

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

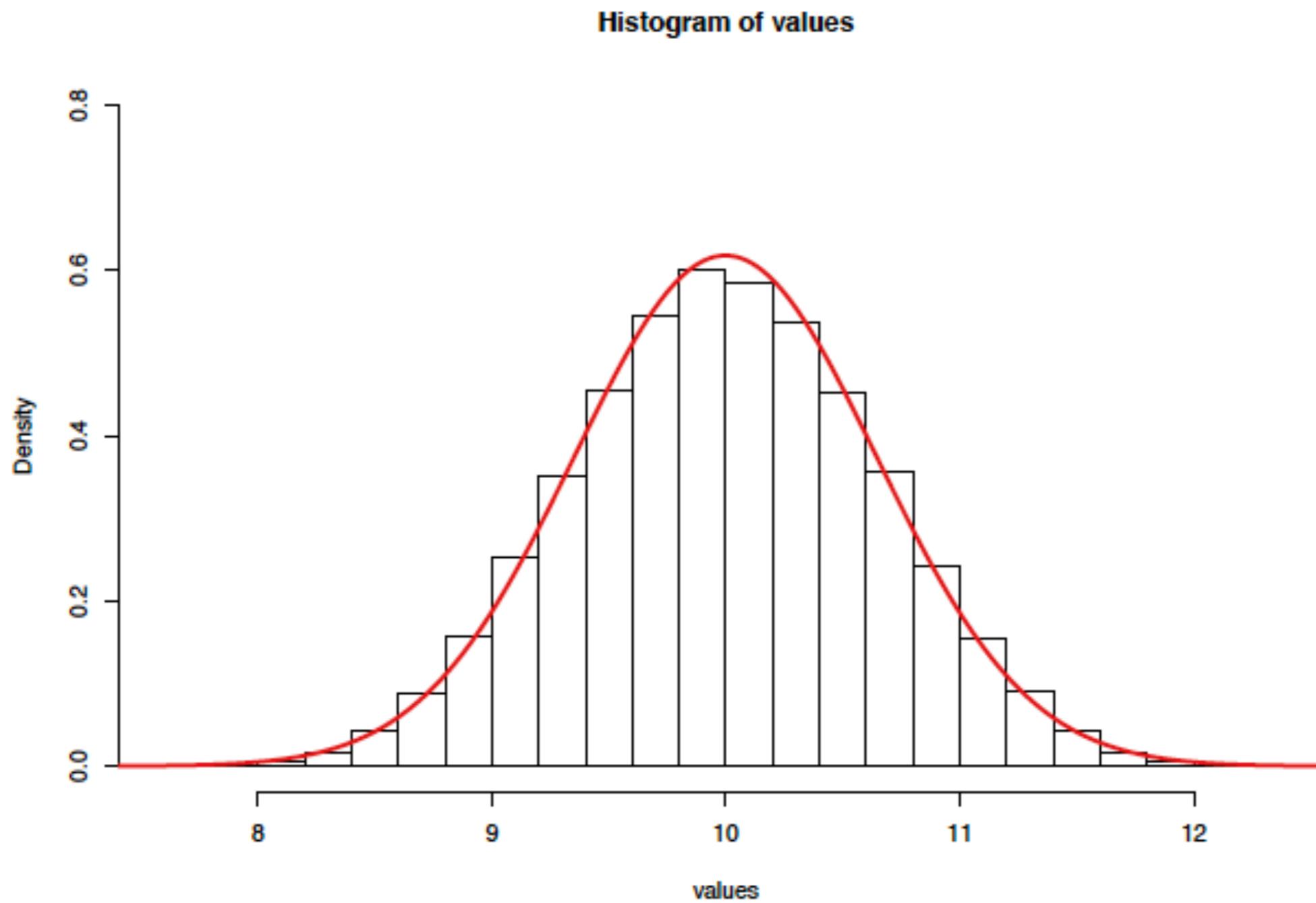
values = true_value + e1 + e2 + e3 + e4 + e5
head(round(values,2))
```

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

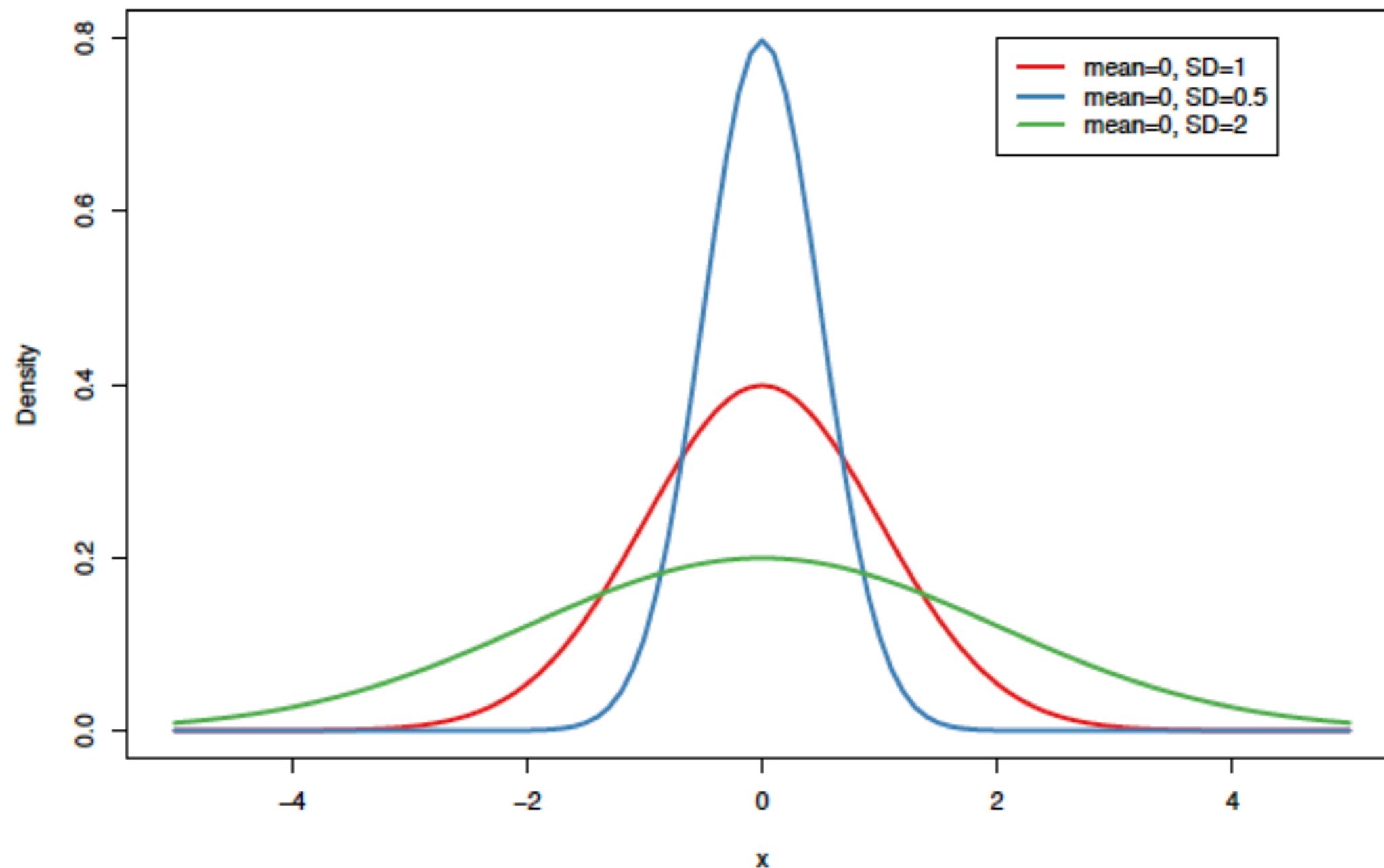
```
tail(round(values,2))
```

```
## [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 THAT
ARE NOT NORMALLY
DISTRIBUTED**

How to Spot Non-Normal Data

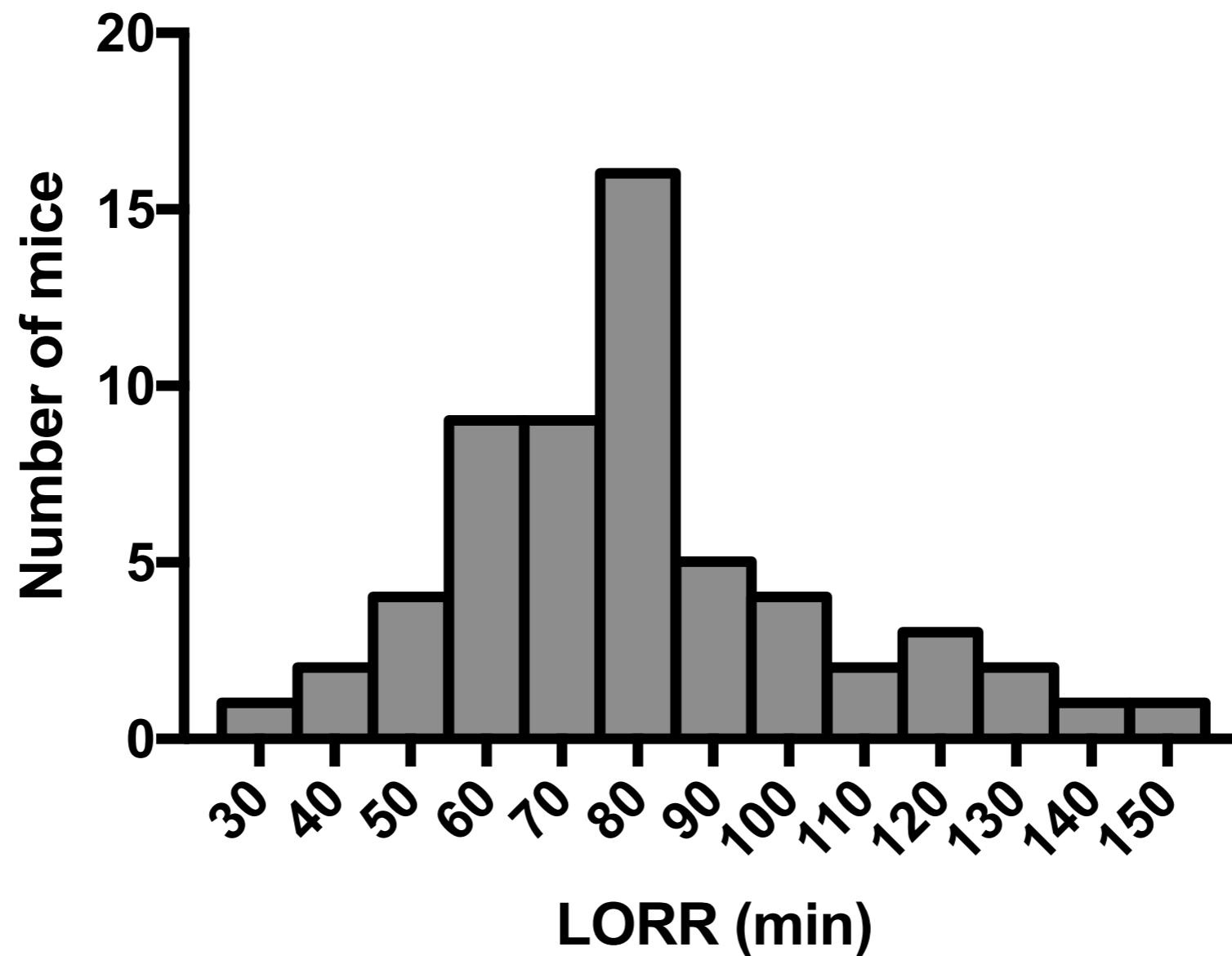
Most often, the easiest/best way to determine if data are normally distributed is to plot them in a histogram

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 the time point when the mouse first loses the ability to right themselves when placed on their back to the time when they can right themselves again.

Histogram of Sleep Time

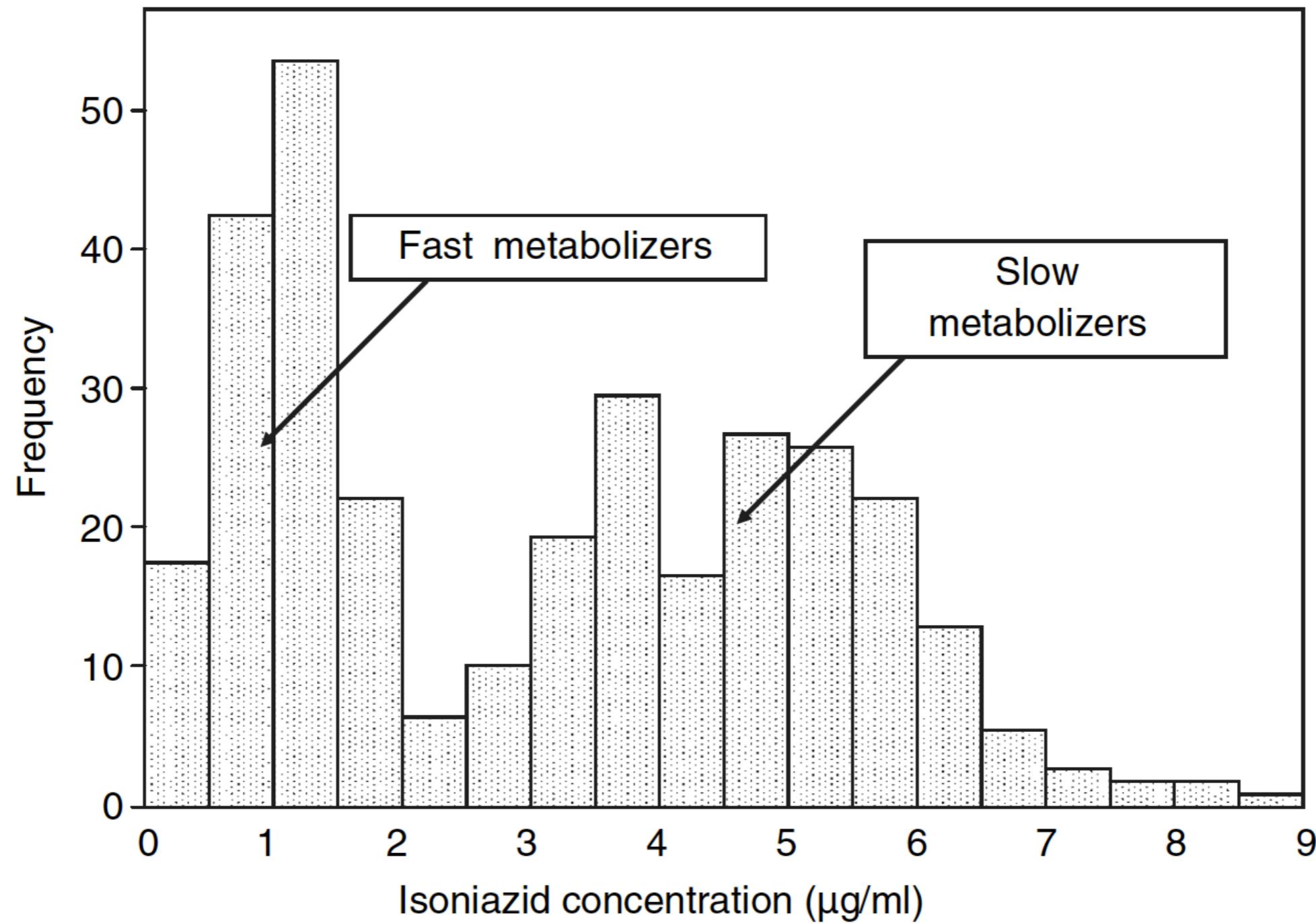


Characteristics of a Normal Distribution

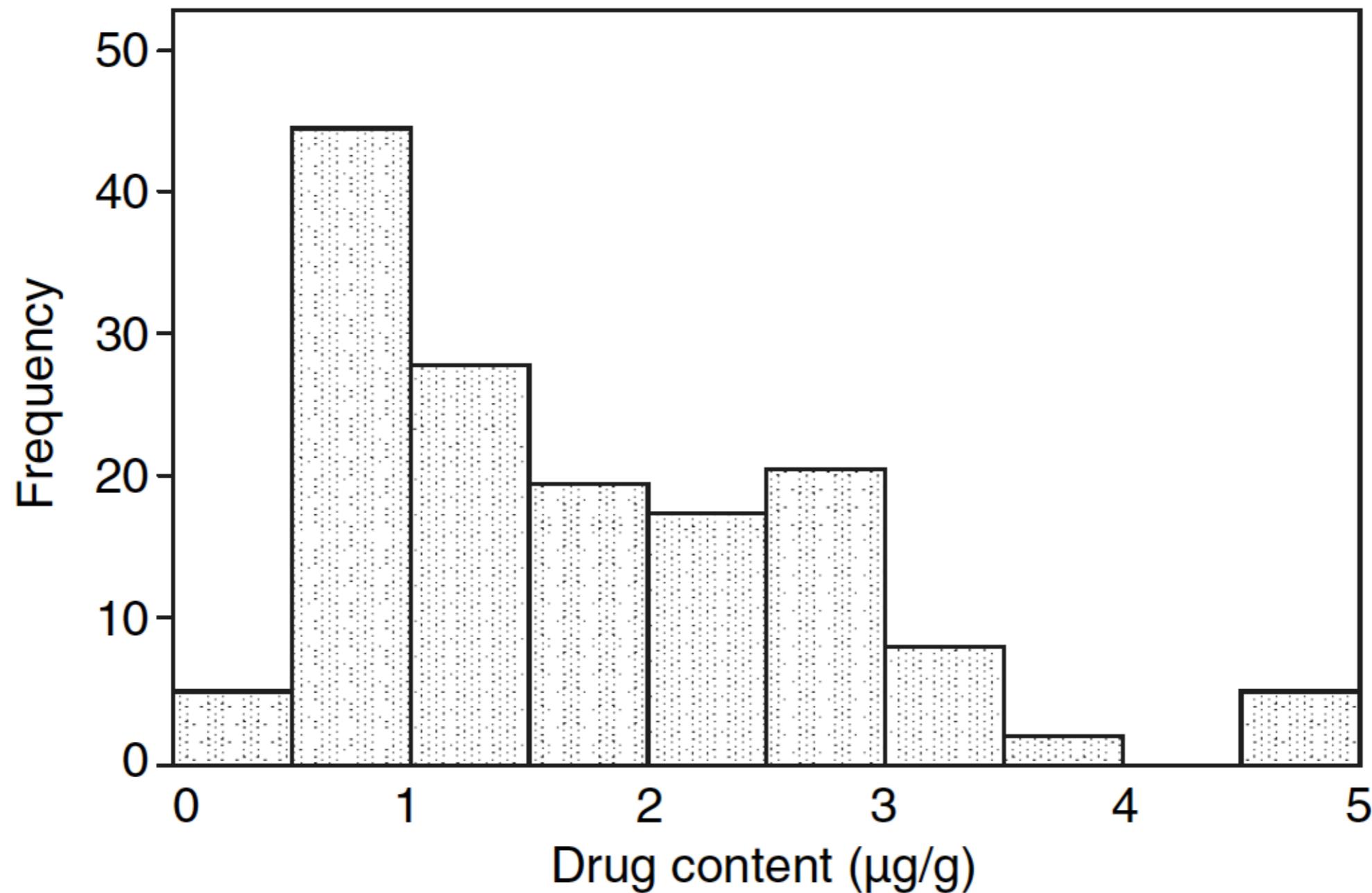
3 visual characteristics that any true normal distribution will possess:

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

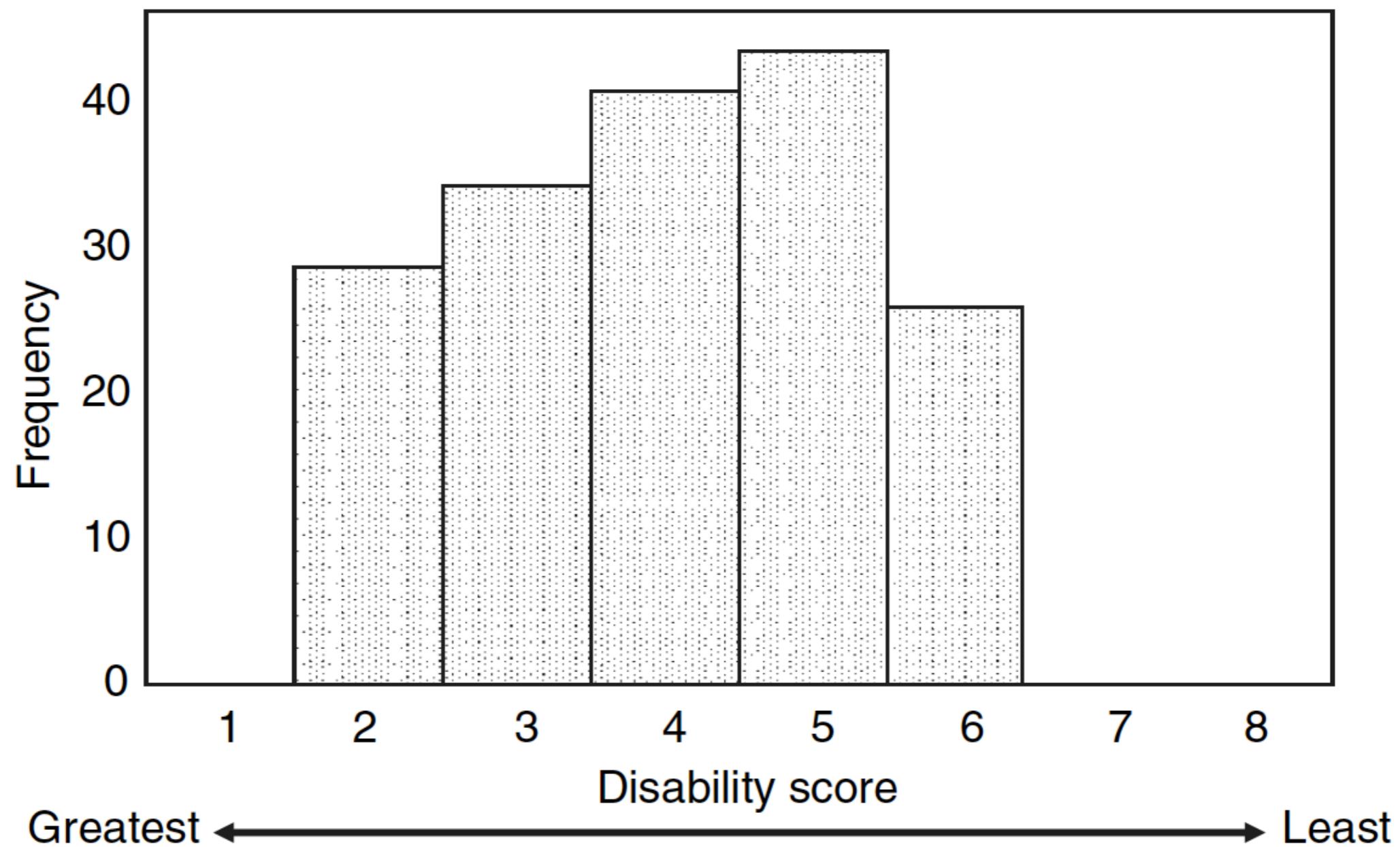
Unimodal vs. Polymodal



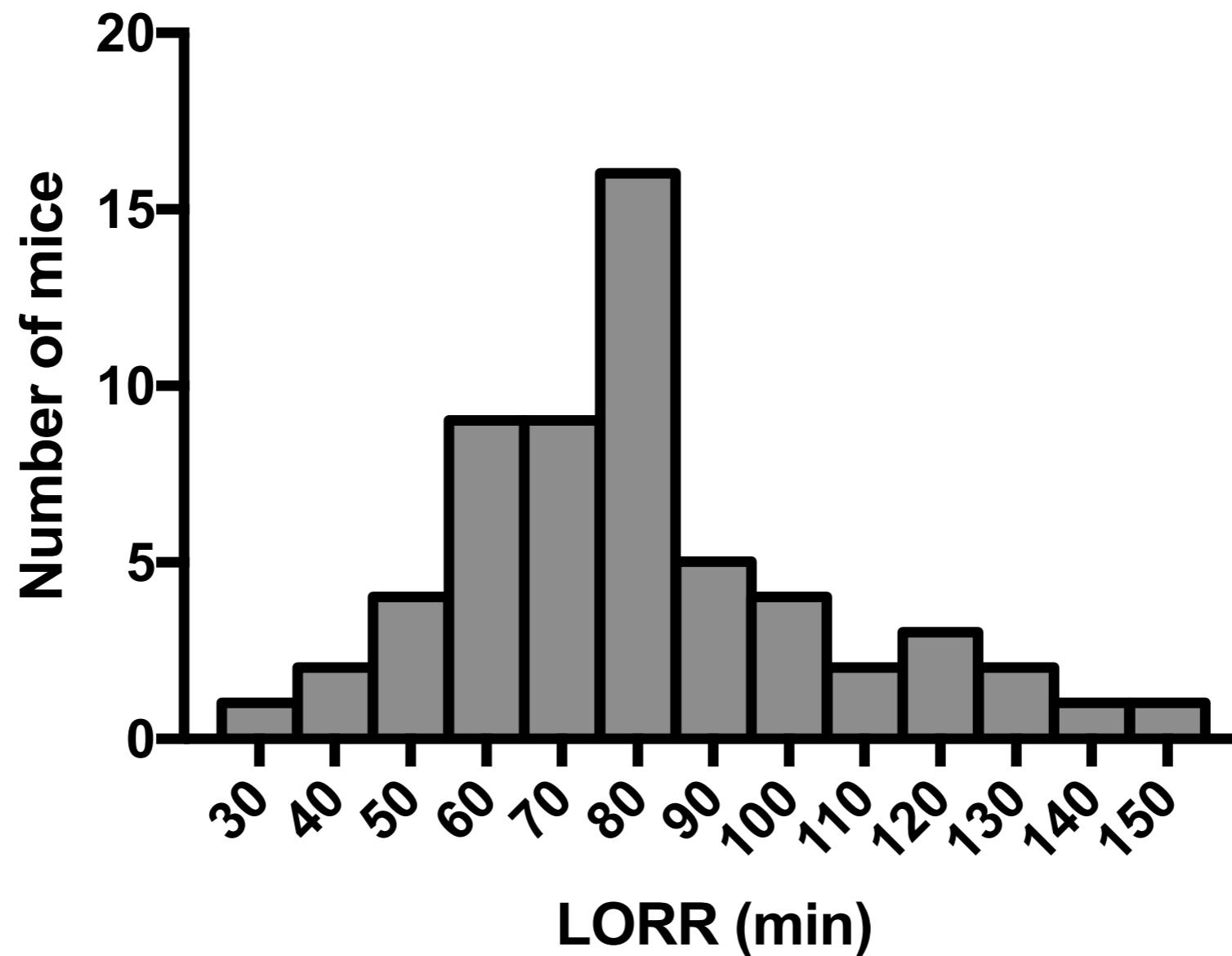
Asymmetrical Distribution



Sharp Cut-Offs



Histogram of Sleep Time

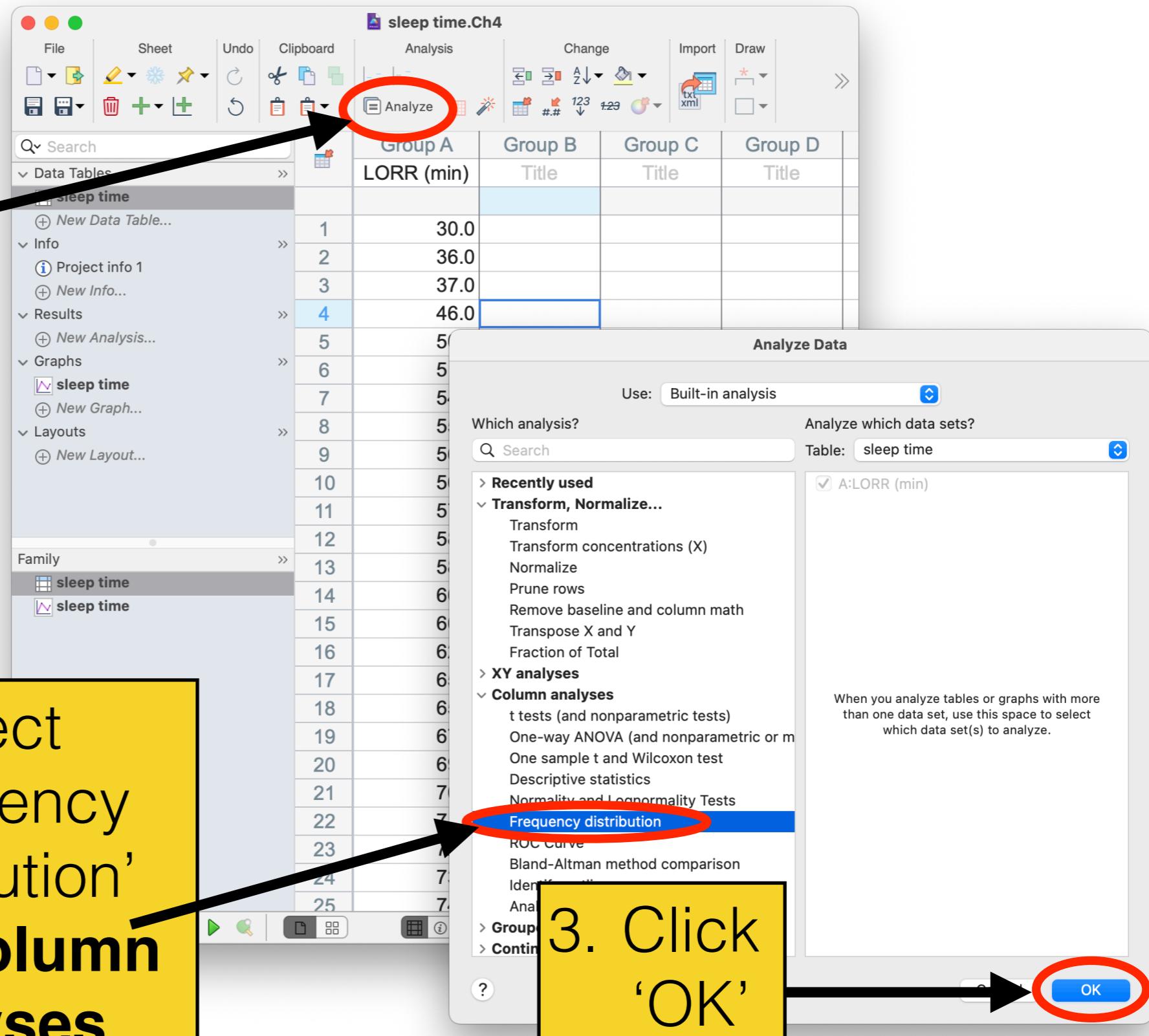


Enter in the LORR data using the 'Column' data table version

	Group A	Group B	Group C	Group D
LORR (min)		Title	Title	Title
1	30.0			
2	36.0			
3	37.0			
4	46.0			
5	50.0			
6	51.0			
7	54.0			
8	55.0			
9	56.0			
10	56.0			
11	57.0			
12	58.0			
13	58.0			
14	60.0			
15	60.0			
16	62.0			
17	65.0			
18	65.0			
19	67.0			
20	69.0			
21	70.0			
22	71.0			
23	72.0			
24	73.0			
25	74.5			

Create a histogram

1. Click the 'Analyze' button



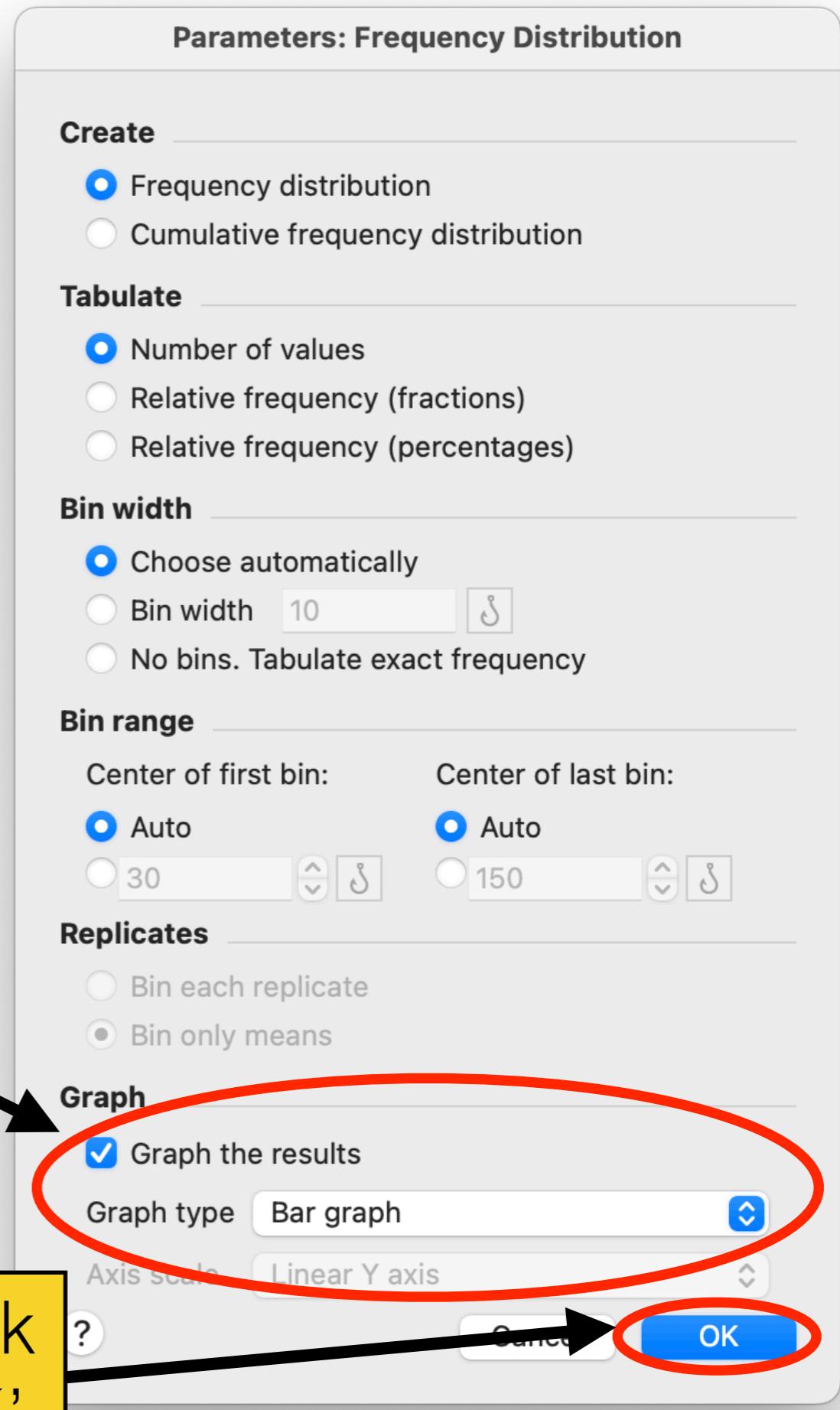
2. Select 'Frequency distribution' from **Column analyses**

3. Click 'OK'

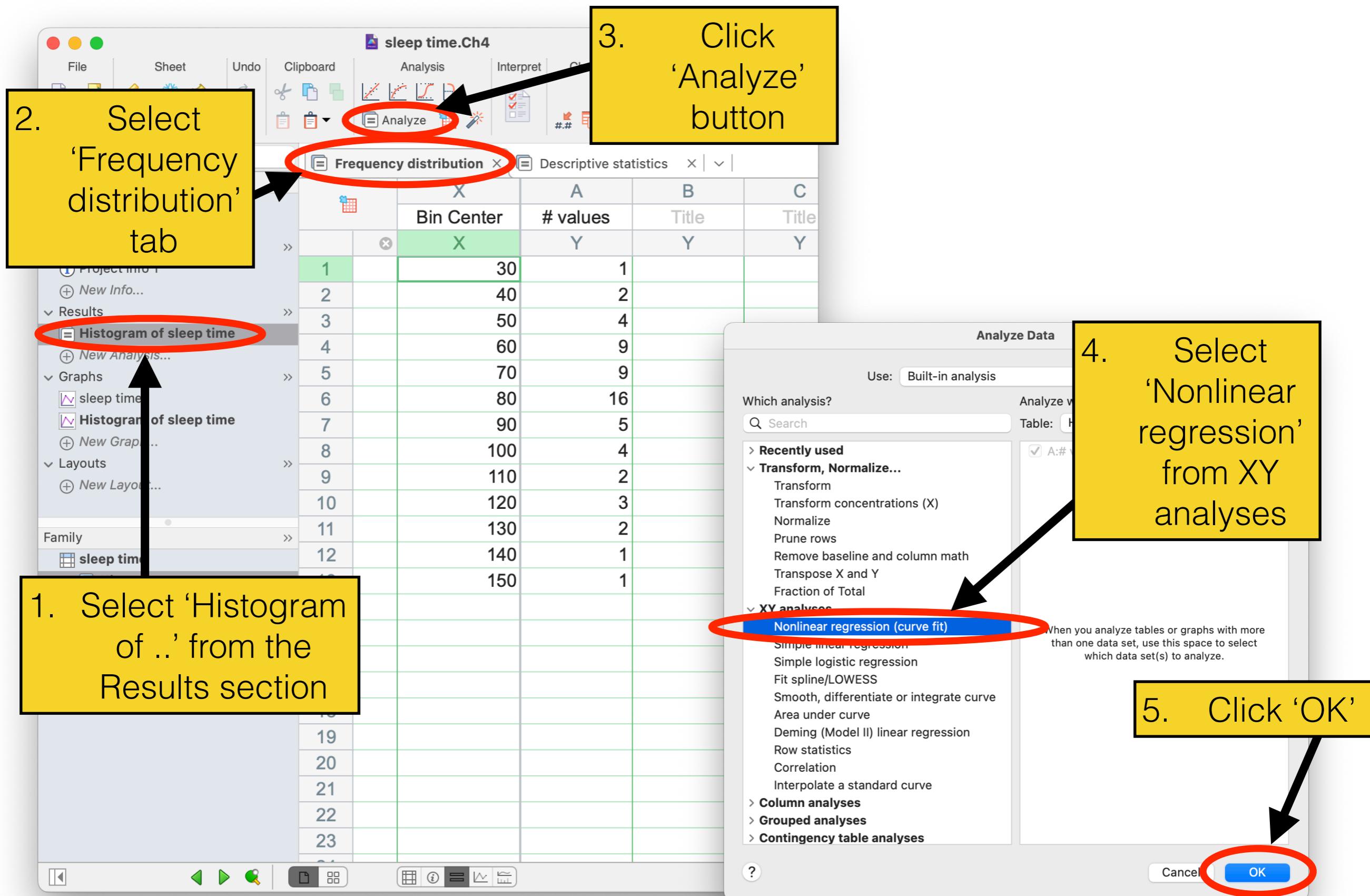
Create a histogram

1. Select 'Graph the results' and pick 'Bar graph' as the Graph type

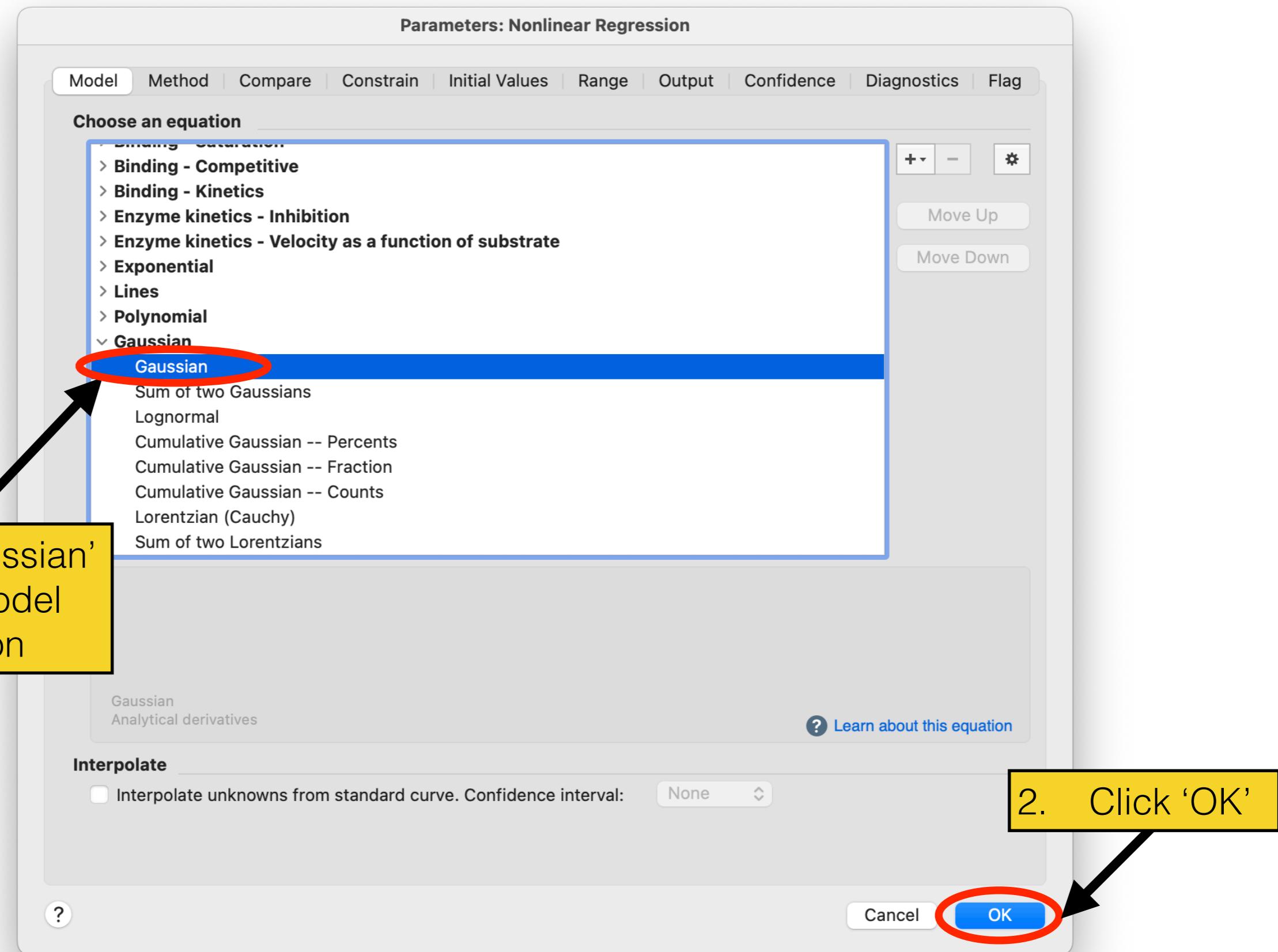
2. Click 'OK'



Adding a Normal Distribution Line - GraphPad

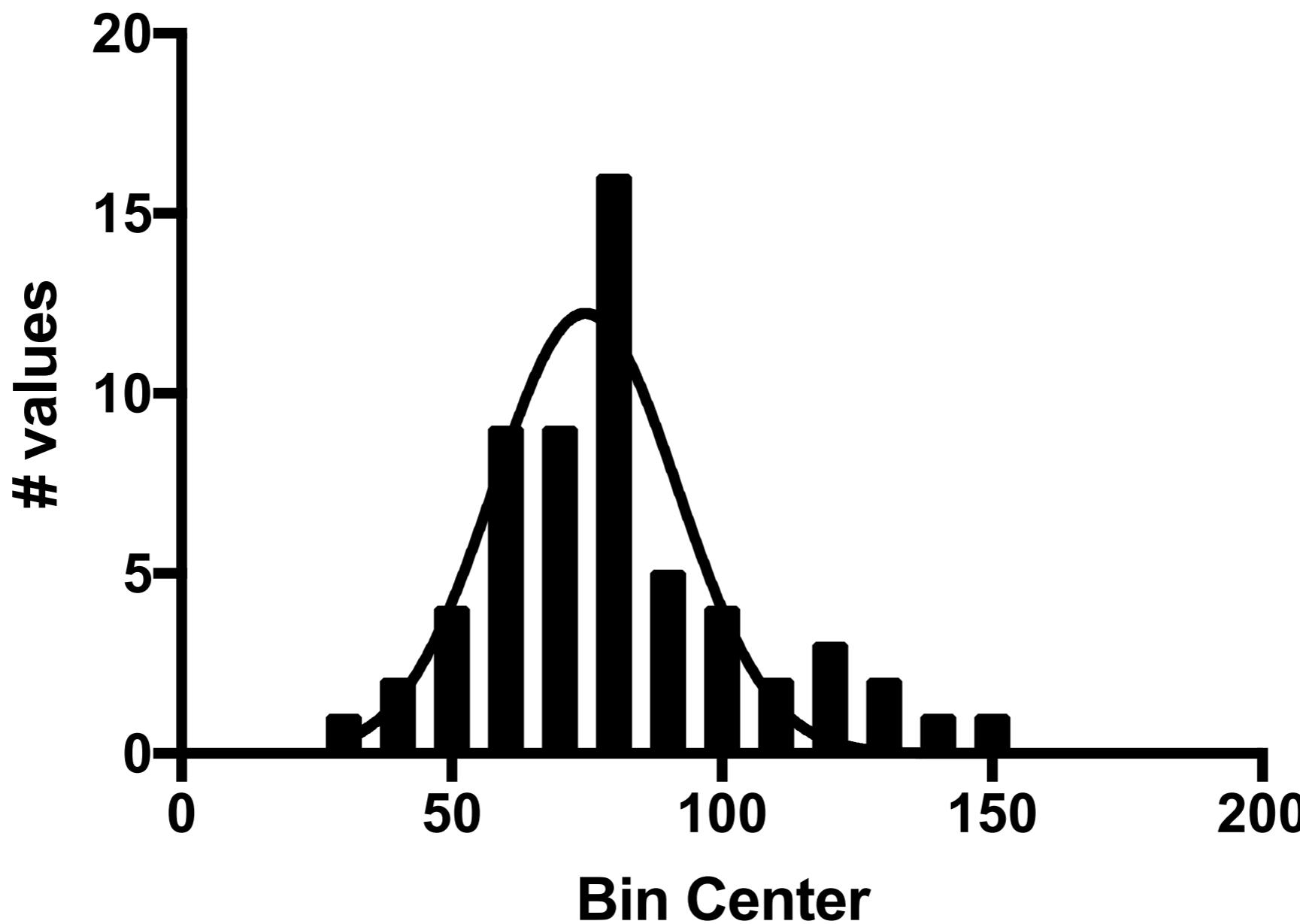


Adding a Normal Distribution Line - GraphPad



Histogram of Sleep Time With Normal Curve

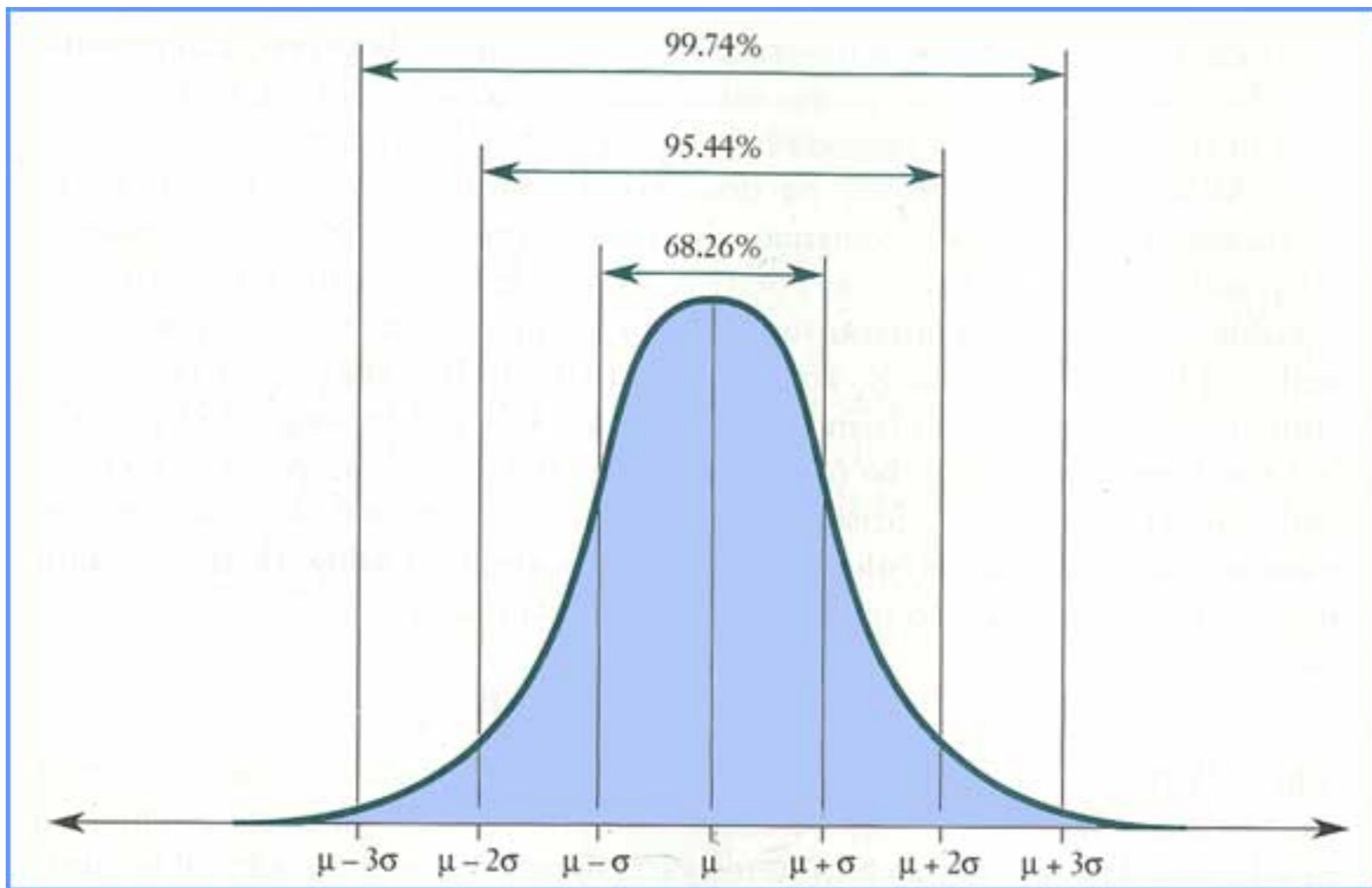
Histogram of sleep time



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

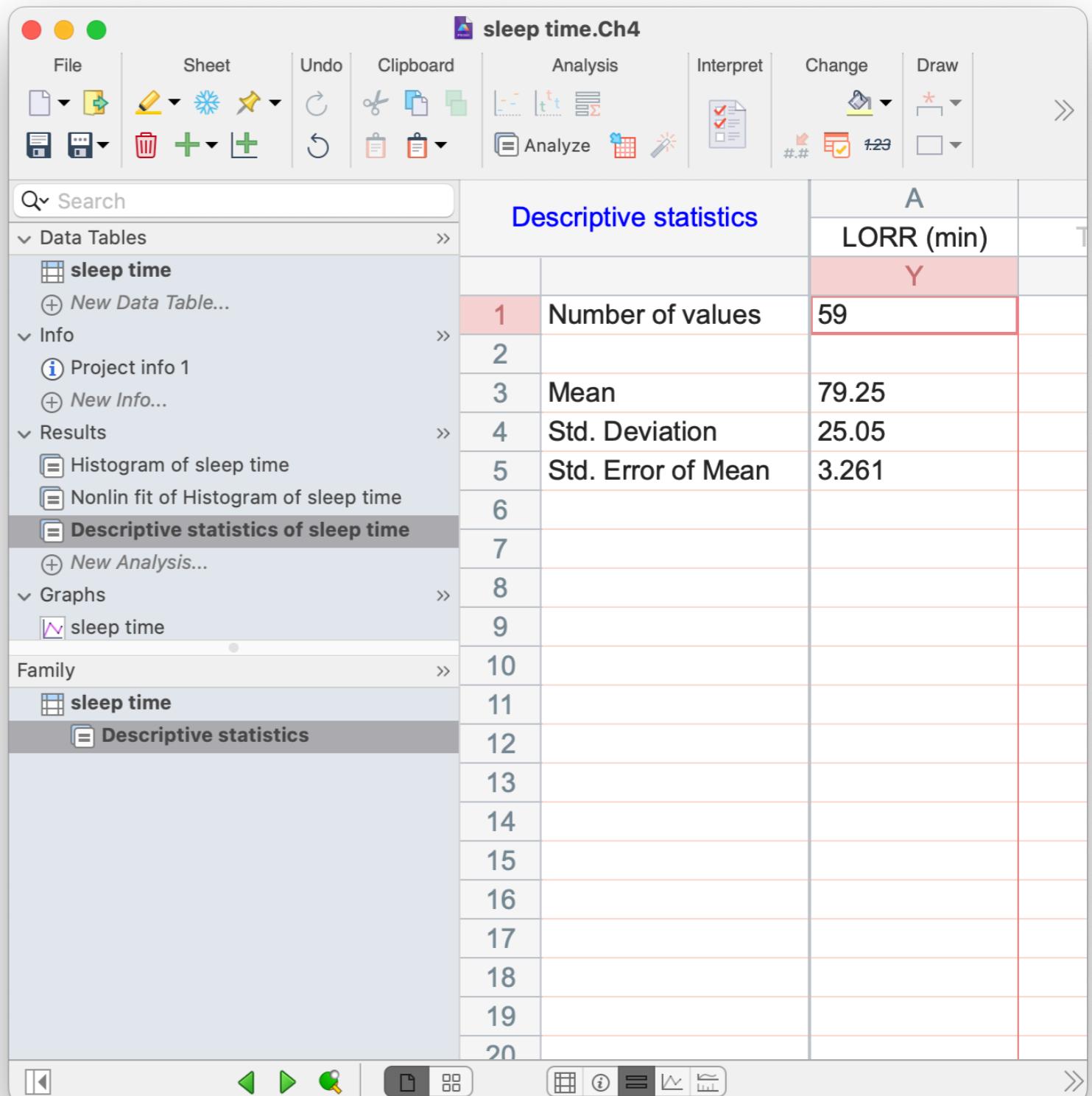
Proportions in Normal Distribution

Ideal Normal Distribution In Terms of SD



Sleep Time Example

- Mean = 79.25
- SD = 25.05
- Mean ‘ \pm ’ 1 SD = (54.2, 104.3)
- Proportion of values within 1 SD = 0.73
- Mean ‘ \pm ’ 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:

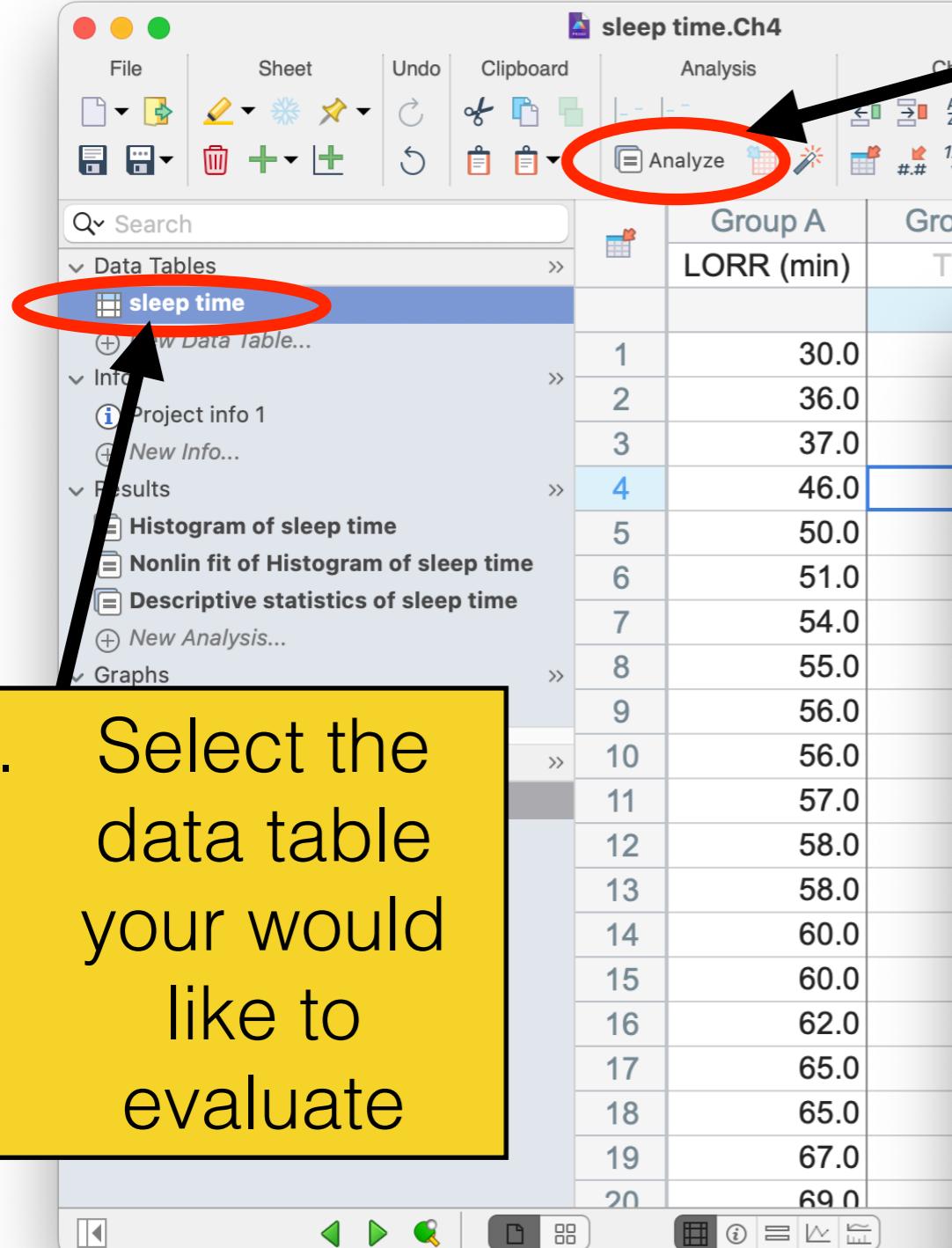
$$Z = \frac{\text{Value} - \text{Mean}}{SD}$$

- 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

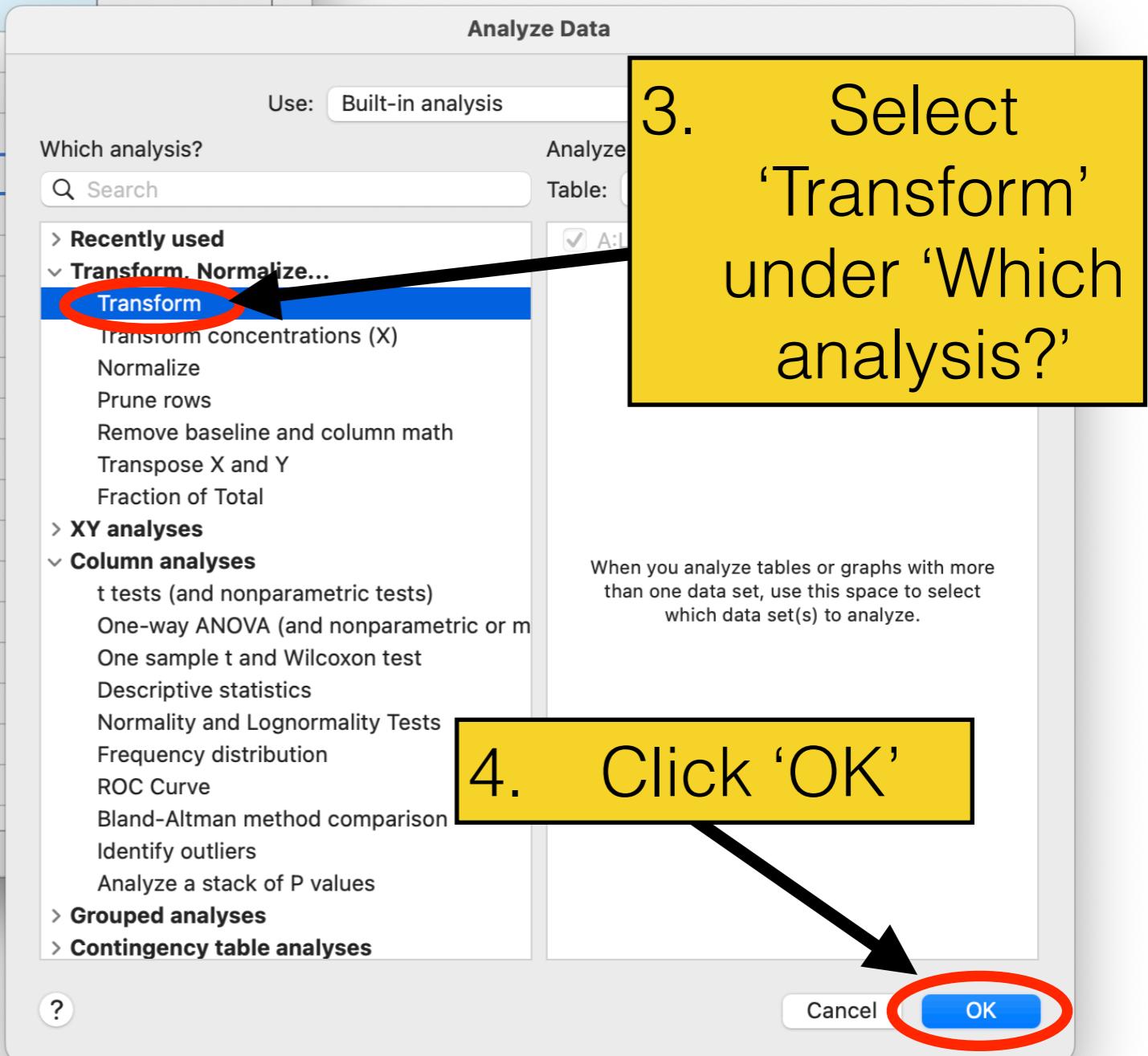
1. Calculate Z scores for each value using 'Transform'
2. Calculate the absolute value of the z scores using 'Transform'
3. Calculate the cumulative frequency based on the absolute values of the z scores

Calculate Z Scores

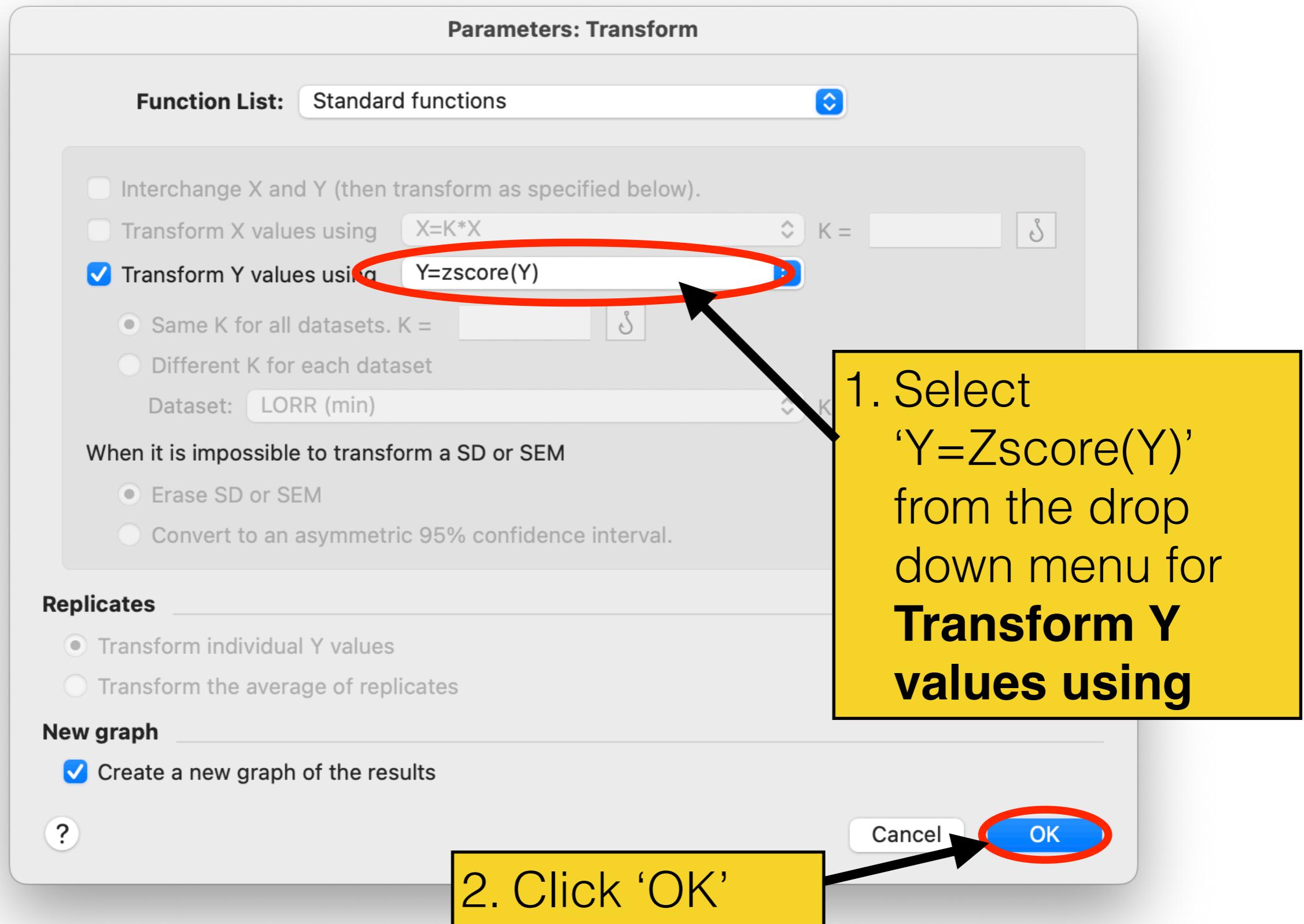


1. Select the data table you would like to evaluate

2. Click on the 'Analyze' Icon



Calculate Z scores



Calculate the absolute value of each Z score

The screenshot shows the QIIME 2 interface with a central data table and three yellow callout boxes containing numbered instructions.

1. Select the Transformed data table under the Results folder

2. Click on the 'Analyze' Icon

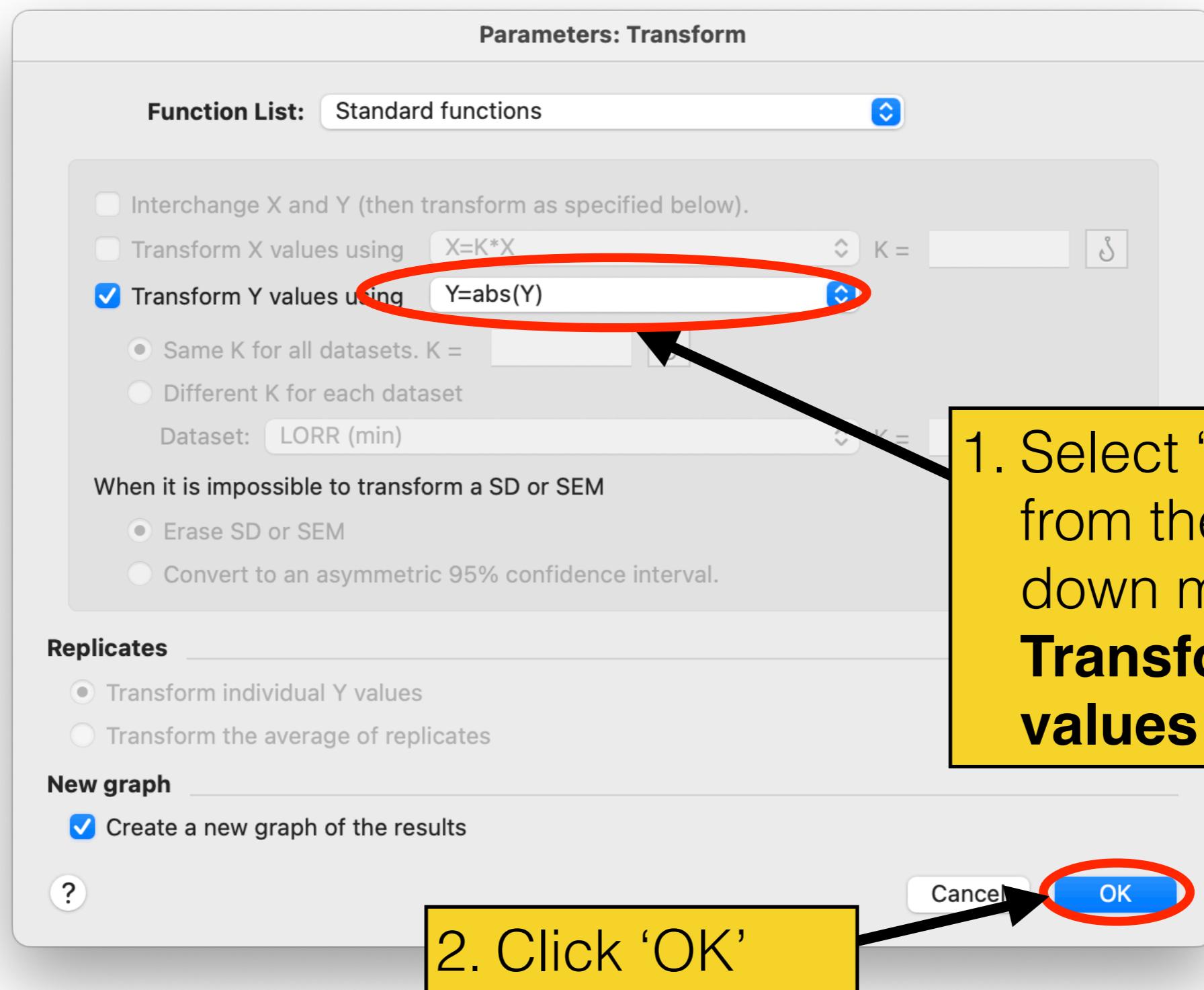
3. Select 'Transform' under 'Which analysis?'

4. Click 'OK'

Data Table Data:

	A	B
1		-1.966
2		-1.727
3		-1.687
4		-1.327
5		-1.168
6		-1.128
7		-1.008
8		-0.968
9		-0.928
10		-0.928
11		-0.888
12		-0.848
13		-0.848
14		-0.768
15		-0.768
16		-0.689
17		-0.569
18		-0.569
19		-0.489
20		-0.409

Calculate the absolute value of each Z score



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

The screenshot shows the JMP software interface with a yellow callout box for each step.

- 1. Select the Transform of the Transform data table under the Results folder**
A red circle highlights the "Transform of Transform of sleep time" item in the "Results" section of the left pane. An arrow points from this item to the first yellow callout box.
- 2. Click on the 'Analyze' icon**
A red circle highlights the "Analyze" icon in the top toolbar. An arrow points from this icon to the second yellow callout box.
- 3. Select 'Frequency distribution' under 'Column analyses'**
The "Analyze Data" dialog is open. A red circle highlights the "Frequency distribution" option under the "Column analyses" section. An arrow points from this option to the third yellow callout box.
- 4. Click 'OK'**
A red circle highlights the "OK" button at the bottom right of the dialog. An arrow points from this button to the fourth yellow callout box.

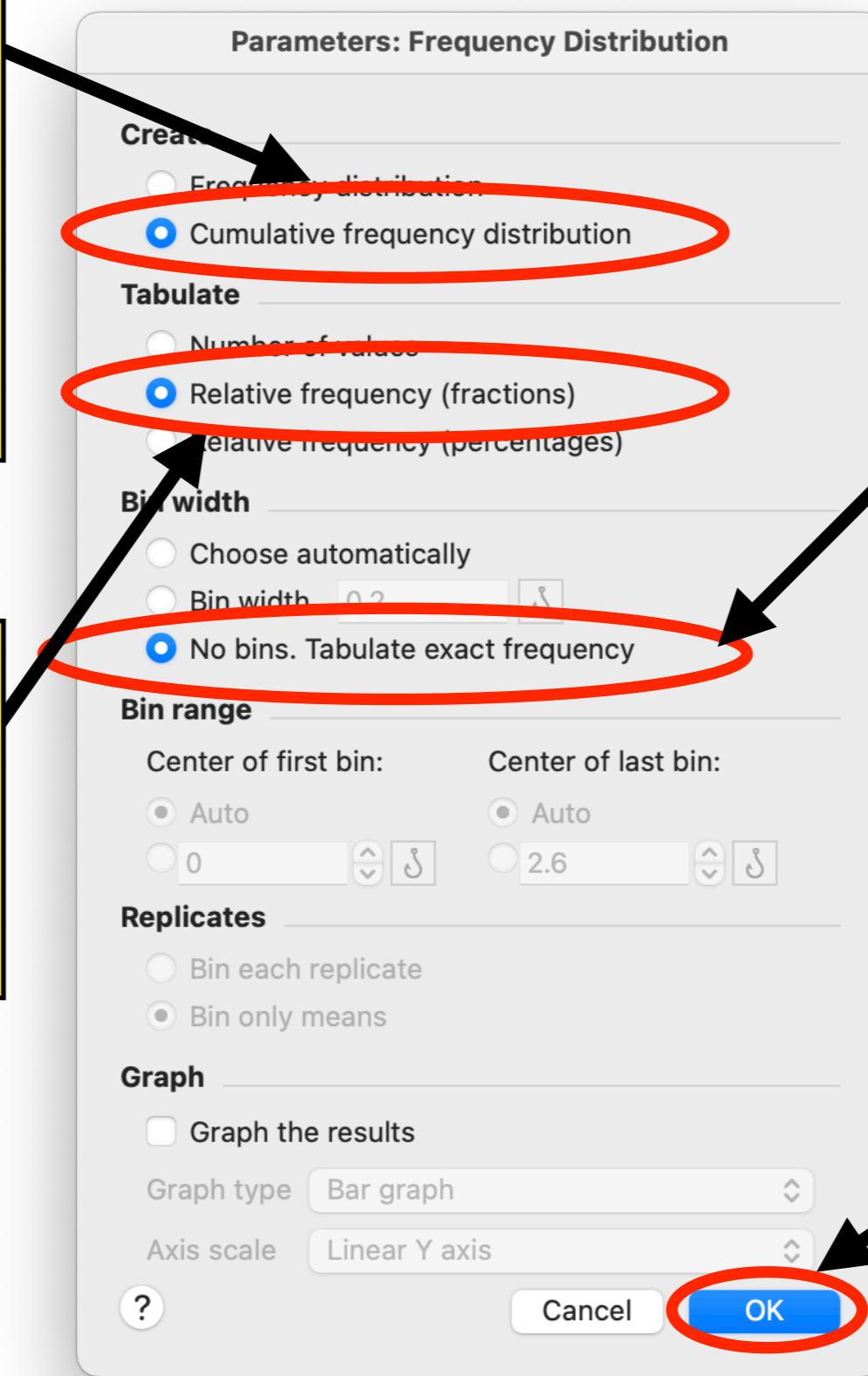
Data Table Preview:

	LORR (min)
1	1.966
2	1.727
3	1.687
4	1.327
5	1.168
6	1.128
7	1.008
8	0.968
9	0.928
10	0.928
11	0.888
12	0.848
13	0.848
14	0.768
15	0.768
16	0.689
17	0.569
18	0.569
19	0.489
20	0.400

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**



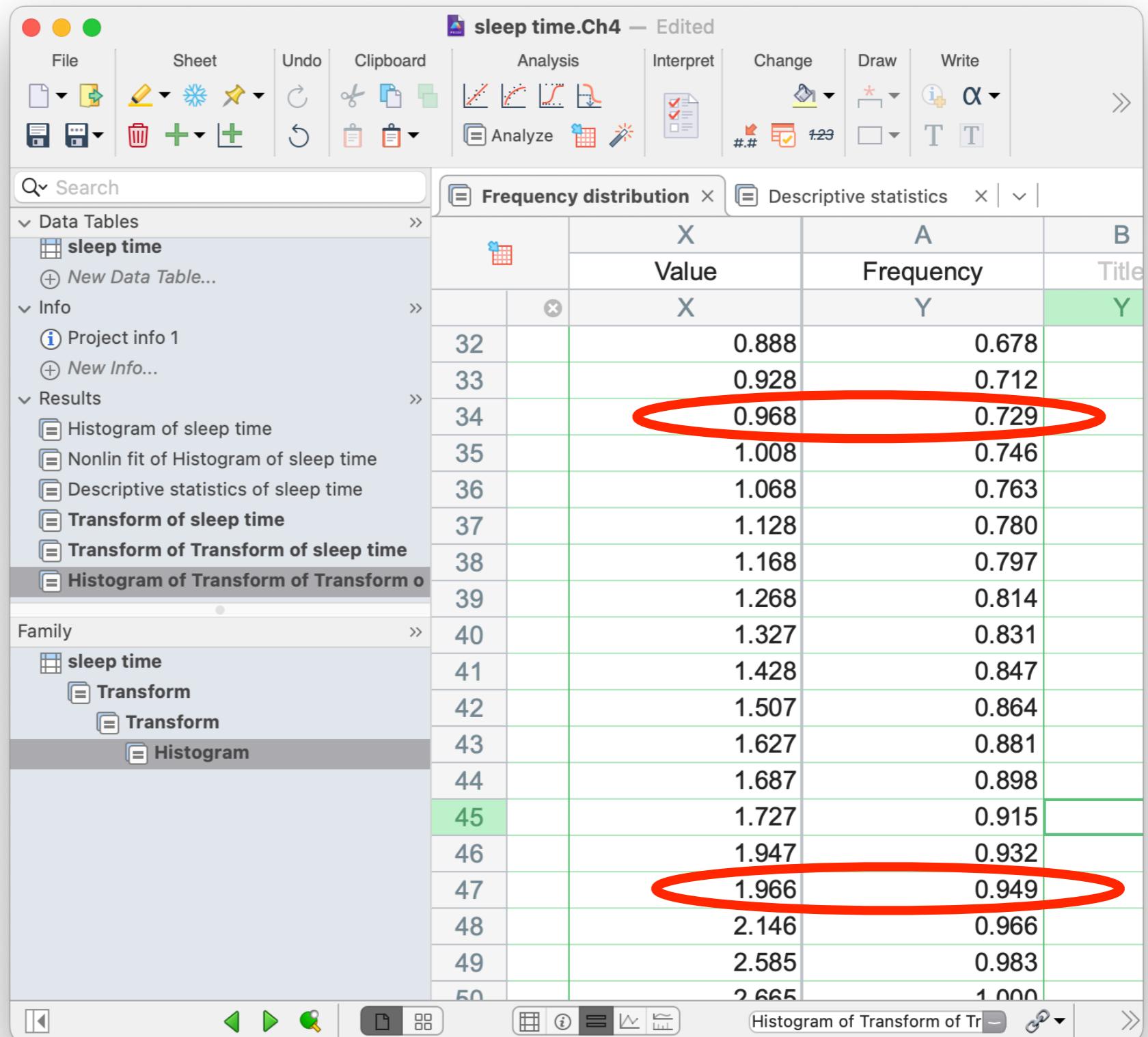
3. Select 'No bins.' under **Bin width**

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

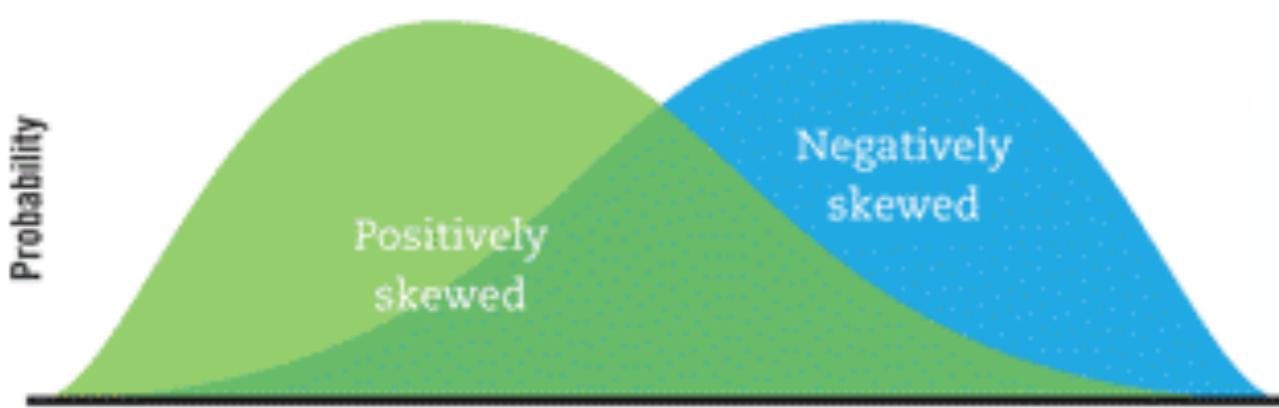


The Normal Distribution Does Not Define Normal Limits

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

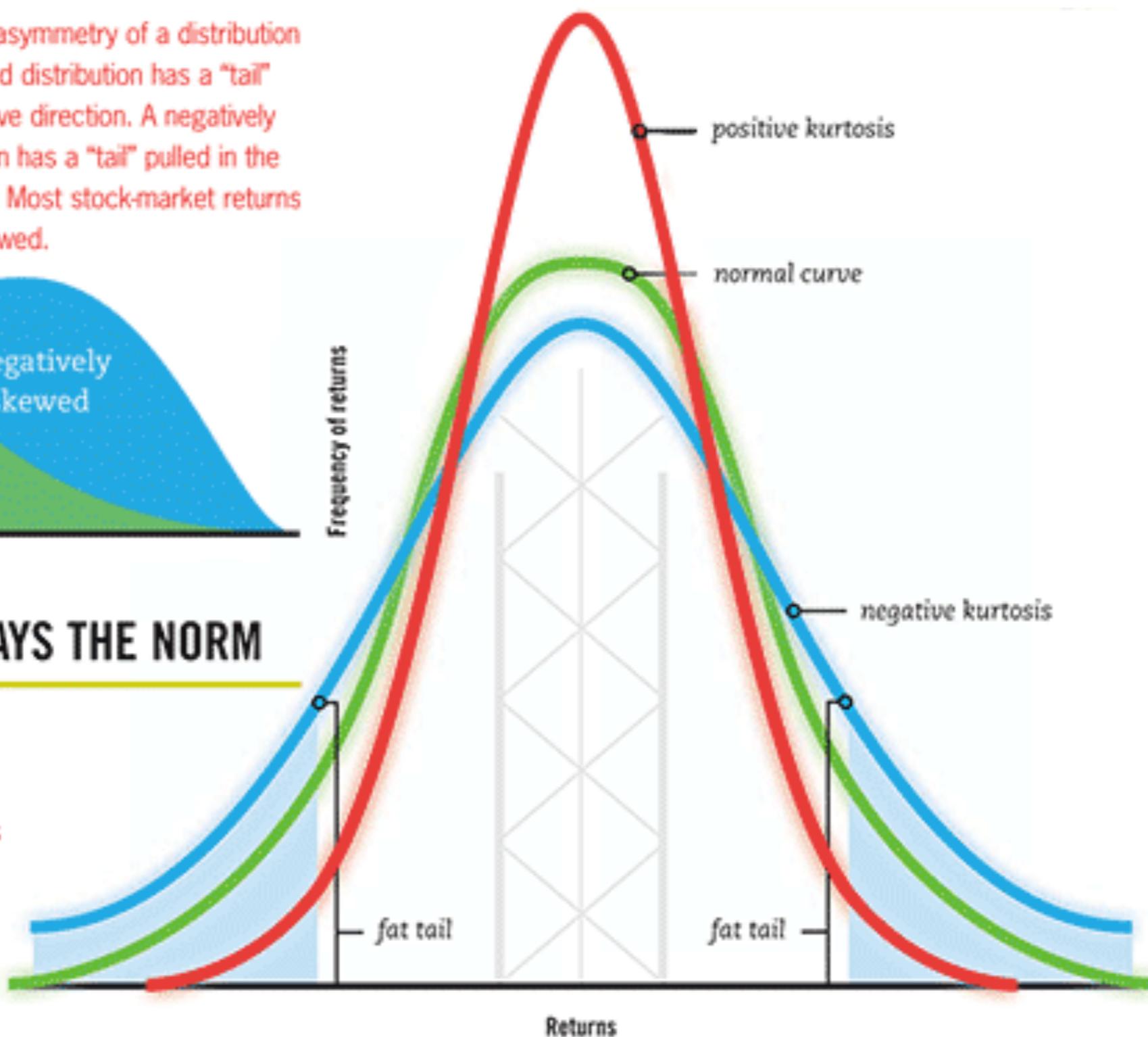
SKEWNESS AND KURTOSIS

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.



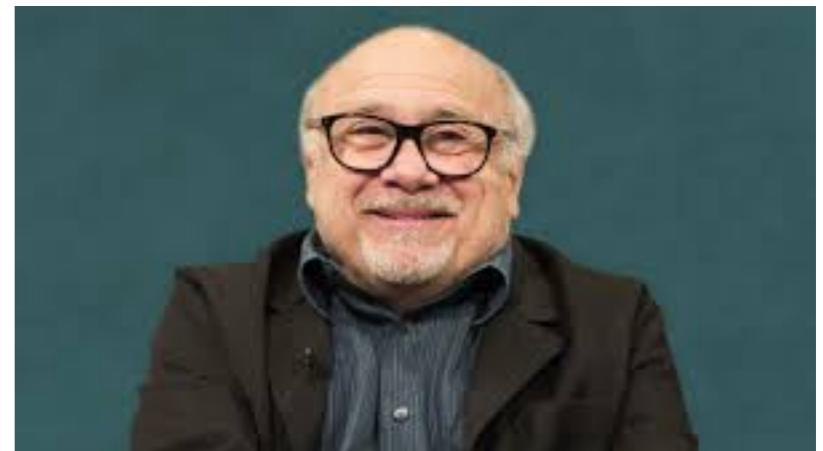
Skewness

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

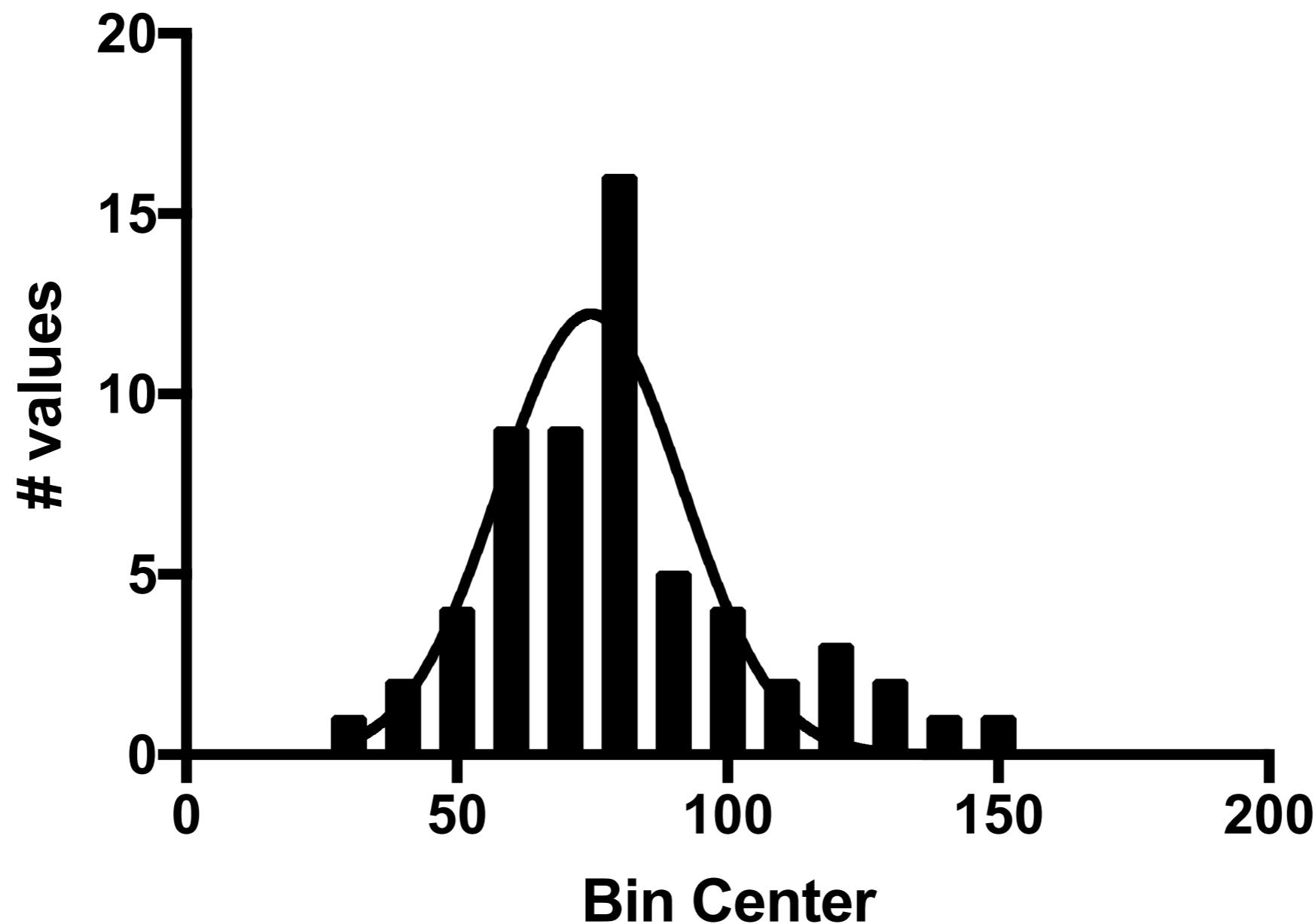


-
- A *positive* kurtosis value indicates that the ‘shoulders’ of the distribution are too low and narrow.



Skewness and Kurtosis in Sleep Time

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



Measuring Skewness and Kurtosis

The screenshot shows a software interface for data analysis. On the left, there's a sidebar with a tree view of data sets and a toolbar with various icons. The main area displays a table titled "sleep time.Ch4" with columns "Group A", "Group B", and "Group C". The "Group A" column contains numerical values from 1 to 20. A red circle highlights the "Analyze" icon in the toolbar, and a black arrow points to it from a yellow box containing the text "1. Click on the 'Analyze' icon". A red box highlights the "sleep time" data set in the sidebar, and a black arrow points to it from another yellow box containing the text "2. Select 'Descriptive statistics'". A red box highlights the "OK" button at the bottom right of the dialog, and a black arrow points to it from a yellow box containing the text "3. Click 'OK'". A red circle also highlights the "Descriptive statistics" option in the list of analyses.

1. Click on the 'Analyze' icon

2. Select 'Descriptive statistics'

3. Click 'OK'

sleep time.Ch4

Analyze Data

Use: Built-in analysis

Which analysis?

Table: sleep time

A:LORR (min)

Recently used

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 or m
- One sample t and Wilcoxon test
- Descriptive statistics**
- Normality and Lognormality Tests
- Frequency distribution
- ROC Curve
- Bland-Altman method comparison
- Identify outliers
- Analyze a stack of P values

Grouped analyses

Contingency table analyses

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

Cancel OK

	Group A	Group B	Group C
1	30.0	Title	Title
2	36.0		
3	37.0		
4	46.0		
5	50.0		
6	51.0		
7	54.0		
8	55.0		
9	56.0		
10	56.0		
11	57.0		
12	58.0		
13	58.0		
14	60.0		
15	60.0		
16	62.0		
17	65.0		
18	65.0		
19	67.0		
20	69.0		

File Sheet Undo Clipboard Analysis Change Import

Search

Data Tables sleep time

Descriptive statistics of sleep time

Transform of sleep time

Transform of Transform of sleep time

Histogram of Transform of Transform o

Histogram

sleep time

Histogram of sleep time

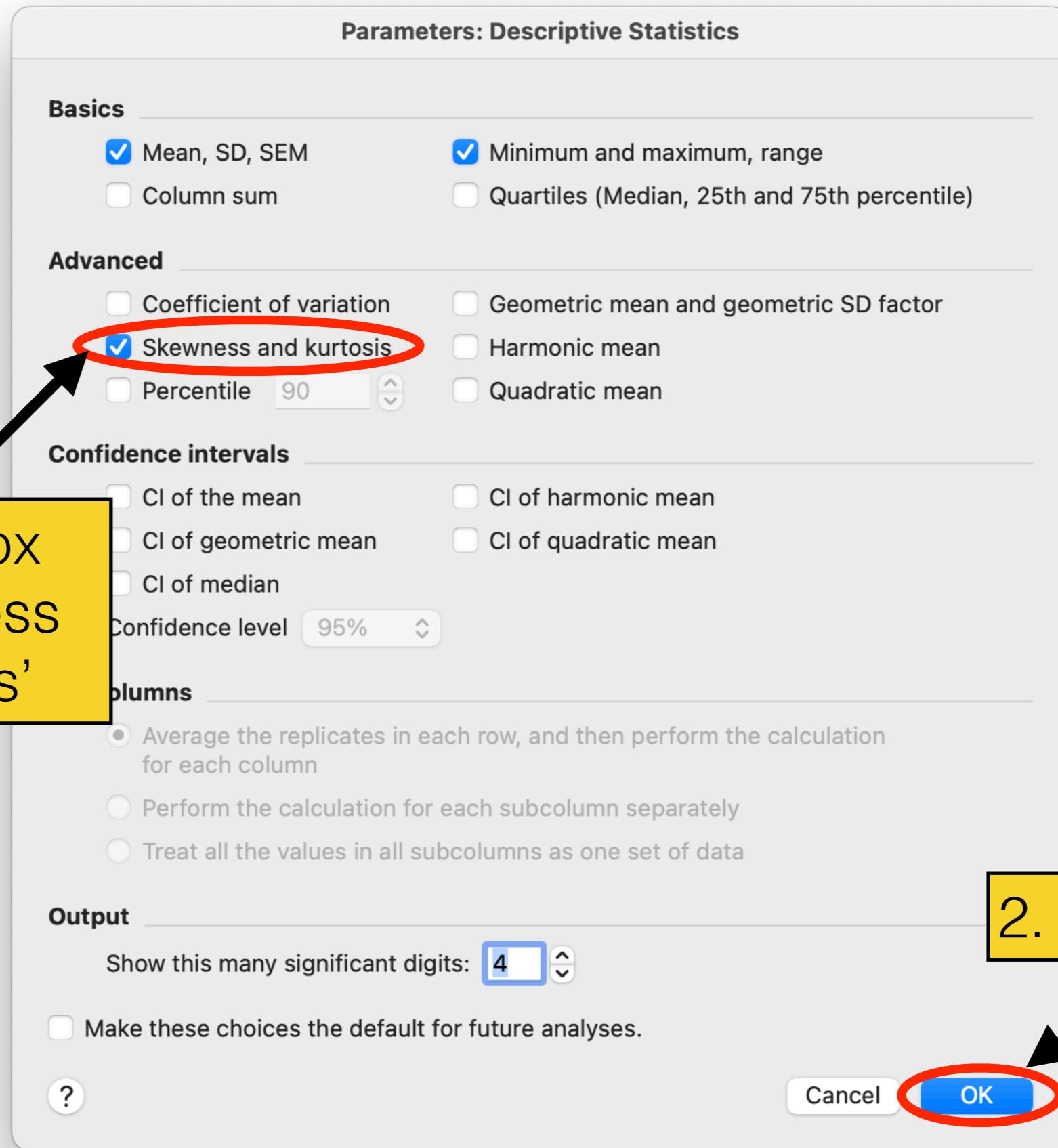
Histogram of sleep time

Transform of sleep time

Transform of Transform of sleep time

Family

Measuring Skewness and Kurtosis



sleep time.Ch4

File Sheet Undo Clipboard Analysis Interpret Change Draw

Search

Data Tables New Data Table... Info Project info 1 New Info... Results Histogram of sleep time Nonlin fit of Histogram of sleep time Descriptive statistics of sleep time Transform of sleep time Transform of Transform of sleep time Histogram of Transform of Transform of sl Descriptive statistics of sleep time

Family sleep time Descriptive statistics

Descriptive statistics

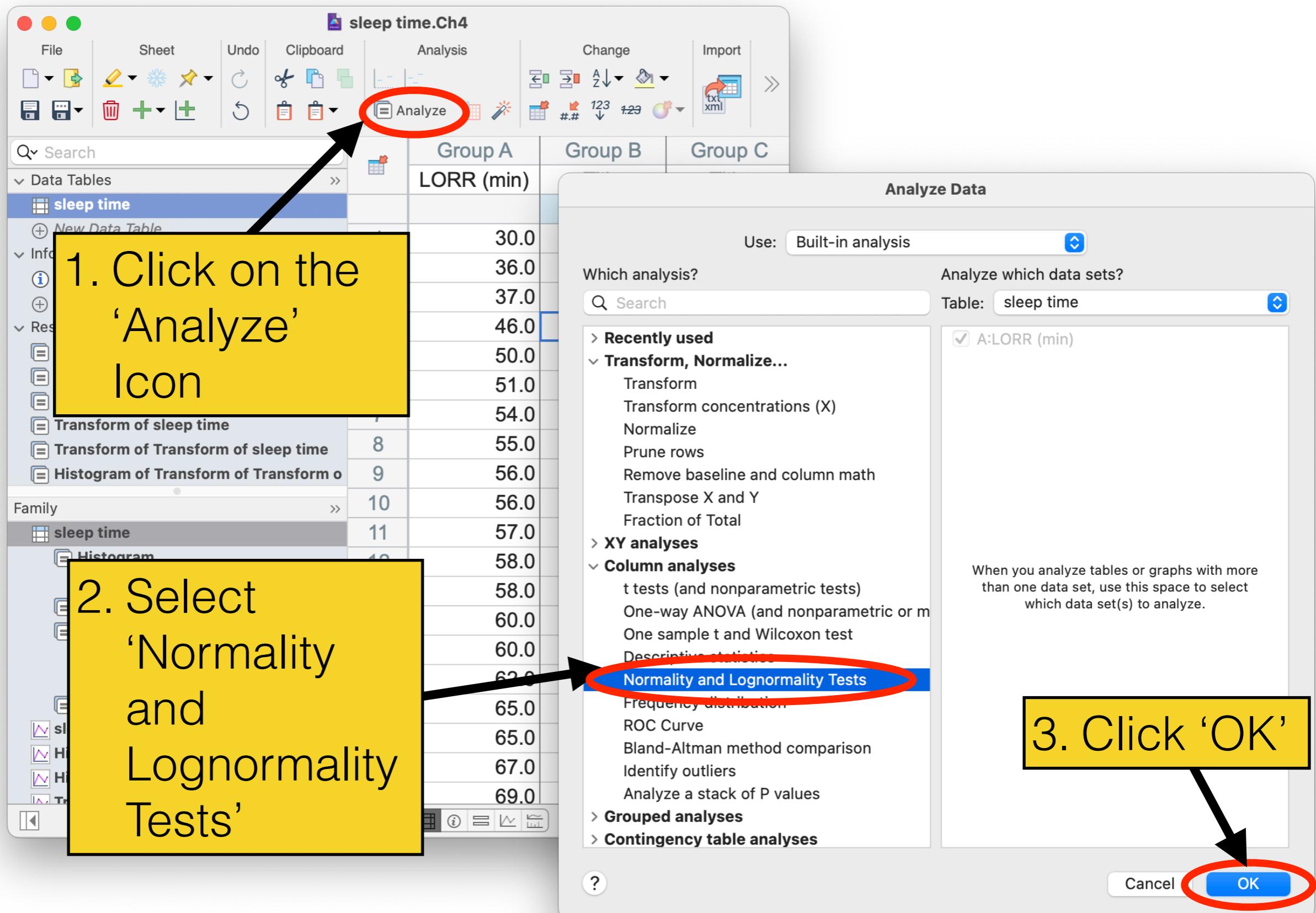
	A	LORR (min)
1	Number of values	59
2		
3	Minimum	30.00
4	Maximum	146.0
5	Range	116.0
6		
7	Mean	79.25
8	Std. Deviation	25.05
9	Std. Error of Mean	3.261
10		
11	Skewness	0.6964
12	Kurtosis	0.6086
13		
14		
15		
16		
17		
18		
19		
20		

TESTS FOR NORMAL DISTRIBUTIONS

D'Agostino-Pearson Omnibus Normality Test

- 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

Normality Test



Normality Test

1. Select
'Normal
(Gaussian)
distribution'

2. Select
'D'Agostino-
Pearson
omnibus
normality
test'



Normality Test

sleep time.Ch4

File Sheet Undo Clipboard Analysis Interpret Change Draw Write

Search

Data Tables Info Results Family

Normality and Lognormality Tests

A	B
LORR (min)	Y
Y	
5.950	
0.0510	
Yes	
ns	
59	

1 Test for normal distribution
2 D'Agostino & Pearson test
3 K2
4 P value
5 Passed normality test (alpha=0.05)?
6 P value summary
7
8 Number of values
9
10
11
12
13
14
15

Normality and Lognormality Test

Row

Is the sleep time data
normally distributed?

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 statistical tests of non-normality with caution.