity:
- don’t use too many tick marks or axis labels
- don’t repeat dollar signs or percent signs in axis labels
- don’t display more decimal places than necessary in axis labels
- don’t use long strings of zero’s in axis labels
- when possible, use direct labeling rather than legends
- consider whether axis titles are necessary
- in graphs showing small multiples, repeat axis labels, axis titles and legends only where needed
- don’t use background shading
- consider whether borders and gridlines are needed
- don’t vary multiple visual dimensions, when varying one will do
- beware of default decoration provided by packages

Clarity:
- don’t let data labels obscure data
- show axis labels in upright position
- include units in axis titles
- consider whether abbreviations and acronyms are familiar to audience
- consider using reference lines
- annotate with text to highlight or explain
- provide context for data
- be consistent both within and between graphs

Above all else: let the data stand out
VI Conclusions

“Graphical excellence is the well-designed presentation of interesting data. It consists of clarity, precision and efficiency.”
- Edward R. Tufte

“Pair the depth and clarity of your data, models and writing with visualizations that are just as clear and compelling.”
- Jonathan A. Schwabish

“Don’t skimp on graphs, they’re worth the investment.”
- Matthew J. Salganik
  (verbal communication)