Chapter 6: 95% Confidence Interval for the Mean and Data Transformation

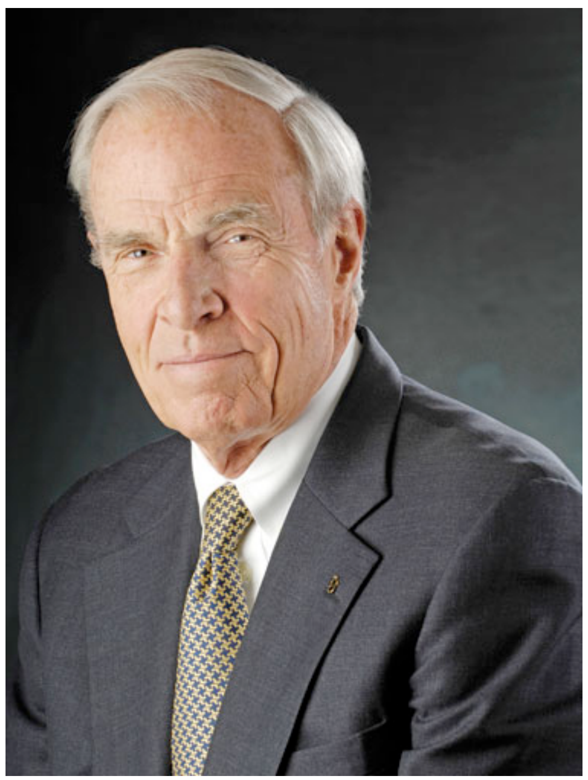
TXCL7565/PHSC7565

What This Chapter Covers

- What is a confidence interval (CI)?
- What do we mean by '95%' confidence?
- Calculating a CI
- Sensitivity of CI to SD, sample size, and level of confidence
- One-sided CIs
- CI for difference between two means
- Normal distribution and CI

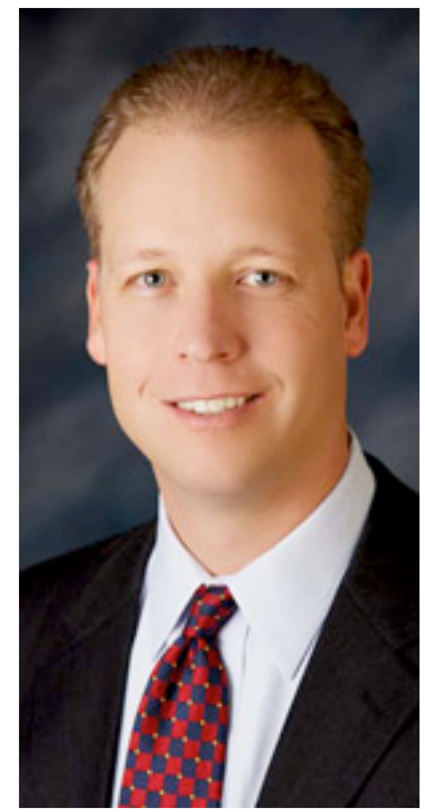
WHAT IS A CONFIDENCE INTERVAL?

Age Guess



Bruce Benson
President of CU

Don Cheadle
Actor
Alumnus of East
High School



Stranger

Confidence Interval

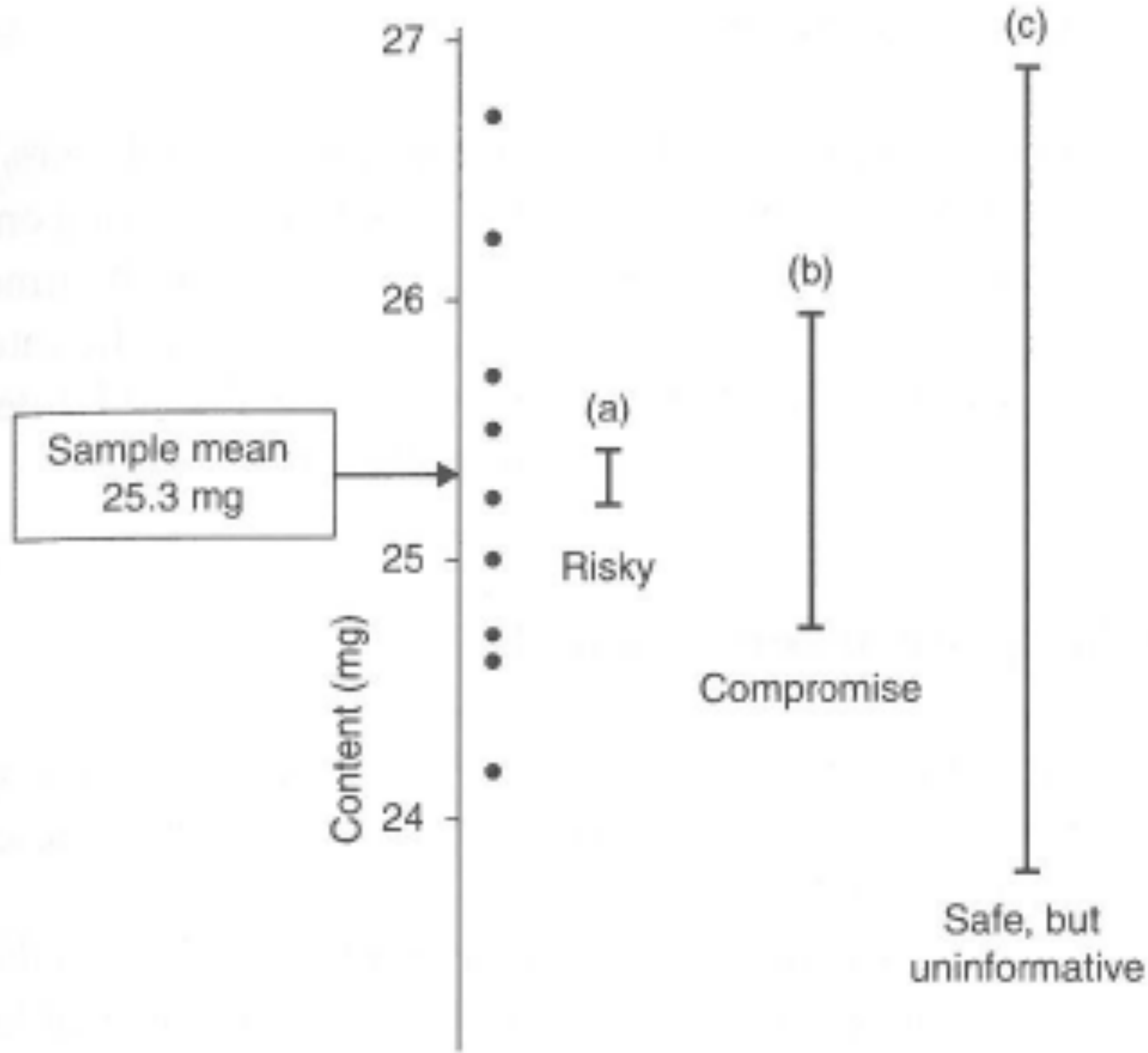
- A mean derived from a sample is unlikely to be a *perfect* estimate of the population mean.
- Therefore, we normally give a range that the population mean is likely to fall within.
- Since we have no reason to believe there is bias in our sample mean estimate, we create a range by subtracting and adding the same amount to the sample mean estimate.

Width of the Interval

The **wider** the interval

→

The **more confident** we are that the true population mean falls within the interval


27
26
25
24
Content (mg)
Sample mean
25.3 mg
(a)
Risky
(b)
Compromise
(c)
Safe, but
uninformative

Norm in the Field

- A ’95% confidence interval’ is the standard confidence interval width.
- It originates from a p-value threshold of 0.05 for significance.
- I’ll explain more when we tackle p-values next week.

WHAT DO WE MEAN BY '95%' CONFIDENCE?

95% Confidence

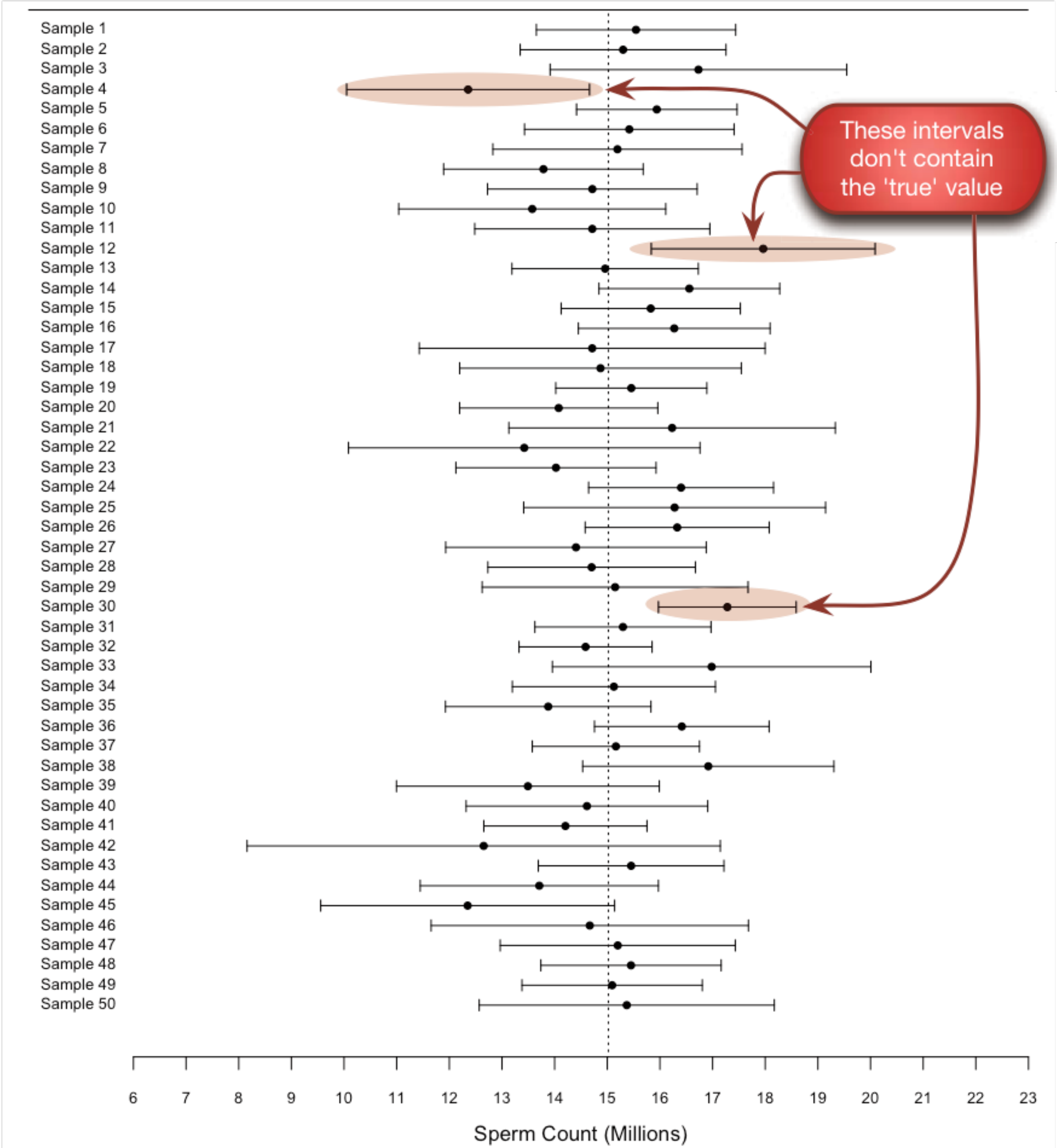
- For a 95% CI, we typically say that we are 95% confident that the interval includes the true population mean

    - CANNOT say that there is a 95% **probability** that the interval includes the true population mean

Statistical Definition of Level of Confidence

Level of Confidence = the proportion of intervals that will cover the true population mean when the intervals are calculated in the same manner and generated from sampling of the same population

FIGURE 2.9 The confidence intervals of the sperm counts of Japanese quail (horizontal axis) for 50 different samples (vertical axis)

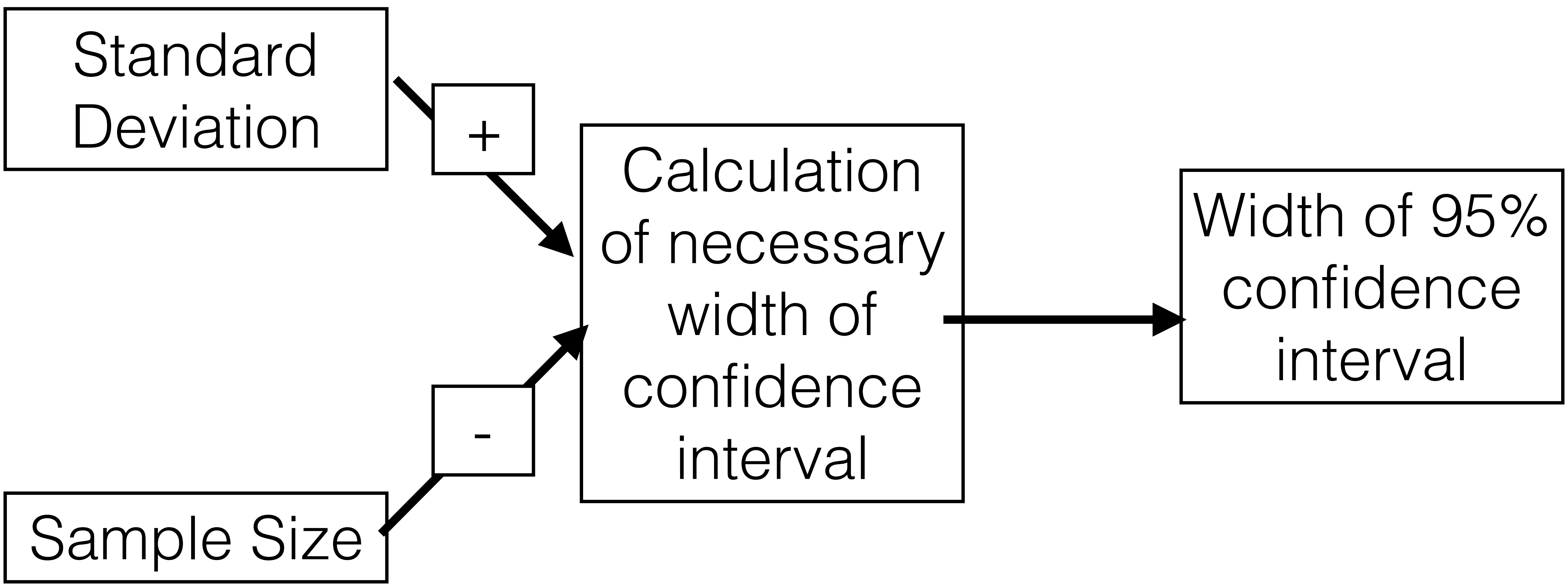


Calculating a CI

The width of confidence intervals are influenced by:

1. Standard Deviation = higher SD, wider interval
2. Sample Size = smaller sample size, wider interval
3. Level of Confidence = higher level of confidence, wider interval

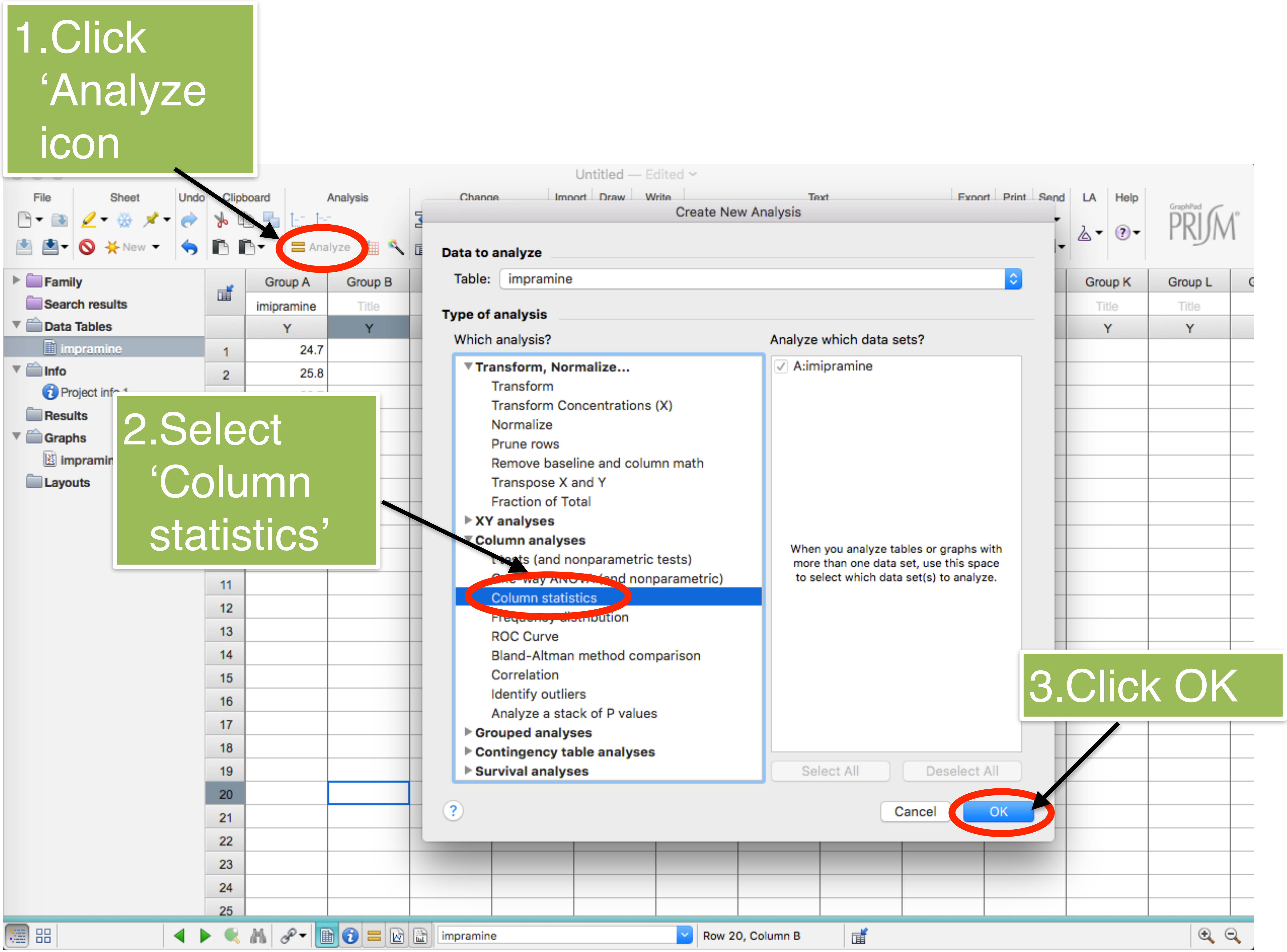
Calculating 95% CI

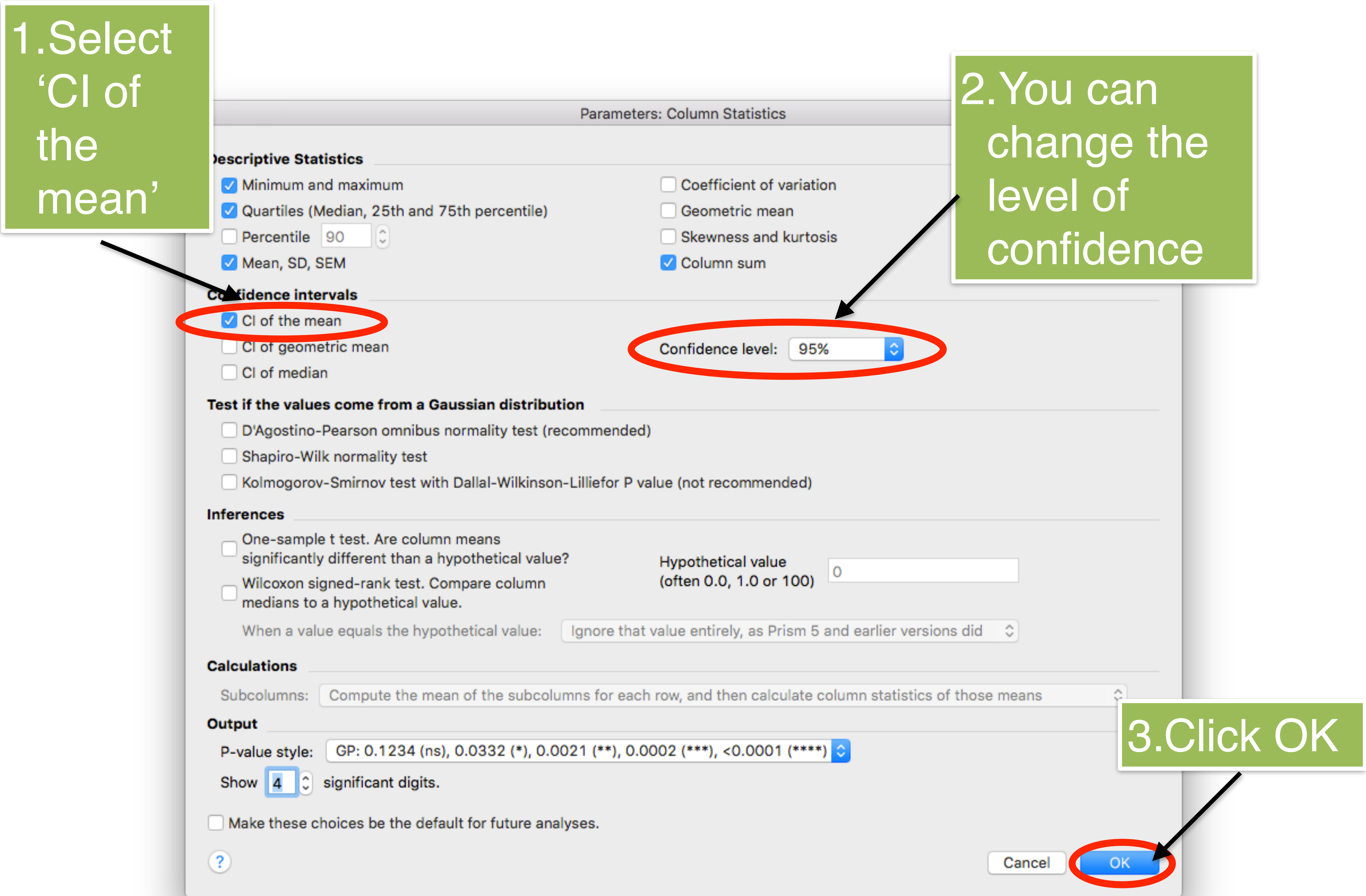


Formula for CI

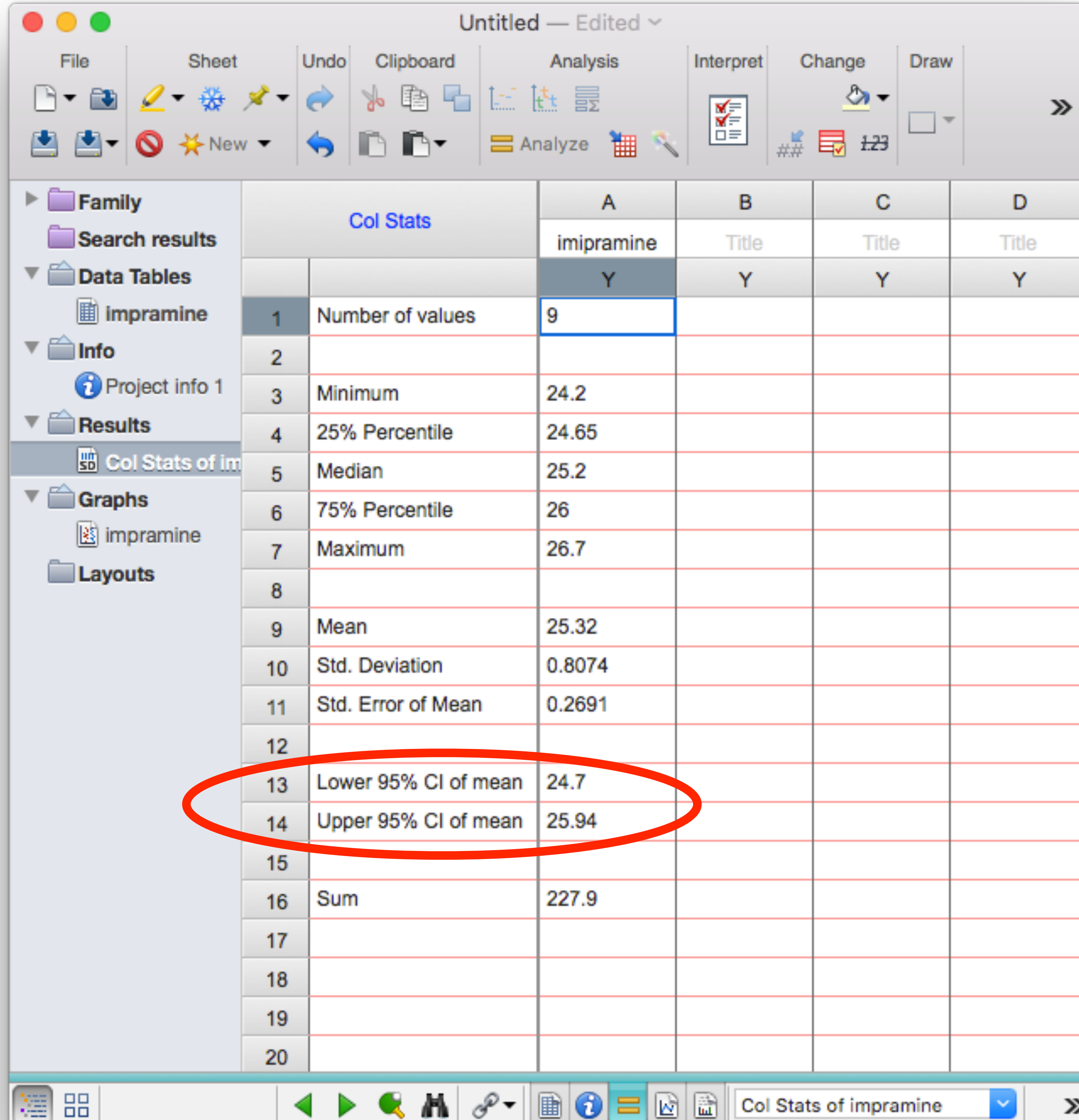
$$\bar{x} \pm t_c \frac{s}{\sqrt{n}}$$

$\bar{x}$ = sample mean
t_c = t-statistic
s = sample standard deviation
n = number of observations in sample


1.Click 'Analyze icon
2.Select 'Column statistics'
3.Click OK
Untitled — Edited
Create New Analysis
Data to analyze
Table: impramine
Type of analysis
Which analysis?
Transform, Normalize...
Transform
Transform Concentrations (X)
Normalize
Prune rows
Remove baseline and column math
Transpose X and Y
Fraction of Total
XY analyses
Column analyses
t tests (and nonparametric tests)
One-way ANOVA (and nonparametric)
Column statistics
Frequency distribution
ROC Curve
Bland-Altman method comparison
Correlation
Identify outliers
Analyze a stack of P values
Grouped analyses
Contingency table analyses
Survival analyses
Analyze which data sets?
A:imipramine
When you analyze tables or graphs with more than one data set, use this space to select which data set(s) to analyze.
Select All
Deselect All
Cancel
OK
Family
Search results
Data Tables
impramine
Info
Results
Graphs
Layouts
Group A
Group B
imipramine
Y
24.7
25.8
Row 20, Column B



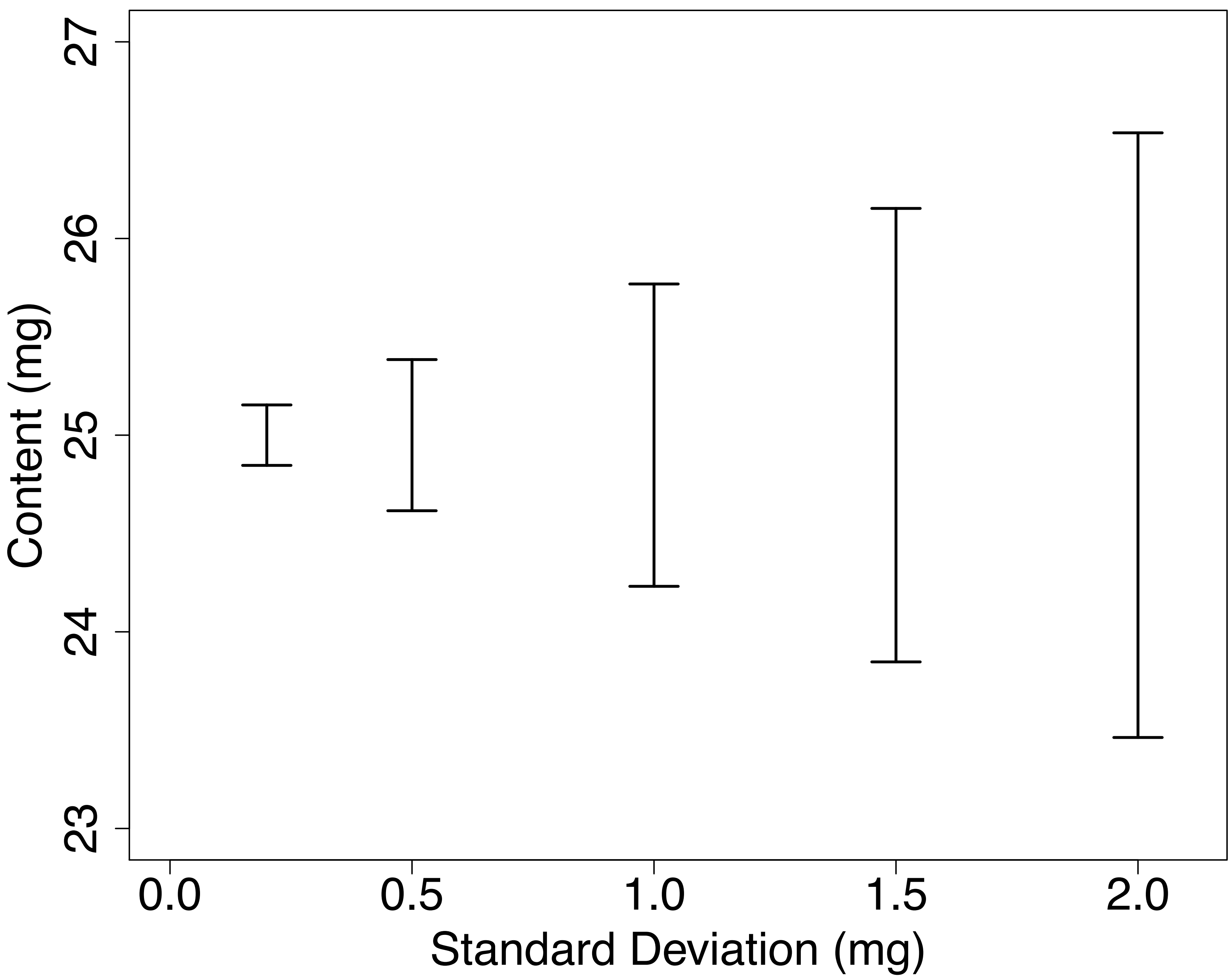
1.Select 'CI of the mean'
2.You can change the level of confidence
3.Click OK
Parameters: Column Statistics
Descriptive Statistics
Minimum and maximum
Quartiles (Median, 25th and 75th percentile)
Percentile 90
Mean, SD, SEM
Coefficient of variation
Geometric mean
Skewness and kurtosis
Column sum
Confidence 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.
?
Cancel
OK



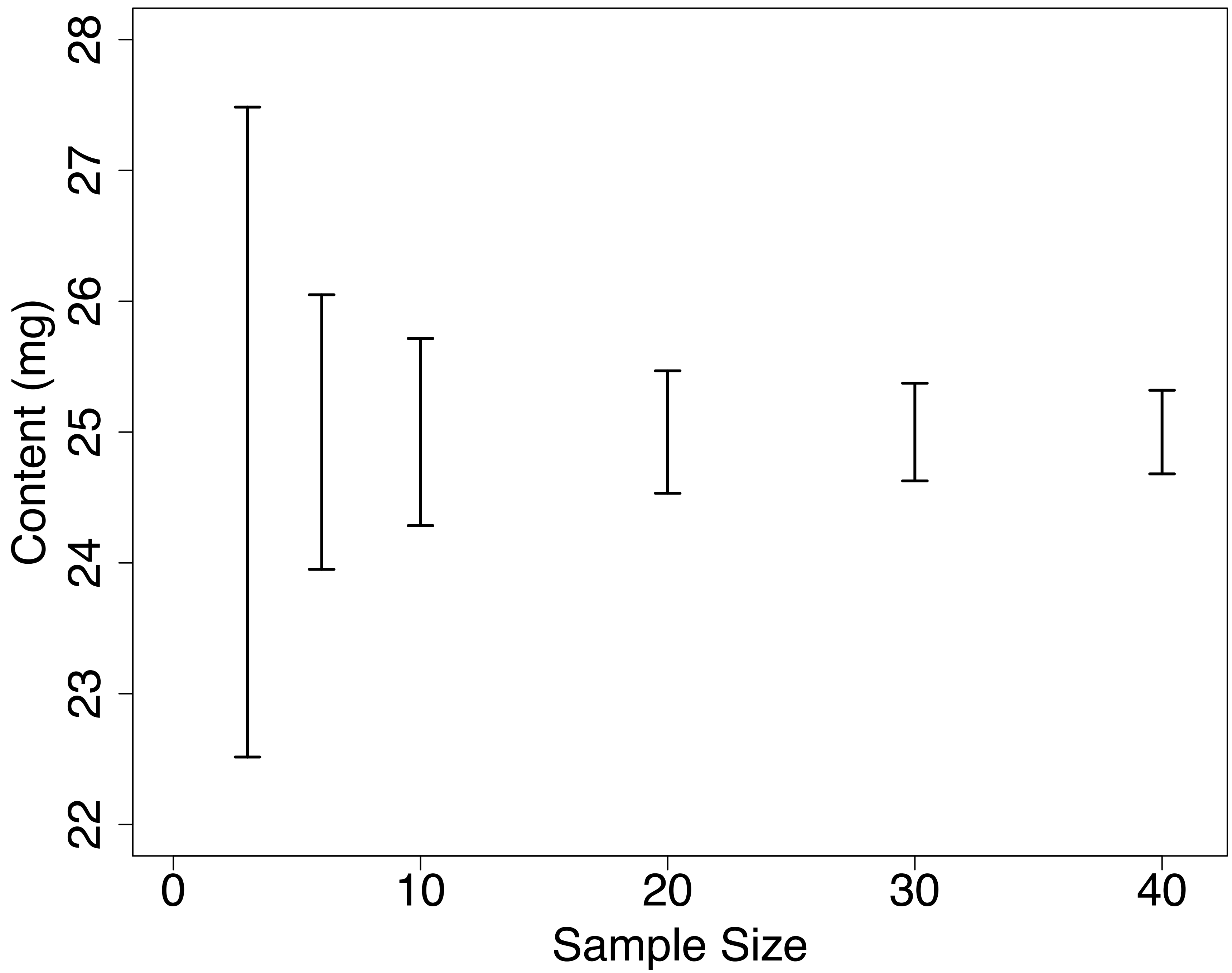
Untitled — Edited
File
Sheet
Undo
Clipboard
Analysis
Interpret
Change
Draw
New
Analyze
Family
Search results
Data Tables
impramine
Info
Project info 1
Results
Col Stats of im
Graphs
impramine
Layouts
Col Stats
A
B
C
D
imipramine
Title
Title
Title
Y
Y
Y
Y
1 Number of values 9
2
3 Minimum 24.2
4 25% Percentile 24.65
5 Median 25.2
6 75% Percentile 26
7 Maximum 26.7
8
9 Mean 25.32
10 Std. Deviation 0.8074
11 Std. Error of Mean 0.2691
12
13 Lower 95% CI of mean 24.7
14 Upper 95% CI of mean 25.94
15
16 Sum 227.9
17
18
19
20
Col Stats of impramine

SENSITIVITY OF CI TO SD, SAMPLE SIZE, AND LEVEL OF CONFIDENCE

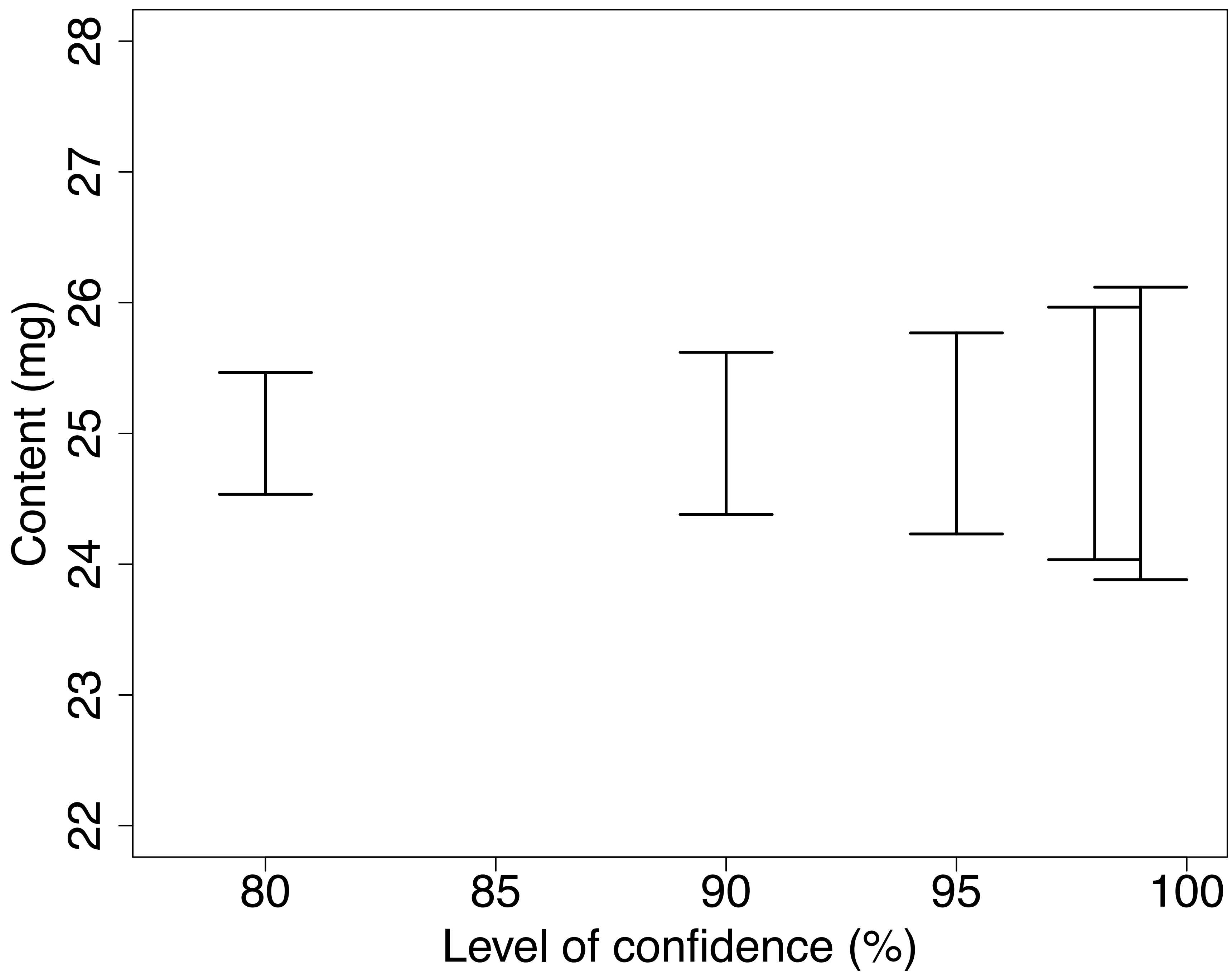
Sensitivity to SD



Sensitivity to Sample Size



Sensitivity to Level of Confidence



Assessing the reliability of our conclusions

Statistics never draws any absolute conclusions. It will offer an opinion and then back that up with a measure of that conclusion's reliability.

One-sided 95% CIs

Two-sided / Two-tailed CIs

- In a two-sided CIs, we specify both a minimum and maximum value to our range.

- The 'hazard', i.e., 1 - level of confidence, is split between the two ends.

1. The true population mean is no *less* than some stated figure (2.5% chance this is false)
2. The true population mean is no *greater* than some stated figure (2.5% chance this is false)

One-sided 95% CIs

- Sometimes we are only interested in whether the mean is above a certain level

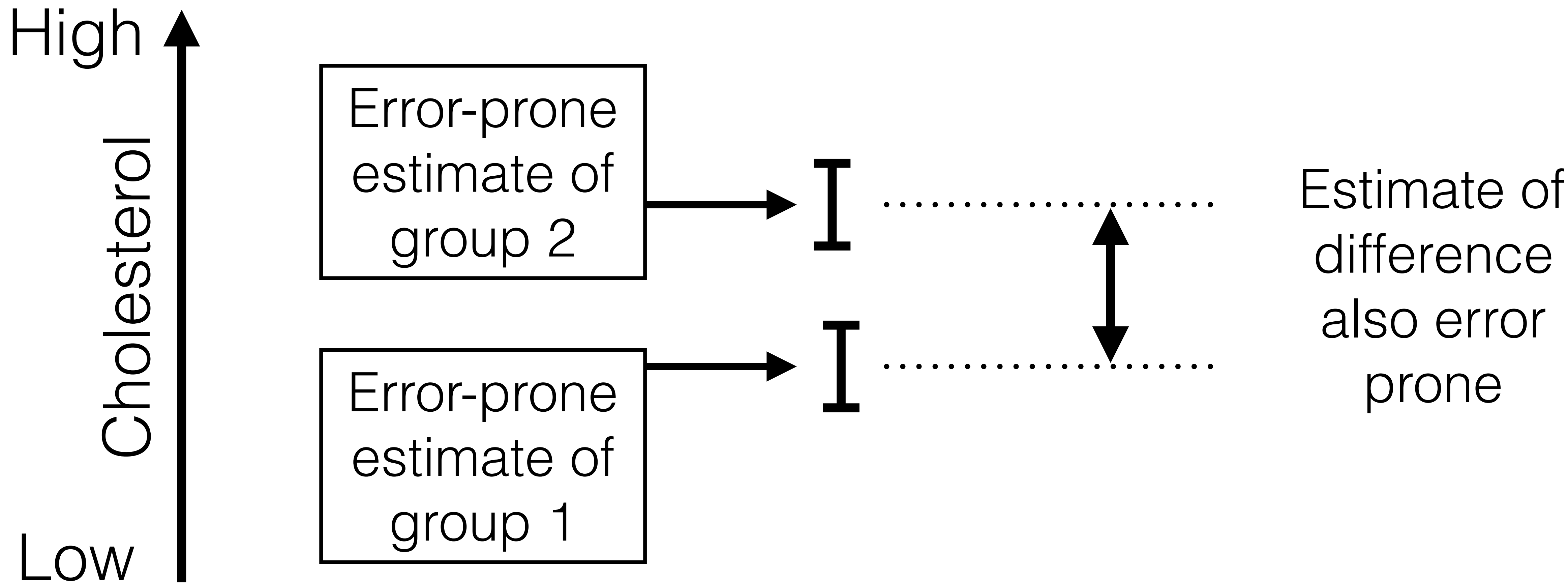
The true population mean is no *less* than some stated figure (5% chance this is false)
OR
The true population mean is no *greater* than some stated figure (5% chance this is false)

One-sided CIs in GraphPad

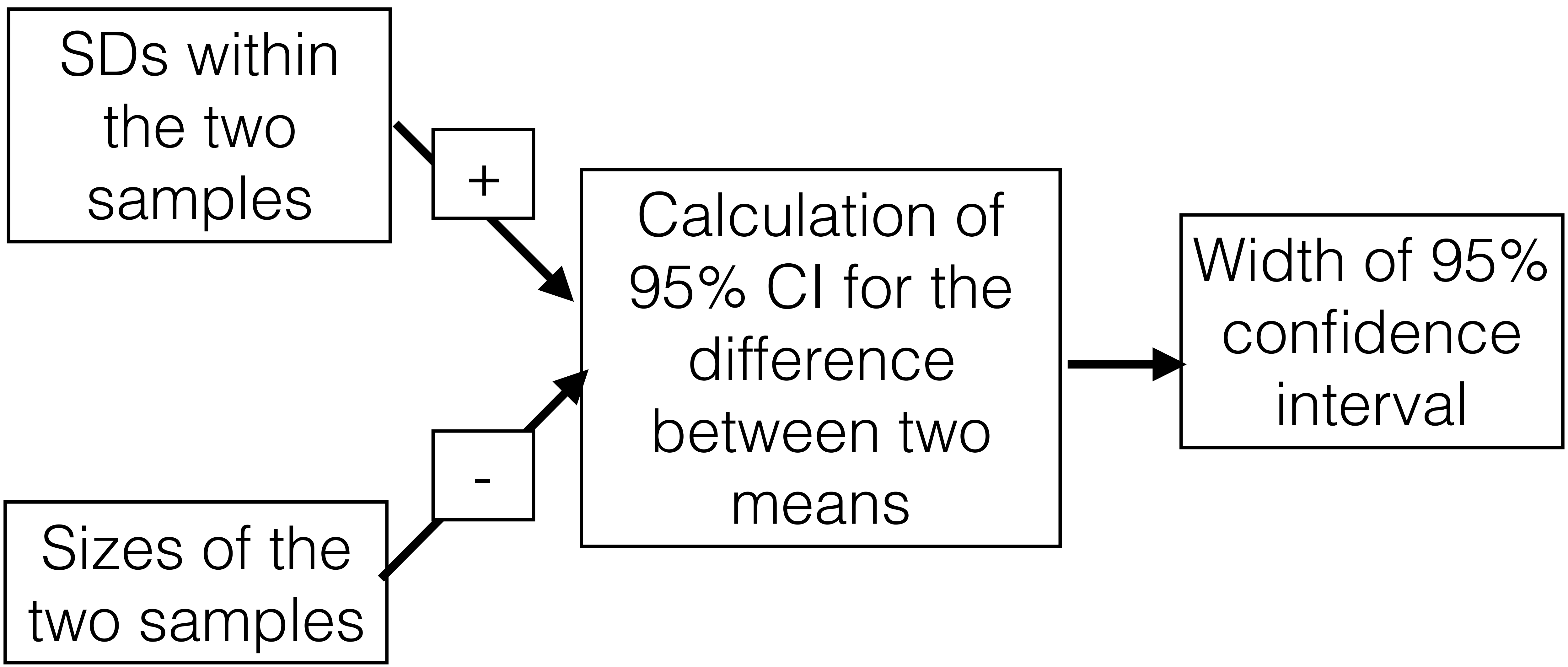
- GraphPad doesn’t offer a one-sided confidence interval options
- Solution - Change the level of confidence by doubling the risk, i.e., to get a 95% one-sided confidence interval, ask for a 90% two-sided confidence interval

95% CI FOR THE DIFFERENCE BETWEEN TWO TREATMENTS

95% CI for Difference Between 2 Means



95% CI of Difference Between 2 Means

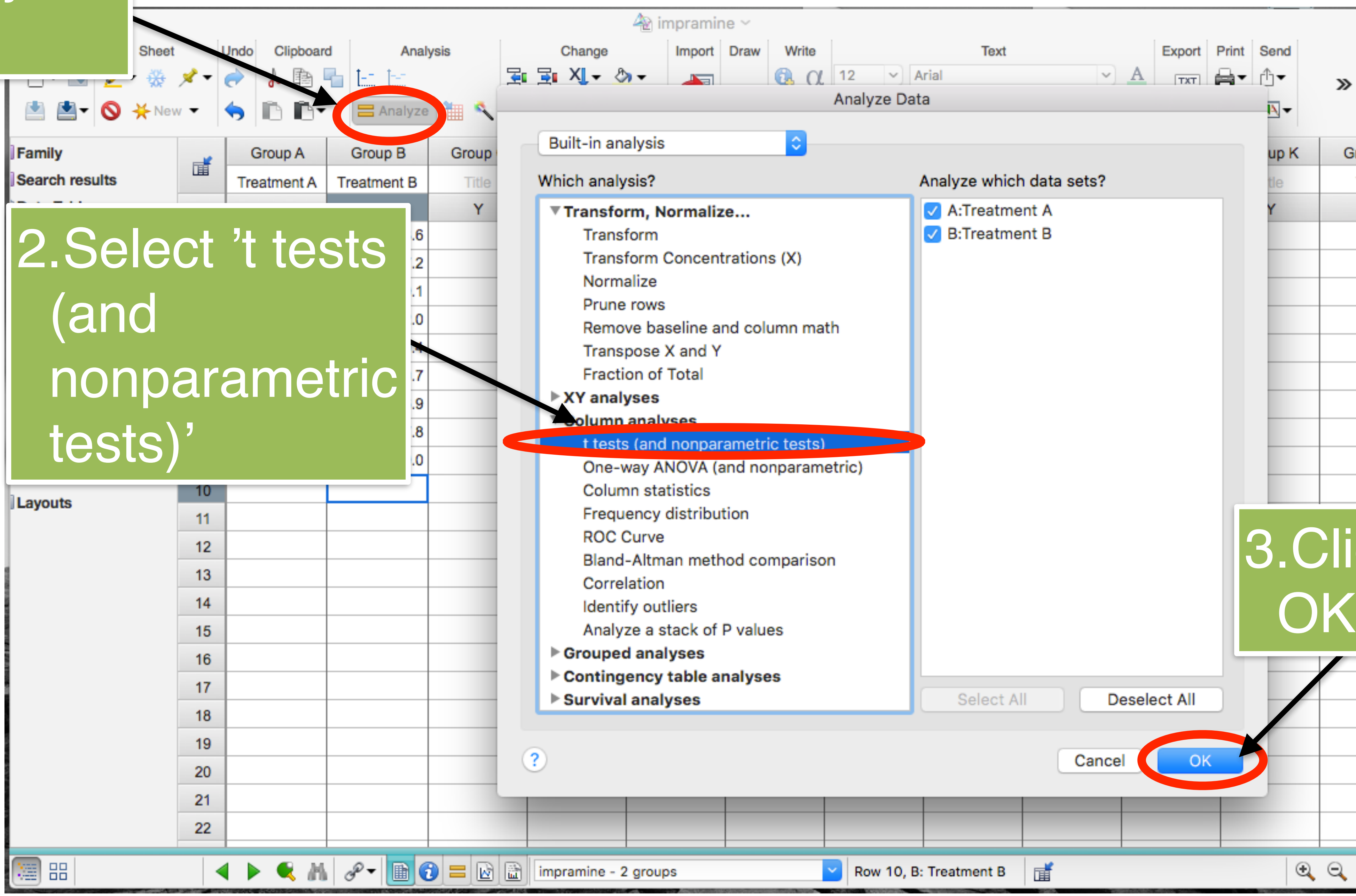


Formula for 95% CI of Difference Between 2 Means

To come in later chapters…

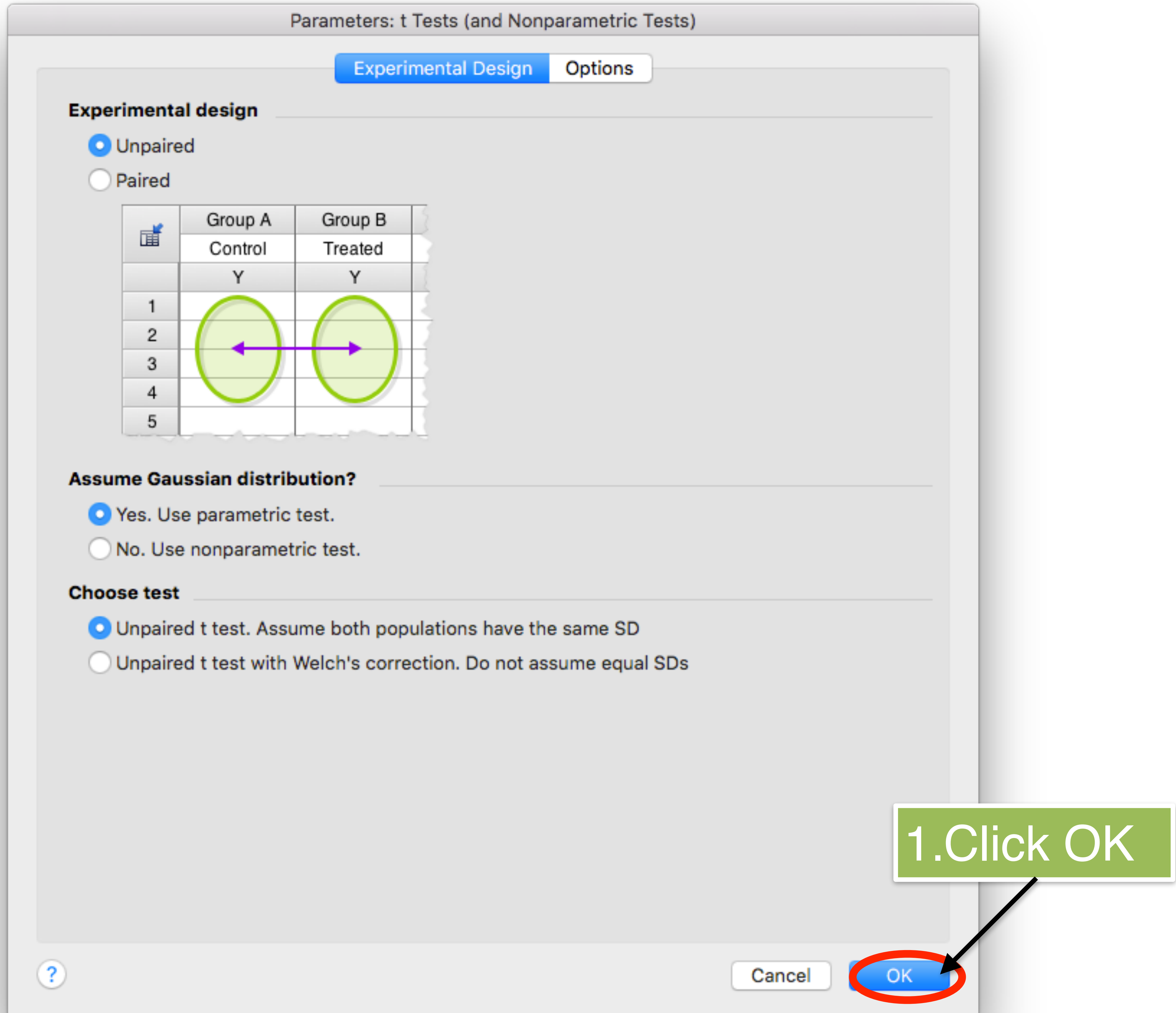
95% Confidence Interval for Difference Between 2 Means - GraphPad

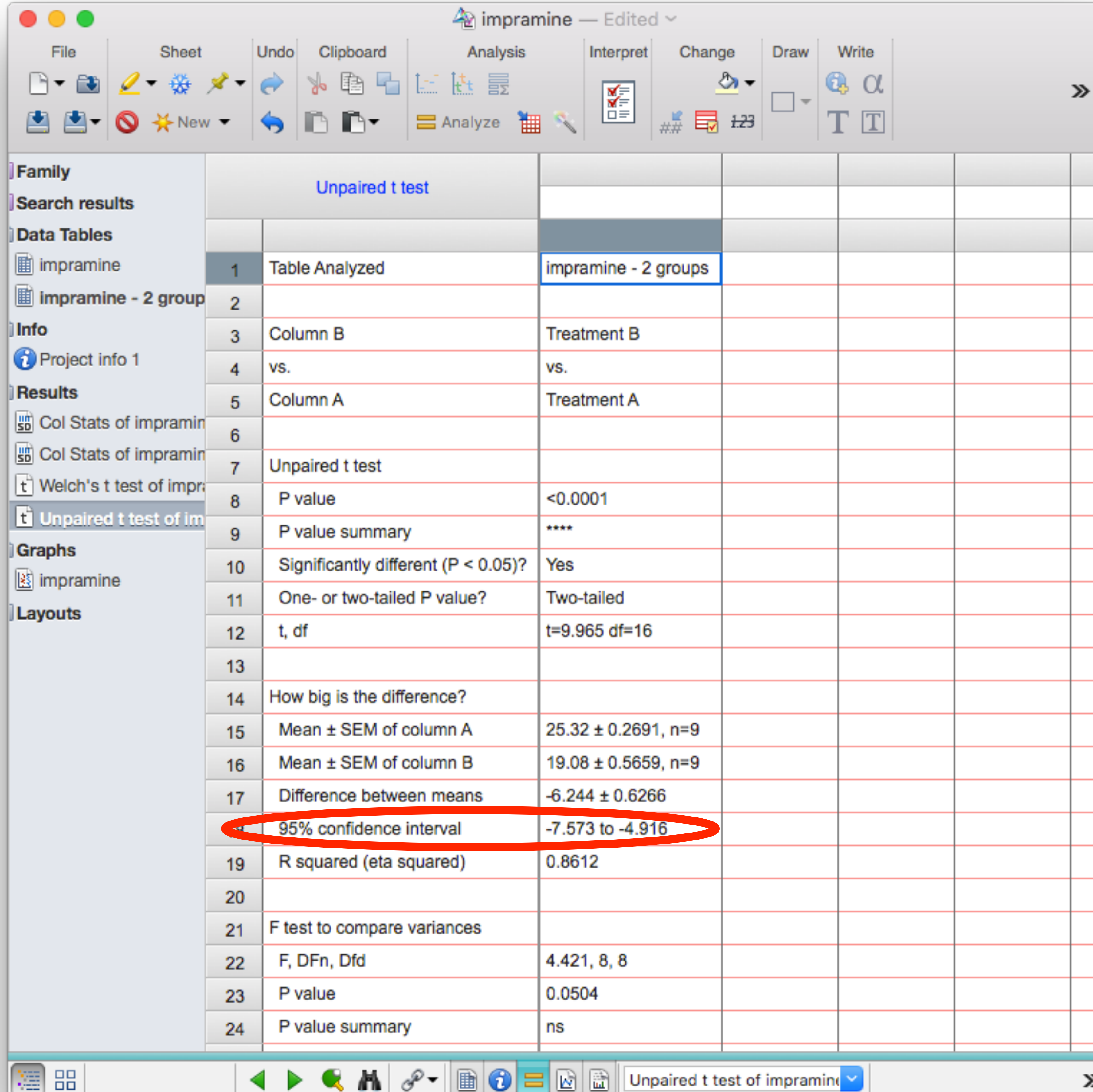
1. Click 'Analyze icon



2. Select 't tests (and nonparametric tests)'

3. Click OK


Parameters: t Tests (and Nonparametric Tests)
Experimental Design
Options
Experimental design
Unpaired
Paired
Group A
Group B
Control
Treated
Y
Y
1
2
3
4
5
Assume Gaussian distribution?
Yes. Use parametric test.
No. Use nonparametric test.
Choose test
Unpaired t test. Assume both populations have the same SD
Unpaired t test with Welch's correction. Do not assume equal SDs
?
Cancel
OK
1.Click OK



	Unpaired t test	
1	Table Analyzed	impramine - 2 groups
2		
3	Column B	Treatment B
4	vs.	vs.
5	Column A	Treatment A
6		
7	Unpaired t test	
8	P value	<0.0001
9	P value summary	****
10	Significantly different (P < 0.05)?	Yes
11	One- or two-tailed P value?	Two-tailed
12	t, df	t=9.965 df=16
13		
14	How big is the difference?	
15	Mean ± SEM of column A	25.32 ± 0.2691, n=9
16	Mean ± SEM of column B	19.08 ± 0.5659, n=9
17	Difference between means	-6.244 ± 0.6266
	95% confidence interval	-7.573 to -4.916
19	R squared (eta squared)	0.8612
20		
21	F test to compare variances	
22	F, DFn, Dfd	4.421, 8, 8
23	P value	0.0504
24	P value summary	ns

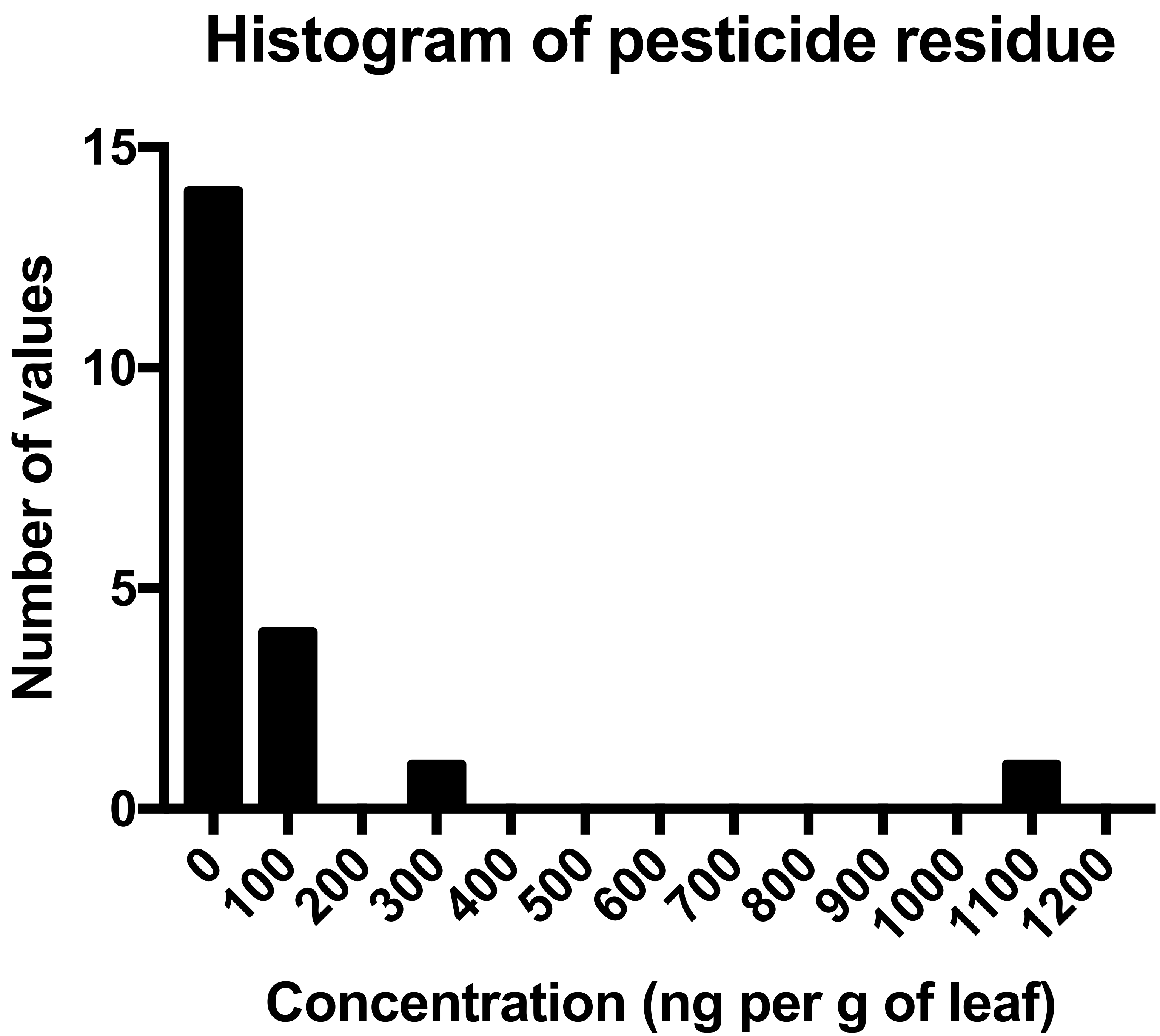
NORMAL DISTRIBUTION AND CI

Normal Distribution and CI

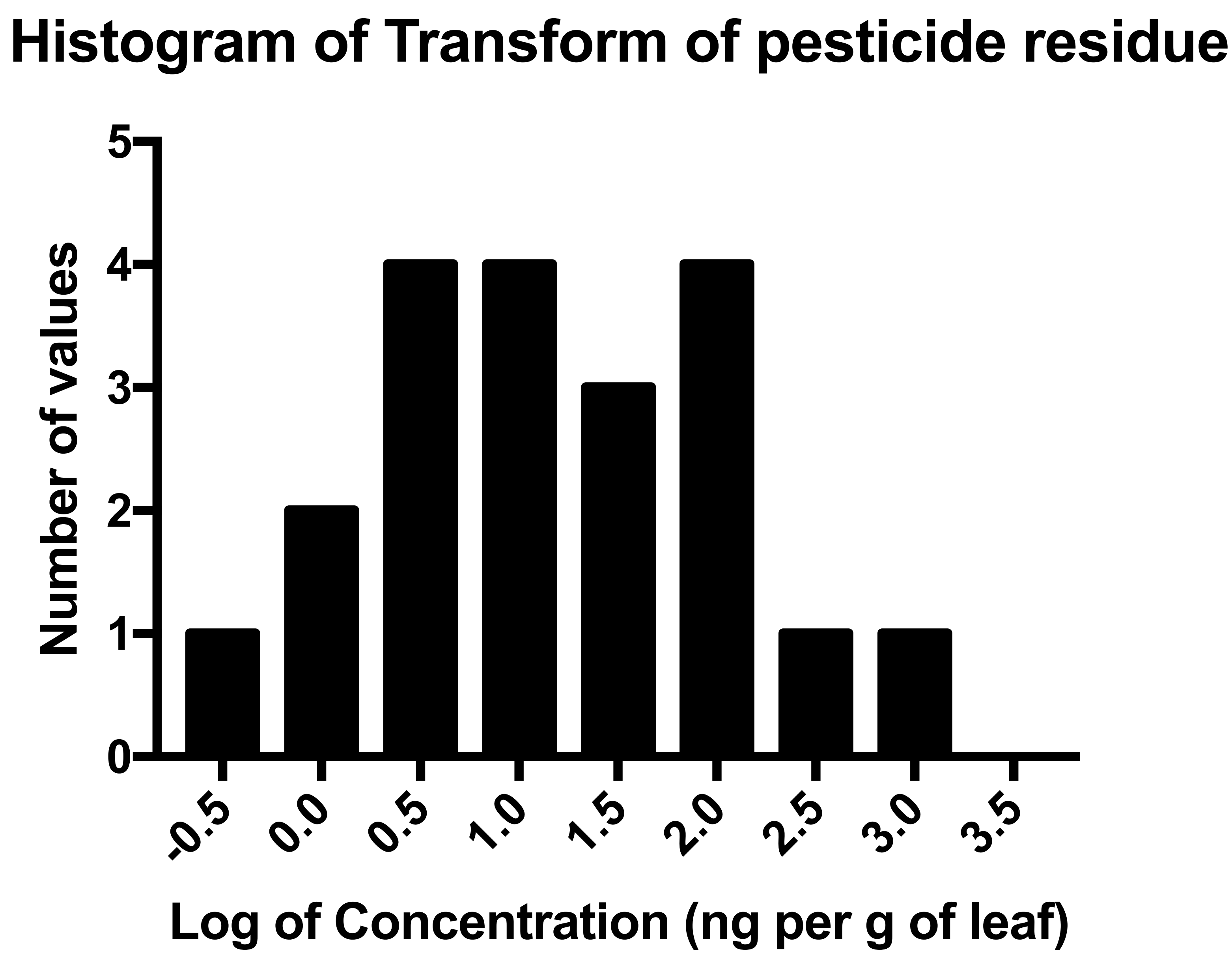
- The methods presented for calculating CI are based on the assumption that the data have a normal distribution.

- However, the CI is pretty robust and only performs poorly when the data are grossly non-normal.

Example of Non-Normal Data

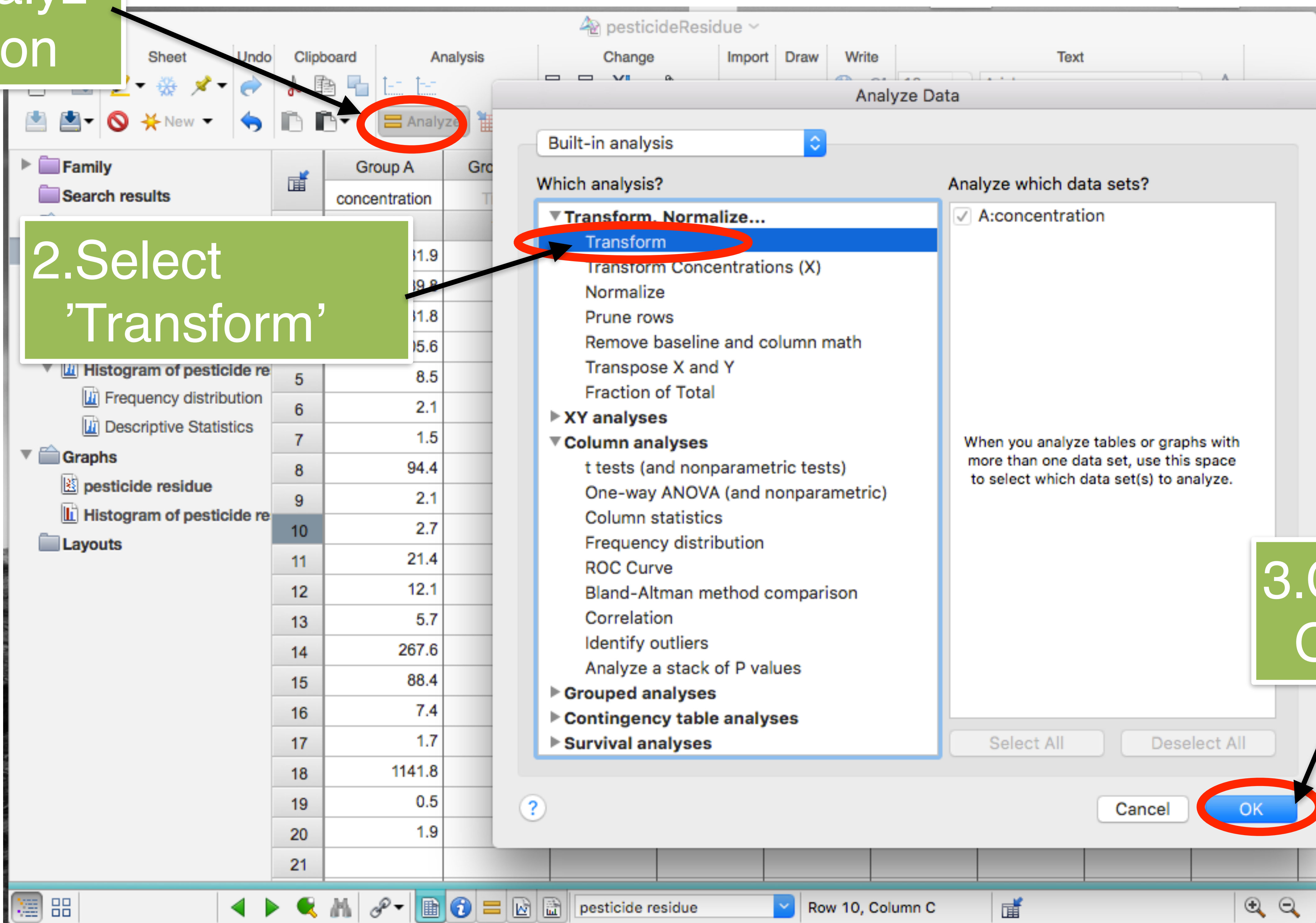


Common Solution to Positive Skewness - Log Transform



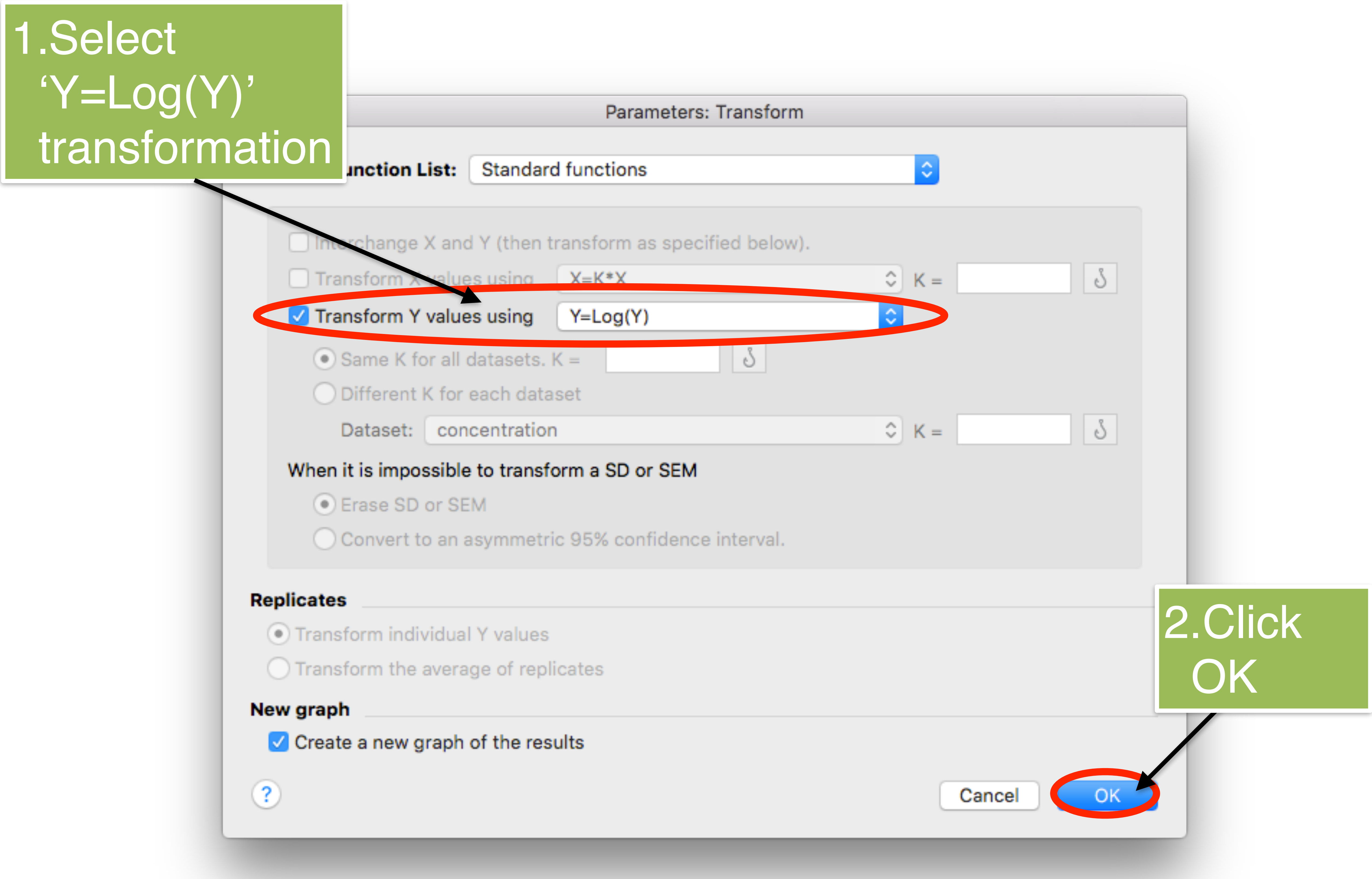
Log Transform Data in GraphPad

1.Click 'Analyze icon



2.Select 'Transform'

3.Click OK


1.Select 'Y=Log(Y)' transformation
Parameters: Transform
Function List:
Standard functions
Interchange X and Y (then transform as specified below).
Transform X values using
X=K*X
K =
Transform Y values using
Y=Log(Y)
Same K for all datasets. K =
Different K for each dataset
Dataset:
concentration
K =
When it is impossible to transform a SD or SEM
Erase SD or SEM
Convert to an asymmetric 95% confidence interval.
Replicates
Transform individual Y values
Transform the average of replicates
New graph
Create a new graph of the results
?
Cancel
OK
2.Click OK

Geometric Mean

To calculate a geometric mean:

1. Log transform data
2. Calculate mean of log transformed data
3. Take the antilog of the mean of the log transformed data

$$\text{geometric mean} = 10^{\left[\frac{1}{n}\sum_{i=1}^{n}\log a_i\right]}$$

Calculate 95% Based on Transformed Data

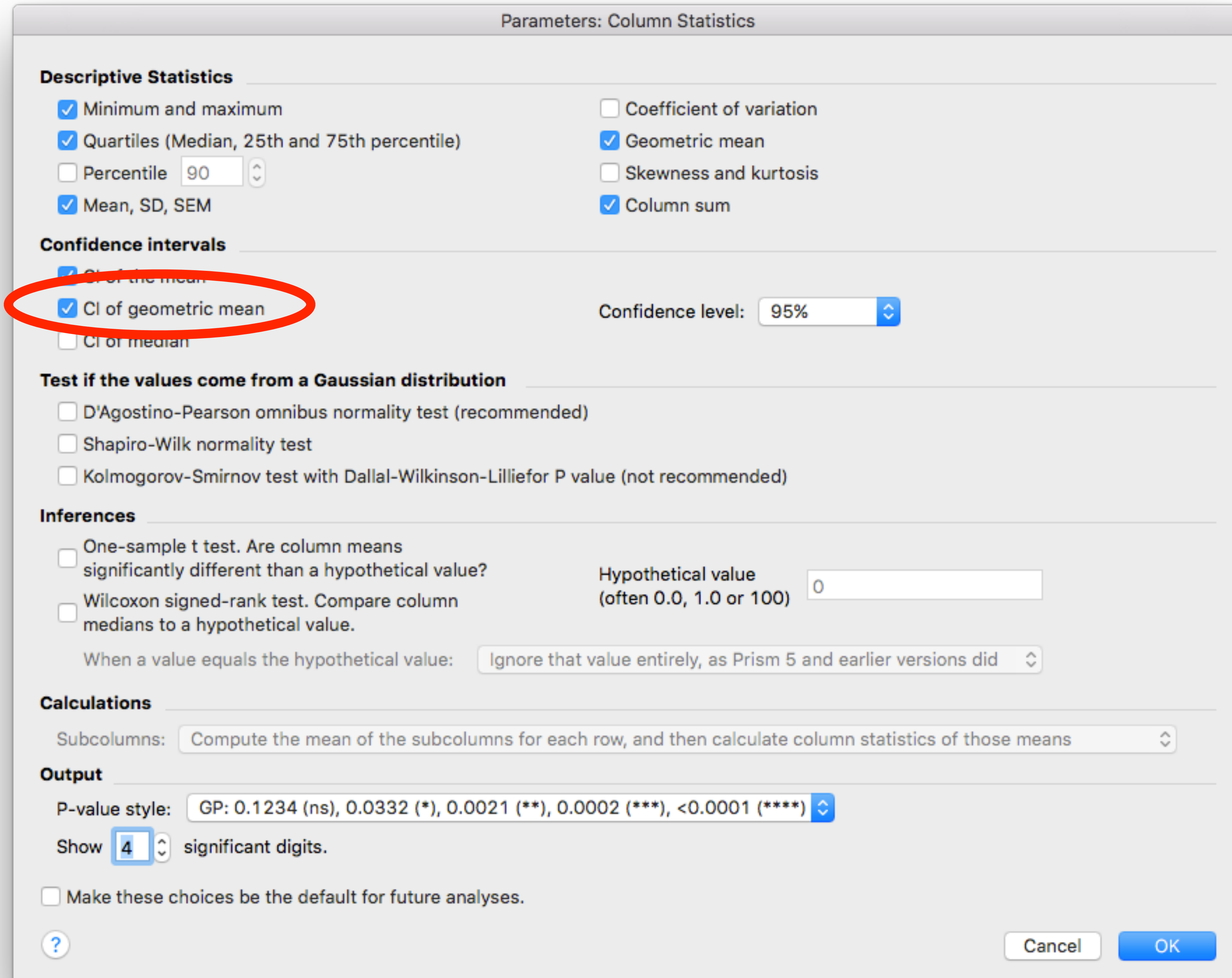
1. Log transform the data

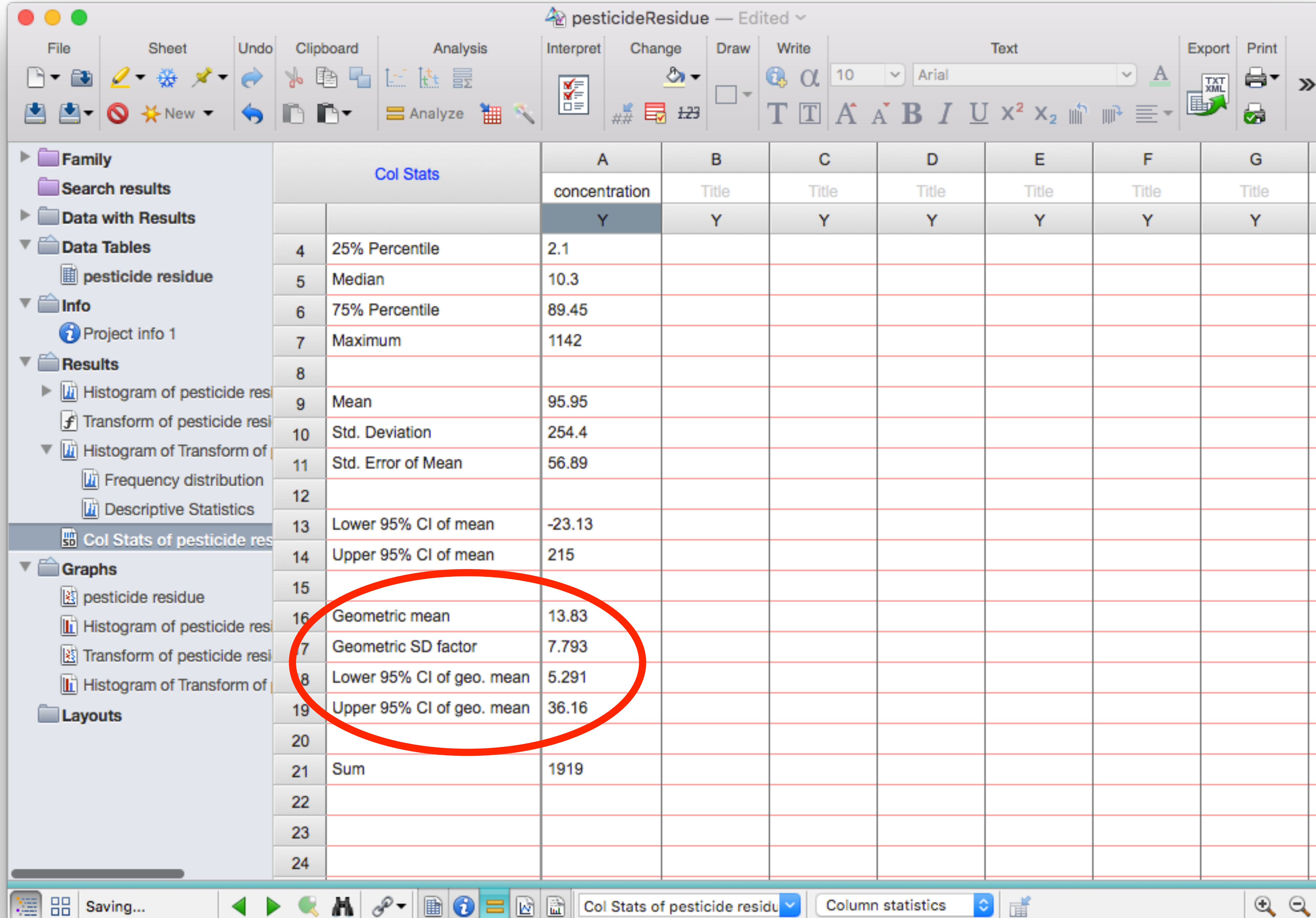
2. Calculate the 95% CI based on the log transformed data

3. Take the antilog of the lower and upper bounds of the 95% CI calculated on the log transformed data

Log Transform Gives Asymmetrical Limits

GraphPad will calculate CI for the geometric mean automatically




pesticideResidue — Edited
File Sheet Undo Clipboard Analysis Interpret Change Draw Write Text Export Print
Family
Search results
Data with Results
Data Tables
pesticide residue
Info
Project info 1
Results
Histogram of pesticide resi
Transform of pesticide resi
Histogram of Transform of
Frequency distribution
Descriptive Statistics
Col Stats of pesticide res
Graphs
pesticide residue
Histogram of pesticide resi
Transform of pesticide resi
Histogram of Transform of
Layouts
Col Stats
A concentration
4 25% Percentile 2.1
5 Median 10.3
6 75% Percentile 89.45
7 Maximum 1142
9 Mean 95.95
10 Std. Deviation 254.4
11 Std. Error of Mean 56.89
13 Lower 95% CI of mean -23.13
14 Upper 95% CI of mean 215
16 Geometric mean 13.83
17 Geometric SD factor 7.793
18 Lower 95% CI of geo. mean 5.291
19 Upper 95% CI of geo. mean 36.16
21 Sum 1919
Saving...
Col Stats of pesticide residu
Column statistics

Other transformations

- **Log transformation with an added constant** - if the data contains zeros or negative values, it is impossible to log transform the data. The usual solution is to add a fixed value to each point prior to log transforming the data.

- **Square-root transformation** - Data that represent counts tend to be positively skewed. Transform the data by taking the square root.

What did we learn?

- Confidence intervals give a sense of how accurately we estimated the population mean from our sample mean.
- Level of confidence refers to the number of intervals generated in the same manner that would cover the true mean.
- Larger SD results in larger CI, smaller sample size results in larger CI, higher level of confidence results in larger CI
- 95% CI for the difference between two means takes into account the error in estimating both means.
- Deviations from normality can cause CI to be incorrect
- Data transformation can alleviate the problems of non-normality