

Graphing Good Practices and a case against the bar chart

Kathleen Torkko
September 11, 2019

TA Office hour(s) and location:

small group room Ed 2 North 2308 (seats 10) on Mondays 1:00-3:00

TAs

Randy Jin (xin.2.jin@cuanschutz.edu)

Kaitlin Olson (Kaitlin.olson@cuanschutz.edu)

Today's Objectives

Learn about different graph types

Learn when to use each graph type

Learn about good graphing technique, and bad

Learn when to use standard deviation, standard error of the mean, and confidence intervals on a graph

Learn that bar graphs and SEM are not necessarily the best representation of your data

Good Statistical Practice

The first step is ALWAYS getting to know your data

★ Summarize and visualize your data (Descriptive statistics and Graphs)

Summarizing and visualizing data can help determine if the distribution is symmetrical, or if there are mistakes in the data (i.e., outliers that are from measurement or data entry errors)

The first step I usually do is calculate summary statistics (mean, median, minimum, maximum values)

If the mean and median are different this is good indication of a skewed distribution or “outliers”

Descriptive Statistics

Numerical Summaries in tables:

- Frequencies

- Proportions

- Measures of central tendency (i.e., mean, median)

- Measures of variability (i.e., standard deviation)

Graphical Summaries:

- Frequency Histograms

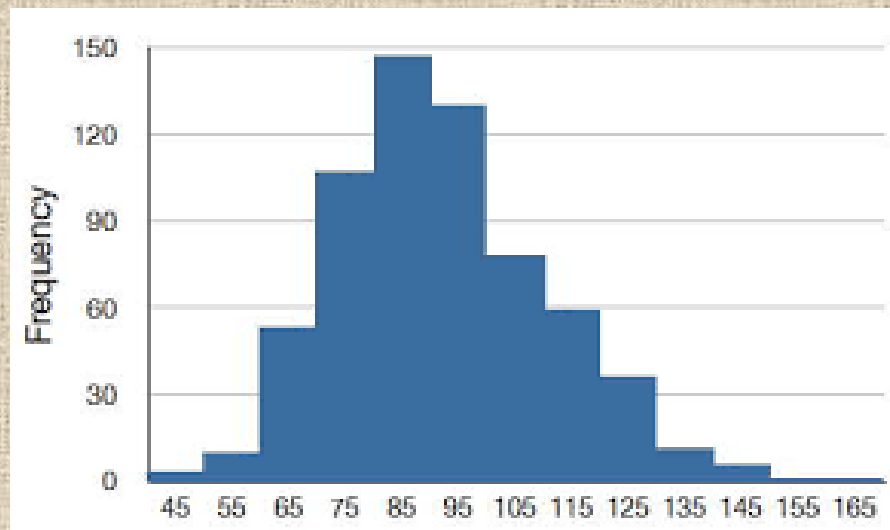
- Box Plots

- Bar graphs or Pie graphs

- Scatterplots

- Line graphs

Frequency Histograms



Boxplots (AKA box and whisker plots)

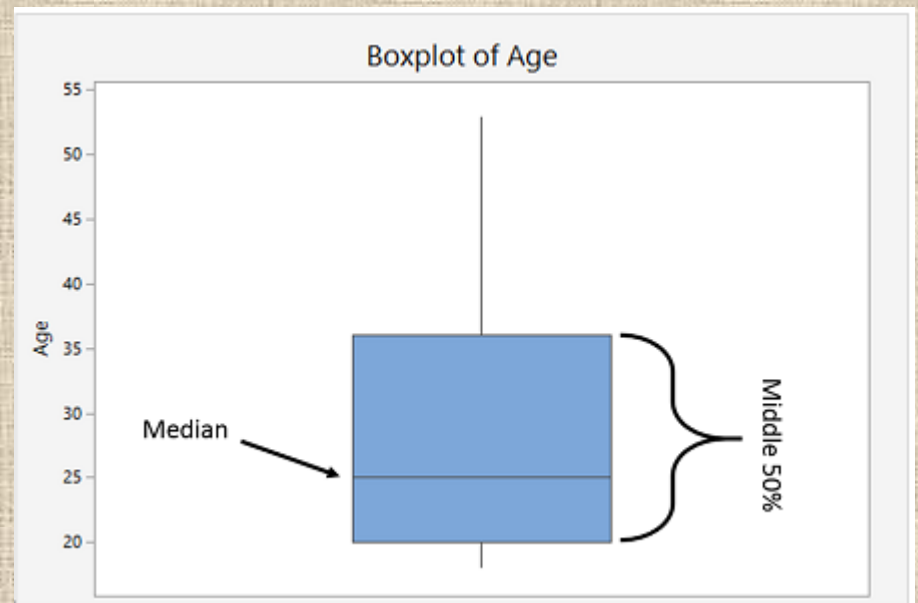
Graphical representation for continuous data in which a box represents the middle 50% of values

The line in the box represents the middle of the distribution (*i.e.*, median)

Lines extend to the highest and lowest values (or other choices)

The box shows the interquartile range from the first to third quartiles (25th percentile to 75th percentile)

Like frequency histograms, the advantage of boxplots is that they are easy to construct. A disadvantage is that individual observations are not usually shown.



Prism offers different ways to create whiskers in box plots

Change Graph Type

Graph family: Column

Individual values **Box and violin** Mean/median & error

Box & whiskers

Plot: **Min to Max**

- Min to Max
- Tukey
- 10-90 percentile
- 5-95 percentile
- 2.5-97.5 percentile
- 1-99 percentile
- Min to Max. Show all points

paired t test data

Male Female

Help OK

How the Tukey method plots whiskers and “outliers”

1. Calculate the inter-quartile range

2. Add the 75th percentile plus 1.5 times IQR. If this value is greater than (or equal to) the largest value in the data set, draw the upper whisker to the largest value. Otherwise stop the upper whisker at the largest value less than the sum of the 75th percentile plus 1.5 IQR, and plot any values that are greater than this as individual points.

3. Calculate the 25th percentile minus 1.5 times IQR. If this value is less than the smallest value in the data set, draw the lower whisker to the smallest value. Otherwise stop the lower whisker at the lowest value greater than the 25th percentile minus 1.5 IQR, and plot any values that are greater than this as individual points.

4. Any data above or below the 1.5 times IQR limits would be considered “outliers” by Tukey

The Tukey Method and Outliers

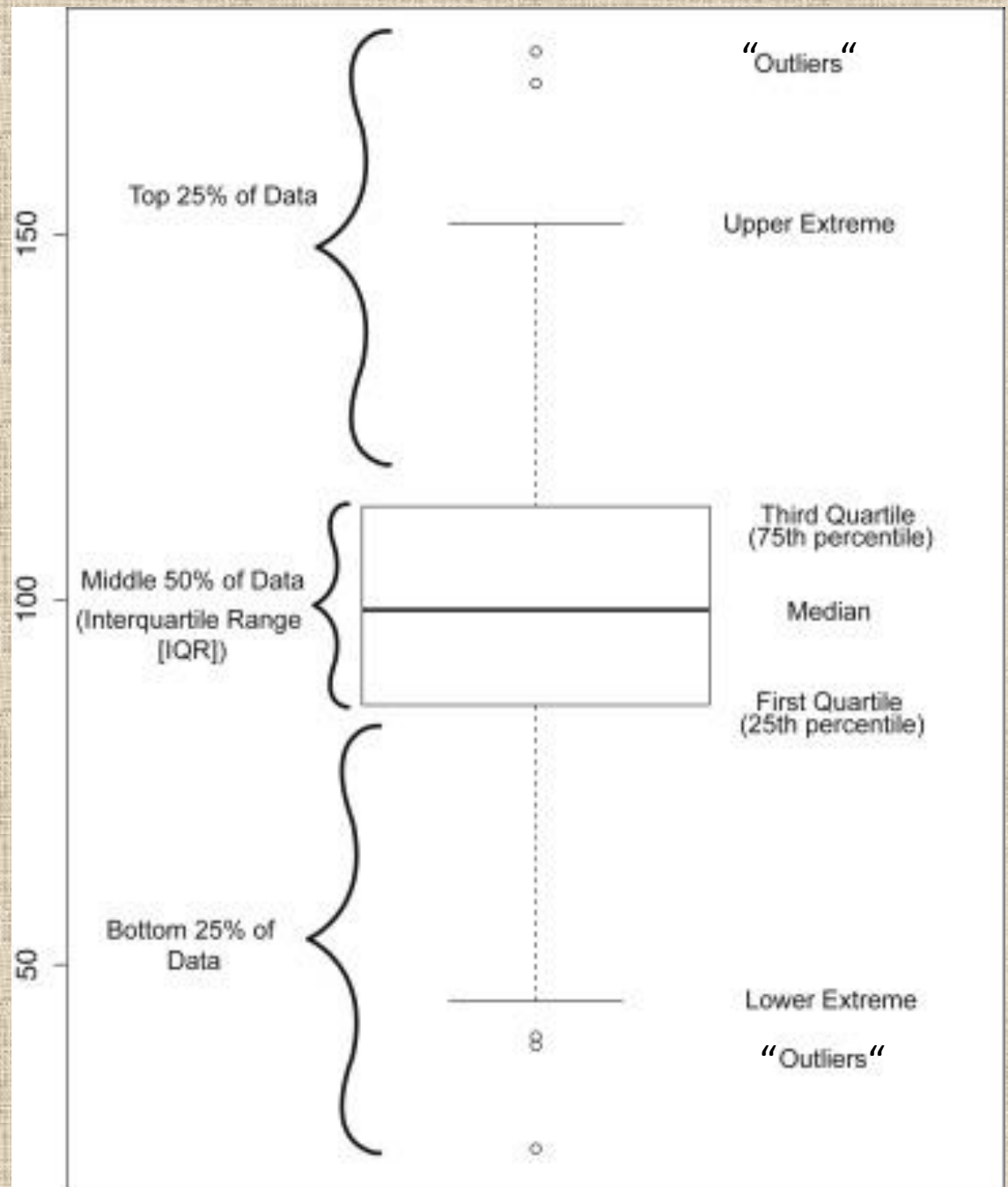
Why 1.5 IQR?

There is no statistical rationale; it is simply how Tukey decided to do it, and he invented the idea of boxplots.

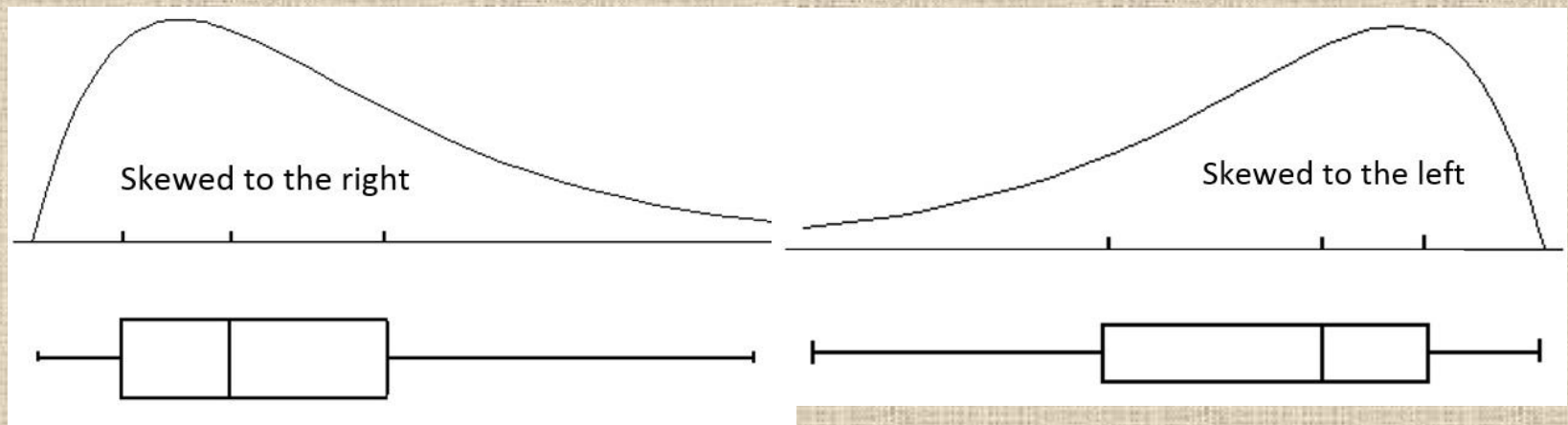
A Closer Look at the Box Plot

Obviously these whiskers are not min to max

Annotated box plot of 1000 points from a normal distribution with a mean of 100 and a standard deviation of 20. Nuzzo RL. PM&R, Volume 8, Issue 3, 2016, 268–272

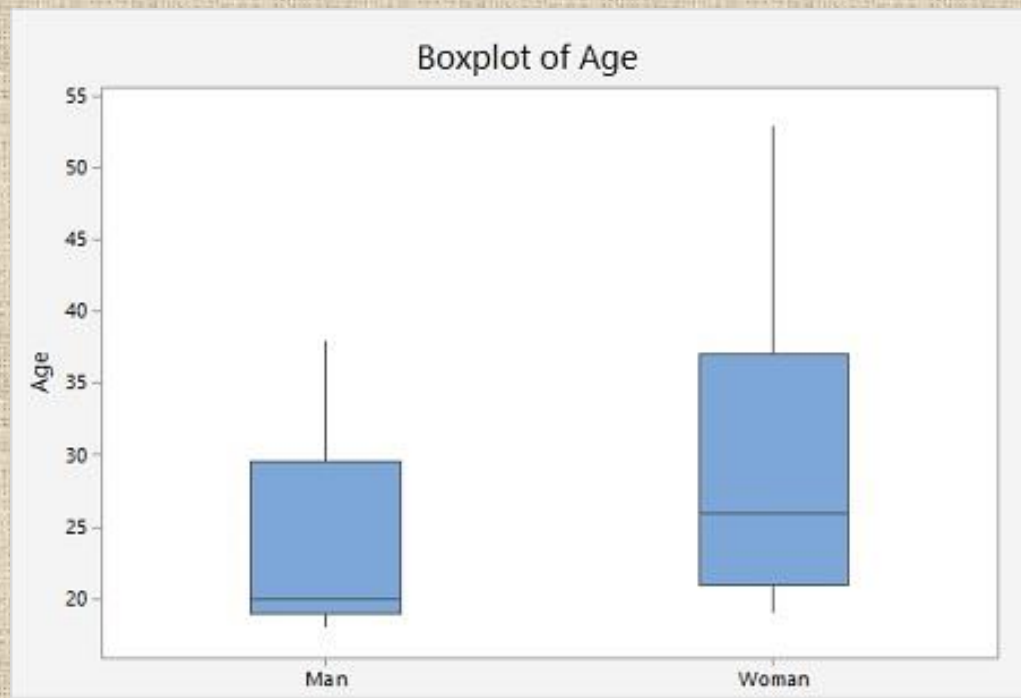


Like histograms, boxplots can be used to evaluate the skewness of a distribution



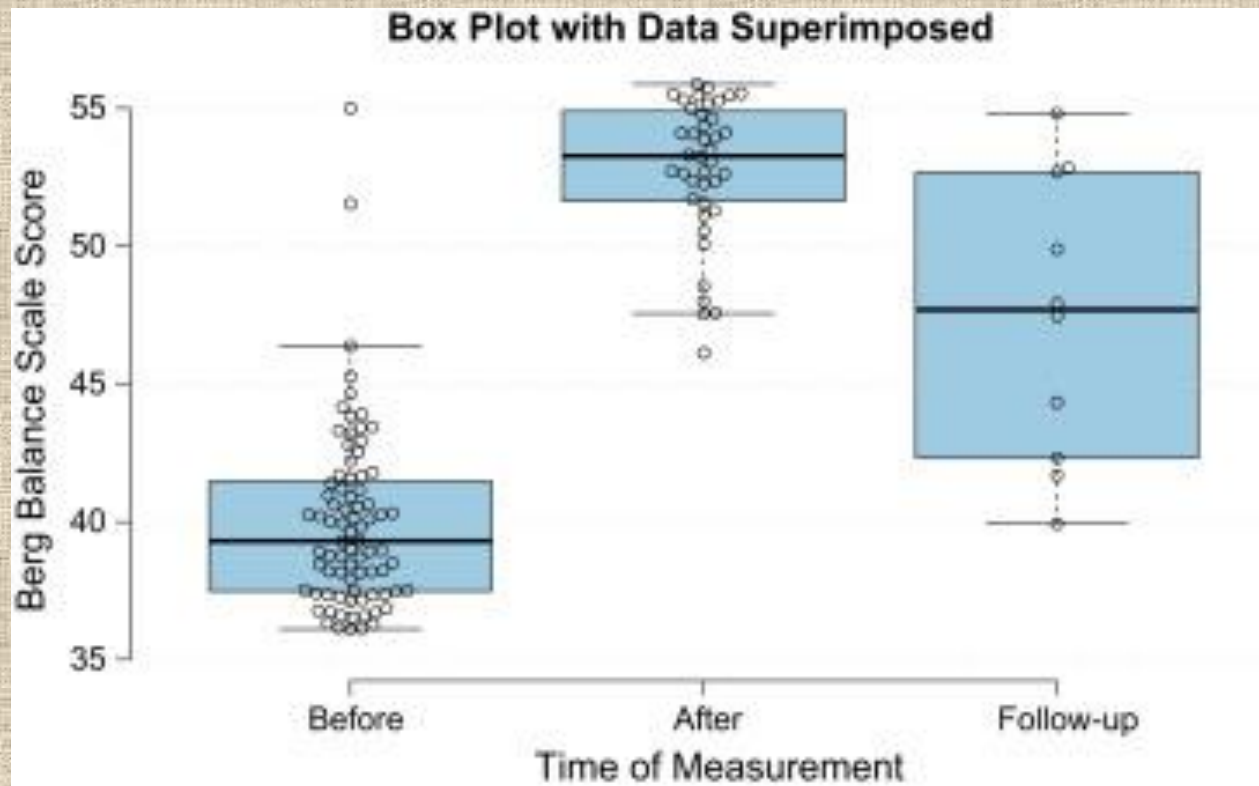
Side-by-side boxplots

The side-by-side box plot below compare men and women on the variable of age. The middle of the distribution for women's ages is higher than that of the distribution for men's ages, thus it appears that the women in the sample tend to be slightly older than the men in the sample.



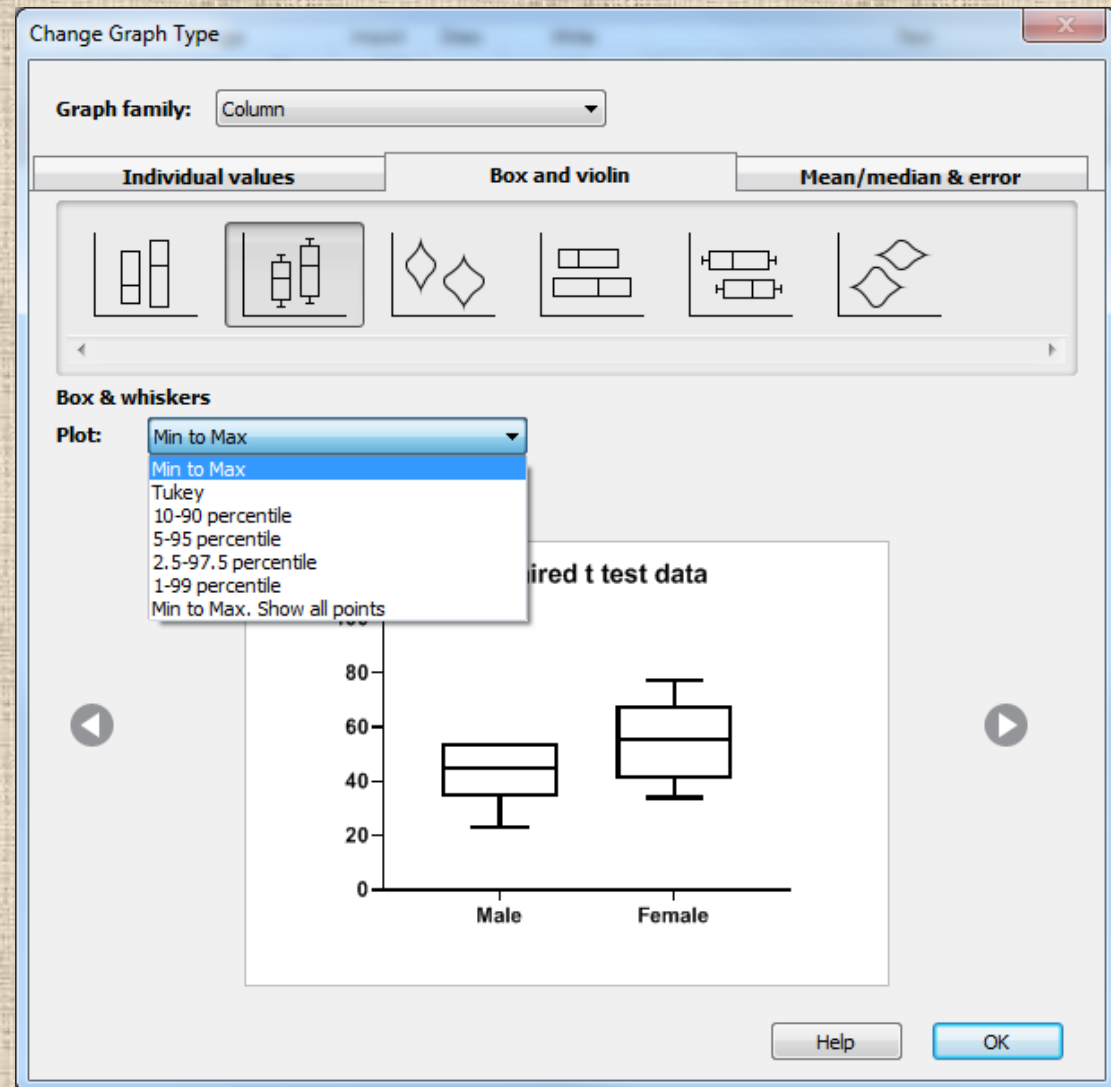
Skew

Many values
clustered
around 20

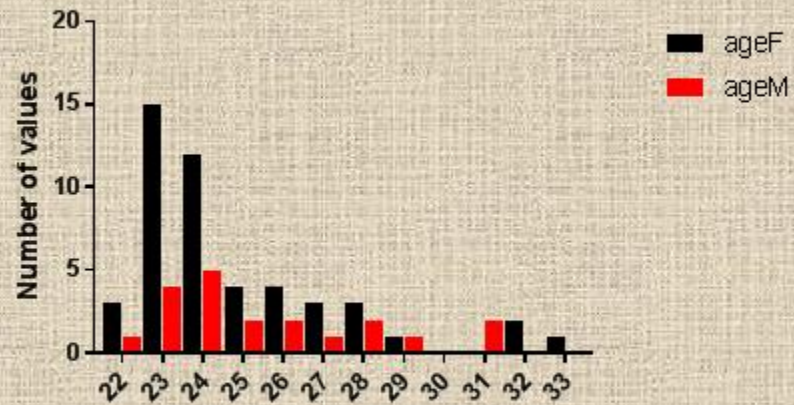


Box Plots in Prism

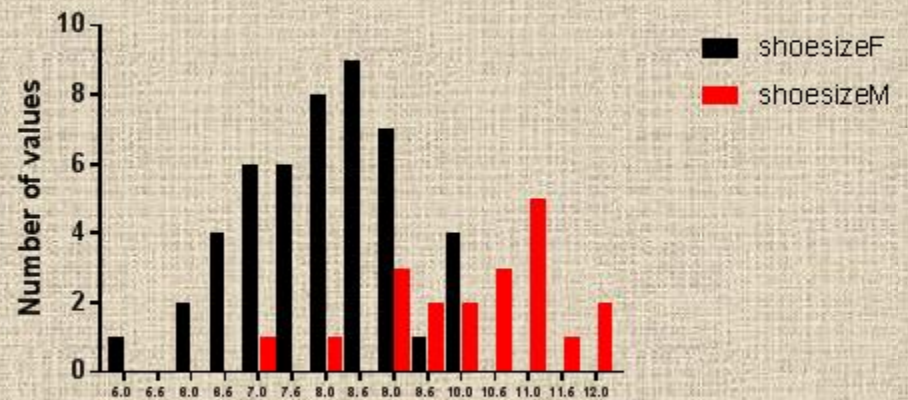
In “Column” Table and Data, once you have created a Data Table, select the graph for that data table



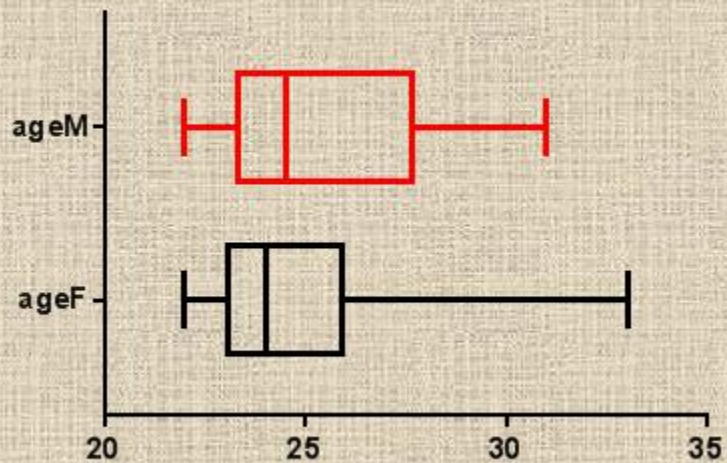
Histogram of Female vs. Male Age



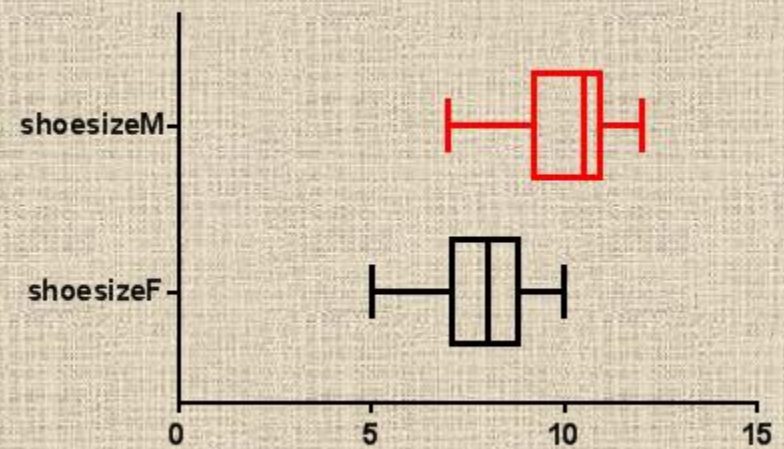
Histogram of Female vs. Male Shoesize



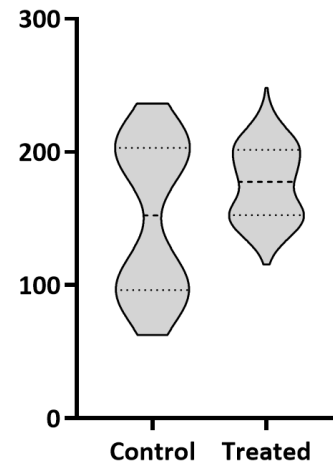
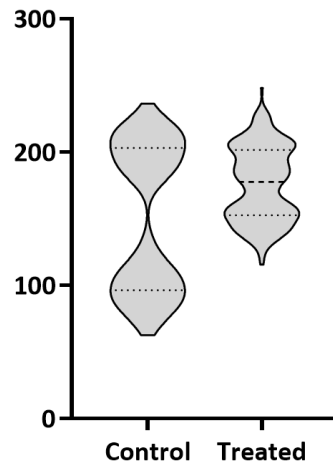
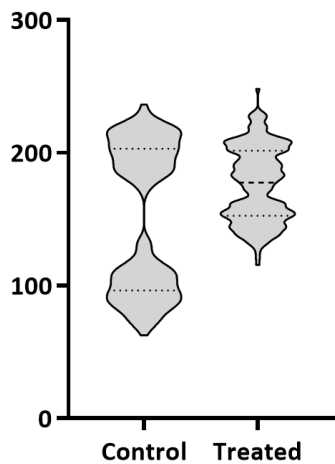
Female vs. Male age



Female vs. Male Shoesize



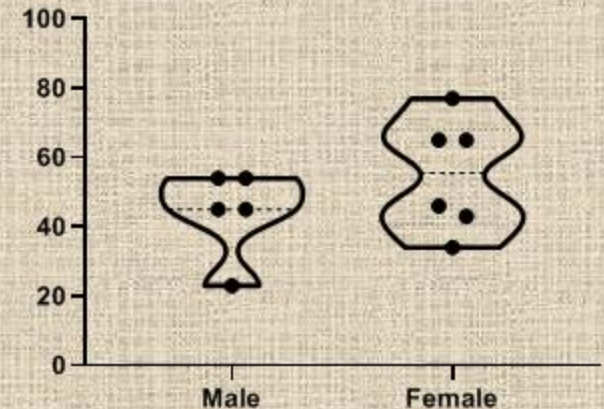
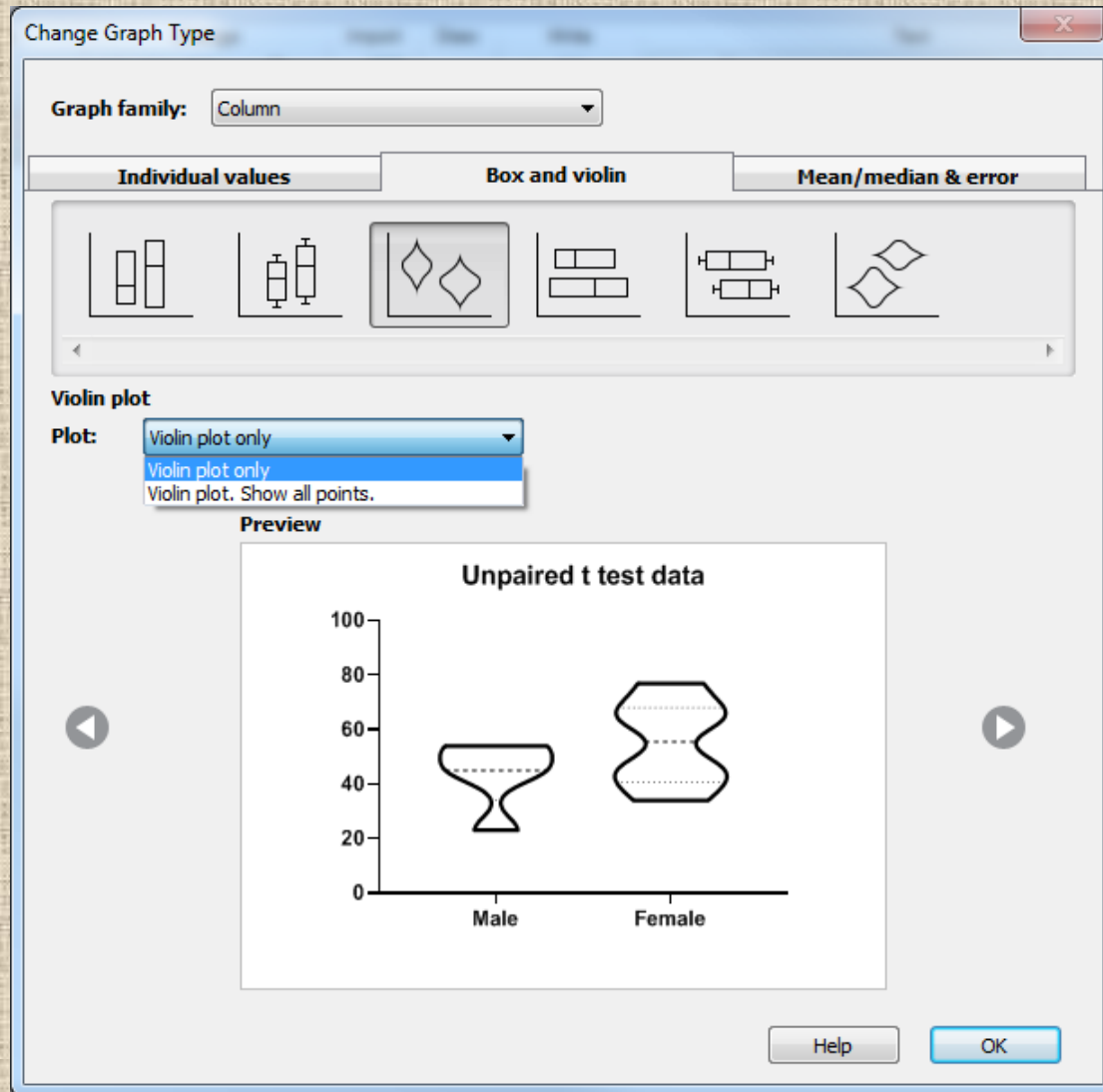
Violin Plots



The thicker part means the values in that section of the “violin” has higher frequency
The thinner part means lower frequency

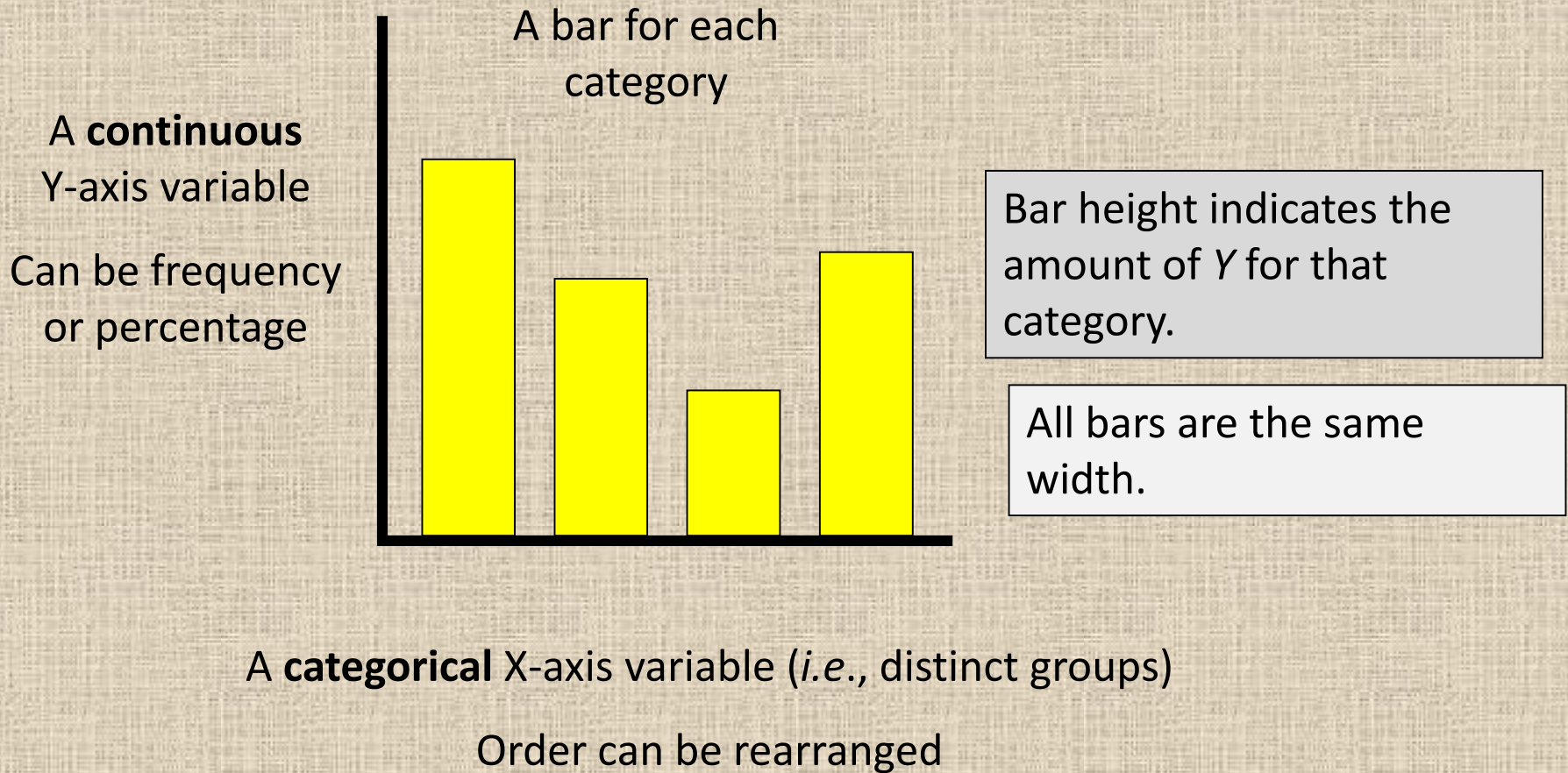
Violin plots contain all data points

In “Column” Table and Data, once you have created a Data Table, select the graph for that data table



Bar Graph

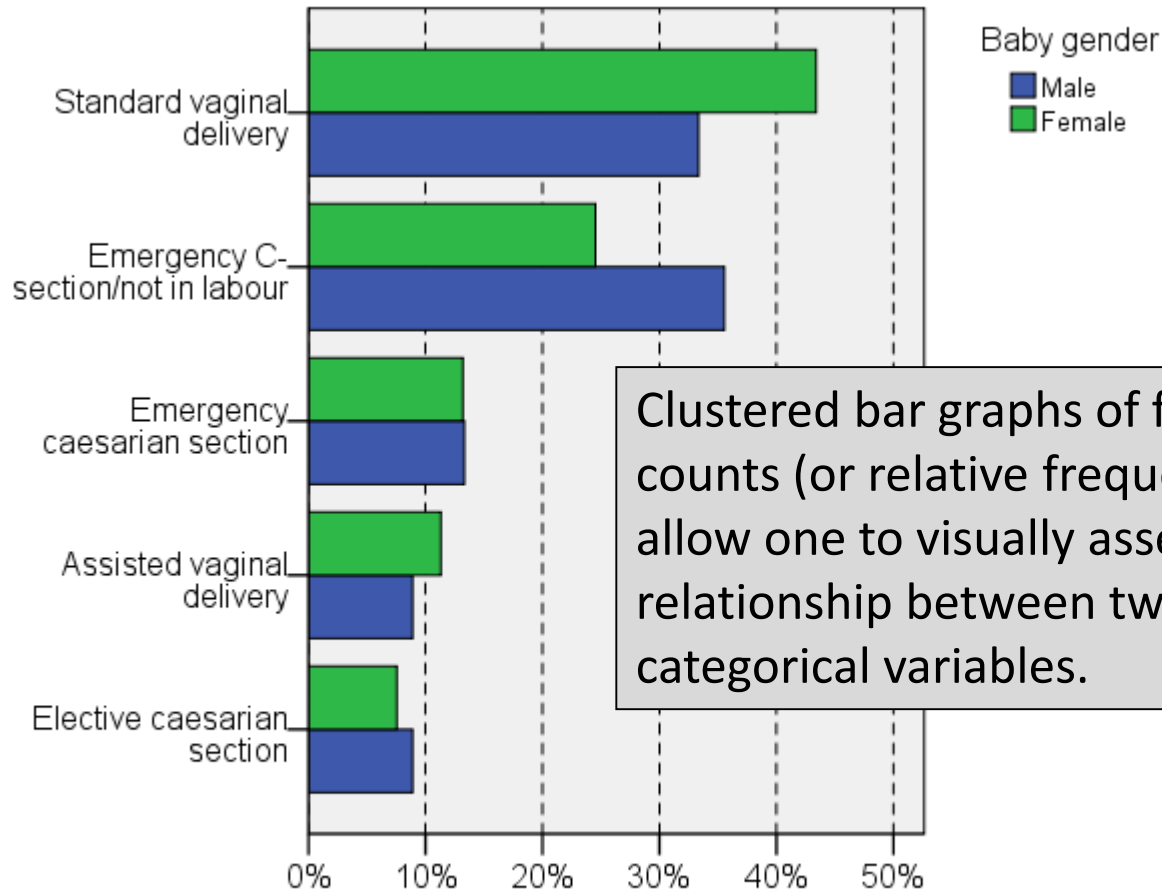
Shows differences in frequencies or percentages among categories of a **nominal** or an **ordinal** variable.



When category labels are too long to display easily on the X-axis, the bar chart may be transposed as shown here.

Clustered Bar Graphs

Type of Delivery



Clustered bar graphs of frequency counts (or relative frequencies) allow one to visually assess the relationship between two categorical variables.

Percent (within gender)

With percentages, always be clear about percentage *of what!*

Good Graphing Practice vs. Bad

Statistical Check List by *Nature Cell Biology*

Graphs

➔ Were effect sizes distorted? (by truncation of y axis, etc.)



Were error bars absent

Were error bars unlabeled?

Were the statistical measures (mean, standard error, standard deviation, etc.) reported, and were they clearly labeled?

Are mean and standard deviation used to describe data sets that may be non-normally distributed

P-value

Specific

Report

Provid

Discus

Rando

Explai

Descri

for ex

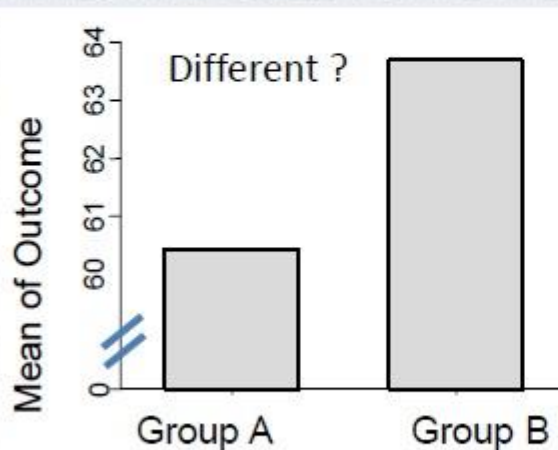
Sampl

Repor

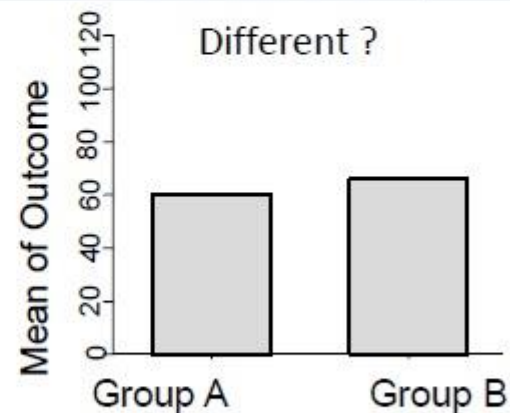
Provid

Explained reasons for any discrepancy between initial n and

Explanation of data exclusions, if any



Bar-graph



Bar-graph

The checklist is no longer used, but has been replaced. Used here for example of proper graphing

Example of a truncated y axis

The Fluctuating Female Vote: Politics, Religion, and the Ovulatory Cycle

Kristina M. Durante¹, Ashley Rae¹, and
Vladas Griskevicius²

¹College of Business, University of Texas, San Antonio, and ²Carlson School of Management, University of Minnesota

Psychological Science
24(6) 1007–1016
© The Author(s) 2013
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/0956797612466416
pss.sagepub.com
SAGE

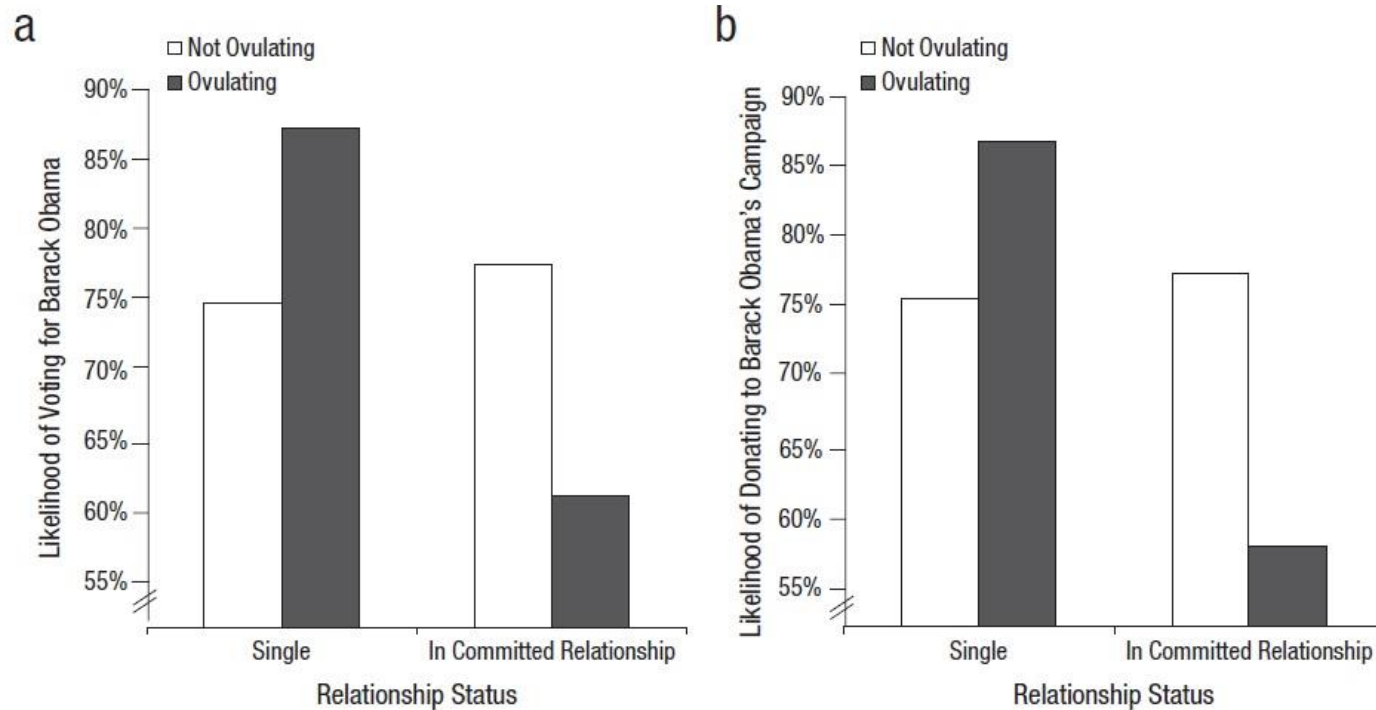
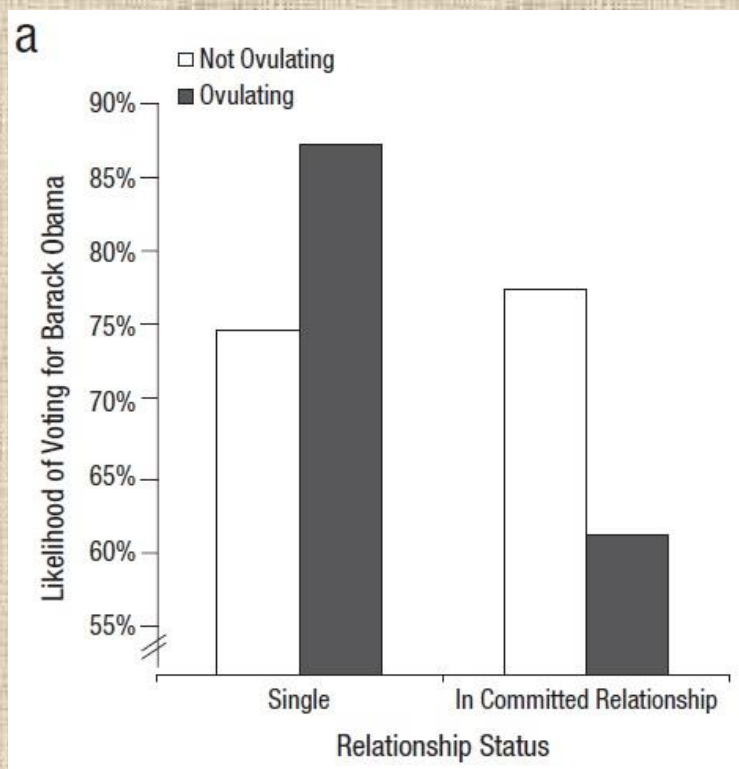


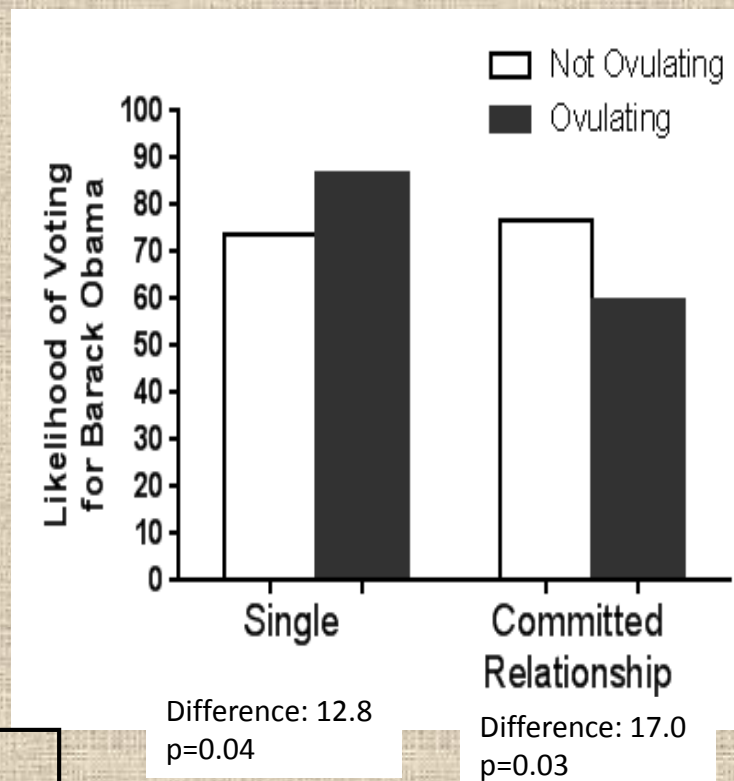
Fig. 4. Results from Study 2: women's (a) likelihood of voting for Barack Obama and (b) likelihood of donating \$1 to Barack Obama's campaign as a function of fertility and relationship status.

Misleading Graphs

Y Axis Truncated



Full Y Axis



They concluded that: "Ovulation led women in committed relationships to become more likely to vote for Mitt Romney." *But 60% still voted for Obama!*

Bar Graphs: The Y-axis Variable and Error Bars

The Y-axis variable is continuous

Common examples include:

Frequency (i.e., number of cases per category)

Relative frequency (% or proportion of all cases per category)

A summary statistic—e.g., the mean, *RR*, *OR*, etc

Error bars of some sort may also be displayed

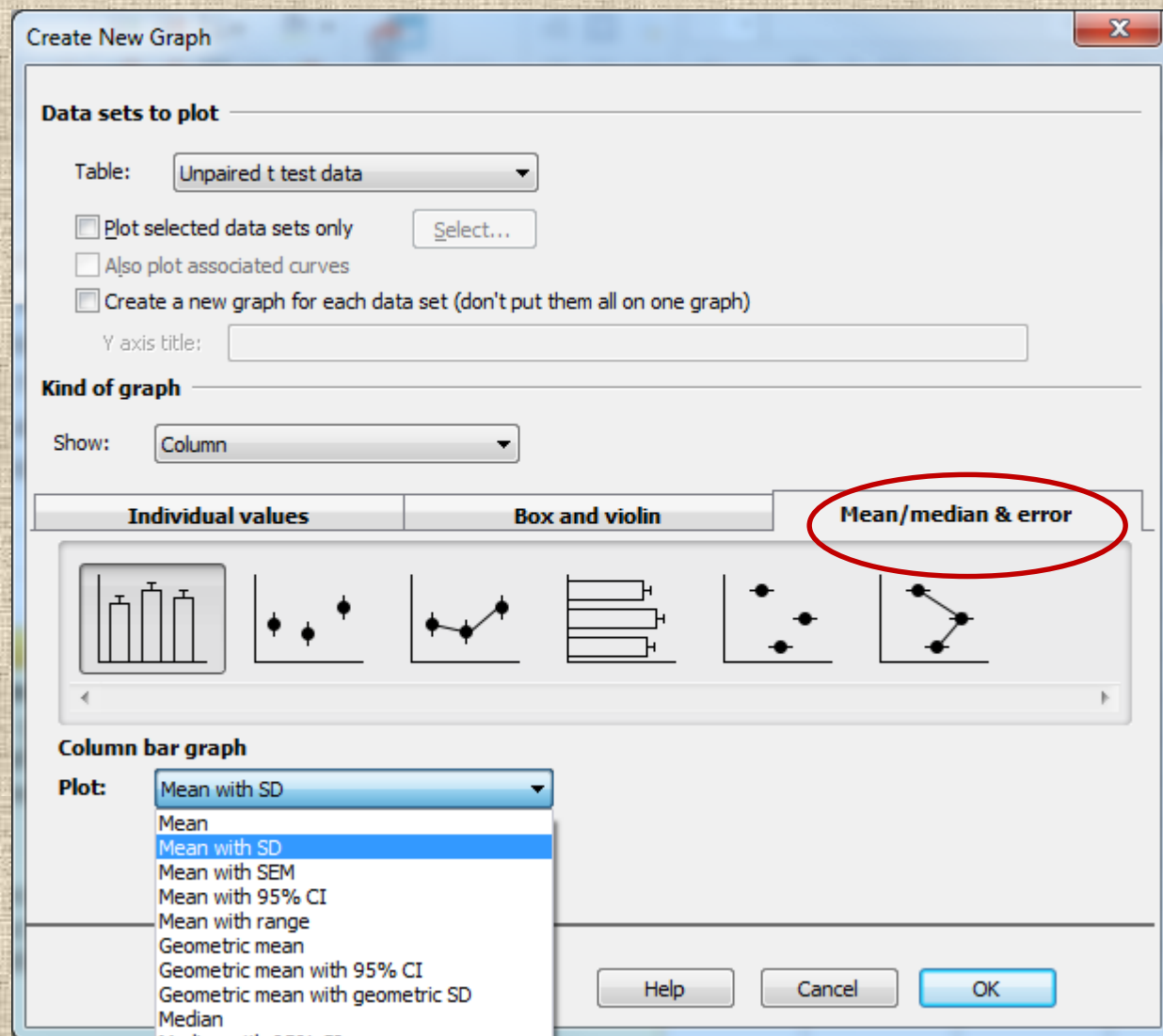
The 95% CI is most common, but *could* be \pm SEM (or SD)

It is important to read the figure legend to make sure you know what the error bars show and the sample size

It is important to put what the error bars represent in a figure legend and what the sample size is

Bar Graphs in Prism

In “Column” Table and Data, once you have created a Data Table, select the graph for that data table



Median with 95% CI
Median with range
Median with interquartile range

Create New Graph



Data sets to plot

Table: Unpaired t test data

☐ Plot selected data sets only

Select...

☐ Also plot associated curves

☐ Create a new graph for each data set (don't put them all on one graph)

Y axis title:

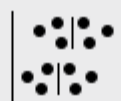
Kind of graph

Show: Column

Individual values

Box and violin

Mean/median & error



Scatter plot with bar

Plot:

Mean

Mean

Mean with SD

Mean with SEM

Mean with 95% CI

Mean with range

Geometric mean

Geometric mean with 95% CI

Geometric mean with geometric SD

Median

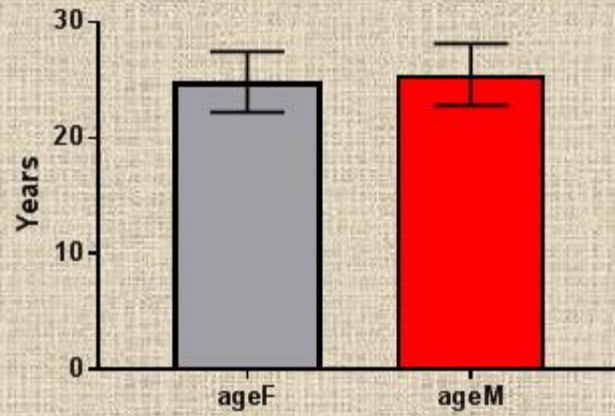
Median with 95% CI

Help

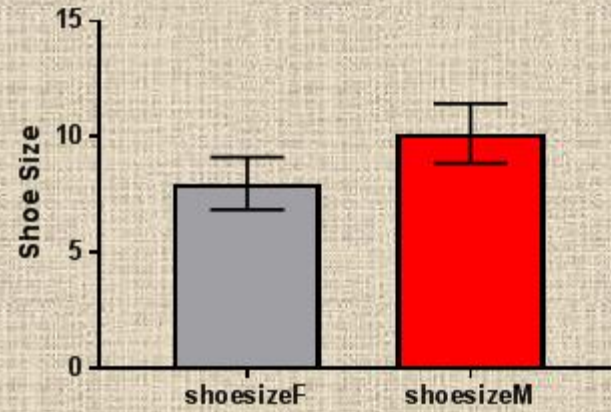
Cancel

OK

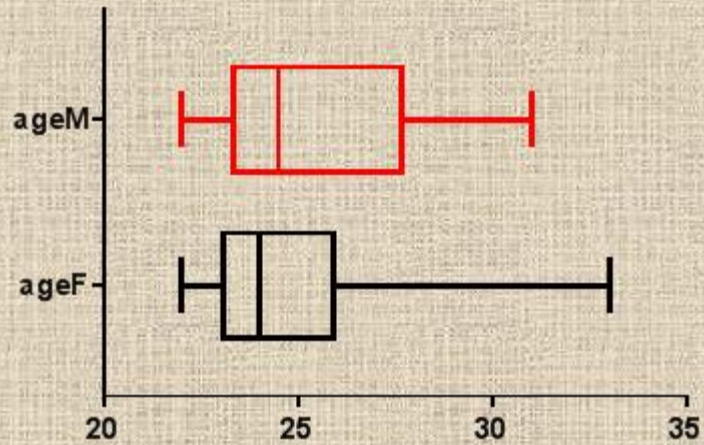
Female vs. Male Bar Mean SD



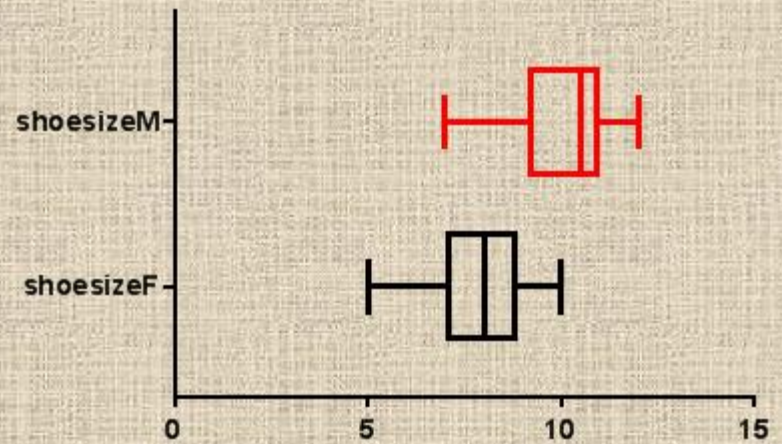
Female vs. Male Shoesize Bar Mean SD



Female vs. Male age



Female vs. Male Shoesize



The Pie Chart

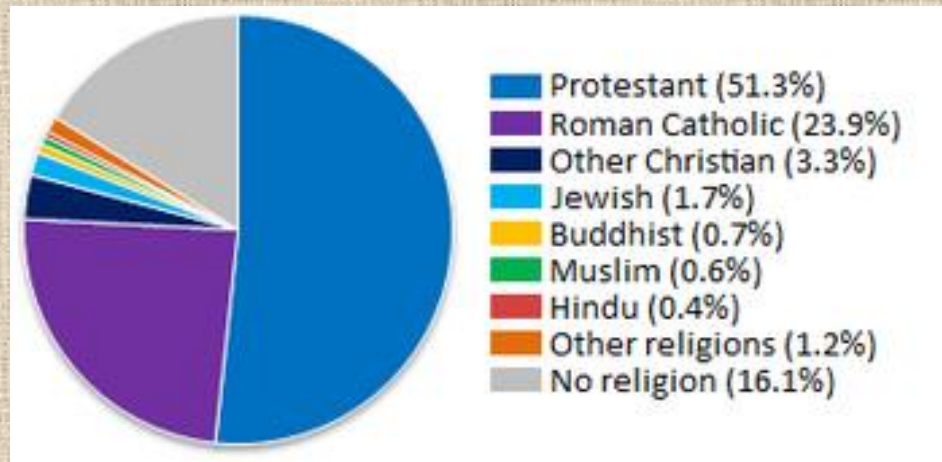


Graph showing the differences in frequencies or percentages among categories of a **nominal** or an **ordinal** variable

The larger the slice, the greater the percentage.

Categories are displayed as segments of a circle whose pieces add up to 100 percent of the total frequencies

Not used much in biomedical sciences

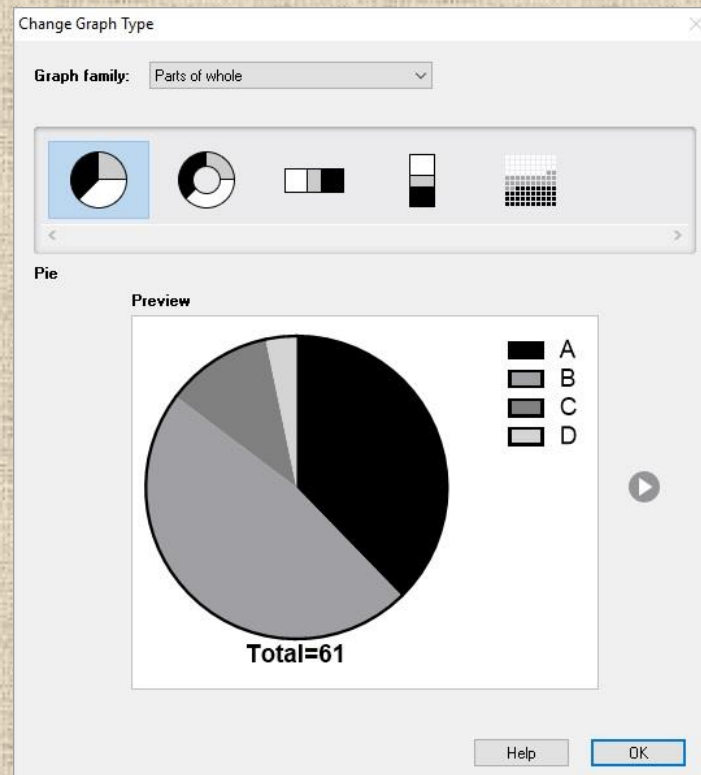


Major religions by overall percentage (2007).
US Census

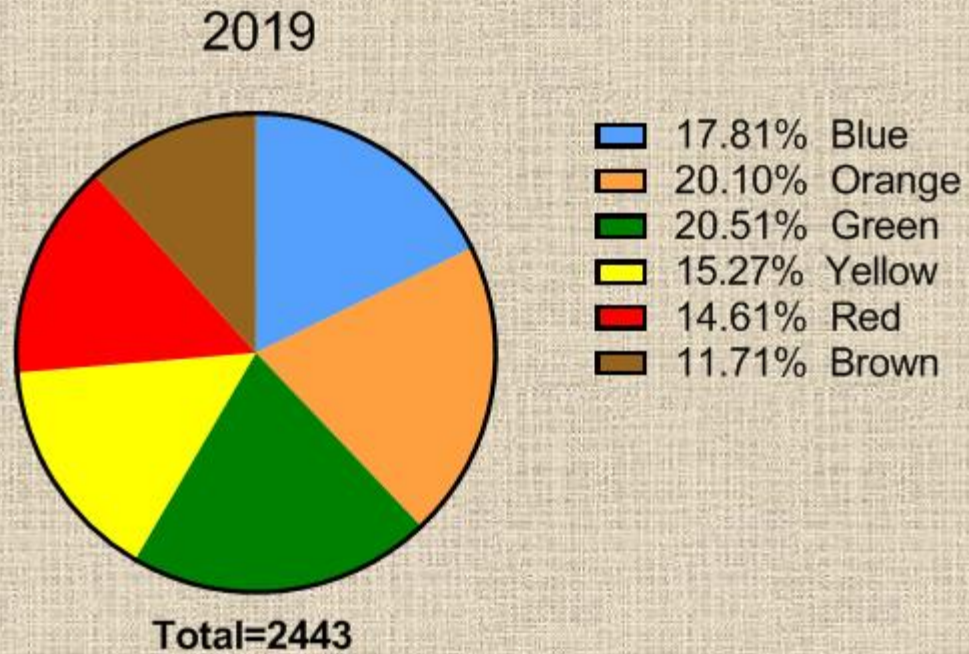
Pie Charts in Prism

Table format: Parts of whole		A	B
		Number of Students	Titl
		Y	Y
1	A	23	
2	B	29	
3	C	7	
4	D	2	
5	E	0	
6	Title		
7	Title		
8	Title		

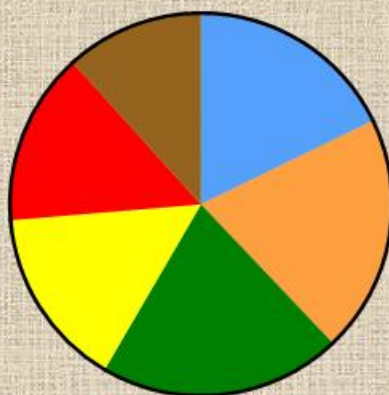
In “Parts of whole” Table and Data, once you have created a Data Table, select the graph for that data table



Your M&M data



2019



- 17.81% Blue
- 20.10% Orange
- 20.51% Green
- 15.27% Yellow
- 14.61% Red
- 11.71% Brown

Total=2443

2017



- 25.43% Blue
- 6.23% Orange
- 15.01% Green
- 12.23% Yellow
- 17.47% Red
- 23.63% Brown

Total=3051

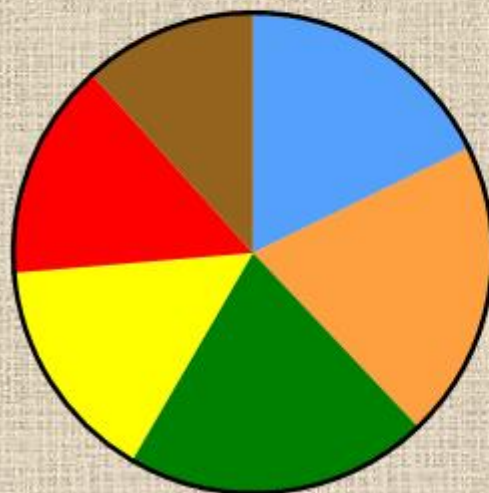
2018



- 20.80% Blue
- 20.32% Orange
- 19.07% Green
- 13.68% Yellow
- 12.10% Red
- 14.03% Brown

Total=3115

2019



17.81% Blue
20.10% Orange
20.51% Green
15.27% Yellow
14.61% Red
11.71% Brown

Factory 1



25.00% Blue
25.00% Orange
12.50% Green
12.50% Yellow
12.50% Red
12.50% Brown

Factory 2



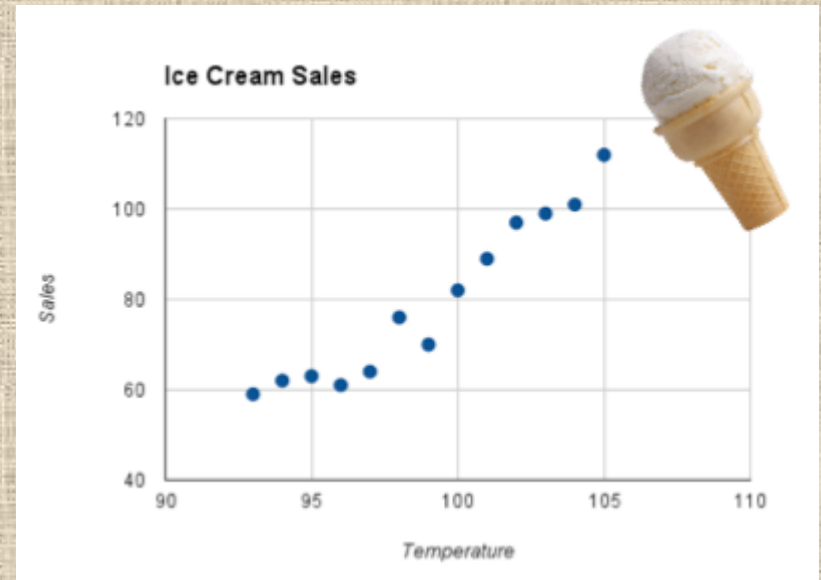
20.70% Blue
20.50% Orange
19.80% Green
13.50% Yellow
13.10% Red
12.40% Brown

Scatter Plots (AKA Scatter Graphs)

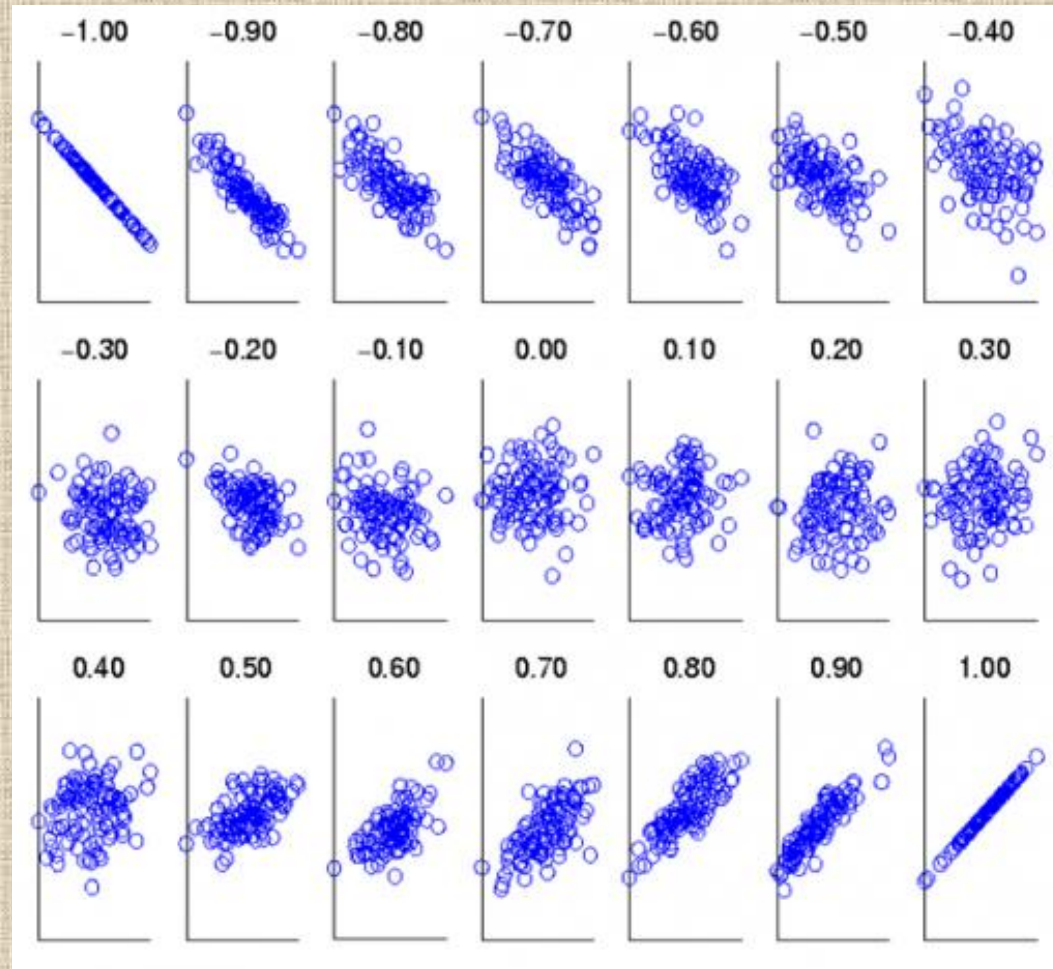
The scatter plot graphs pairs of continuous data to look for a relationship between them

Shows how one variable is correlated (related) with another. Used as part of correlation coefficient analyses

Can identify outliers, assess linearity



Scatterplots of x and y: Correlation Coefficients at Different Levels



Scatter Plots in Prism

In “XY” Table and Data, once you have created a Data Table, select the graph for that data table

Create New Graph ✕

Data sets to plot

Table: Ozone correlations ▾

☐ Plot selected data sets only Select...

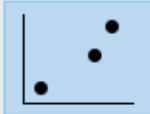




☐ Also plot associated curves

☐ Create a new graph for each data set (don't put them all on one graph)

Y axis title:

Kind of graph

Show: XY ▾

Points only

Plot: ▾ ▾

☐ Set as default for Points only

Help Cancel OK

Univariate Scatter Plots

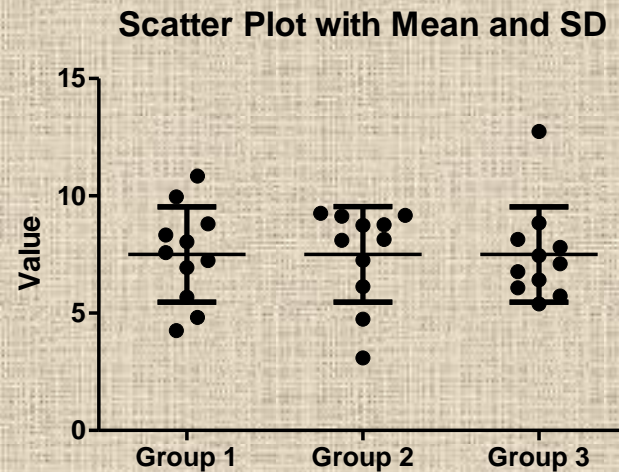
Similar to bar charts in that the X axis is categorical; y is continuous

Individual data points are shown for each category

Often a summary statistic (mean or median) is shown as a line through the dots

Error bars can also be added

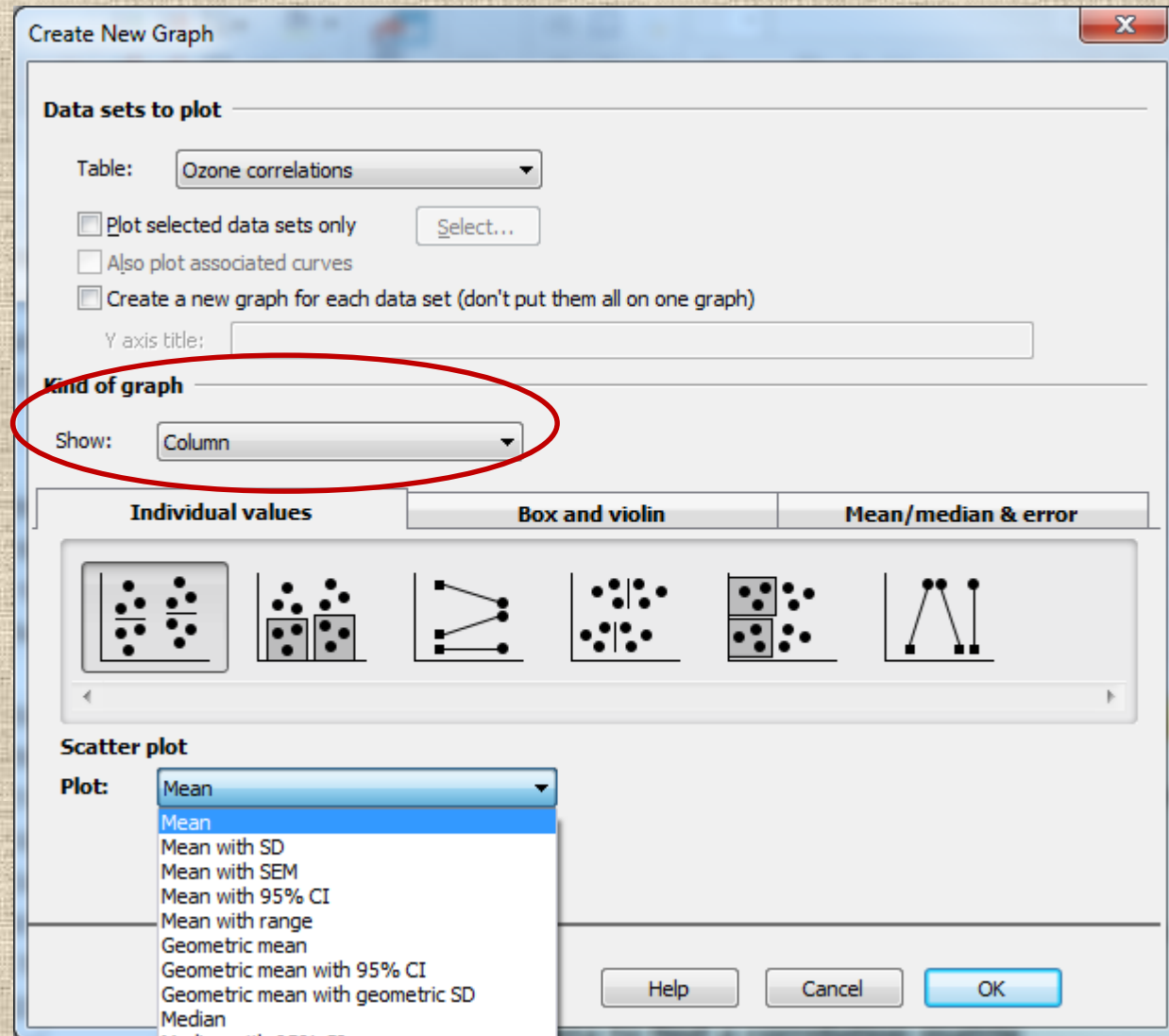
Useful and better for smaller sample sizes



Univariate Scatter Plots in Prism

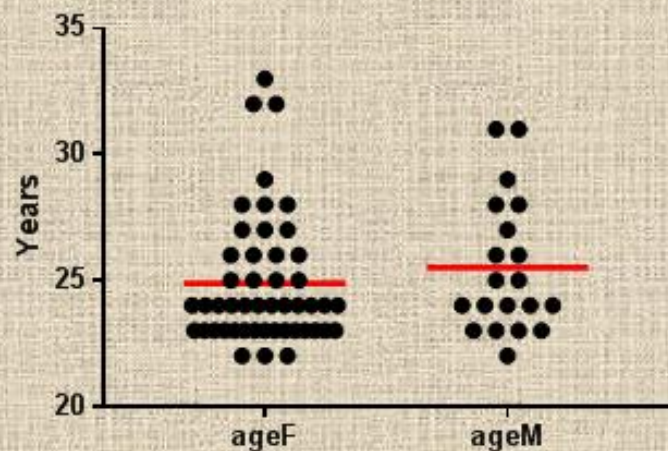
In “Column” Table and Data, once you have created a Data Table, select the graph for that data table

Can also choose Column from an XY table

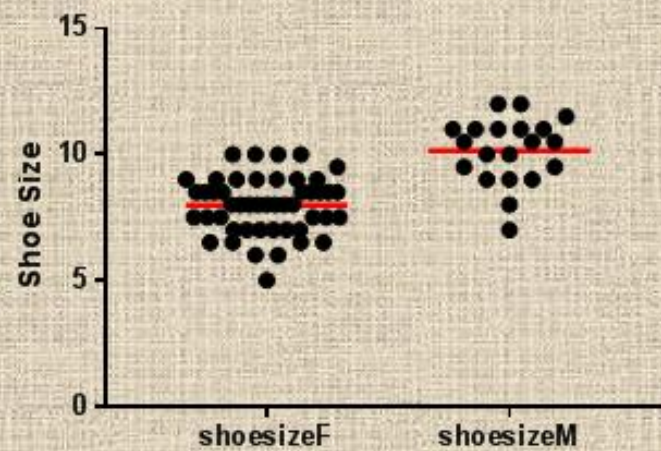


Median with 95% CI
Median with range
Median with interquartile range
No line or error bar

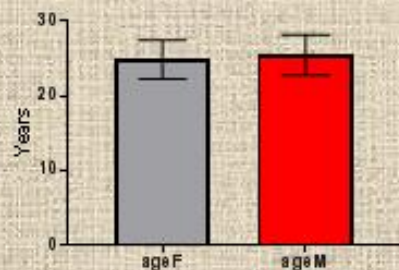
Female vs. Male Age Scatter Mean



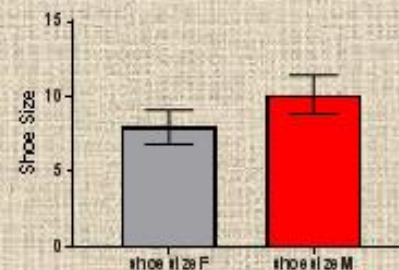
Female vs. Male Shoesize Scatter Mean



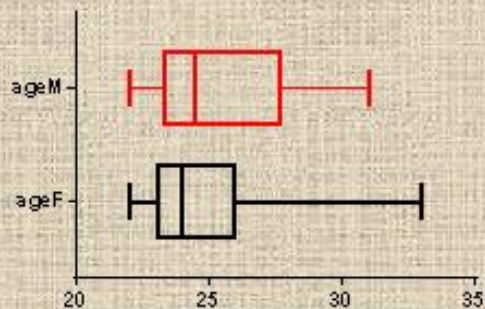
Female vs. Male Bar Mean SD



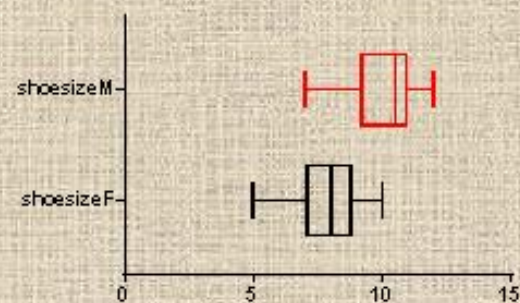
Female vs. Male Shoesize Bar Mean SD



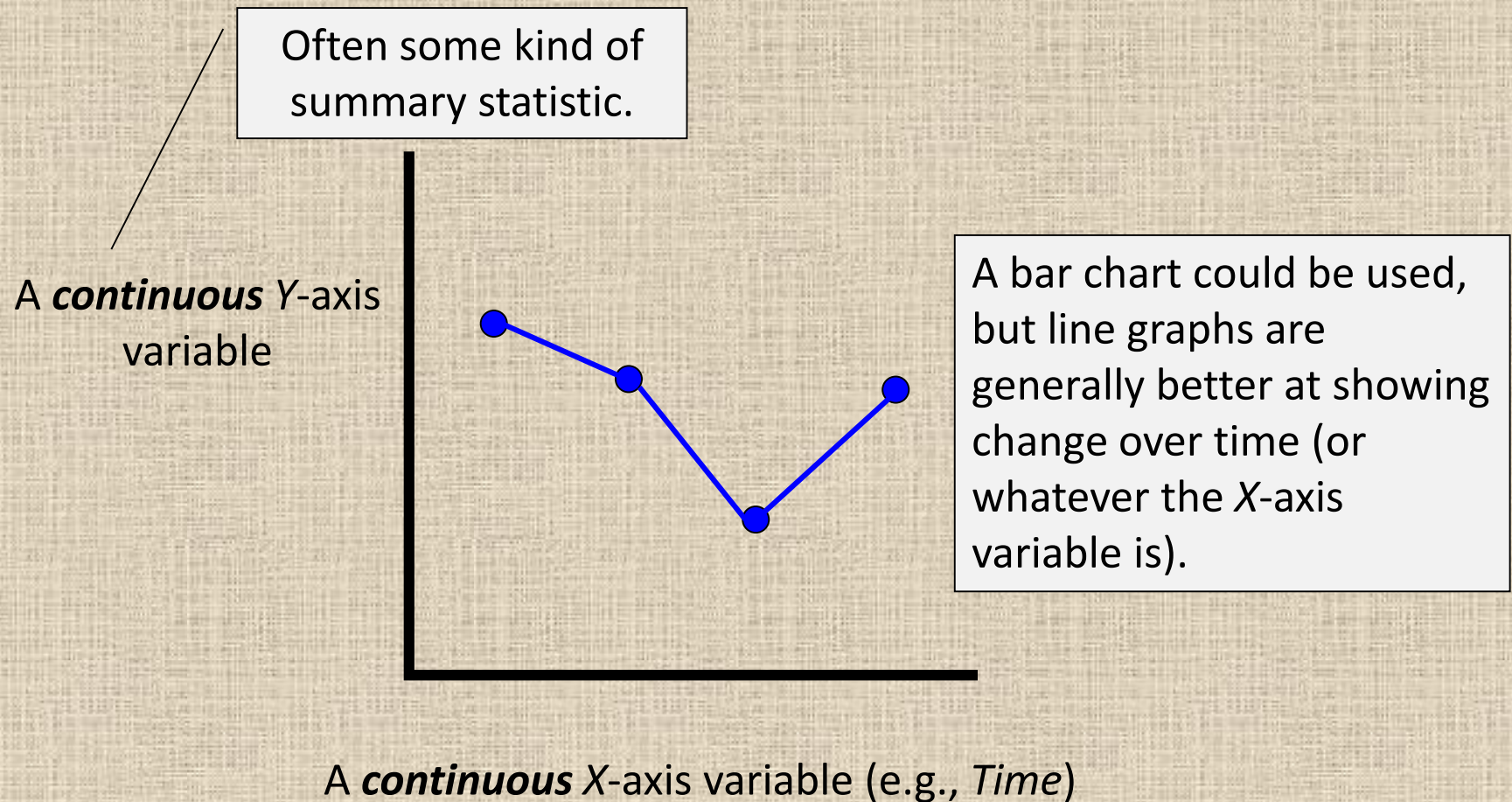
Female vs. Male age



Female vs. Male Shoesize



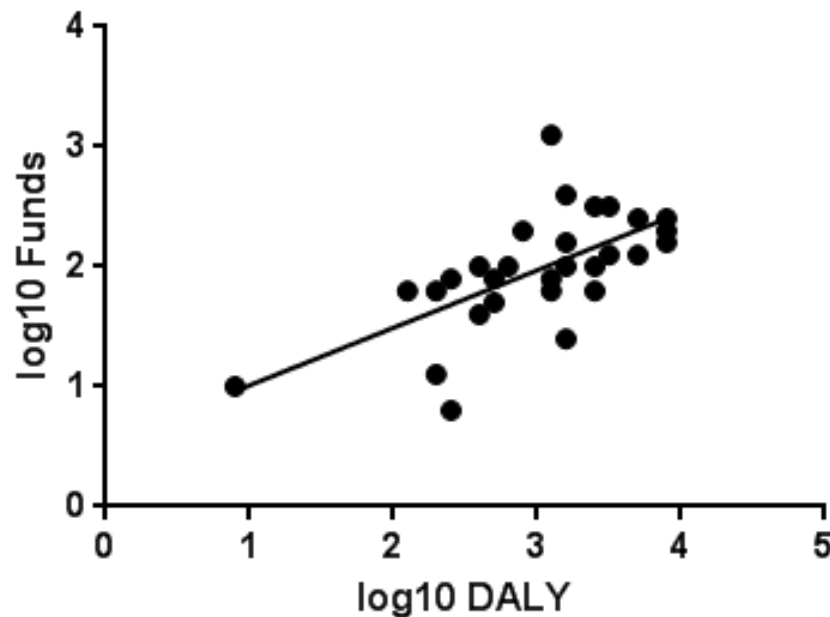
Basic Layout of a Line Graph



A Line Graph Should Not Be Confused with a Linear Regression Plot

Linear Regression

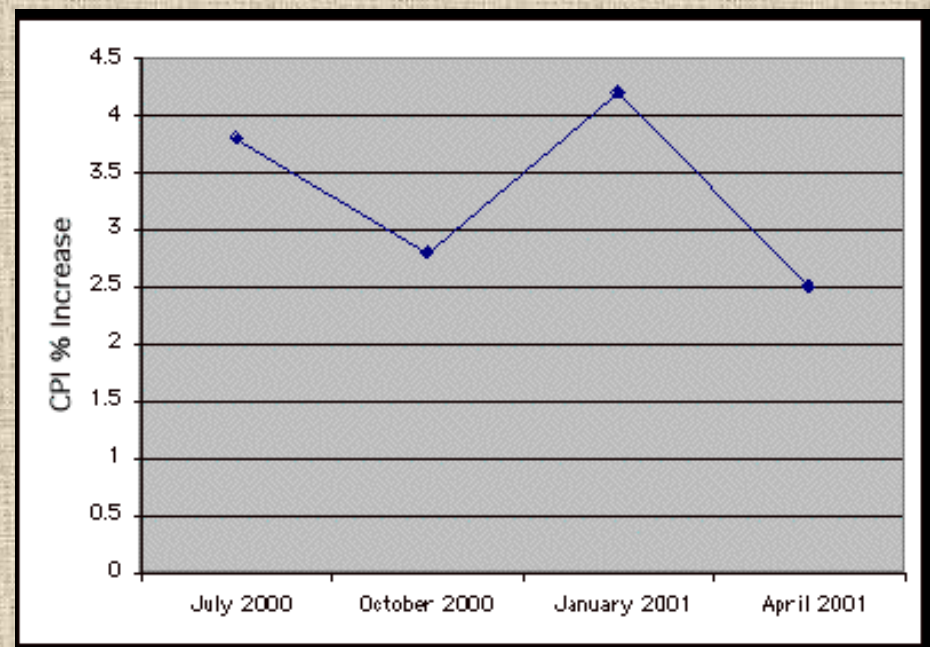
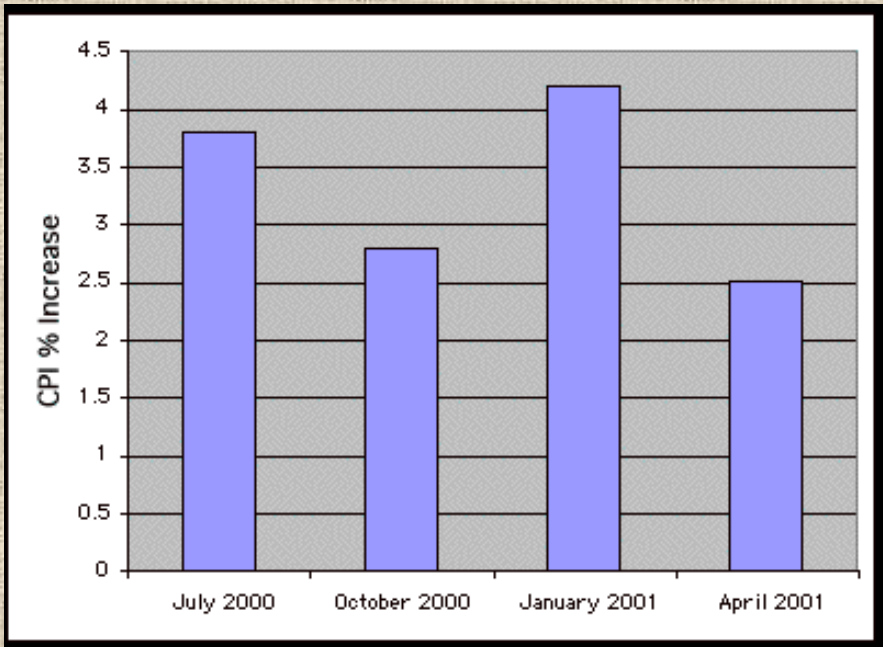
Log₁₀ All Data



A line plot connects successive data points

Linear regression tries to fit the best line through the data

Bar Chart vs. Line Graph

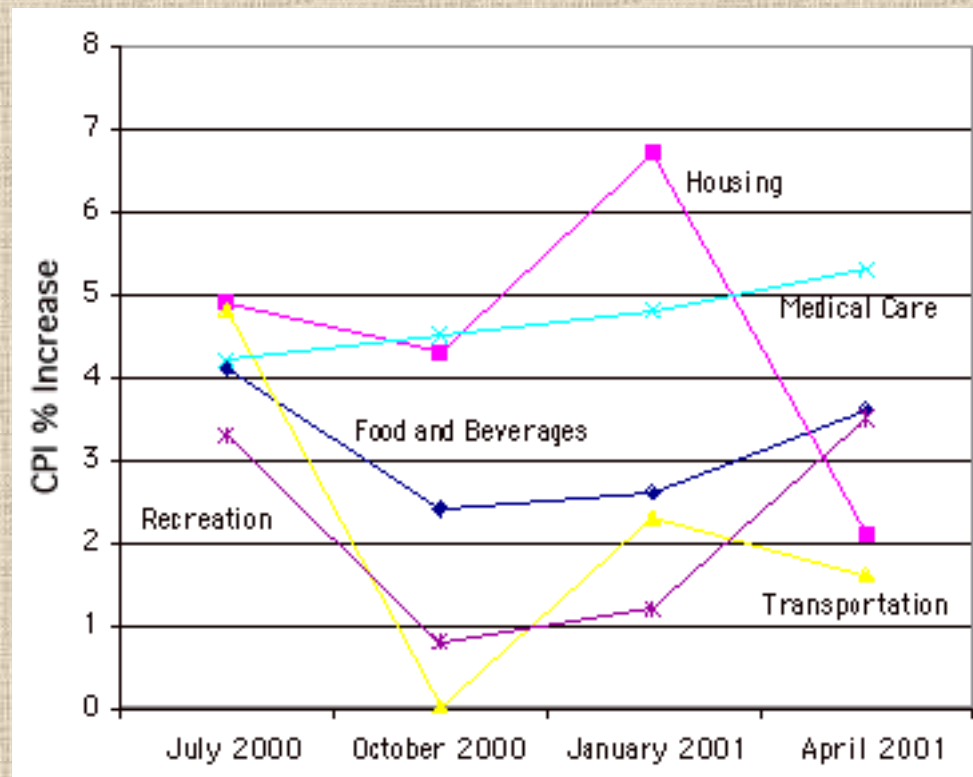


Two graphs showing Consumer Price Index (CPI) as a function of time

Changes over time are easier to see in the line graph

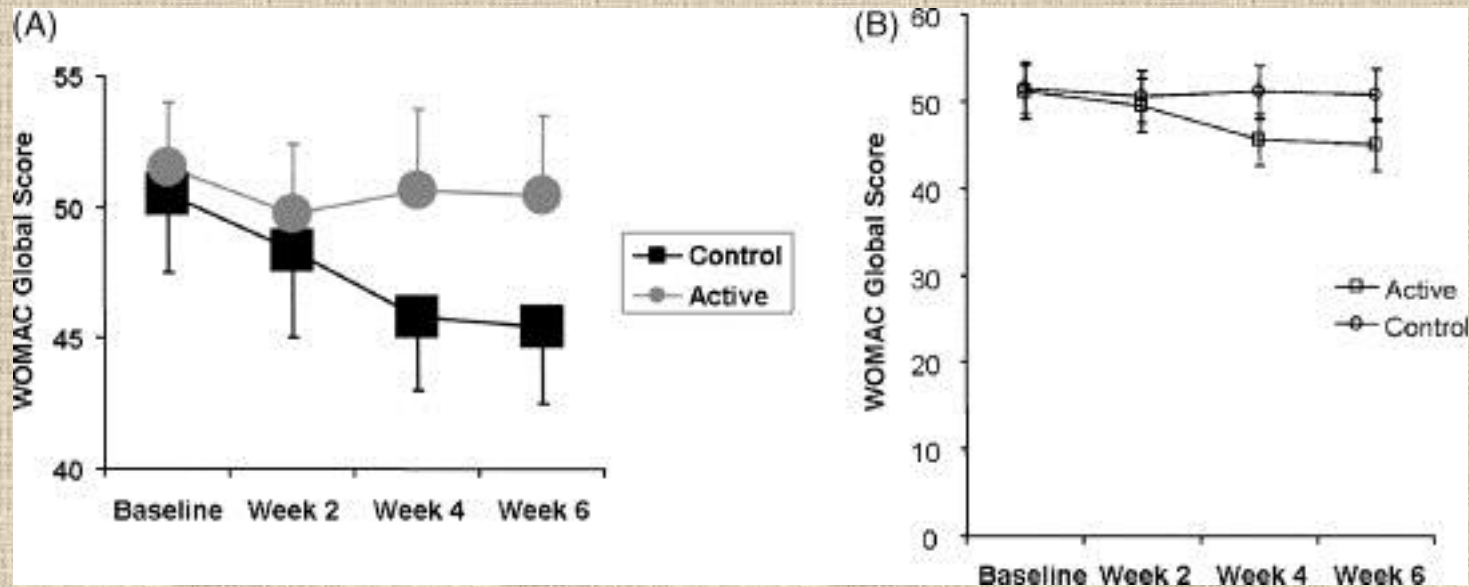
People tend to focus on differences between groups with a bar graph

Line Graph with Multiple Lines



A clustered bar chart showing these same data might be cluttered and difficult to read

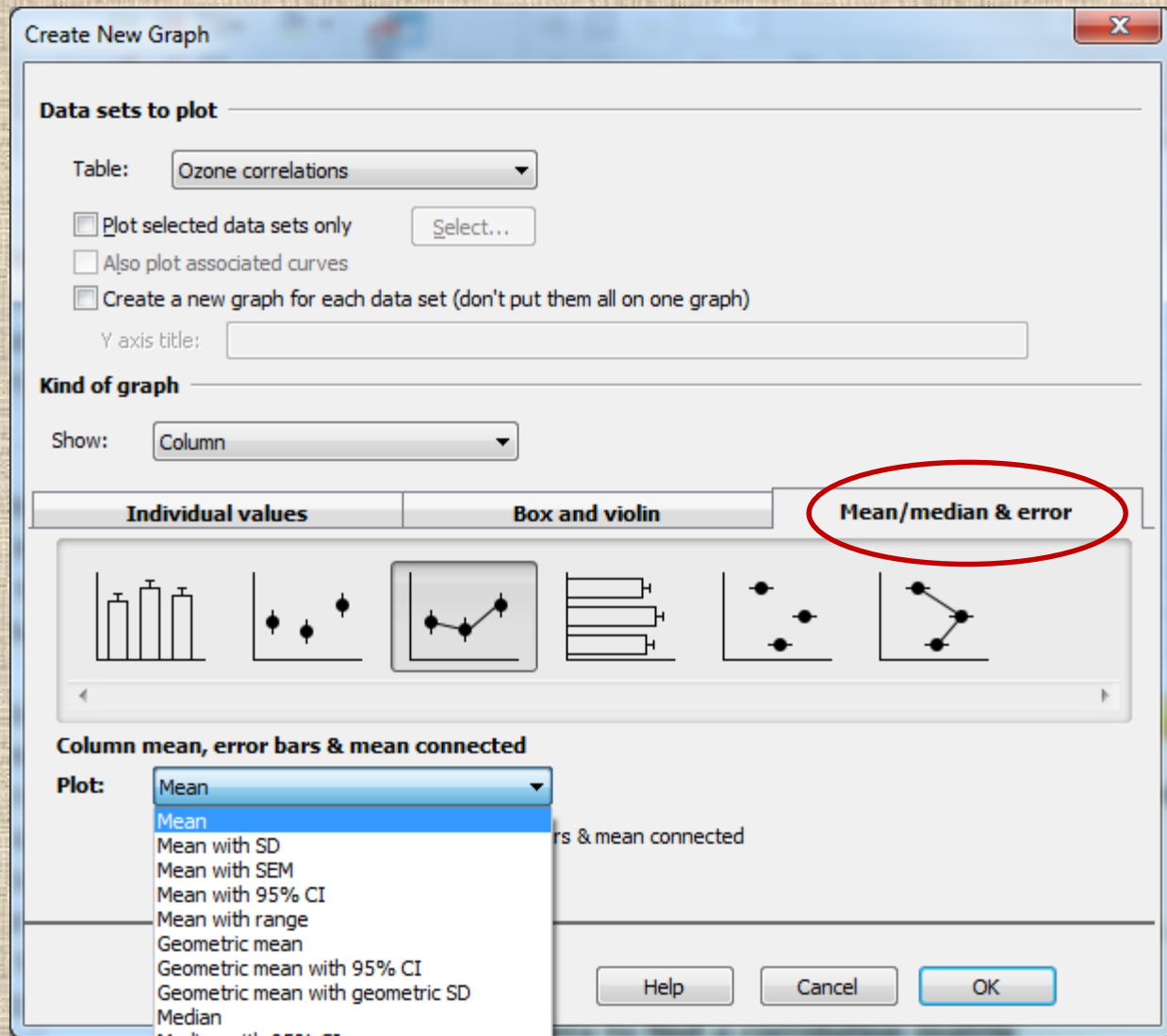
Misleading Line Graph



Misguidance of readers on the effectiveness of magnetic pulse treatment compared with placebo for knee osteoarthritis (A). The WOMAC (Western Ontario and McMaster University OA Index) may range from 0 (perfect health) to 96 (maximum pain, stiffness and functional impairment). The original y-axis depicts only the interval from 40 to 55, thereby visually inflating the effect size. The artificial difference is pronounced by one-sided error bars (standard deviations, as indicated in the statistics section of the manuscript). After rescaling (B), the ineffectiveness of magnetic pulse treatment becomes obvious.

Line Plots in Prism

In “Column” Table and Data, once you have created a Data Table, select the graph for that data table



In “XY” Table and Data, once you have created a Data Table, select the graph for that data table

Create New Graph

Data sets to plot

Table:

Ozone correlations

☐ Plot selected data sets only

Select...

☐ Also plot associated curves


☐ Create a new graph for each data set (don't put them all on one graph)


Y axis title:


Kind of graph


Show:


XY











Points only

Plot:

☐ Set as default for Points only

Help

Cancel

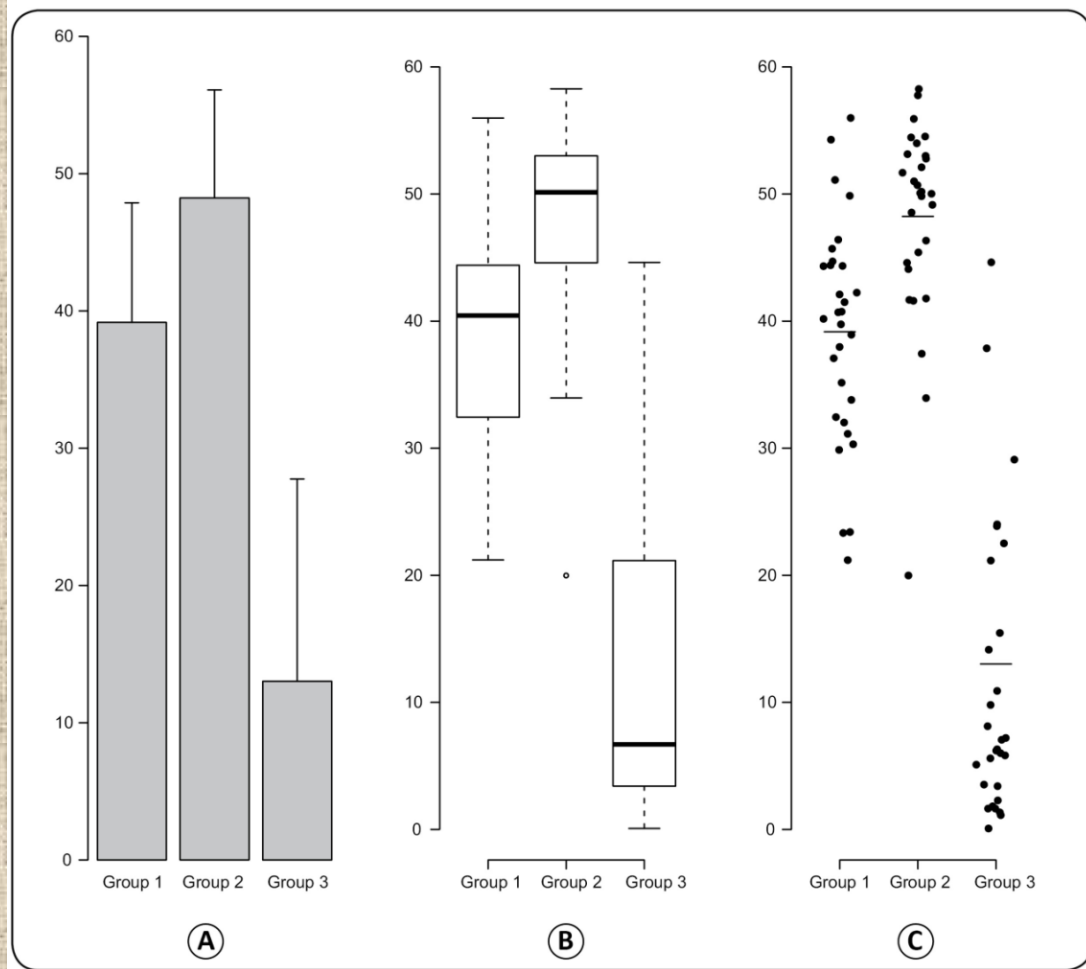
OK

Which Do I Use to Best Represent My Results?

Bar, Box, or Scatter?

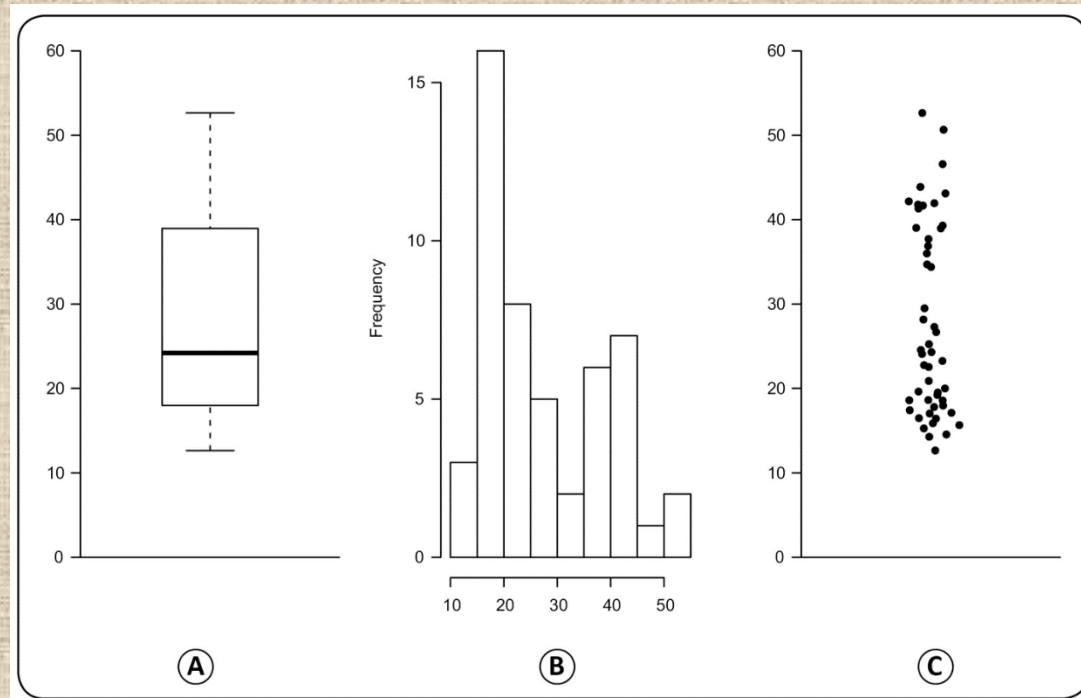


Which Do I Use to Best Represent My Results? Bar, Box, or Univariate Scatter Plots?



Data are shown for three simulated samples ($n = 30$) from “normal” (Groups 1 and 2) and non-normal (Group 3) distributions. A: Bar graphs with SD. B: Box plots (potential outlier seen). C: Univariate scatter plots with horizontal lines representing the means.

Box, Bar, or Univariate Scatter Plots for Bimodal Distributions



A: Box plot for a sample from a random variable that follows a mixture of two normal distributions. Box plots cannot clearly describe multimodal distributions

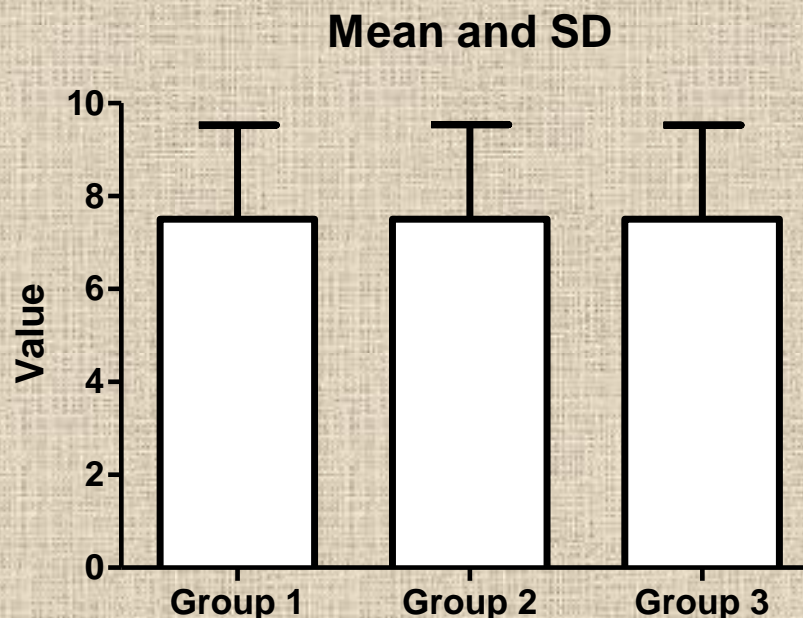
B: A histogram (not a bar graph) for these data. The bimodality is now visible in this graph.

C: A scatter plot for these data. The display of two clouds of points in this figure suggests a bimodal distribution.

Continuing the Case Against Bar Graphs

An Example: Differences of Marker Y in Three Groups

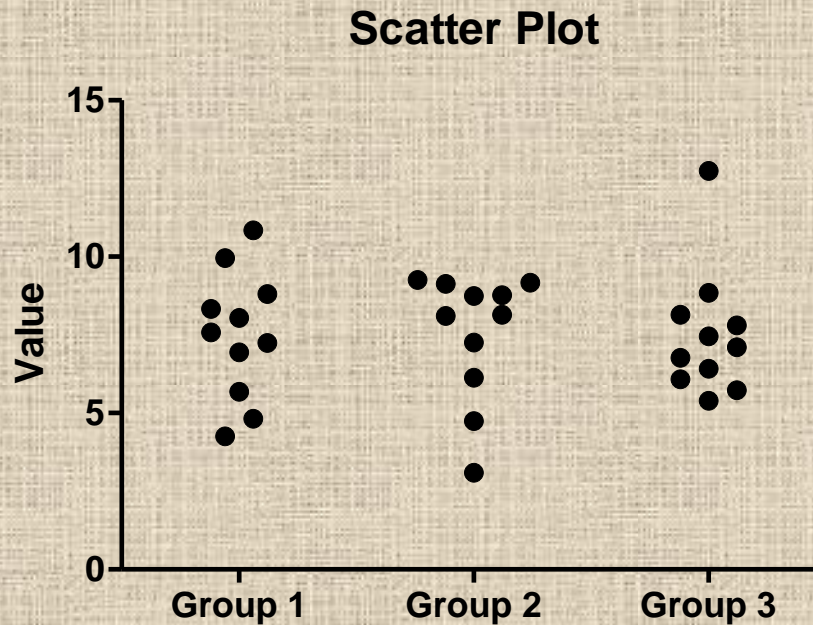
Group 1	Group 2	Group 3
4.26	3.10	5.39
5.68	4.74	5.73
7.24	6.13	6.08
4.82	7.26	6.42
6.95	8.14	6.77
8.81	8.77	7.11
8.04	9.17	7.46
8.33	9.26	7.81
10.84	9.13	8.15
7.58	8.74	12.74
9.96	8.10	8.84



N	11	11	11
Mean	7.50	7.50	7.50
SD	2.03	2.03	2.03

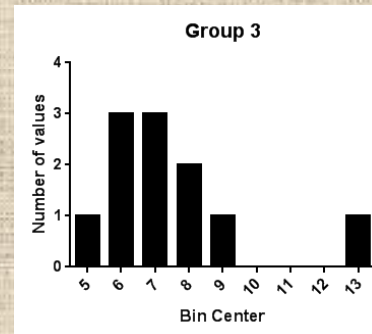
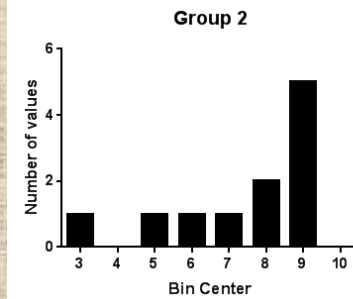
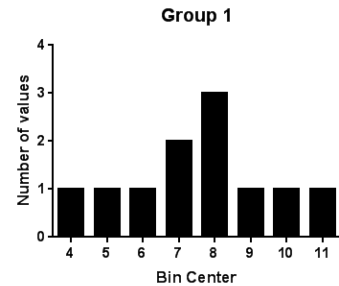
The Case Against Bar Charts

Example: Examine the Underlying Data

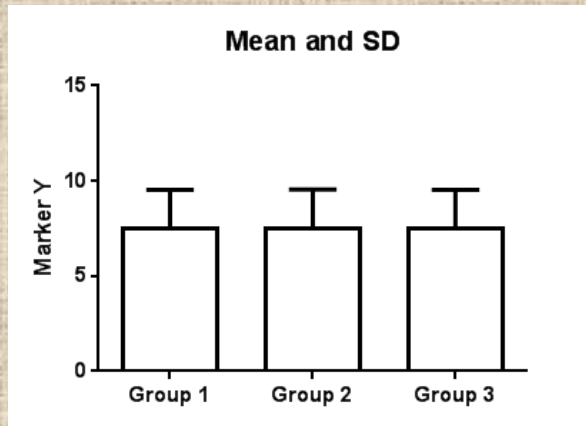


Mean	7.50	7.50	7.50
SD	2.03	2.03	2.03
Median	7.58	8.14	7.11
Sym?	Yes?	Left skew	Outlier?

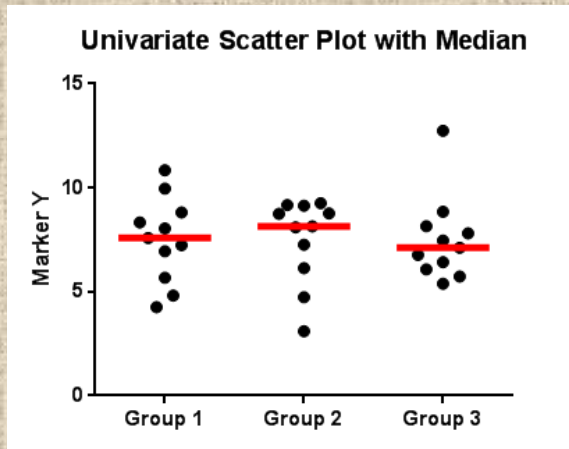
Frequency Distributions



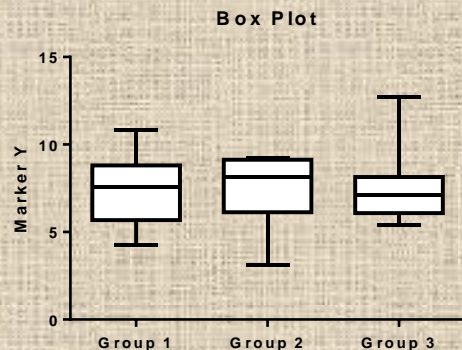
Same Data. Different Graphs.



This bar graph is not appropriate for the data. It masks data variability, distribution, and a possible outlier in Group 3. Means may not be the appropriate summary measure for all groups



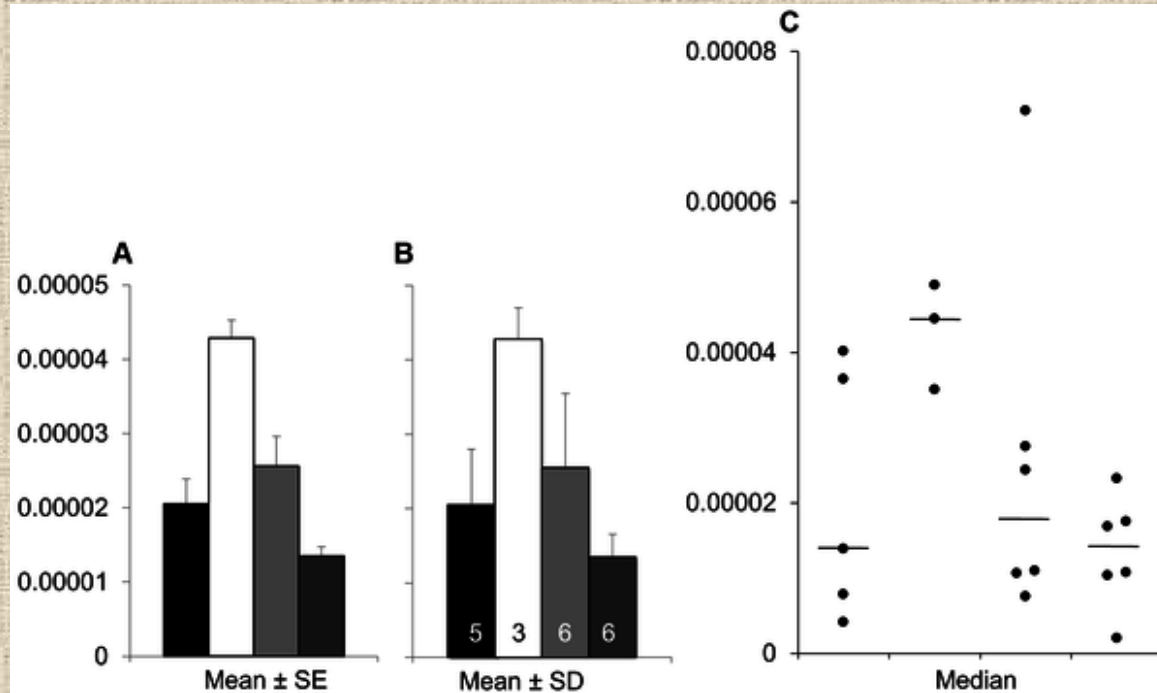
This graph is better as it shows the individual data and the variability. Putting the medians in the graph shows that they are slightly different. There also appears to be a potential “outlier.” Because of the data skew and outlier, a mean would not be appropriate for Groups 2 and 3.



This graph is better than the bar as it clearly shows the skewed distribution for Group 2 and the “outlier” in Group 3 looks like skew; the medians are different. Use this one for larger sample sizes (>20 or 30)

Bar graphs and scatterplots convey very different information

Placental endothelin 1 (*EDN1*) mRNA data for four different groups



Showing SEM rather than SD magnifies the apparent visual differences between groups, and this is exacerbated by the fact that SEM obscures any effect of unequal sample size. The univariate scatterplot (Panel C) clearly shows that the sample sizes are small, group one has a much larger variance than the other groups, and there is an outlier in group three. These problems are not apparent in the bar graphs shown in Panels A and B.

Good Graphing Practice

nature
cell biology

EDITORIAL

An update on data reporting standards

We discuss editorial policies that aim to facilitate transparency and reproducibility, and their impact on the research content published in *Nature Cell Biology*.

NATURE CELL BIOLOGY VOLUME 16 | NUMBER 5 | MAY 2014

we continue to strongly encourage authors to show the full spread of individual data points in the figures where possible, particularly when sample size is small. Although bar graphs are the norm for data representation in most cell biology papers, we are pleased to see a small proportion of papers beginning to explore alternative formats for data display, including plotting the individual data points (see for example: *Nat. Cell Biol.* 15, 1351–1361 (2013); *Nat. Cell Biol.* 15, 1294–1306 (2013)). We join our sister journal, *Nature Methods*, in urging our authors to use box plots when sample size is greater than 5, and invite readers to explore BoxPlotR, an online tool for generating box plots developed as a collaboration between Nature Publishing Group and the community (*Nat. Methods* 11, 113; 2014).

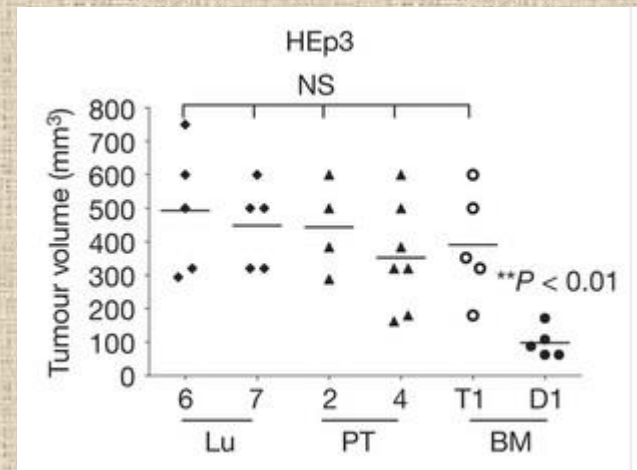
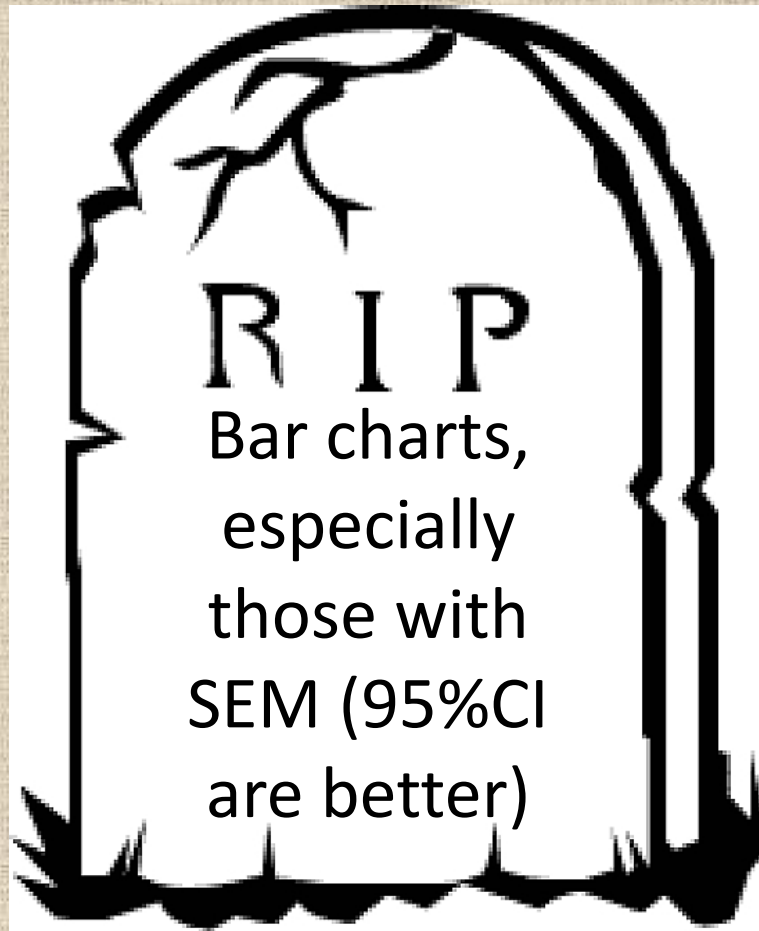


Fig. 1C: *Nat. Cell Biol.* 15, 1351–1361 (2013);



OK, bar graphs are not always the best way to display my data, but what are the best error bars to use?

That depends...

The Sample SEM vs. the Sample SD

In biomedical journals, SEM and SD are often used interchangeably to express the variability

But they measure different things

SEM quantifies uncertainty in the estimate of the population mean from your sample

is an estimate of how far the sample mean is likely to be from the unknown *population* mean

SD describes variability of the sample data from the sample mean
the degree to which individual values within the sample differ from the calculated *sample* mean

A Better Error Bar: Confidence Intervals (CI) of a Mean

A case against the SEM

A confidence interval gives an *estimated range of values that is likely to include the unknown population mean*

The estimated range is calculated from your sample.

The confidence interval (CI) of a point estimate (*i.e.*, mean) from a sample describes the precision of this estimate

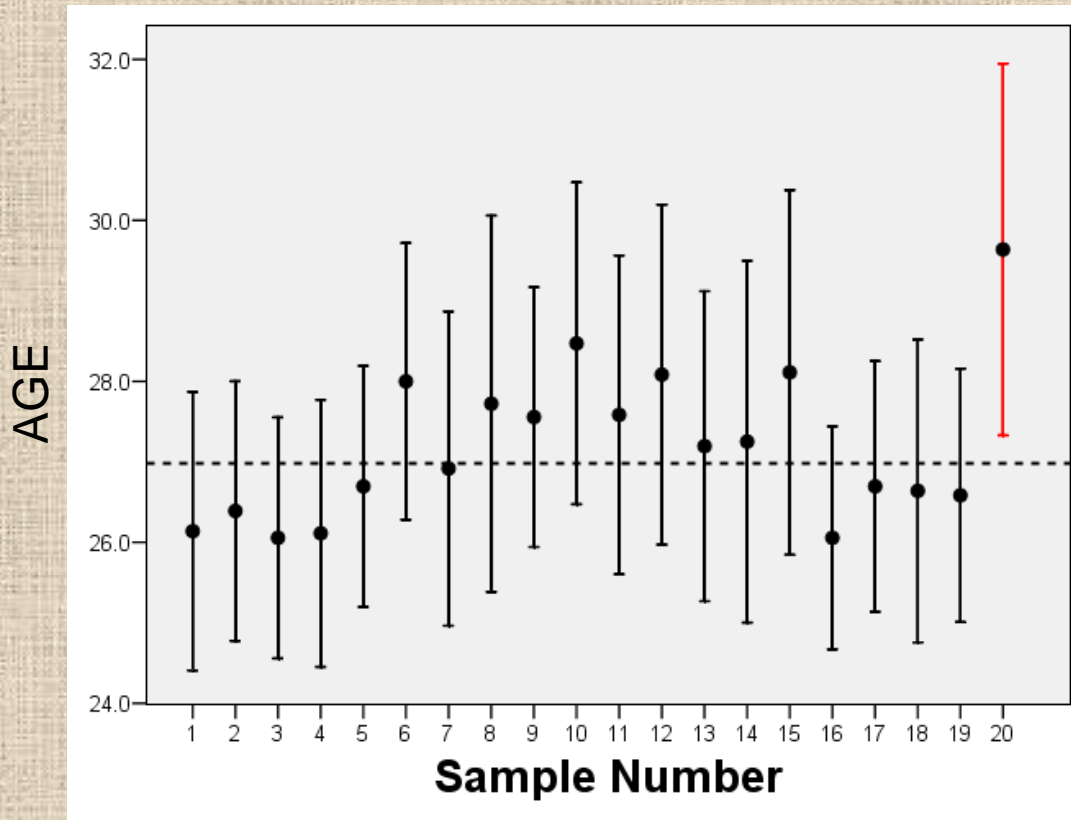
The CI represents a range of values on either side of the estimate. The narrower the CI, the more precise the point estimate

A 95% Confidence Interval is expected to contain the population mean 95% of the time (*i.e.*, of 95% CIs from 100 samples, 95 will contain the population mean)

$$\bar{X} \pm t_{95\%, n-1} SEM$$

95% CI for the Mean from Random Samples

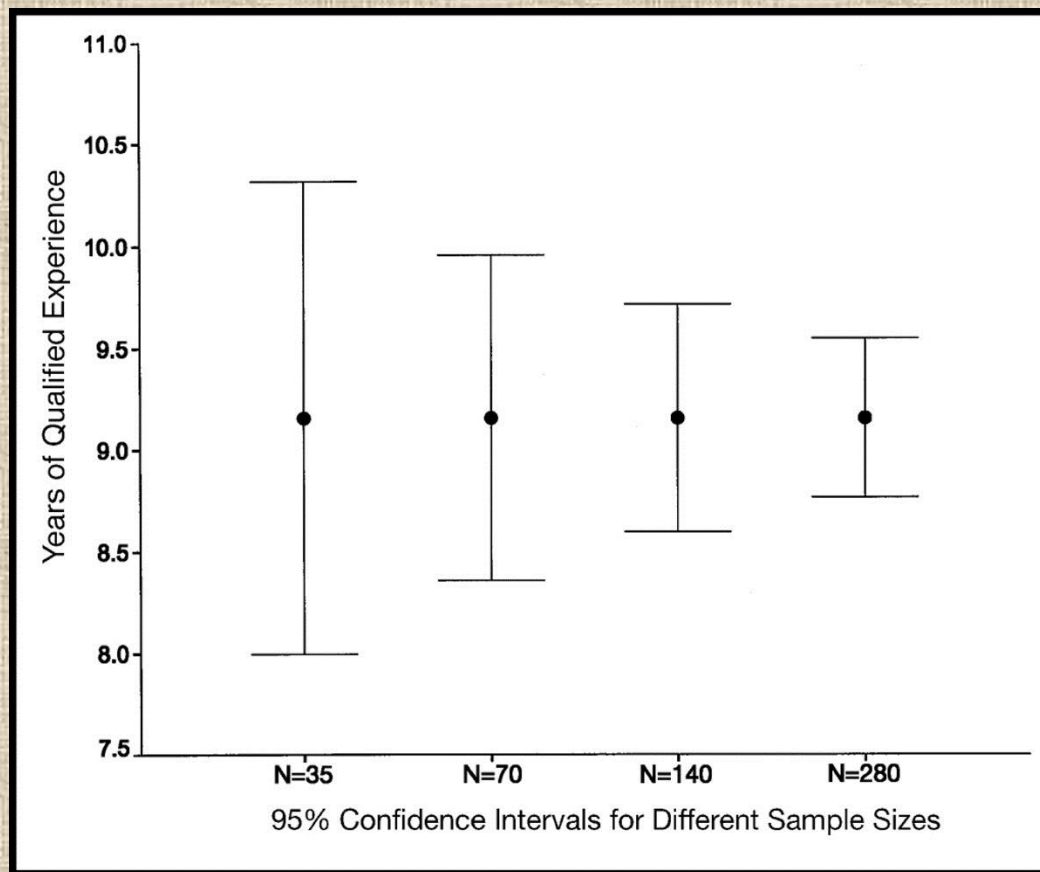
95% CIs for 20 random samples from the population of medical students



Notice that 19 out of 20, or 95% of the 95% confidence intervals, contain the population mean.

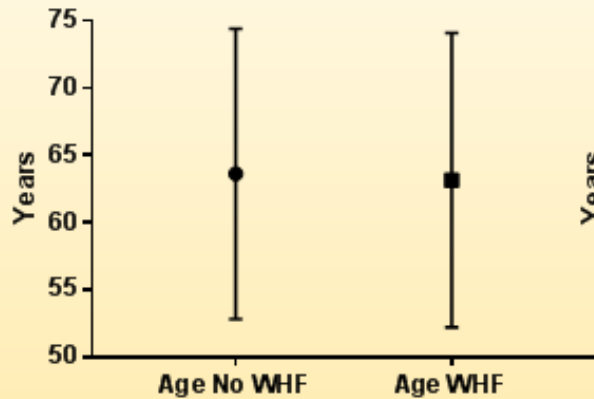
← Population Mean (26.98)

Confidence intervals and mean for progressively larger samples drawn from a single population



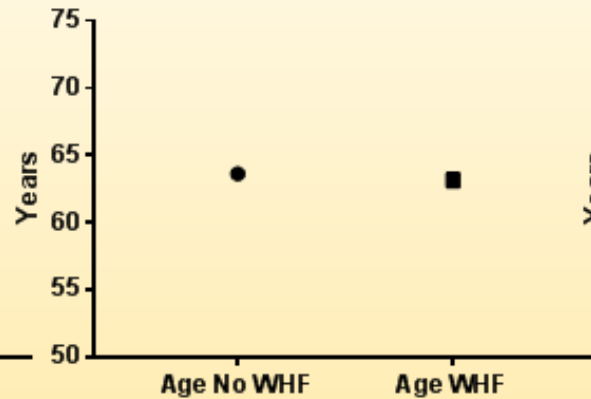
Descriptive

Age (Mean with SD)



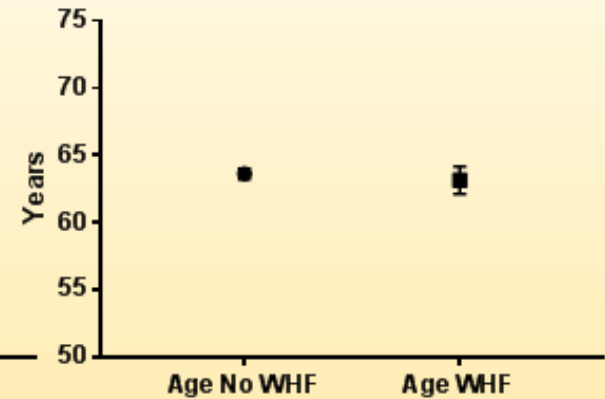
Inferential

Age (Mean with SEM)



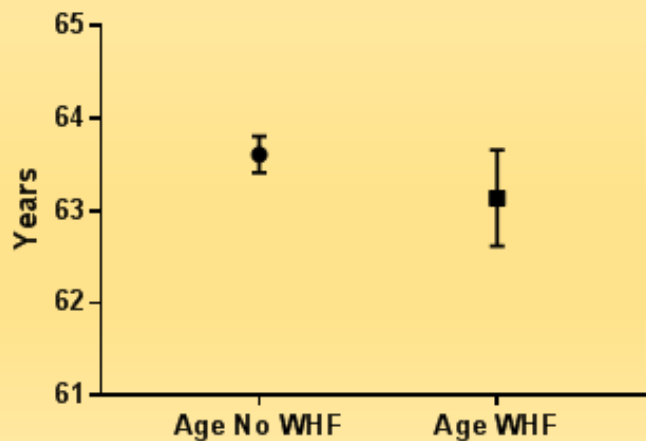
Inferential

Age (Mean with 95% CI)

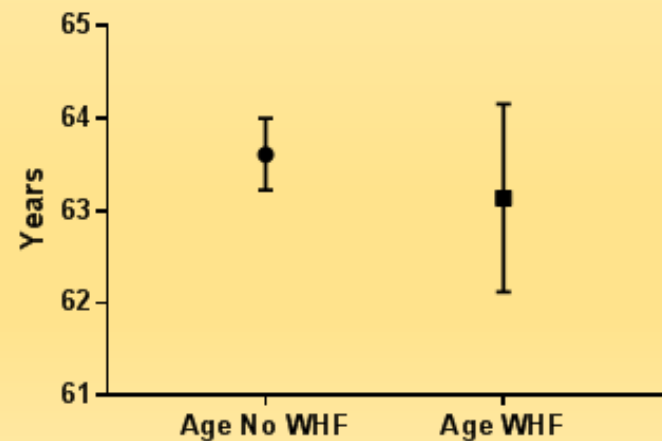


Expanding the y-axis to see SEM and CI bars

Age (Mean with SEM)

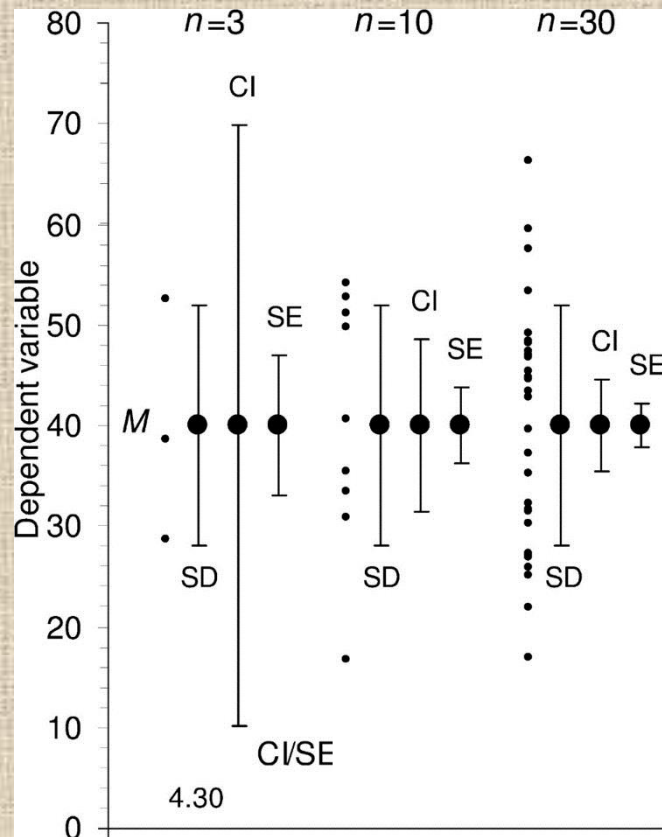


Age (Mean with 95% CI)



95%CI are wider than SEM, but you are more likely to “capture” the true population value.

Descriptive (SD) and Inferential (CI, SEM) Error Bars



SDs provide descriptive information, while SEMs and CIs are both relevant to inference and provide information about precision.

Notice that with increasing sample size, SD remain about the same size, but CI and SEM (SE) get smaller.

Which do I Use: SD or SEM (or 95%CI)?

It depends what your questions are

“What is the variability of my sample data?”

use SD or just graph all data points if n is small (<20)

“What is the likely mean of the population?”

95%CI are valuable

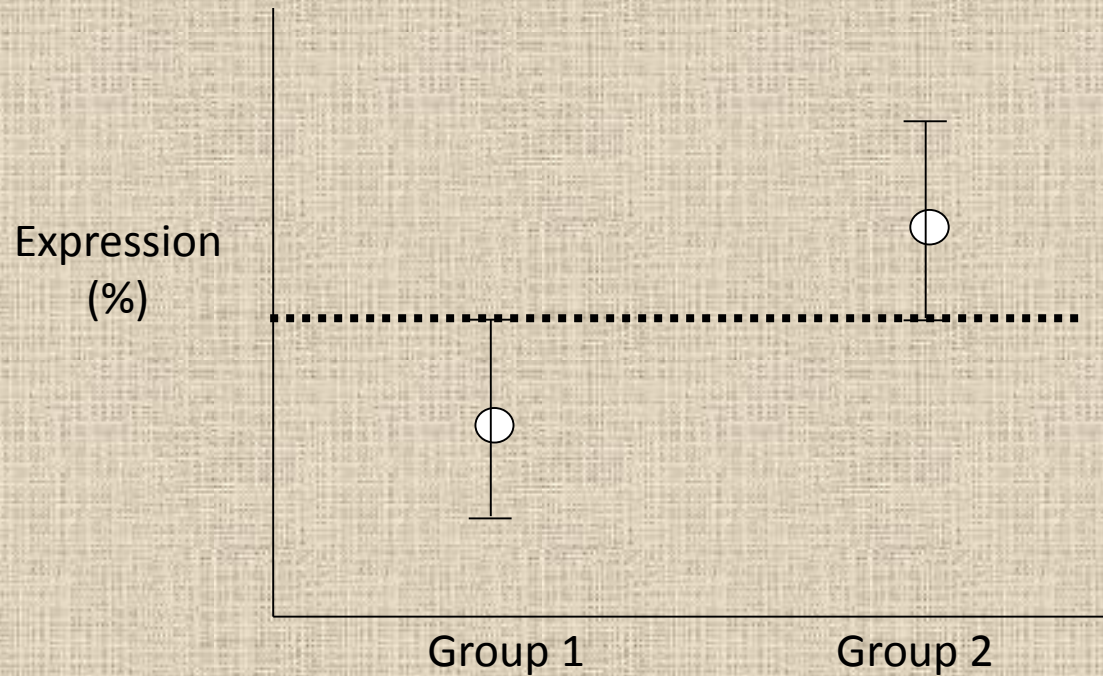
gives a sense of how accurately you have determined the population mean from your sample

“Are groups different determined by a statistical test (do bars overlap)?”

SEM and 95%CI can be informative if you know how to interpret them

Depends on sample size

What Do Scientists Think Error Bars Tell Them?



Most scientists follow the just abutting rule regardless of what the error bars are.

They conclude $p < 0.05$

Using Error Bars to Compare Groups of Data

SD

Demonstrates data variability, but no comparison possible

SEM

If bars overlap, any difference in means $p > 0.05$

If they don't overlap, the results might be $p \leq 0.05$ or they may not be (depending on n and distance between bars)

95% CI

If bars overlap, p could be ≤ 0.05 or > 0.05

Useful rule of thumb: If two 95% CI error bars do not overlap, and the sample sizes are nearly equal, the difference is statistically significant with a p -value much less than 0.05

More on this later....

Assignment

Use same paper as last week

In Week03 folders

Due next Wednesday before class