Chapter 4: The Normal Distribution

TXCL7565/PHSC7565

What This Lecture Covers

- What is a normal distribution
- Identifying data that are not normally distributed
- Proportions of individuals within 1 SD or 2 SD of the mean
- Skewness and kurtosis
- Tests for normal distributions

WHAT IS A NORMAL DISTRIBUTION

Normal Distribution

- Many of the things we measure show a characteristic distribution, with the bulk of the data points clustered around the mean and data points become steadily rarer as we move further from the mean.
- When many (independent) random factors contribute to an observed value, the observed values tend to follow a normal distribution.
- Normal Distribution = Gaussian Distribution

Many Random Factors

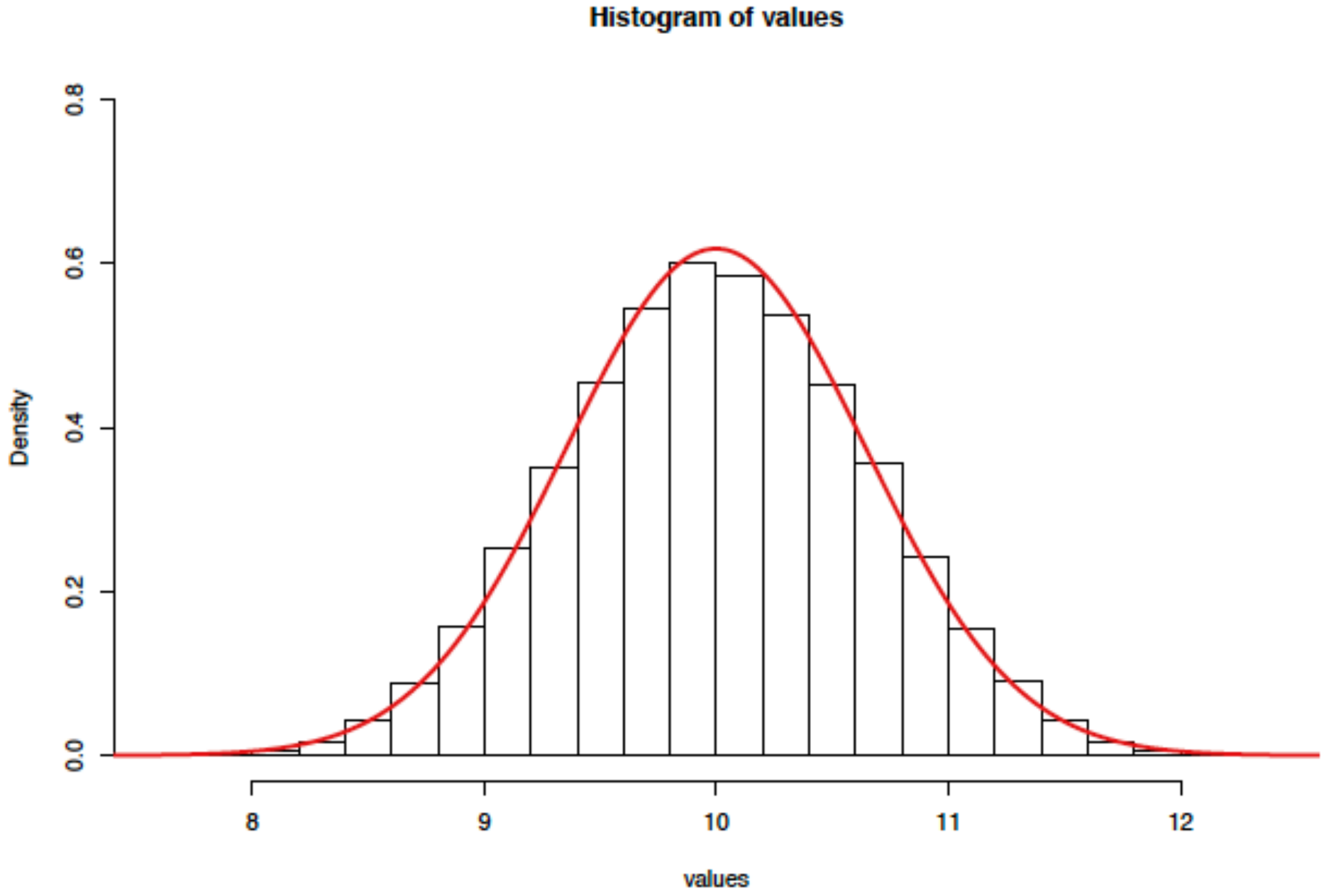
```
true_value = 10
e1 = runif(100000) - 0.5
e2 = runif(100000) - 0.5
e3 = runif(100000) - 0.5
e4 = runif(100000) - 0.5
e5 = runif(100000) - 0.5

values = true_value + e1 + e2 + e3 + e4 + e5
head(round(values,2))
```

```
## [1] 10.31  9.25  9.61  9.20  9.74 10.60
```

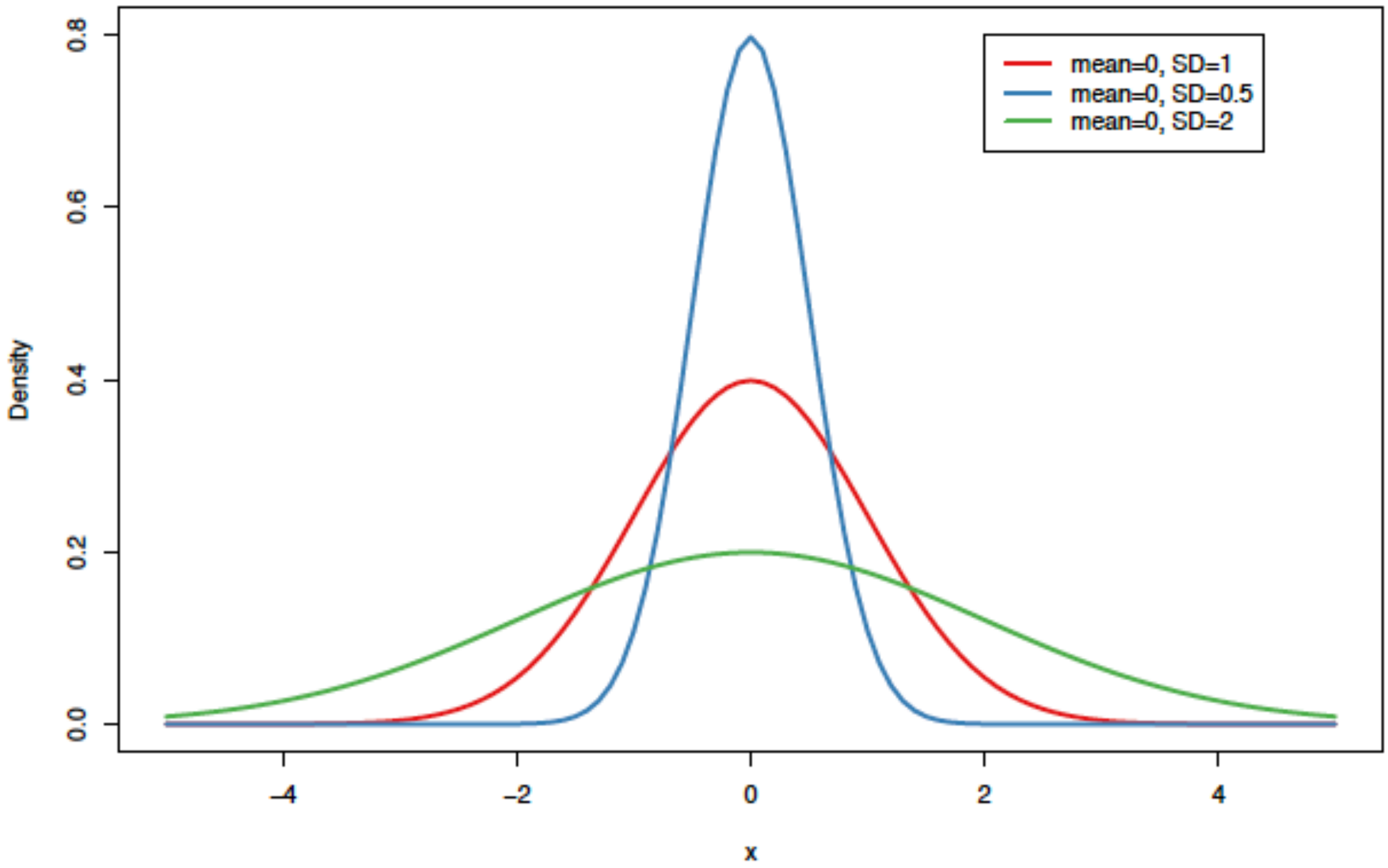
```
tail(round(values,2))
```

```
## [1]  8.57  9.06 10.63  9.87 10.81  9.57
```

Results of many random factors contributing



The Nature of the Normal Distribution



Why Does a Normal Distribution Matter?

Many commonly used statistical tests rely on the assumption that the data have been sampled from a population that follows a normal distribution.

This is often a reasonable assumption.

IDENTIFYING DATA TAHT ARE NOT NORMALLY DISTRIBUTED

How to Spot Non-Normal Data

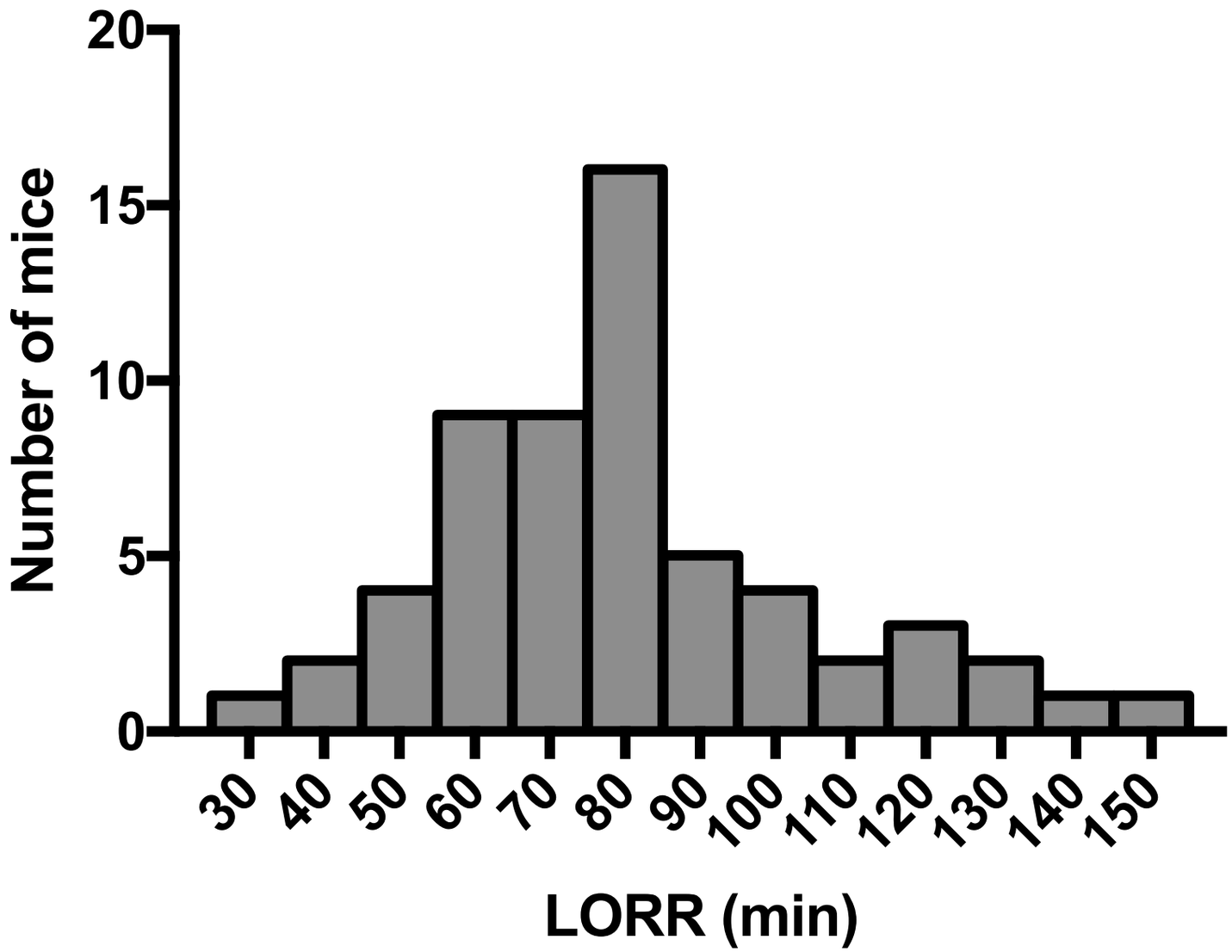
Most often, the easiest/best way to determine if data are normally distributed is to plot them in a histogram

Example Data Set - Sleep Time

One characteristic related to an individual's likelihood of developing an alcohol use disorder is how sensitive they are to alcohol. In the Radcliffe lab, how sensitive a particular mouse is to the sedative effects of alcohol is measured by giving them a large enough dose of alcohol to cause them to 'fall asleep' and then measuring the number of minutes that pass before they wake up.

Sleep time or **Loss of Righting Reflex (LORR)** is the number of minutes between when the mouse first looses the ability to right themselves when placed on their back to when they can right themselves again.

Histogram of Sleep Time



Characteristics of a Normal Distribution

3 visual characteristics that any true normal distribution will possess:

1. The data are unimodal
2. The distribution is symmetrical
3. The frequencies decline steadily as we move towards higher or lower values, without any sudden, sharp cut-off

Unimodal vs. Polymodal

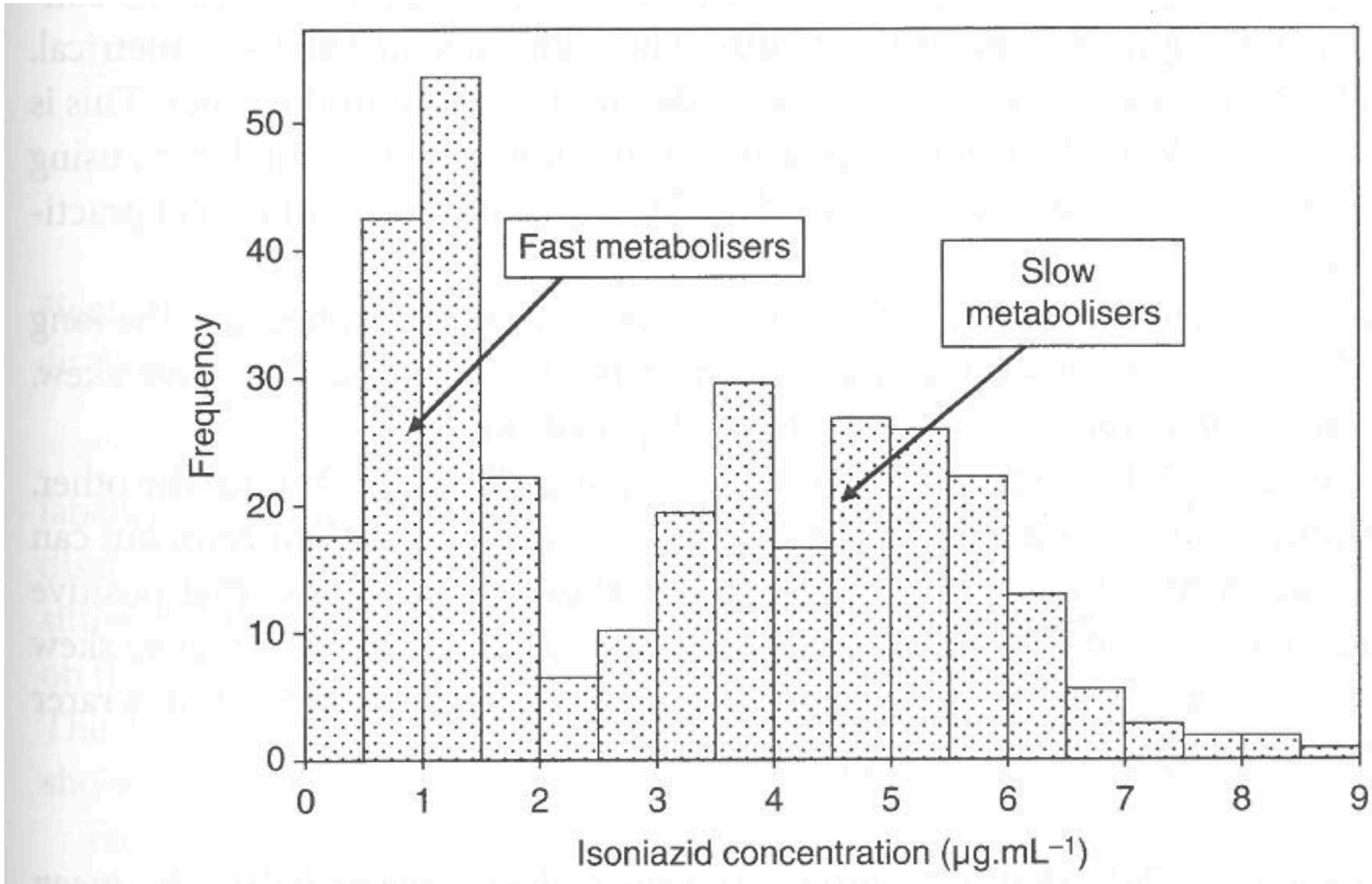


Figure 4.1 Bimodal data. Isoniazid concentrations (μg.mL^{-1}) six hours after a standard oral dose

Symmetrical Distribution

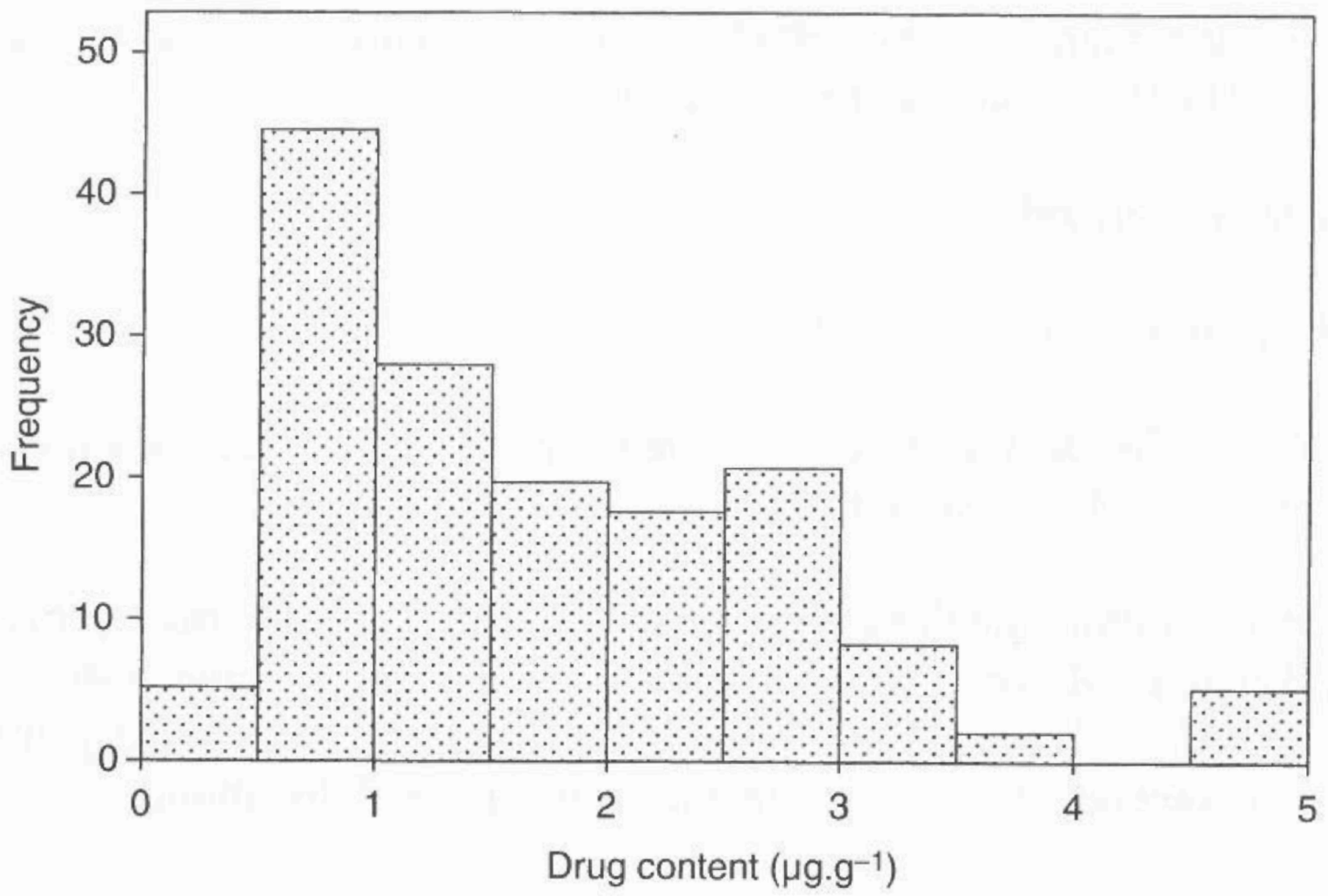


Figure 4.2 Skewed data. Drug content (mg.g^{-1} of dried plant tissue) of a series of individual plants

Sharp Cut-Offs

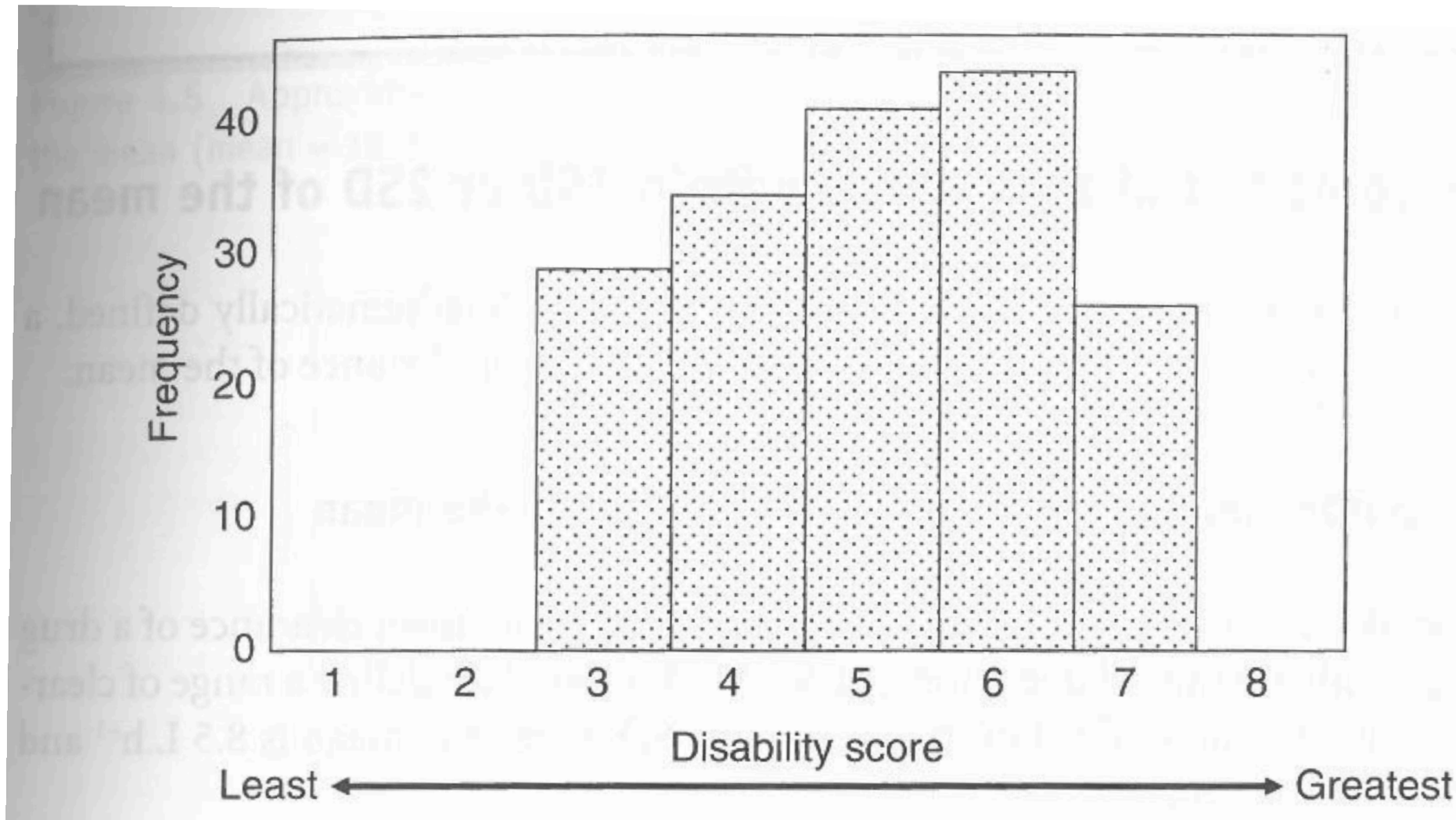
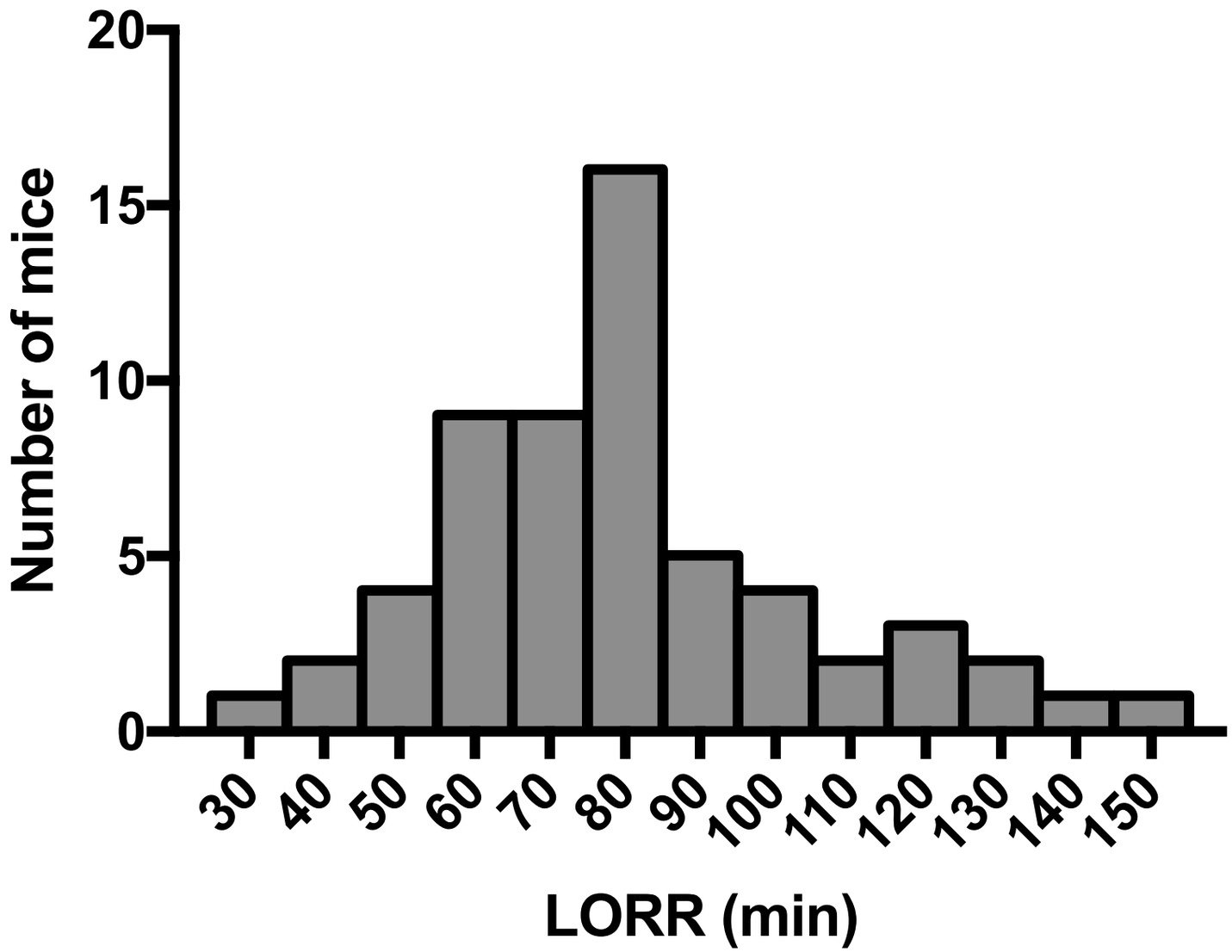
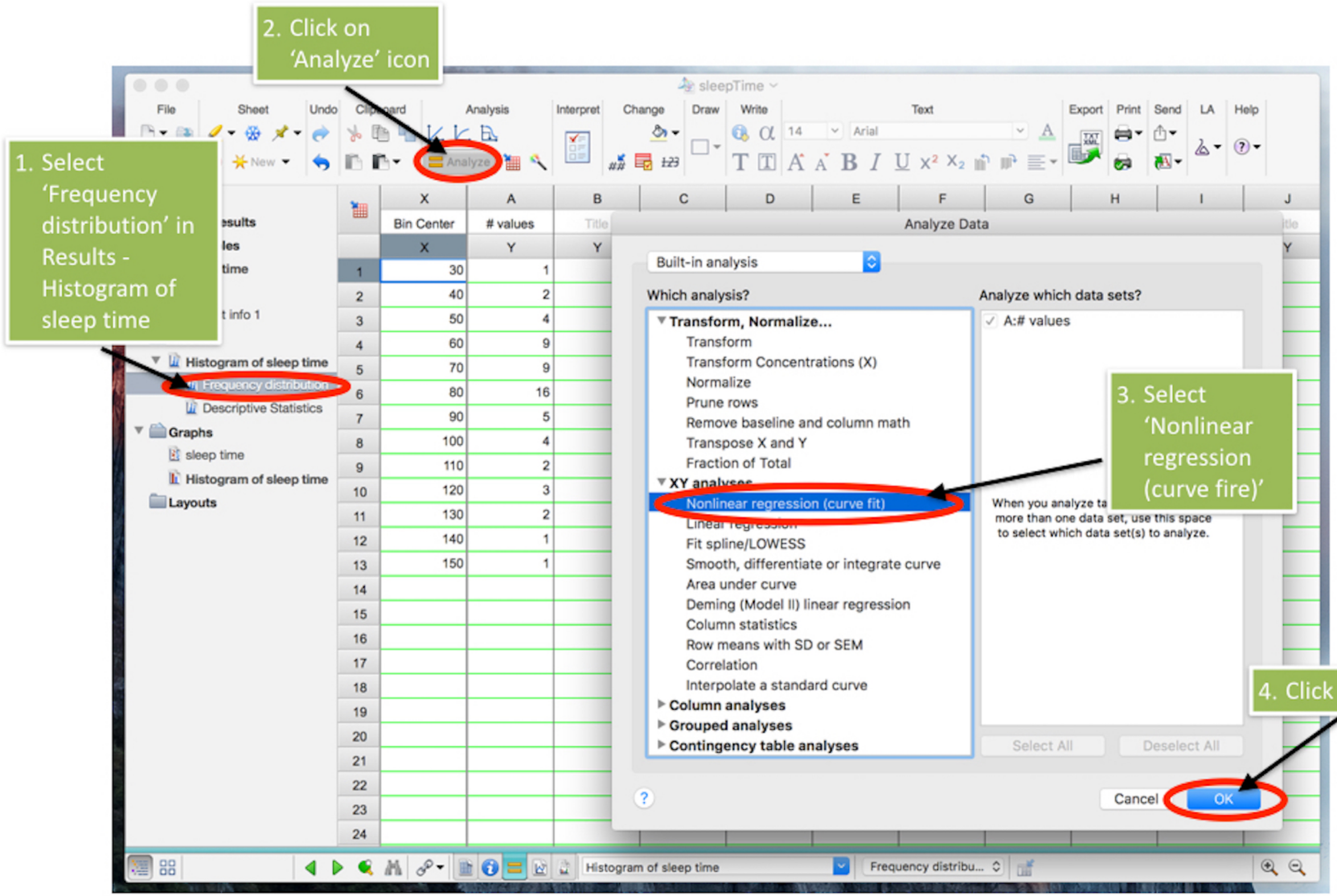


Figure 4.4 Data with sudden cut offs. Patients' self assessment scores for degree of disability due to gout

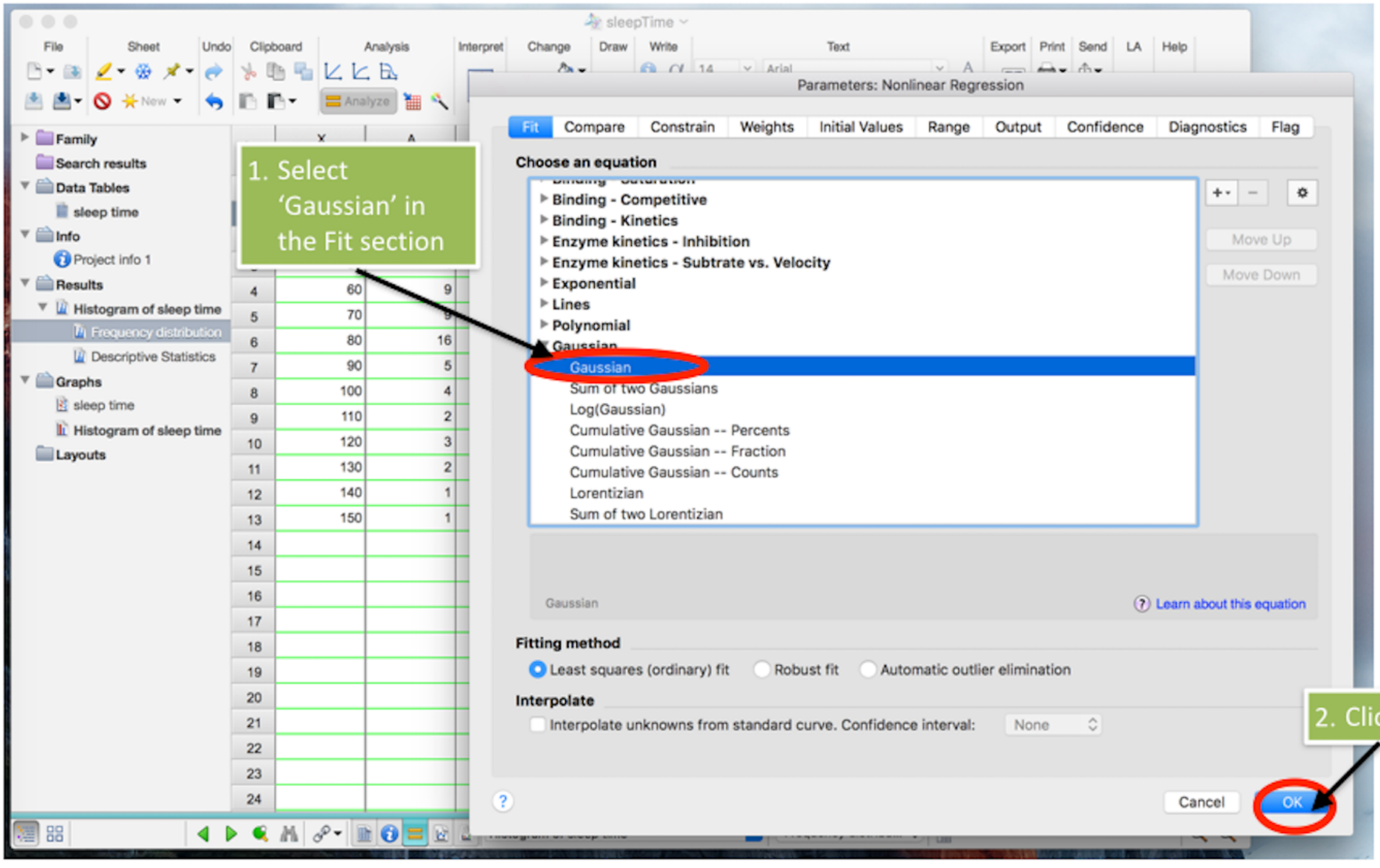
Histogram of Sleep Time



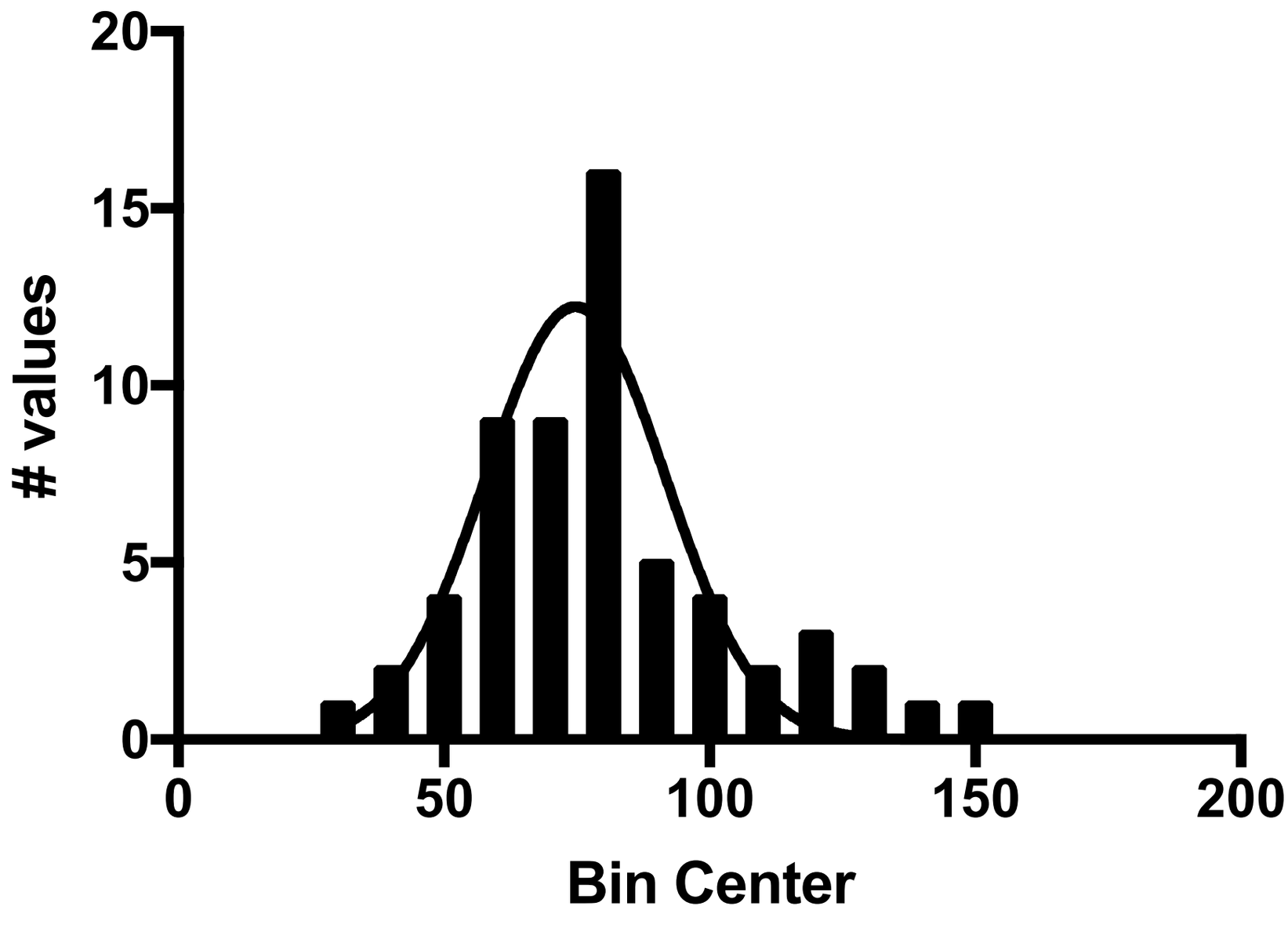
Adding a Normal Distribution Line - GraphPad



Adding a Normal Distribution Line - GraphPad



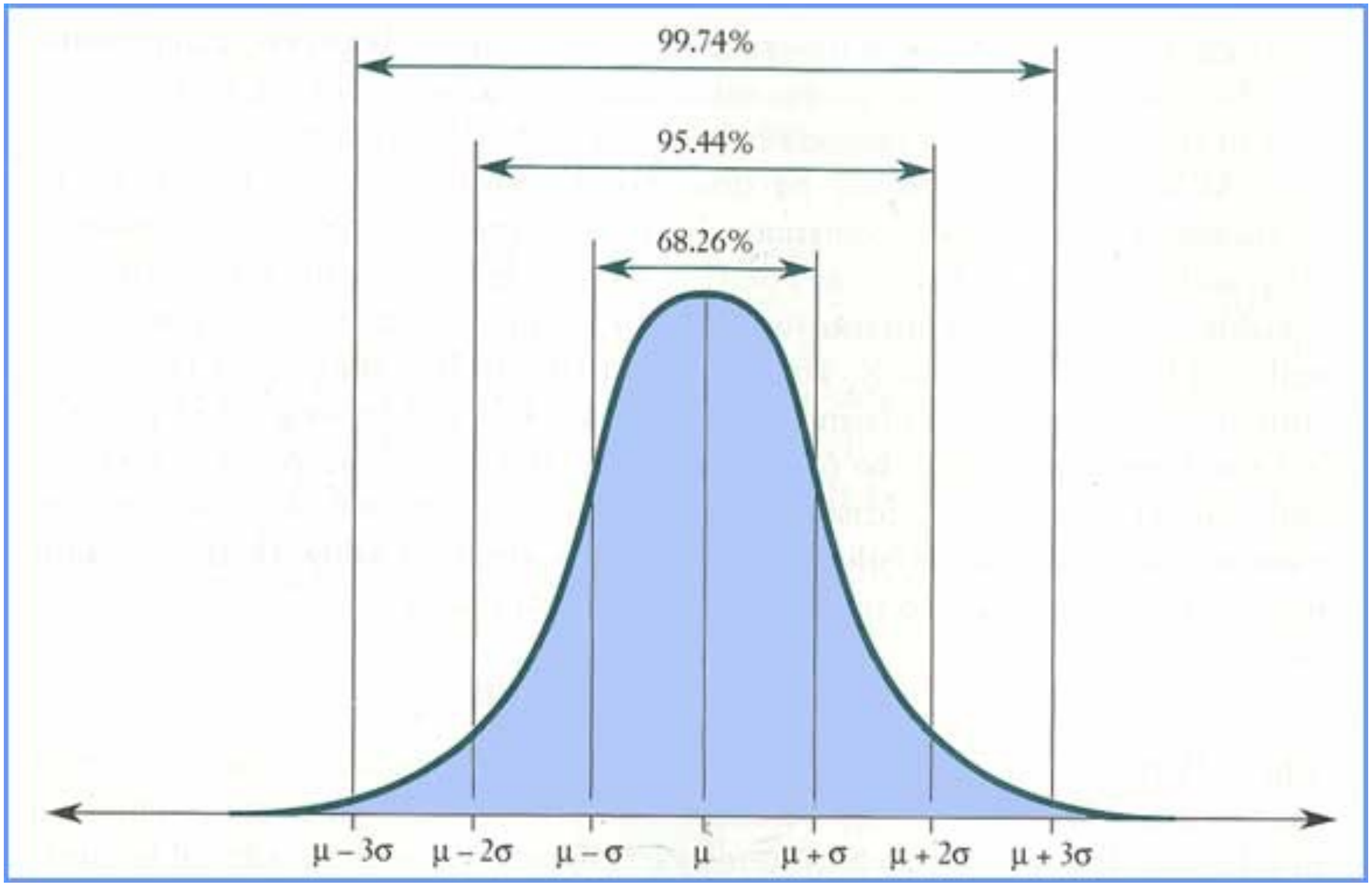
Histogram of Sleep Time With Normal Curve



PROPORTIONS OF INDIVIDUALS WITHIN 1 SD OR 2 SD OF THE MEAN

Proportions in Normal Distribution

Ideal Normal Distribution In Terms of SD



Sleep Time Example

- Mean = 79.2
- SD = 25.05
- Mean ‘±’ 1 SD = (54.2, 104.3)
- Proportion of values within 1 SD = 0.73
- Mean ‘±’ 2 SD = (29.2, 129.3)
- Proportion of values within 2 SD = 0.95

The Standard Normal

- Standard Normal - Normal distribution with mean = 0 and SD =1

- All normal distributions can be converted to a standard normal distribution using the following formula:

$$Z = \frac{Value - Mean}{SD}$$

- Z is the number of SD the value is away from the mean

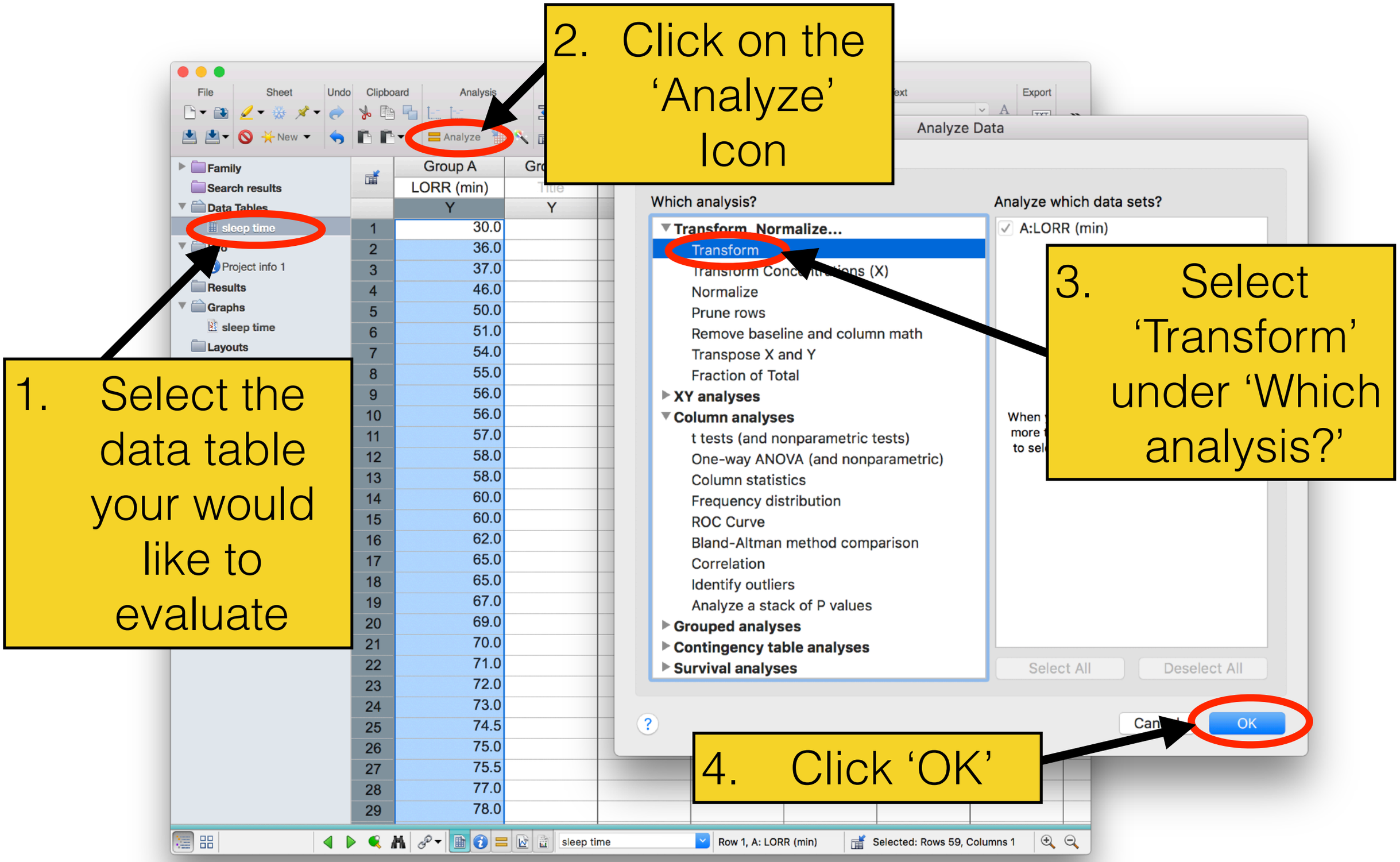
Calculate the proportion of samples within 1 SD and 2 SD - GraphPad

1. Calculate Z scores for each value using ‘Transform’

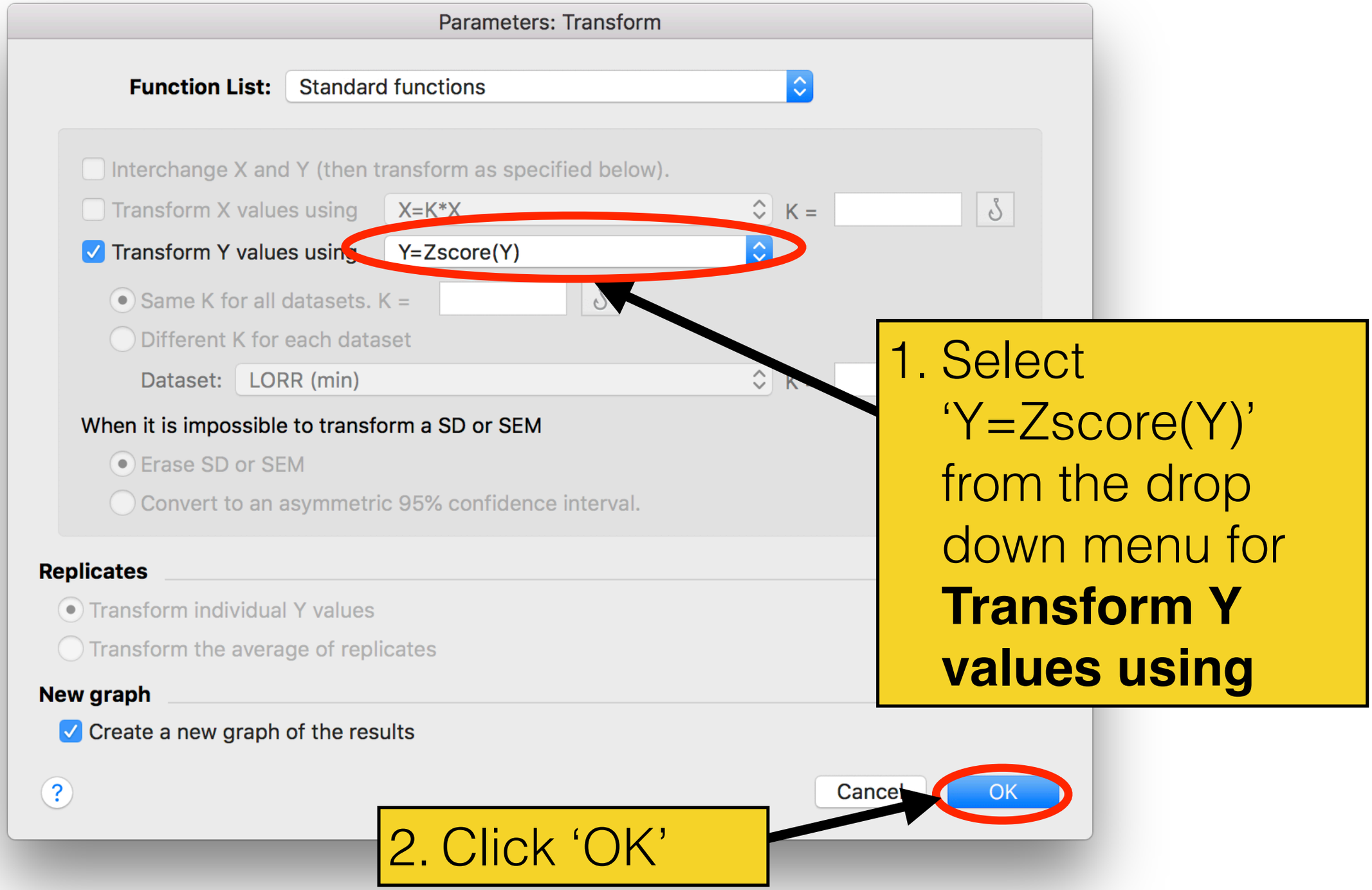
2. Calculate the absolute value of the z scores using ‘Transform’

3. Calculate the cumulative frequency based on the absolute values of the z scores

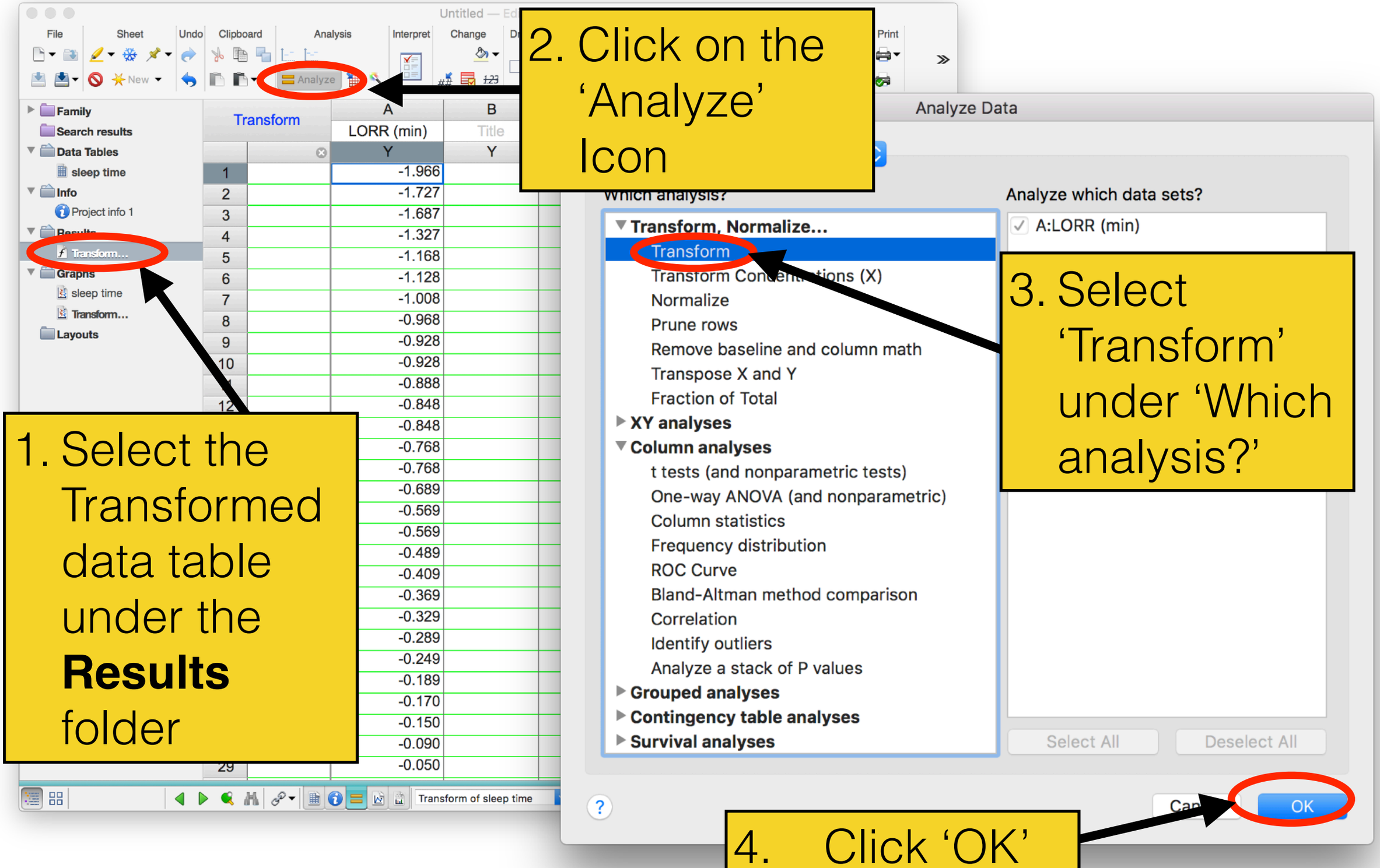
Calculate Z Scores



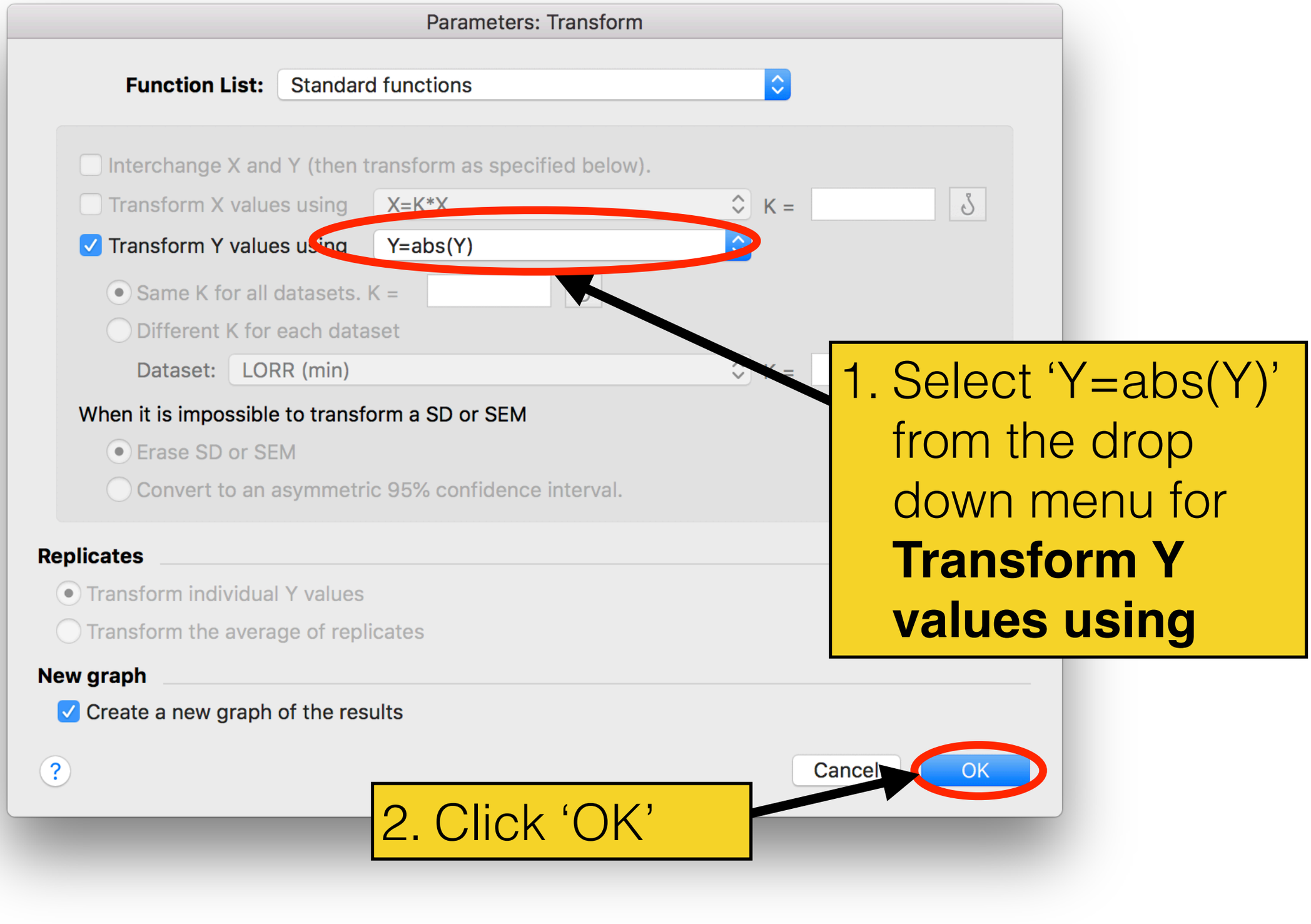
Calculate Z scores



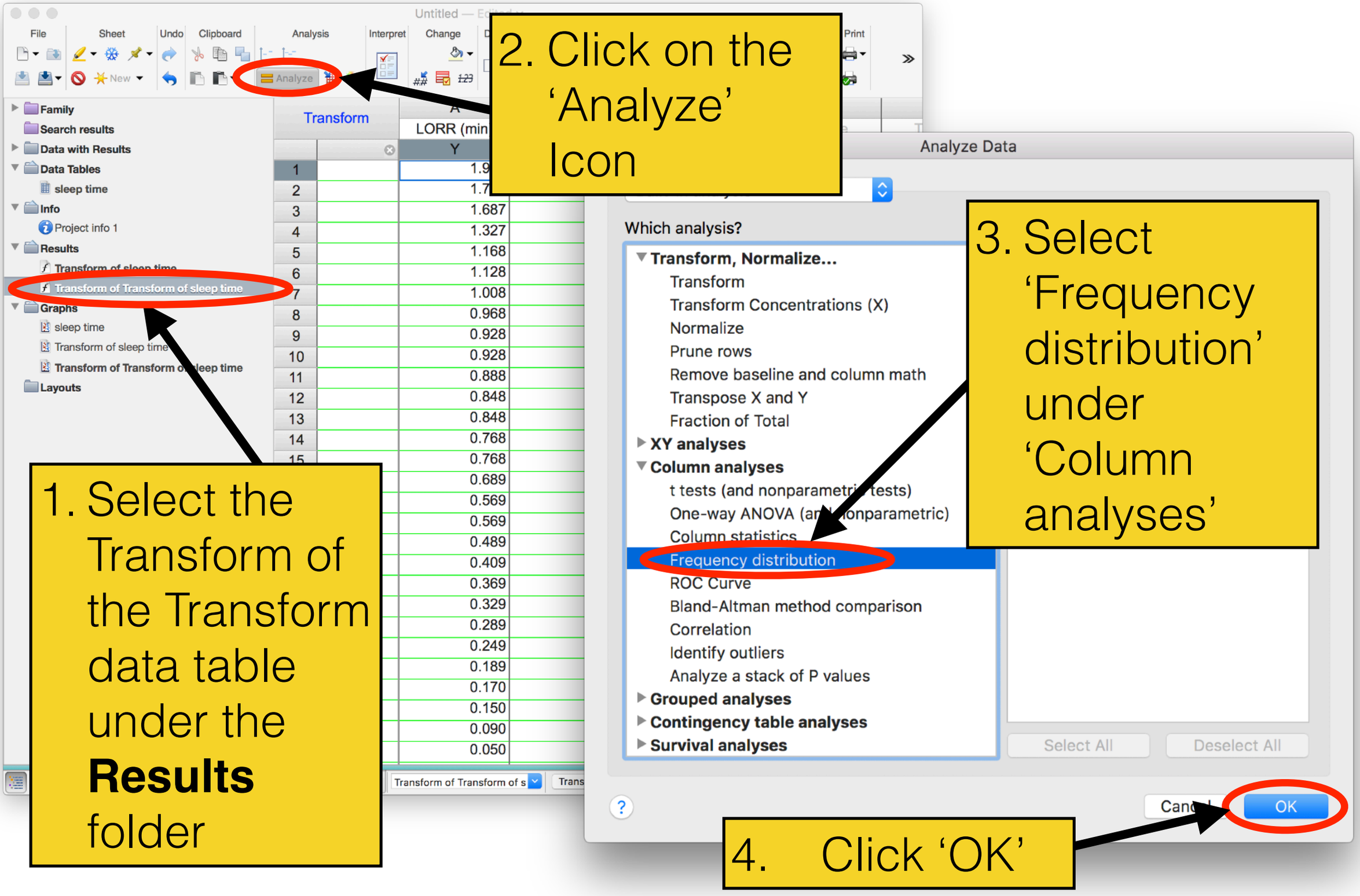
Calculate the absolute value of each Z score



Calculate the absolute value of each Z score



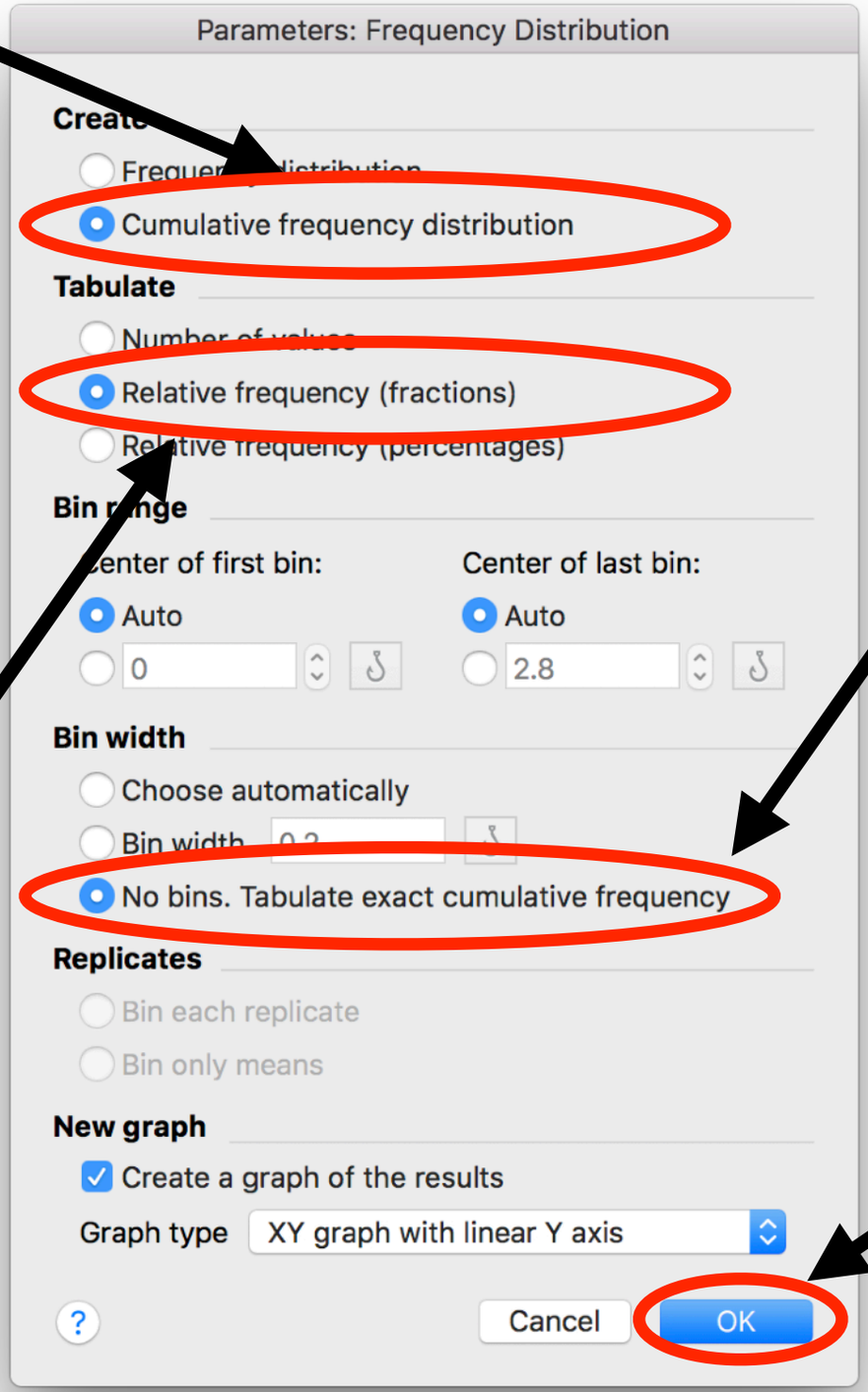
Calculate the cumulative frequency based on the absolute values of the z scores



Calculate the cumulative frequency based on the absolute values of the z scores

1. Select ‘Cumulative frequency distribution’ under **Create**

2. Select ‘Relative frequency (fractions)’ under **Tabulate**



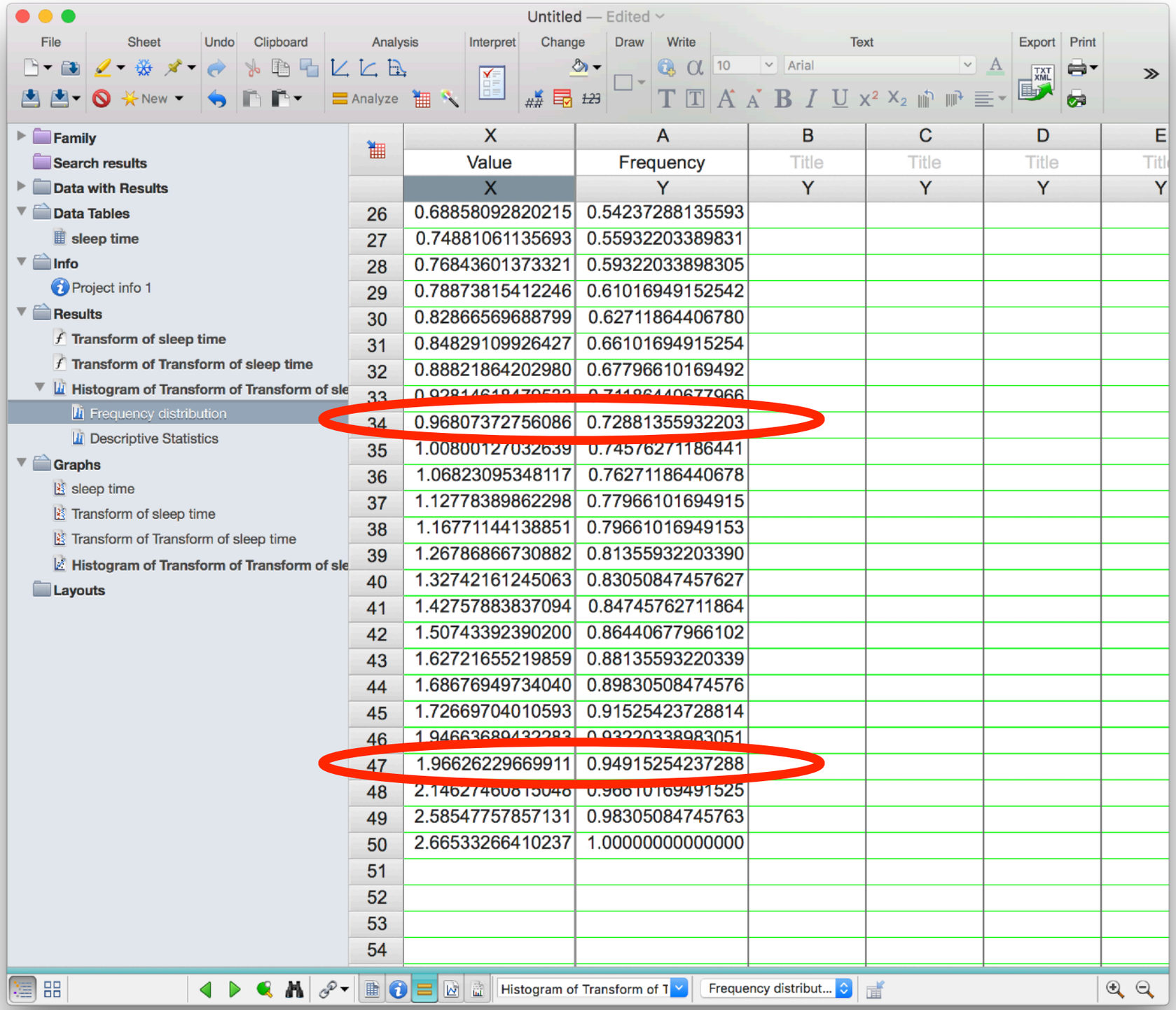
3. Select ‘No bins.’ under **Bin width**

4. Click ‘OK’

Calculate the proportion of samples within 1 SD and 2 SD - GraphPad

73% of the data points are less than 1 SD from the mean

95% of the data points are less than 2 SD from the mean



The Normal Distribution Does Not Define Normal Limits

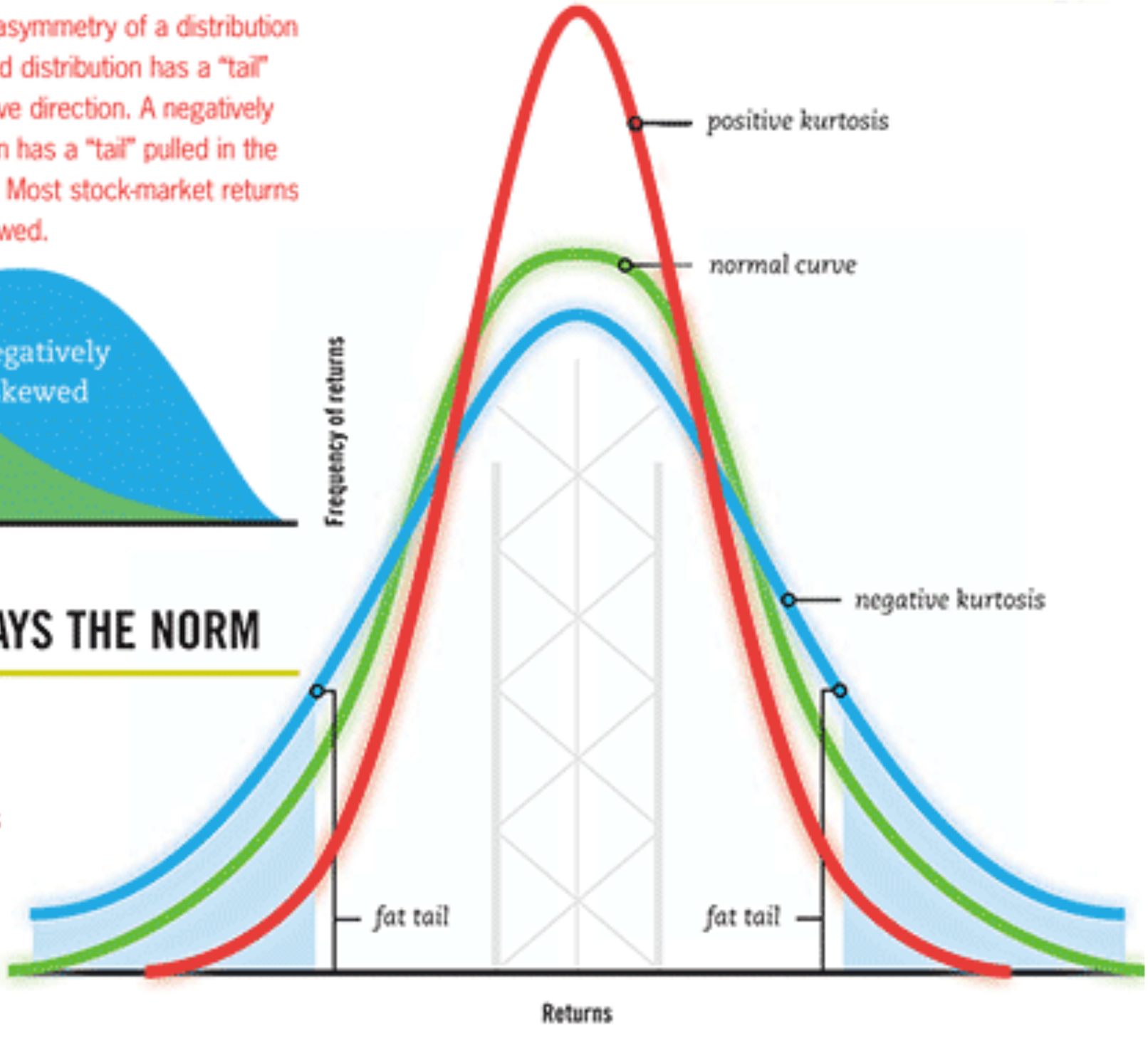
- Normal distribution doe not equate to ‘normal’ range.
- Defining the normal limits of a clinical measurement is not straightforward and requires clinical thinking, not just statistics.

SKEWNESS AND KURTOSIS

Skewness is the asymmetry of a distribution
A positively skewed distribution has a "tail" pulled in the positive direction. A negatively skewed distribution has a "tail" pulled in the negative direction. Most stock-market returns are negatively skewed.

NORMAL NOT ALWAYS THE NORM

Kurtosis refers to how peaked the curve is: steeper means positive kurtosis and flatter means negative kurtosis. Fat tails occur when there are more outsize returns on the downside or upside, or both, than the normal curve suggests.



https://anotherbloodybullshitblog.wordpress.com/2016/01/18/statistics-how-to-test-if-your-data-follows-the-normal-distribution/

Skewness

Skewness can be quantified numerically by a number that ranges from negative infinity to positive infinity, with a value of 0 indicating that no skewness is present

Kurtosis

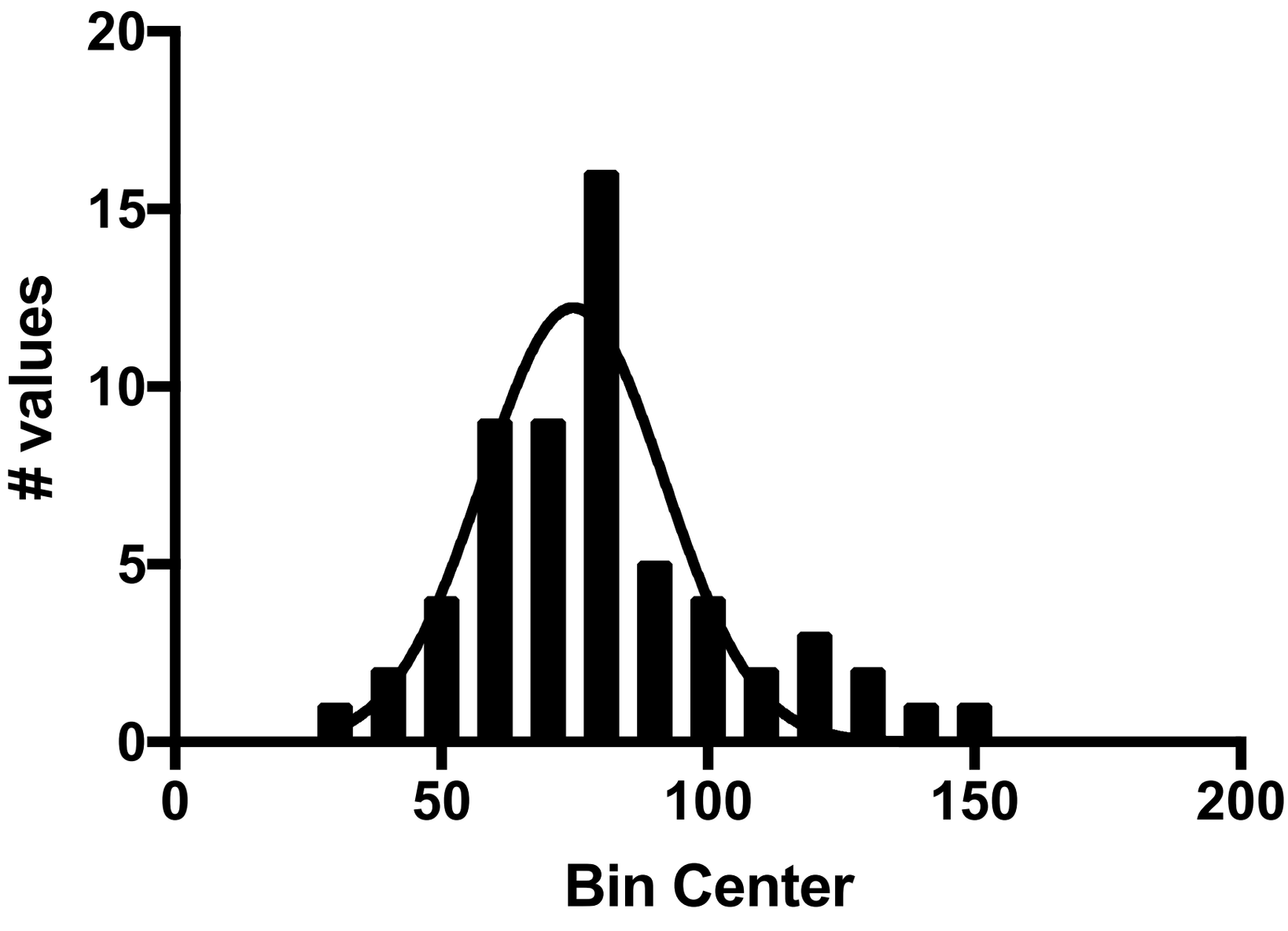
Like skewness, **kurtosis** can also be quantified numerically with a range of negative infinity to positive infinity, with a value of 0 indicating that no kurtosis is present.

- A *negative* kurtosis value indicates that the ‘shoulders’ of the distribution are too high and wide

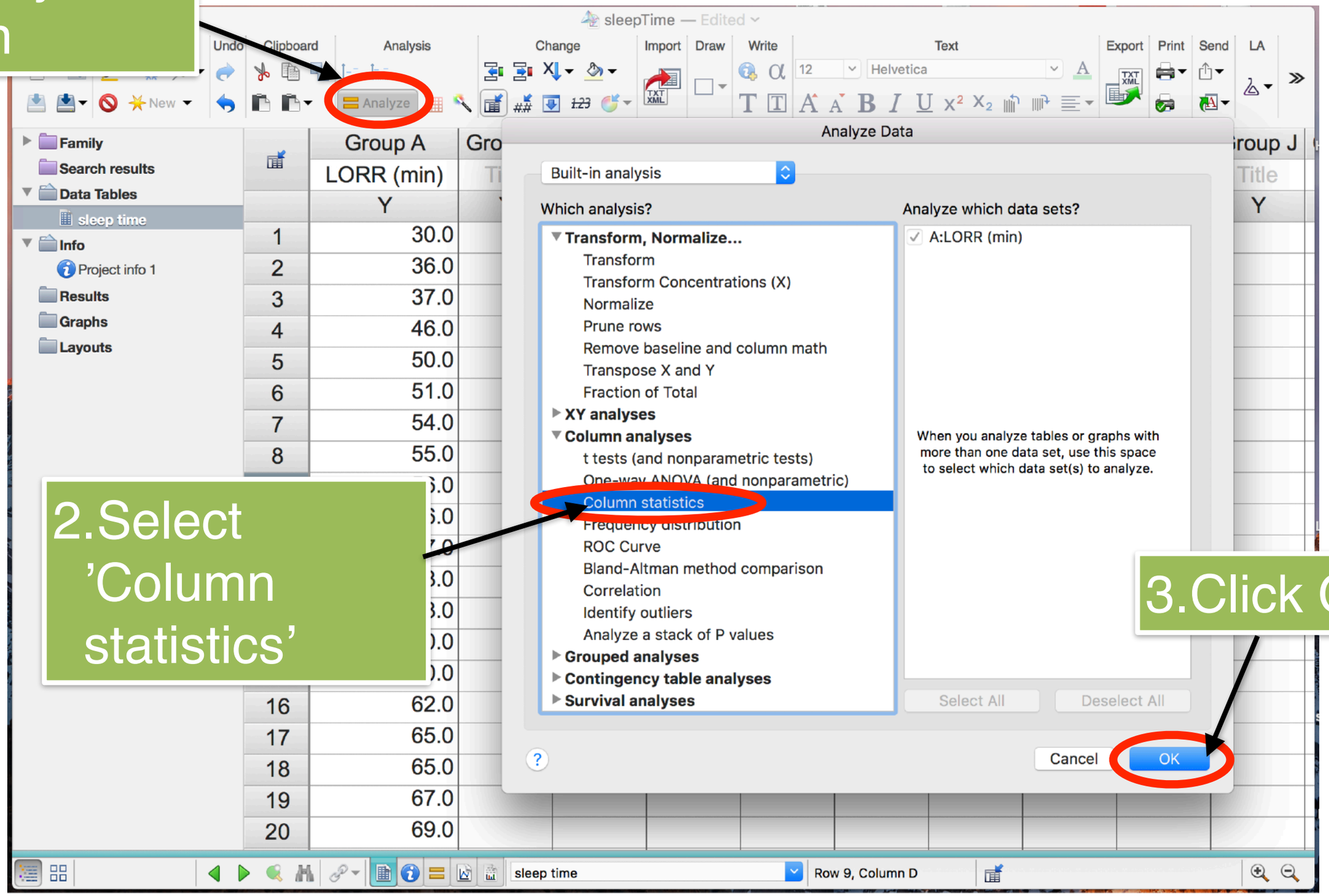
- A *positive* kurtosis value indicates that the ‘shoulders’ of the distribution are too low and narrow.

Skewness and Kurtosis in Sleep Time

Skewness = 0.70 (positive skew); kurtosis = 0.61 (narrow peak)



1.Click ‘Analyze icon



2.Select ’Column statistics’

3.Click OK

Parameters: Column Statistics

Descriptive Statistics

- [x] Minimum and maximum
- [x] Quartiles (Median, 25th and 75th percentile)
- [ ] Percentile 90
- [x] Mean, SD, SEM
- [ ] Coefficient of variation
- [x] Geometric mean
- [x] Skewness and kurtosis
- [x] Column sum

Confidence intervals

- [x] 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

1.Click ‘Skewness and kurtosis’

2.Click OK

sleepTime — Edited

File Sheet Undo Clipboard Analysis Interpret Change Draw Write Text Export Print Send LA Help

New Analyze 12 Helvetica

GraphPad PRISM

- Family
- Search results
- Data Tables
  - sleep time
- Info
  - Project info 1
- Results
  - Col Stats of sleep time
- Graphs
- Layouts

Col Stats	A	B	C	D	E	F
	LORR (min)	Title	Title	Title	Title	Title
	Y	Y	Y	Y	Y	Y
1	Number of values	59				
2						
3	Minimum	30				
4	25% Percentile	60				
5	Median	78				
6	75% Percentile	89				
7	Maximum	146				
8						
9	Mean	79.25				
10	Std. Deviation	25.05				
11	Std. Error of Mean	3.261				
12						
13	Lower 95% CI of mean	72.72				
14	Upper 95% CI of mean	85.77				
15						
16	Geometric mean	75.41				
17	Geometric SD factor	1.381				
18						
19	Skewness	0.6964				
20	Kurtosis	0.6086				
21						
22	Sum	4676				
23						
24						

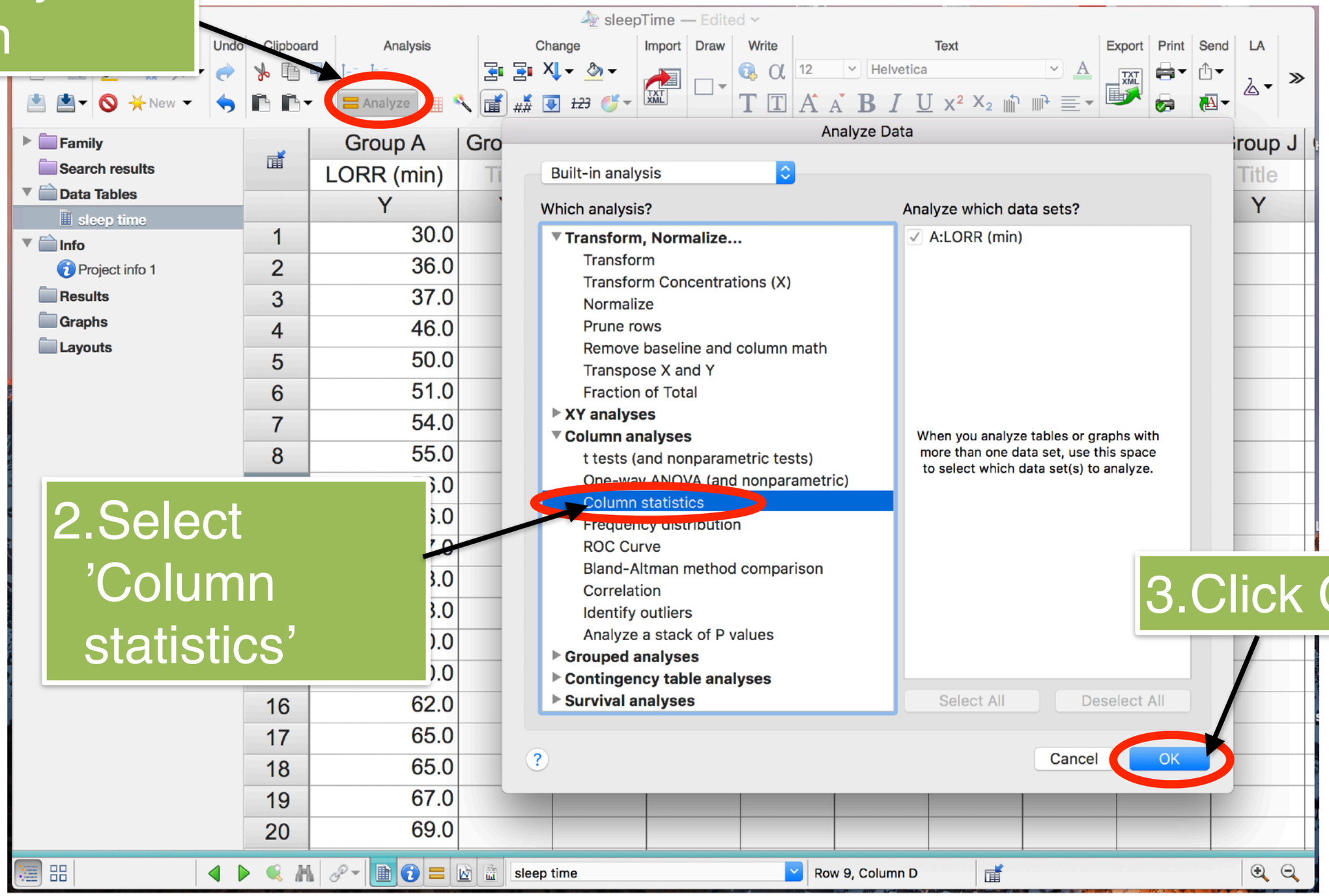
Col Stats of sleep time | Column statistics

TESTS FOR NORMAL DISTRIBUTIONS

D'Agostino-Pearson Omnibus Normality Test

- Recommended in GraphPad
- Test for both skewness and kurtosis
- A significant p-value ($p<0.05$) indicates that the distribution is significantly different from normal, i.e., the data are non-normal.
- CAVEATS:
  - With a small sample size even large departures from normality may not be detected
  - With a large sample size even small departures from normality will be detected

1.Click ‘Analyze icon



2.Select ’Column statistics’

3.Click OK

Parameters: Column Statistics

Descriptive Statistics

- [x] Minimum and maximum
- [x] Quartiles (Median, 25th and 75th percentile)
- [ ] Percentile 90
- [x] Mean, SD, SEM
- [ ] Coefficient of variation
- [ ] Geometric mean
- [ ] Skewness and kurtosis
- [x] Column sum

Confidence intervals

- [x] CI of the mean
- [ ] CI of geometric mean
- [ ] CI of median

Confidence level: 95%

Test if the values come from a Gaussian distribution

- [x] D'Agostino-Pearson omnibus normality test (recommended)
- [ ] Shapiro-Wilk normality test
- [ ] Kolmogorov-Smirnov test with Dallal-Wilkinson-Lilliefor P value (not recommended)

1.Click 'D'Agostino-Pearson omnibus normality test'

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.

2.Click OK

Cancel OK

sleepTime — Edited

File Sheet Undo Clipboard Analysis Interpret Change Draw Write Text Export Print Send LA Help

New Analyze 12 Helvetica

GraphPad PRISM

- Family
- Search results
- Data Tables
  - sleep time
- Info
  - Project info 1
- Results
  - Col Stats of sleep time
  - Col Stats of sleep time
- Graphs
- Layouts

Col Stats		A	B	C	D	E
		LORR (min)	Title	Title	Title	Title
		Y	Y	Y	Y	Y
1	Number of values	59				
2						
3	Minimum	30				
4	25% Percentile	60				
5	Median	78				
6	75% Percentile	89				
7	Maximum	146				
8						
9	Mean	79.25				
10	Std. Deviation	25.05				
11	Std. Error of Mean	3.261				
12						
13	Lower 95% CI of mean	72.72				
14	Upper 95% CI of mean	85.77				
15						
16	Sum	4676				
17						
18	D'Agostino & Pearson normality test					
19	K2	5.95				
20	P value	0.0510				
21	Passed normality test (alpha=0.05)?	Yes				
22	P value summary	ns				
23						
24						

Col Stats of sleep time | Column statistics

What did we learn?

- The normal distribution often results from many random factors creating variability and therefore, is common in a research setting.
- 3 main characteristics of a normal distribution: 1) the data are unimodal, 2) the distribution is symmetric, and 3) the frequencies decline steadily as we move towards higher and lower values, without any sudden sharp cut-off.
- If the data are normally distributed, 68% of observations will fall within 1 SD of the mean and 95% of observations will fall within 2 SD.
- Skewness is a quantitative measure of the lack of symmetry in the data distribution and kurtosis is a quantitative measure of how closely the peak, shoulders, and tails of the distribution match a normal distribution.
- Use tests of non-normality with caution.