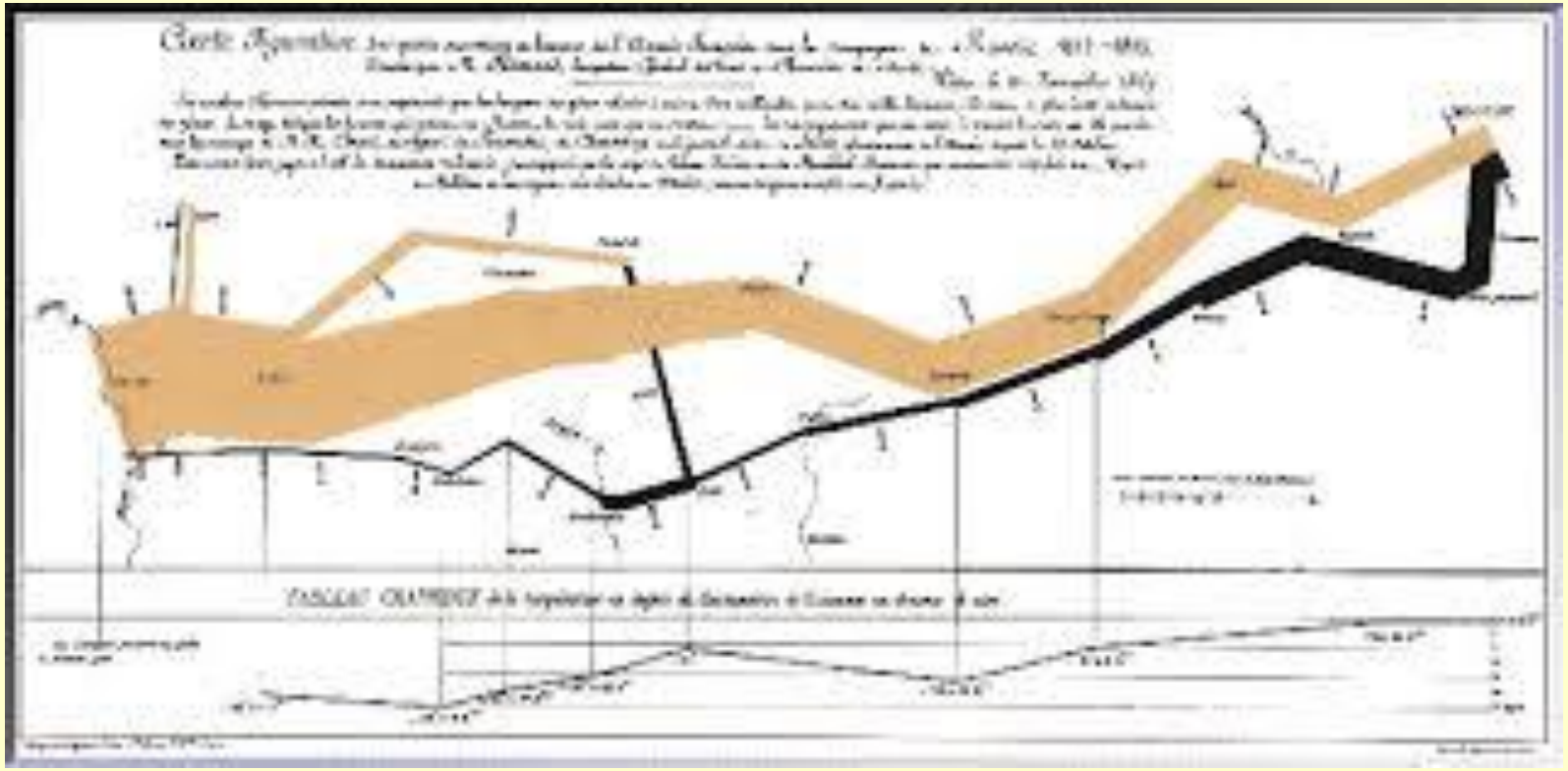
MARS 4910: Feb 25 / 27

Plan for This Week:

- Statistics Workshop – Categorical Variables
  - Work on a list of tests (for your report)
- Graphing Clinic
  - Work on a list of figures (for your report)

Exploring Data With Graphs

- What Makes a Graph Effective ?



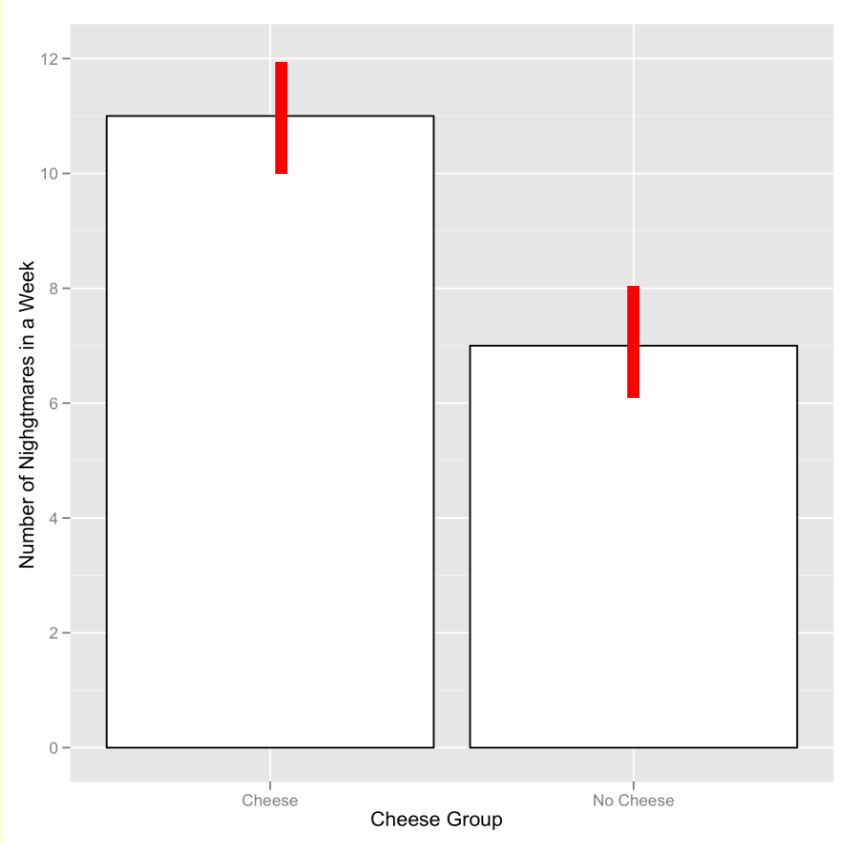
(Tufte, 2001)

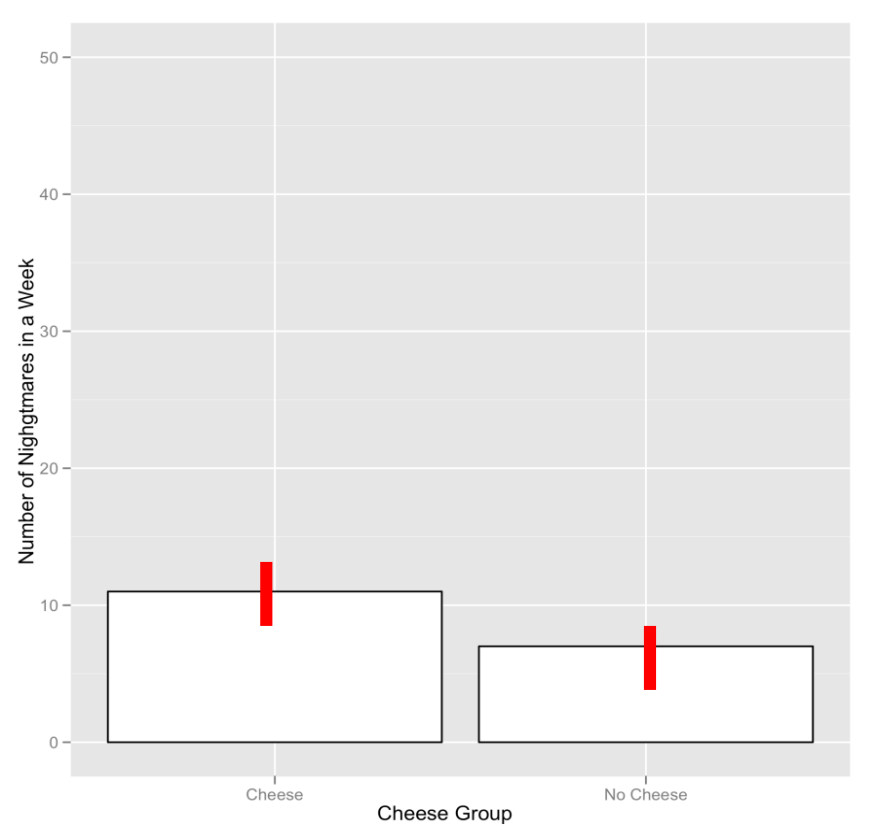
The Art of Presenting Data

- Graphs should (Tufte, 2001):
  - Reveal the data: show the data, highlighting the stories.
  - Induce reader to think about data being presented (rather than some other aspect of the graph).
  - Avoid distorting the data.
  - Present many numbers with minimum ink.
  - Make large dataset coherent (assuming you have one).
  - Encourage reader to compare different data pieces.

What is the Message ?

Two graphs about cheese





Mean +/- S.D.

Use the SE or the SD ?

The "standard error" and "standard deviation" are often confused. The contrast between these two terms reflects the important distinction between data description and inference, one that all researchers should appreciate.

- If you want to describe how scattered the measurements are, use the standard deviation.
- If you want to indicate the uncertainty around the estimate of the mean, use the standard error.
- The standard error also provides a way to calculate a confidence interval around the mean, usually 95%.

When to use the SD ?

The standard deviation (SD) is a measure of variability.

When we calculate the standard deviation of a sample, we are using it as an estimate of the variability of the population from which the sample was drawn.

For data with a normal distribution, 95% of individuals will have values within 1.96 SD units of the mean.

The other 5% of the data, are equally scattered above and below these limits.

When to use the SE ?

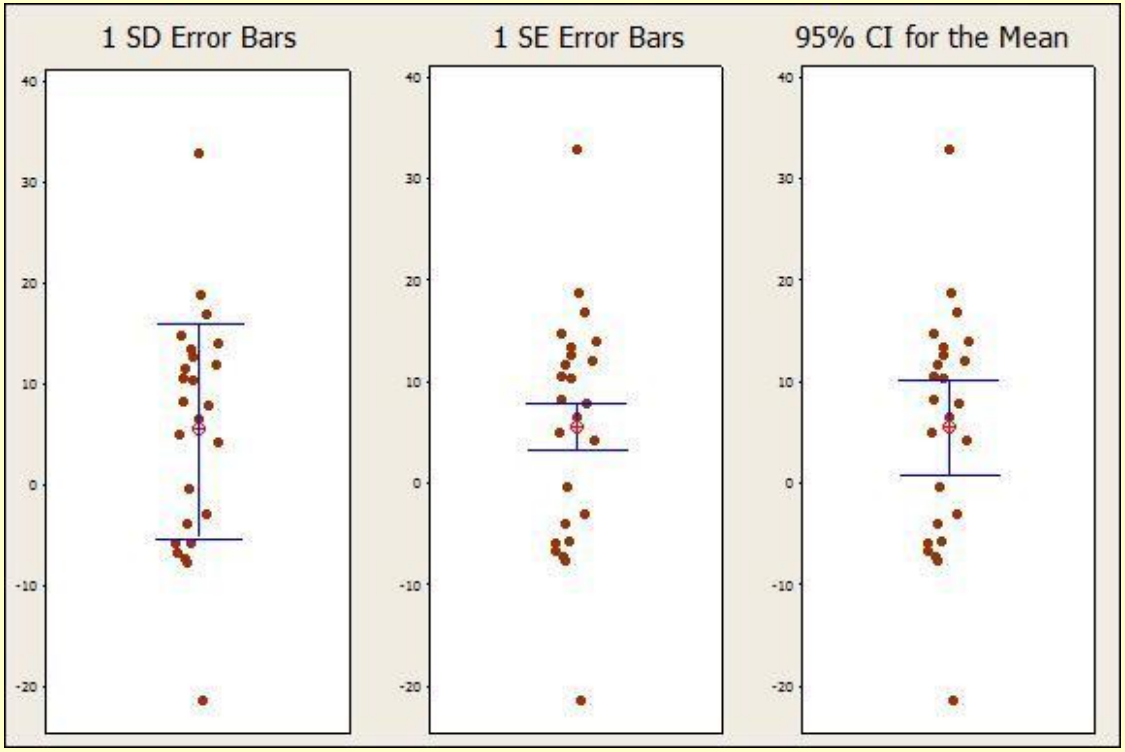
When we calculate the mean of a sample, we are interested in the mean for the biological population — in statistical terms, for the population from which the sample comes.

We use the sample mean as an estimate of the mean for the whole population. Because the sample mean will vary from sample to sample, we described this variability using the "sampling distribution" of the mean.

We estimate how much sample means will vary using the standard deviation of this sampling distribution, which we call the standard error (SE) of the estimate of the mean.

Thus, the SE measures the precision of the sample mean.

When to use SD or SE or CI ?



Recommendations:

Use SD to compare samples

Use CI to compare estimates

Ten Simple Rules for Better Figures

Rougier NP, Droettboom M, Bourne PE (2014) Ten Simple Rules for Better Figures. PLoS Comput Biol 10(9): e1003833. https://doi.org/10.1371/journal.pcbi.1003833

Rule 1: Know Your Audience
Rule 2: Identify Your Message
Rule 3: Adapt the Figure to the Support Medium
Rule 4: Captions Are Not Optional
Rule 5: Do Not Trust the Defaults
Rule 6: Use Color Effectively
Rule 7: Do Not Mislead the Reader
Rule 8: Avoid "Chartjunk"
Rule 9: Message Trumps Beauty
Rule 10: Get the Right Tool

Ten Simple Rules for Better Figures

Rule 7: Do not mislead the reader

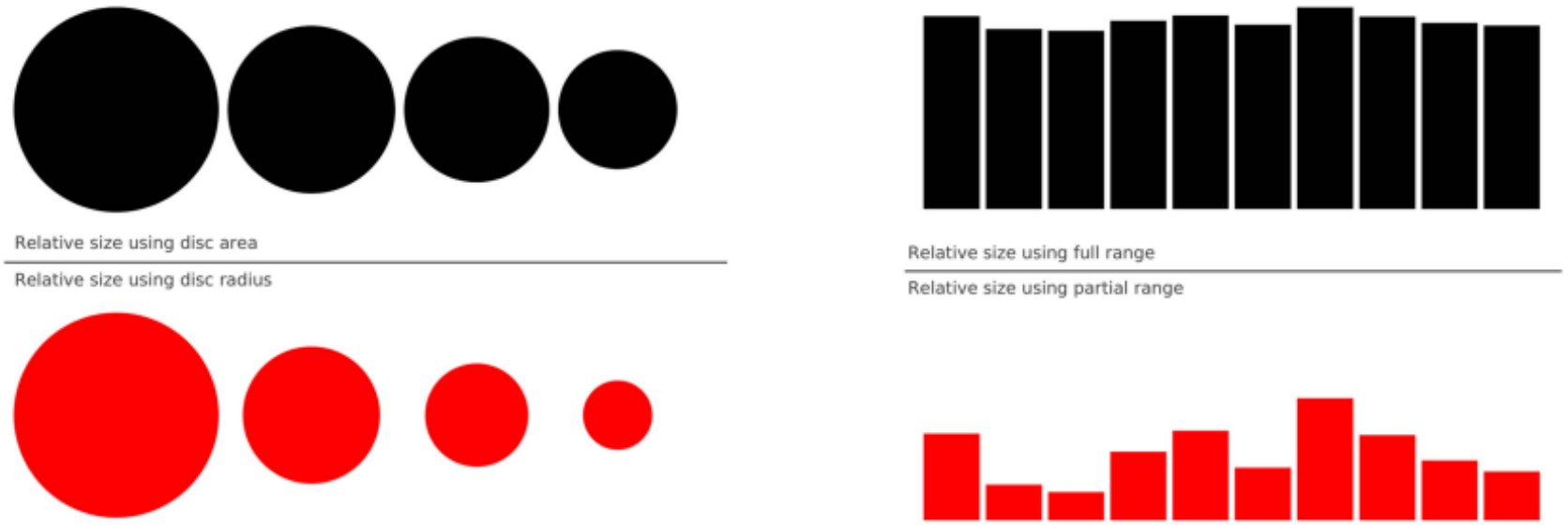


Figure 6. Do not mislead the reader. On the left part of the figure, we represented a series of four values: 30, 20, 15, 10. On the upper left part, we used the disc area to represent the value, while in the bottom part we used the disc radius. Results are visually very different. In the latter case (red circles), the last value (10) appears very small compared to the first one (30), while the ratio between the two values is only 3:1. This situation is actually very frequent in the literature because the command (or interface) used to produce circles or scatter plots (with varying point sizes) offers to use the radius as default to specify the disc size. It thus appears logical to use the value for the radius, but this is misleading. On the right part of the figure, we display a series of ten values using the full range for values on the top part (y axis goes from 0 to 100) or a partial range in the bottom part (y axis goes from 80 to 100), and we explicitly did not label the y-axis to enhance the confusion. The visual perception of the two series is totally different. In the top part (black series), we tend to interpret the values as very similar, while in the bottom part, we tend to believe there are significant differences. Even if we had used labels to indicate the actual range, the effect would persist because the bars are the most salient information on these figures.
doi:10.1371/journal.pcbi.1003833.g006

Ten Simple Rules for Better Figures

Rule 8: Avoid "Chartjunk"

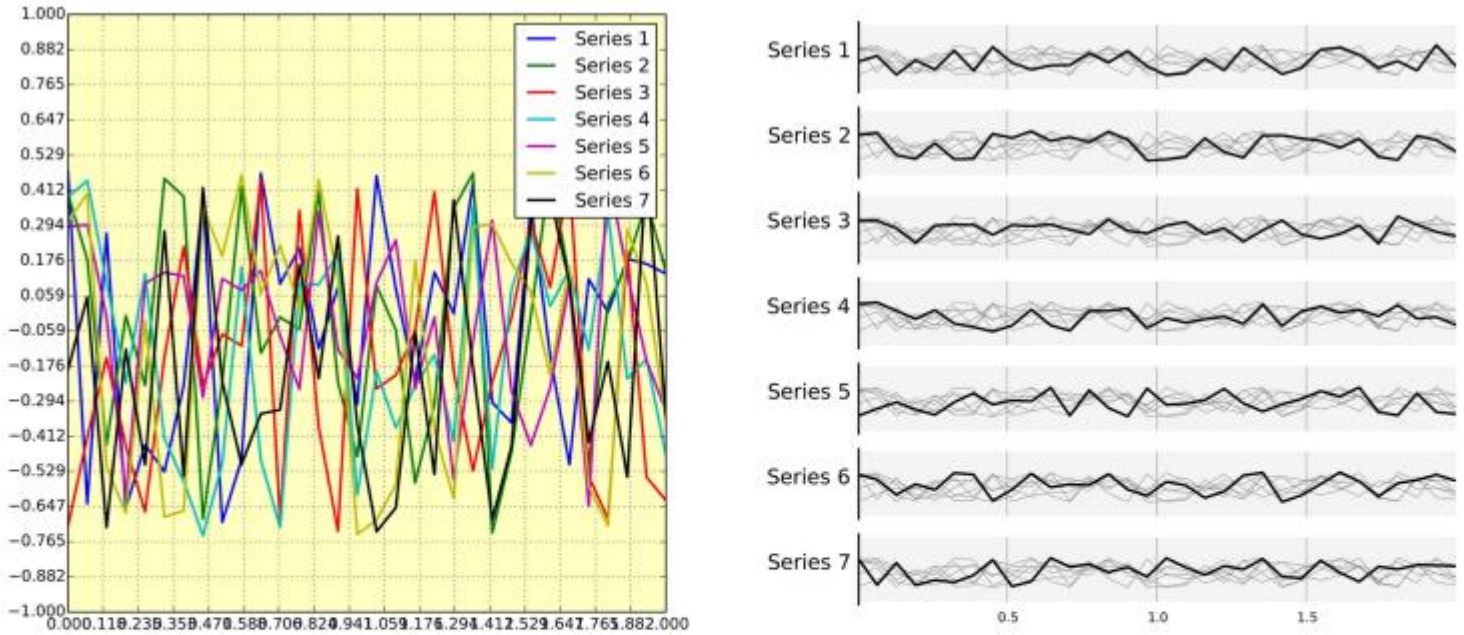


Figure 7. Avoid chartjunk. We have seven series of samples that are equally important, and we would like to show them all in order to visually compare them (exact signal values are supposed to be given elsewhere). The left figure demonstrates what is certainly one of the worst possible designs. All the curves cover each other and the different colors (that have been badly and automatically chosen by the software) do not help to distinguish them. The legend box overlaps part of the graphic, making it impossible to check if there is any interesting information in this area. There are far too many ticks: x labels overlap each other, making them unreadable, and the three-digit precision does not seem to carry any significant information. Finally, the grid does not help because (among other criticisms) it is not aligned with the signal, which can be considered discrete given the small number of sample points. The right figure adopts a radically different layout while using the same area on the sheet of paper. Series have been split into seven plots, each of them showing one series, while other series are drawn very lightly behind the main one. Series labels have been put on the left of each plot, avoiding the use of colors and a legend box. The number of x ticks has been reduced to three, and a thin line indicates these three values for all plots. Finally, y ticks have been completely removed and the height of the gray background boxes indicate the [−1,+1] range (this should also be indicated in the figure caption if it were to be used in an article).
doi:10.1371/journal.pcbi.1003833.g007

Conceptualize the Figure

Example 1:

You have measured the arachnophobia of two groups (males and females) to real spiders / photo spiders.

How would you plot these data:

GROUP	FEMALE PHOTO	MALE PHOTO	FEMALE REAL	MALE REAL
MEAN	10	11	18	19
SD	1	2	2	4

Conceptualize the Figure

Example 2:

You have measured the arachnophobia of two groups (males and females) to real spiders / photo spiders.

How would you plot these data:

GROUP	FEMALE PHOTO	MALE PHOTO	FEMALE REAL	MALE REAL
MEAN	10	18	11	19
SD	1	2	2	4

Today's First Task

Assemble list of statistical tests for reporting your results (for your final report):

Test Title: descriptive name for test

Test Goal: what is the goal of this test

Test Elements: what variable types are compared, how many datasets are compared

Today's Second Task

Assemble list of figures (and tables) for reporting your results (for your final report):

Figure Title: descriptive name of figure

Figure Message: what are the take-home messages reader takes away from this figure

Figure Elements: how many datasets are compared; what statistics are used to summarize the data

Selecting Your Statistical Tests

Identify Predictors for Each Response Variable (Driver OR Independent Variables)

NOTE: consider one response variable at a time

Hypothesis Number and Verbal Description	# Continuous Predictors	# Categorical Predictors	Paired OR Unpaired Data ?	Do Data Meet Parametric Assumptions

Are the predictors Continuous OR Categorical ?

Coming Up

- **Proposal Peer Review – Due Feb 27**
- **Next Week: SPRING BREAK**

Week After (March 11 / 13):

- Feedback from Graphing and Statistics Workshop
- Presentation Workshop

Starting Next Week:

- Research Teams provide brief presentations on their preliminary data (using figures / tables)

- Research Teams provide written report of their preliminary data (normality, transformations)