Chapter 3: Descriptive Statistics

TXCL7565/PHSC7565

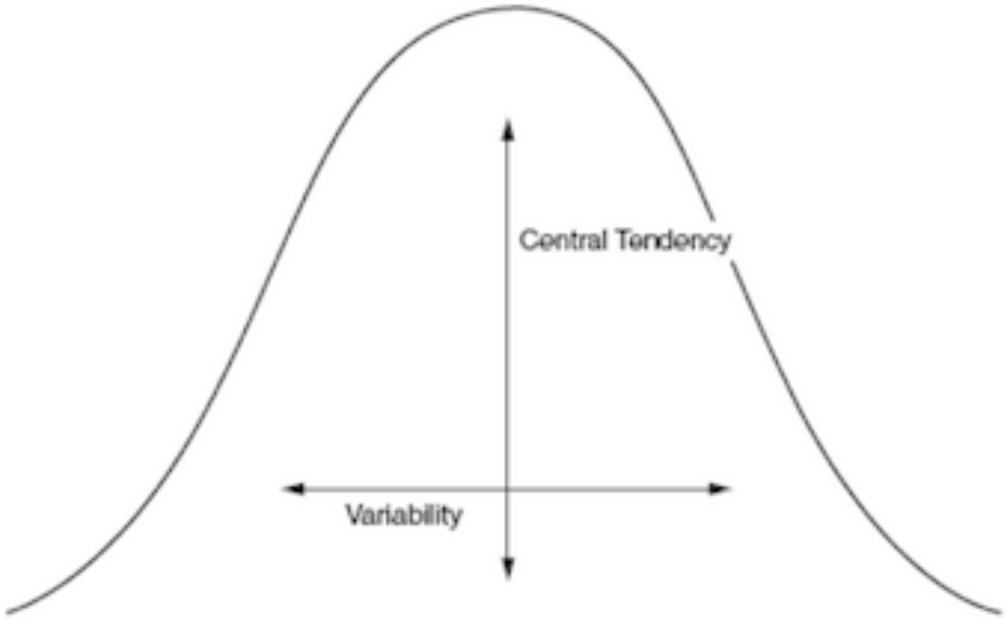
What This Lecture Covers

- Indicators of central tendency: mean, median, mode
- Describing variability: standard deviation and coefficient of variation
- Quartiles
- Describing ordinal data
- Descriptive statistics in GraphPad

Descriptive Statistics

With large data sets, we often ask:

1. How large are the values?
   - indicators of central tendency
2. How variable are the values?
   - Indicators of variability/dispersion



INDICATORS OF CENTRAL TENDENCY: MEAN, MEDIAN, MODE

Indicators of Central Tendency

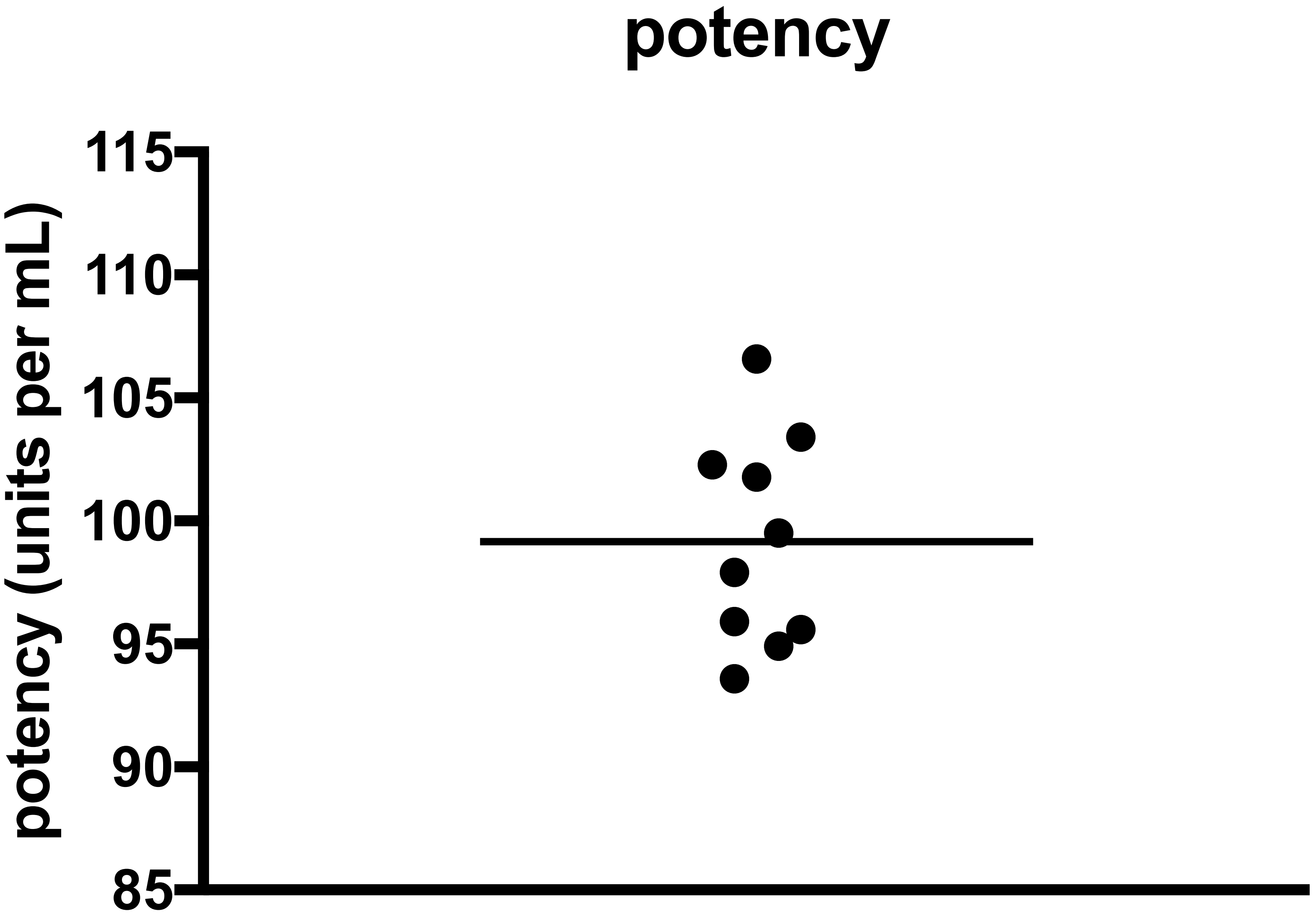
Indicator of Central Tendency - any statistic used to indicate a value around which the data are clustered

Mean

- One of the most common statistics calculated on any data set
- Arithmetic mean = Average

$$mean = \frac{\sum_{i=1}^{N} x_i}{N}$$

Mean of Potency



Median

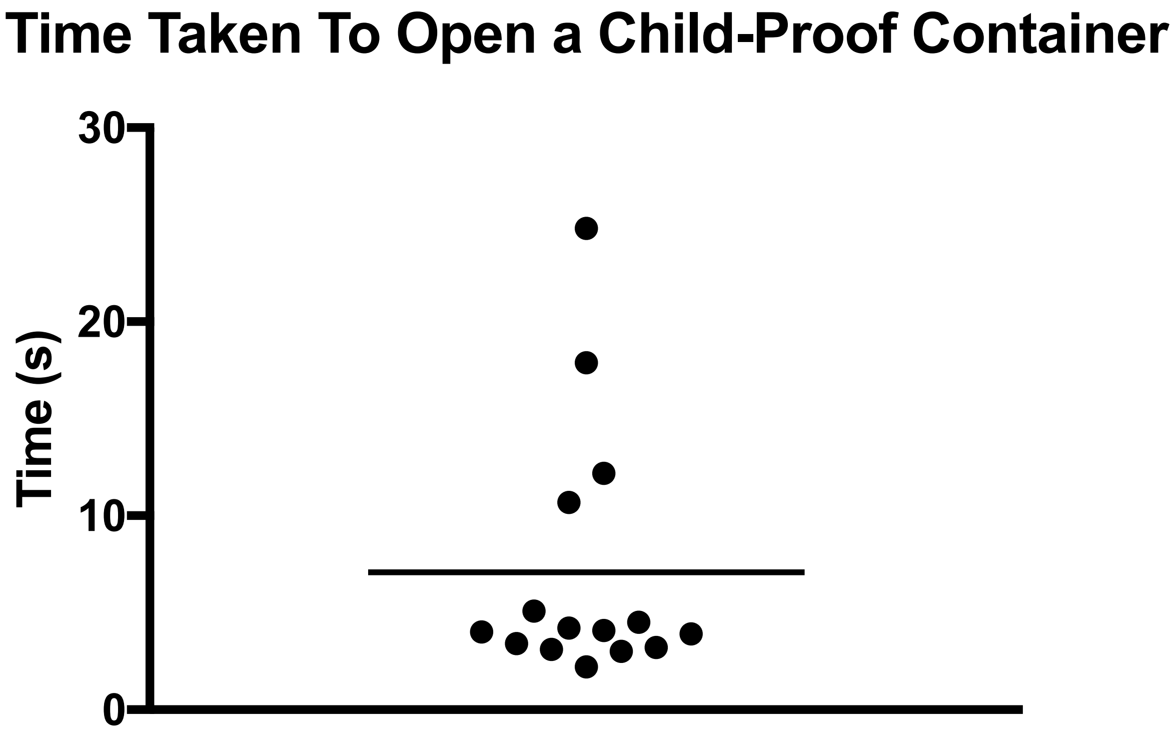
Often, we want a measure of the central tendency that is more **robust** to outliers than the mean.

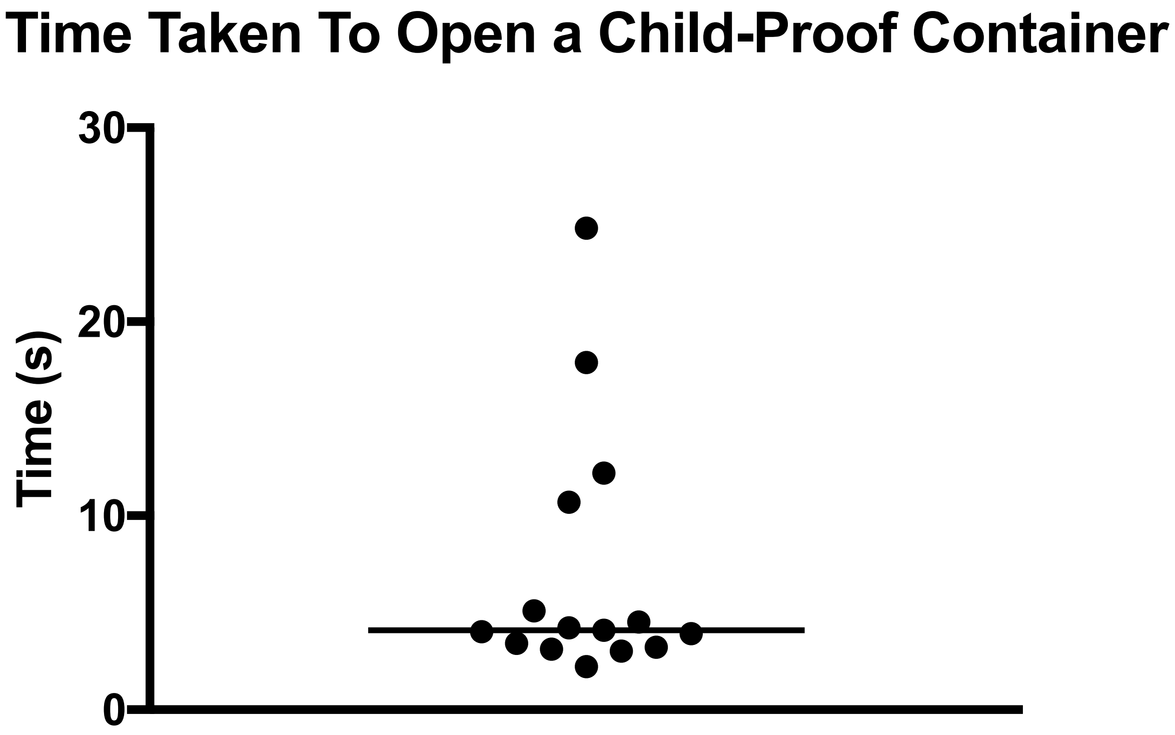
- median = middle value = 50th percentile
- If even number of samples, take the average of the two middle ones
    - e.g., for potency (N=10), take the average of the 5th and 6th ranked values

Median of Time Takes to Open a Child-Proof Container

Rank	Time
1	2.2
2	3.0
3	3.1
4	3.2
5	3.4
6	3.9
7	4.0
8	4.1
9	4.2
10	4.5
11	5.1
12	10.7
13	12.2
14	17.9
15	24.8

Median of Time Taken to Open Child-Proof Container



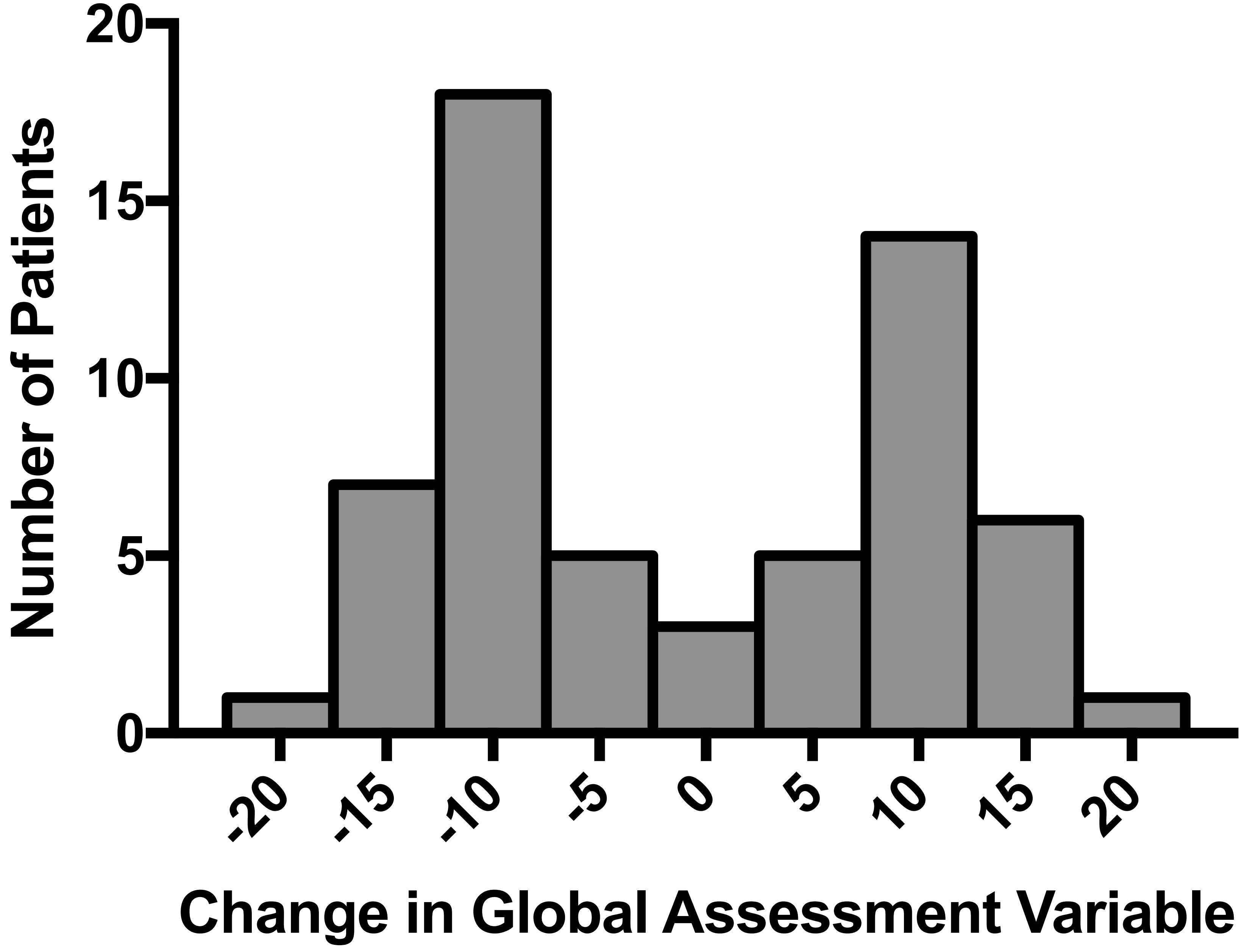


Mode

Rather than the mean, we might want to know what is the most likely value. In that case, you would report the **mode** instead.

- Mode - the value(s) that occurs most often
- Usually only relevant if the values are integers or a number with only one or two significant digits
- A data set can have more than one mode

Distribution of Change in Global Assessment



Unimodal and Polymodal Data

- **Unimodal** - in a single cluster
- **Polymodal** - in more than one cluster (a general term)
- **Bimodal** - specifically in two clusters
- **Trimodal** - in three clusters
- etc.

Comparison of Measures of Central Tendency

x
2
2
6
7
10
21

DESCRIBING VARIABILITY - STANDARD DEVIATION AND COEFFICIENT OF VARIATION

Error

With respect to data values, error is often used to describe the variation between data values. This can include:

- Biological variation
- Experimental error, i.e., imprecision
- Technical error, e.g., typos

Statisticians tend to prefer the terms **scatter** or **variability** rather than error

Standard Deviation

Standard deviation (SD) - is a measure of variation among values that has the same units as the original data and is a summary of the ‘deviation’ of each value from the mean.

It is the most commonly accepted indicator of dispersion.

$$SD = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \bar{x})^2}{N - 1}}$$

Example Calculation of Standard Deviation

x
2
2
6
7
10
21

Reporting the SD - The ± symbol

The ± symbol - reasonably interpreted as meaning ‘more or less’ and is used to indicate variability

Since it is conceivable that some statistic other than the SD has been quoted, it is useful to state this explicitly.

Units of SD

SD is often preferred over variance (SD^2) because it has the same units as the mean and as the original values.

SD and Sample Size

SD estimates the variation within a population. Therefore:

- The estimate of SD does not differ based on sample size
- However, the estimate of SD will be more accurate with a larger sample size.

Coefficient of Variation

Coefficient of variation (CV) = SD/mean

- Used to ‘normalize’ the standard deviation
- Example:
  - Mean = 10 cm, sd = 2 cm
  - Mean = 100 mm, sd = 20 mm
  - CV = 2/10 = 0.2 or CV = 20/100 = 0.2
- NOTICE: the CV is unitless and is usually expressed as either a fraction (0.2) or a percentage (20%)
- Often used to standardize SD across different units of measurement
- ONLY valid if all values are greater than zero

QUARTILES - ANOTHER WAY TO DESCRIBE THE DATA

QUARTILES

3 quartiles split the data points into four equal-sized groups

- i.e., one fourth of the data points fall below Q1, one fourth fall between Q1 and Q2, one fourth fall between Q2 and Q3, and one fourth are greater than Q3
- Q1 = 25th percentile
- Q2 = 50th percentile = median
- Q3 = 75th percentile

Inter-Quartile Range As a ROBUST Indicator of Dispersion

Inter-quartile range is the difference between the upper and lower quartiles (Q3 - Q1)

Example of inter-quartile range

Rank	Time
1	2.2
2	3.0
3	3.1
4	3.2
5	3.4
6	3.9
7	4.0
8	4.1
9	4.2
10	4.5
11	5.1
12	10.7
13	12.2
14	17.9
15	24.8

Other Quantiles

Quantile systems divide ranked data sets into groups with equal numbers of observations in each group. Specifically:

- 3 Quartiles divide data into 4 equal groups
- 4 Quintiles divide data into 5 equal groups
- 9 Deciles divide data into 10 equal groups
- 99 Centiles divide data into 100 equal groups

DESCRIBING ORDINAL DATA

Using the Mean - Ordinal Data

- Using a mean to summarize ordinal data is most people's first instinct
- Using a mean can be misleading:

1. The mean of an ordinal variable is not likely to be a whole number or an actual value observed in the original data set
2. The step-sizes between the available scores are not necessarily of equal significance

Example - Using Mean

Outcome	Control	Active Treatment
1 - Died	0	4
2 - Deteriorated	8	7
3 - Unchanged	11	7
4 - Moderate improvement	19	8
5 - Great improvement	5	18

Control Mean = 3.5

Active Treatment Mean = 3.7

Using Median - Ordinal Data

- Median may be a more realistic value
- It may also be harder to distinguish between groups (frequent ties)

Using Mode - Ordinal Data

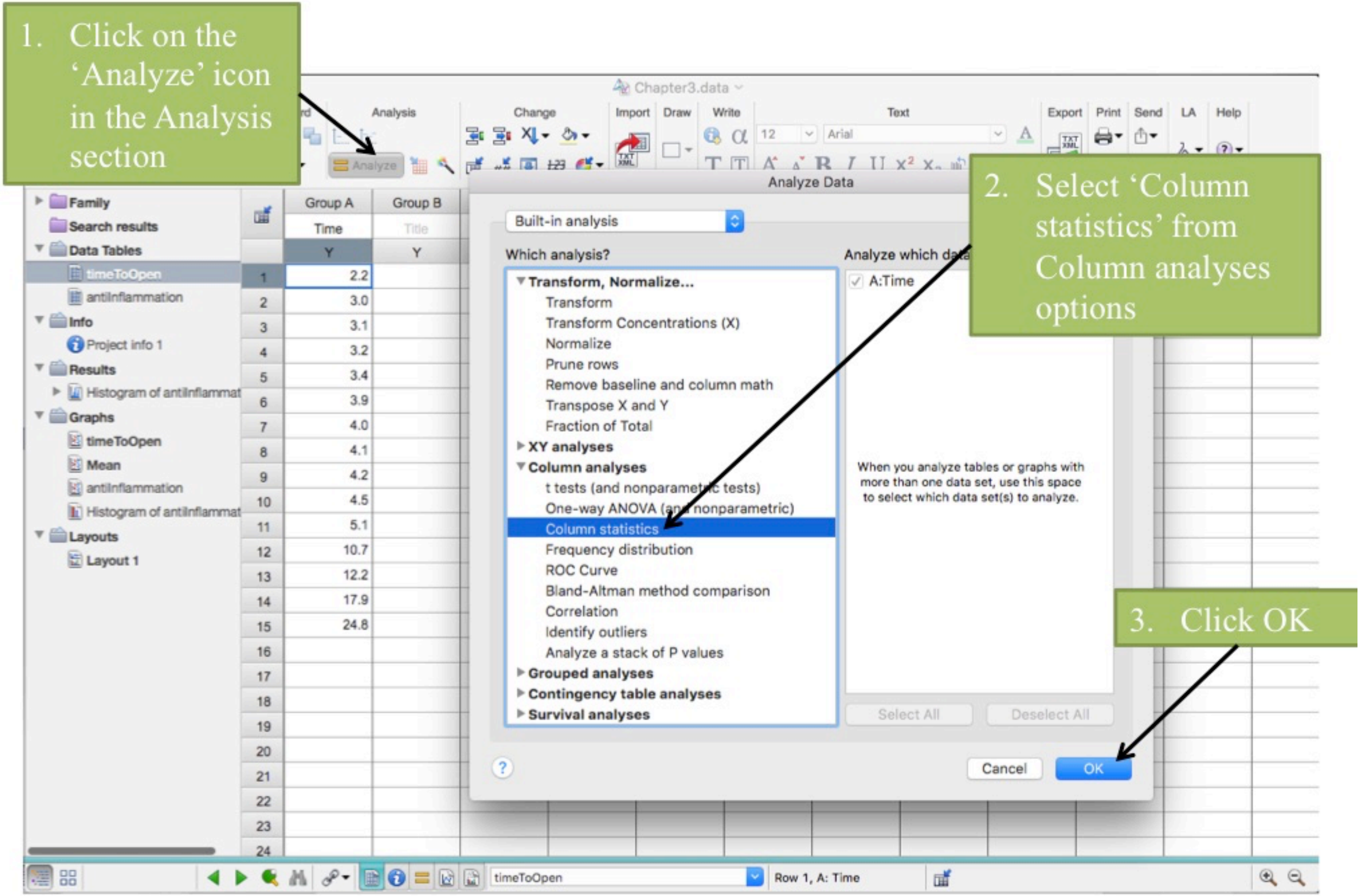
Modes tend to be very unstable:

- It only takes a difference of one observation to change a mode.

How Can We Describe Ordinal Data

- No universal solution
- Histograms can answer many different questions
- In general, median is better than mean
- Inter-quartile range can describe dispersion/ variability

DESCRIPTIVE STATISTICS IN GRAPHPAD



	Group A	Group B
	Time	Title
	Y	Y
1	2.2	
2	3.0	
3	3.1	
4	3.2	
5	3.4	
6	3.9	
7	4.0	
8	4.1	
9	4.2	
10	4.5	
11	5.1	
12	10.7	
13	12.2	
14	17.9	
15	24.8	

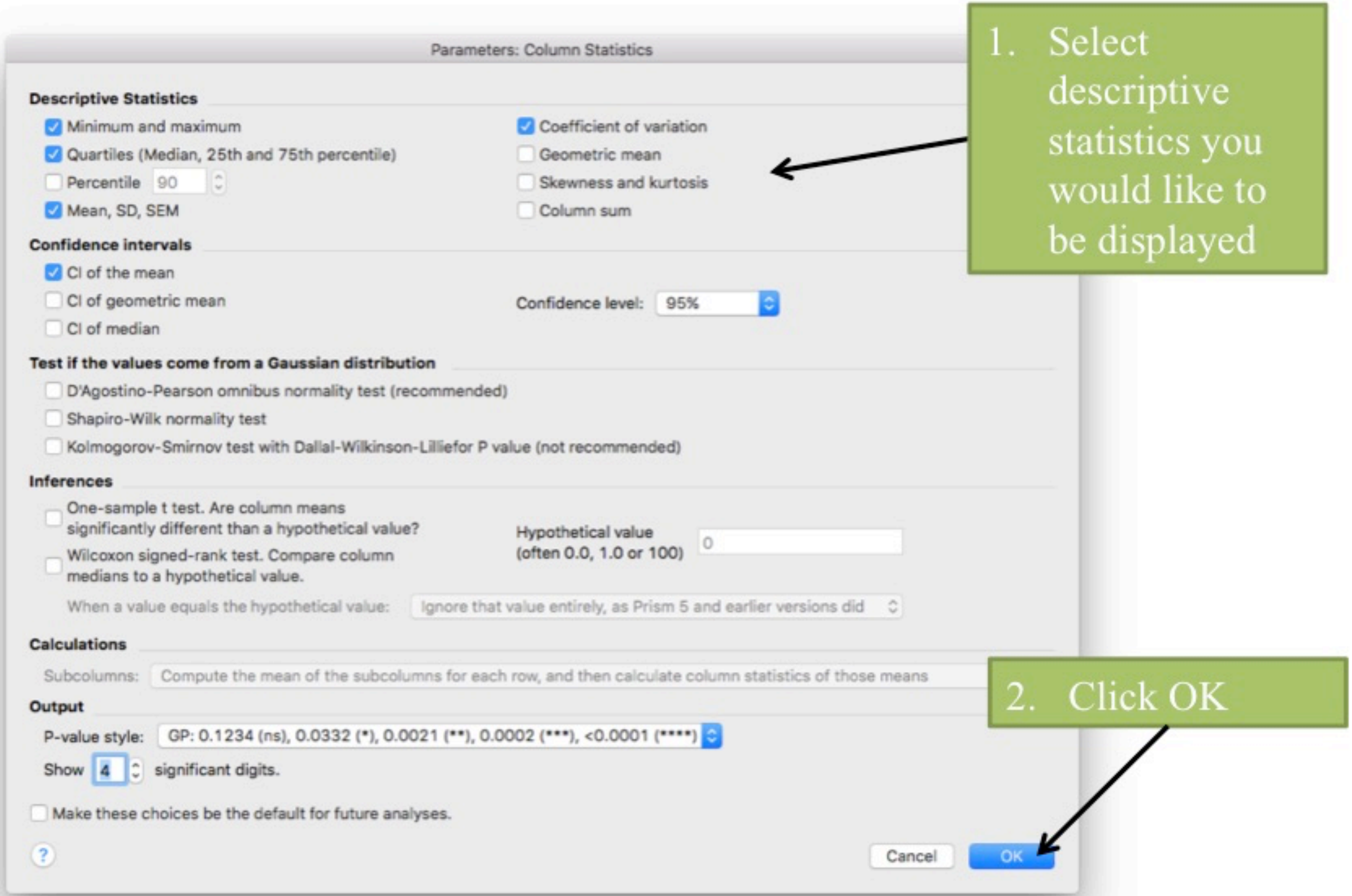
☑ A:Time

When you analyze tables or graphs with more than one data set, use this space to select which data set(s) to analyze.

1. Click on the ‘Analyze’ icon in the Analysis section

2. Select ‘Column statistics’ from Column analyses options

3. Click OK



Parameters: Column Statistics

Descriptive Statistics

☑ Minimum and maximum

☑ Quartiles (Median, 25th and 75th percentile)

□ Percentile 90

☑ Mean, SD, SEM

☑ Coefficient of variation

□ Geometric mean

□ Skewness and kurtosis

□ Column sum

Confidence intervals

☑ CI of the mean

□ CI of geometric mean

□ CI of median

Confidence level: 95%

Test if the values come from a Gaussian distribution

□ D'Agostino-Pearson omnibus normality test (recommended)

□ Shapiro-Wilk normality test

□ Kolmogorov-Smirnov test with Dallal-Wilkinson-Lilliefor P value (not recommended)

Inferences

□ One-sample t test. Are column means significantly different than a hypothetical value?

□ Wilcoxon signed-rank test. Compare column medians to a hypothetical value.

Hypothetical value (often 0.0, 1.0 or 100) 0

When a value equals the hypothetical value: Ignore that value entirely, as Prism 5 and earlier versions did

Calculations

Subcolumns: Compute the mean of the subcolumns for each row, and then calculate column statistics of those means

Output

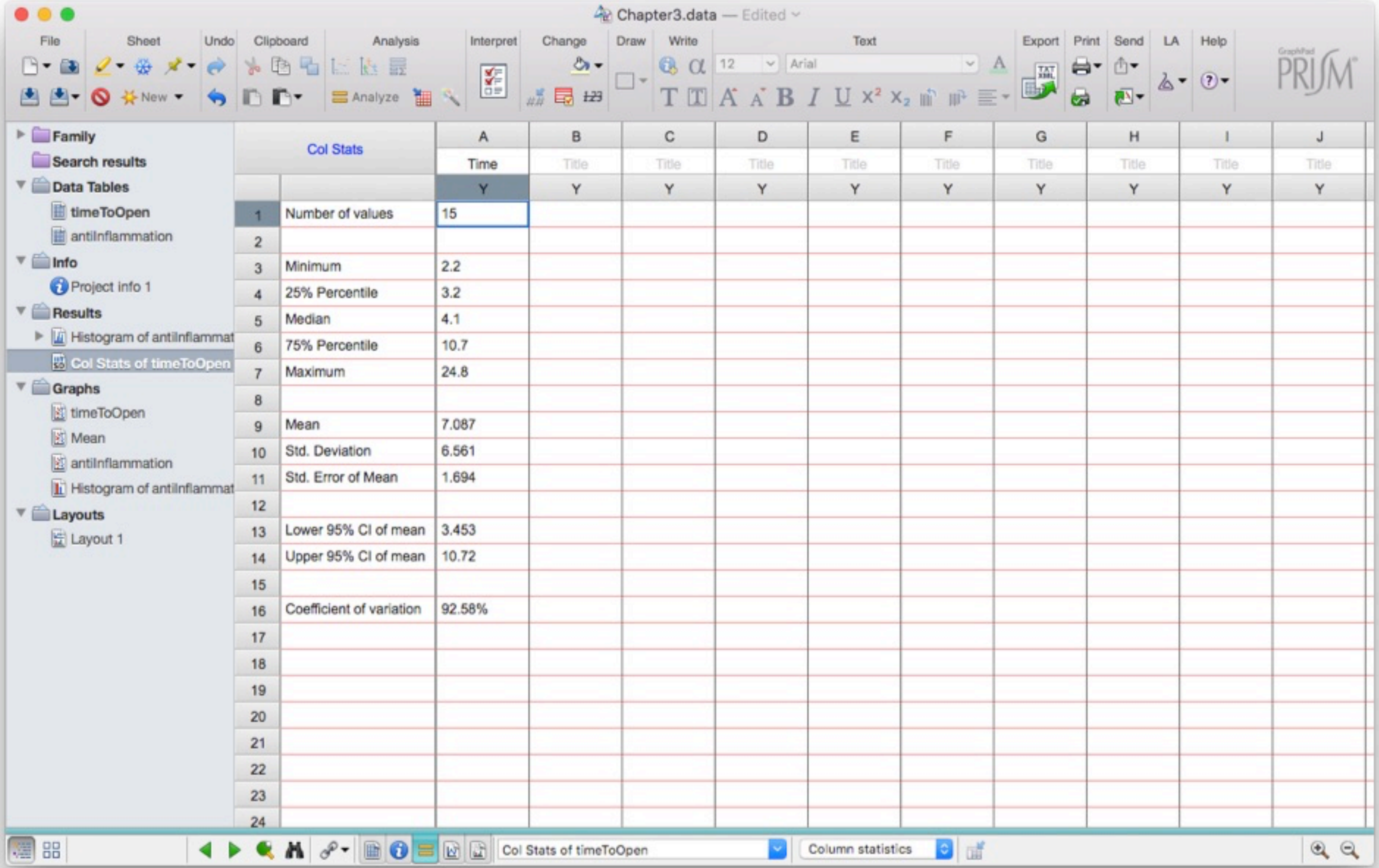
P-value style: GP: 0.1234 (ns), 0.0332 (*), 0.0021 (**), 0.0002 (***), <0.0001 (****)

Show 4 significant digits.

□ Make these choices be the default for future analyses.

1. Select descriptive statistics you would like to be displayed

2. Click OK



	Col Stats	A	B	C	D	E	F	G	H	I	J
		Time	Title	Title	Title	Title	Title	Title	Title	Title	Title
		Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
1	Number of values	15									
2											
3	Minimum	2.2									
4	25% Percentile	3.2									
5	Median	4.1									
6	75% Percentile	10.7									
7	Maximum	24.8									
8											
9	Mean	7.087									
10	Std. Deviation	6.561									
11	Std. Error of Mean	1.694									
12											
13	Lower 95% CI of mean	3.453									
14	Upper 95% CI of mean	10.72									
15											
16	Coefficient of variation	92.58%									
17											
18											
19											
20											
21											
22											
23											
24											

Col Stats of timeToOpen

Column statistics

What did we learn?

- When choosing a descriptive statistic be aware if:
  - Data contain outliers
  - Data contain a single cluster or are polymodal
- Means and SD summarize the central tendency and the dispersion, respectively
- Median and IQR, as measures of central tendency and dispersion, are more robust to outliers
- Coefficient of variance is the relative variability
- With ordinal data, bar charts need to be used to give a general impression of the central tendency and the dispersion of results