Chapter 17: Describing categorized data and the goodness of fit chi-square test

TXCL7565/PHSC7565

What This Lecture Covers

- Proportions and their confidence intervals
- Comparing observed proportion to some pre-determined figure

PROPORTIONS AND THEIR CONFIDENCE INTERVALS

Nominal outcome data

- Nominal indicates that the variable is categorical
- For nominal outcome data, subjects are often placed into categories such as ‘success or failure’.

Summary measure of nominal data

- Most often, nominal outcomes are summarized by the proportion of subjects in each category.
  - proportion - a fraction, e.g., 20 out of 100 = 0.20
  - percentage - percent of subjects, e.g., 20 out of 100 = 20%
- When there are only 2 categories (dichotomization), we usually refer to one proportion as p and the other as q.

Precision of sample estimates of a proportion

The precision with which proportions within the population can be estimated from samples, depends on:

- sample size
- presence of rare categories

Sample Size Effect on Precision

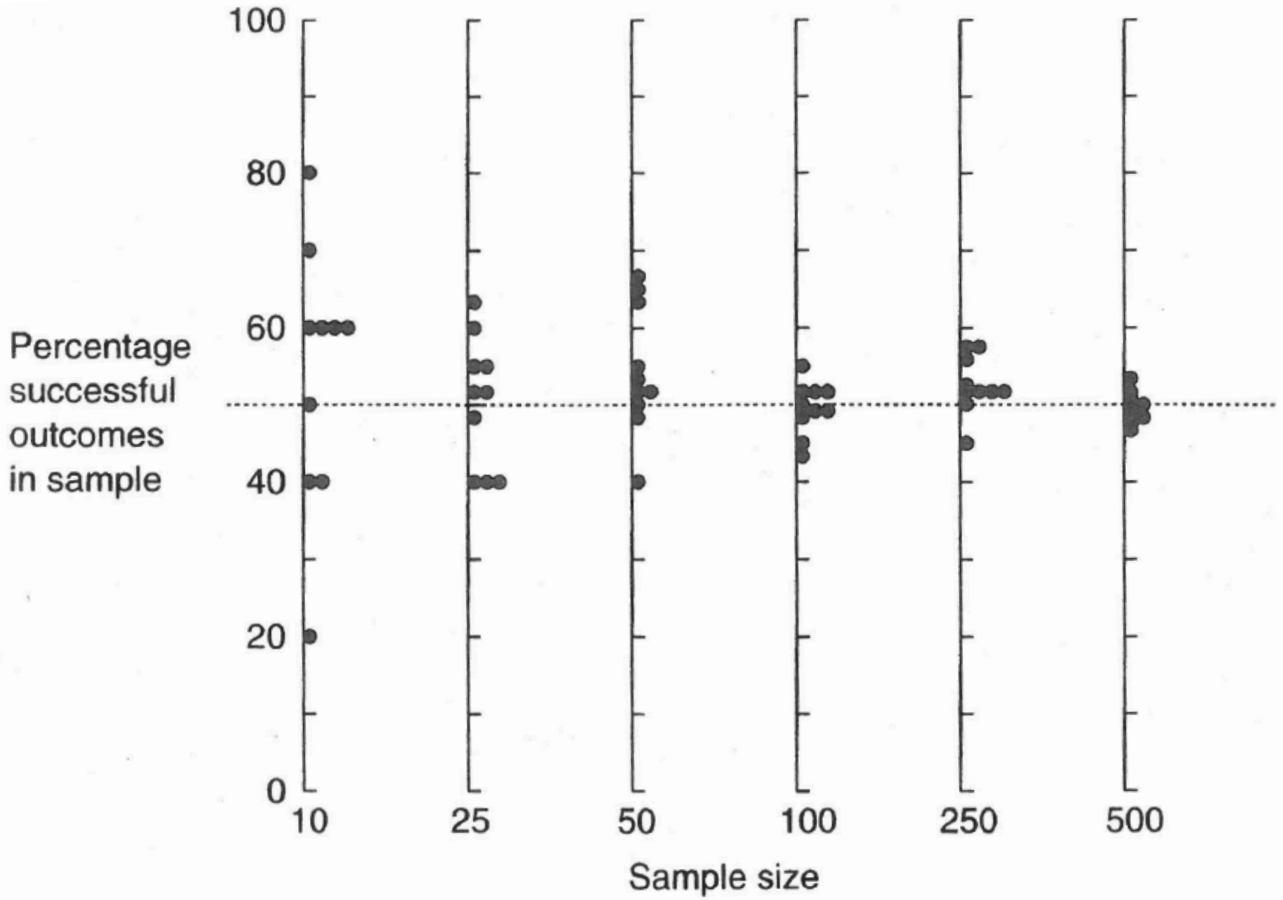
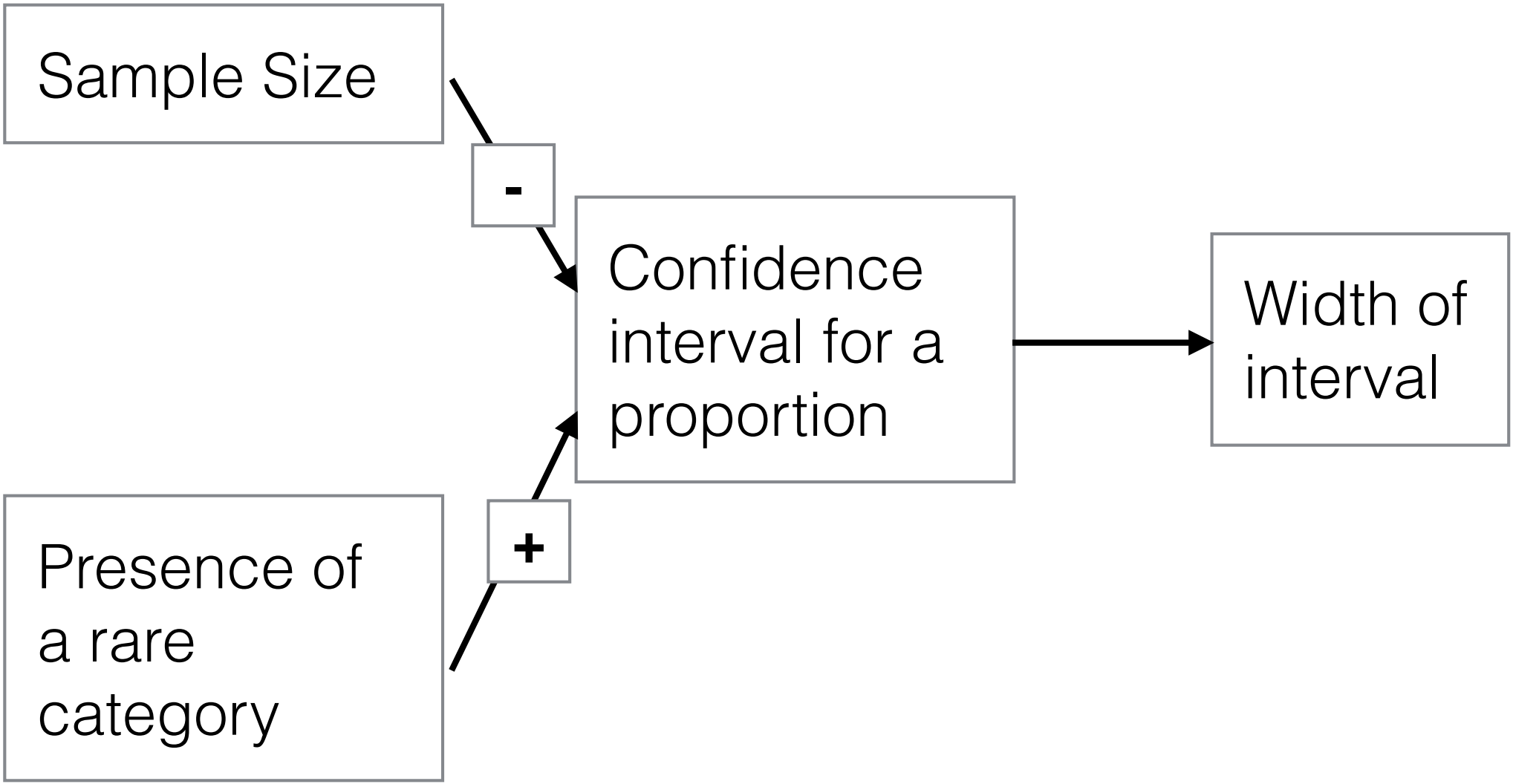


Figure 17.1 Ten sample estimates of the proportion of successful outcomes – Various sample sizes. (True proportion = 50%)

Rare category effect on precision

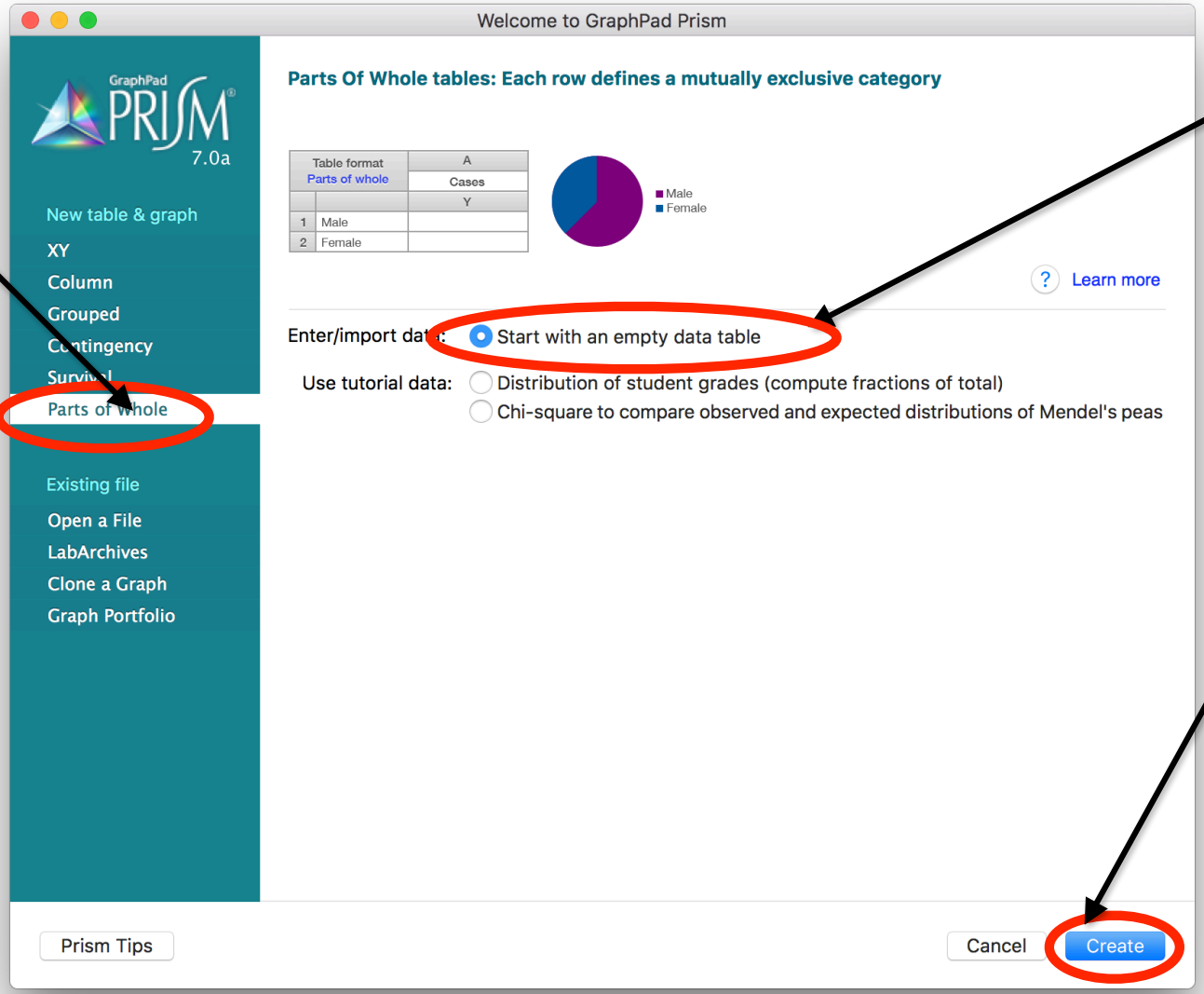
	Without ADR	With ADR	Percent with ADR
Scenario 1A	949	51	5.1%
Scenario 1B	960	40	4.0%
Scenario 1C	940	60	6.0%
Scenario 2A	996	4	0.4%
Scenario 2B	998	2	0.2%
Scenario 2C	994	6	0.6%

95% confidence interval for a proportion



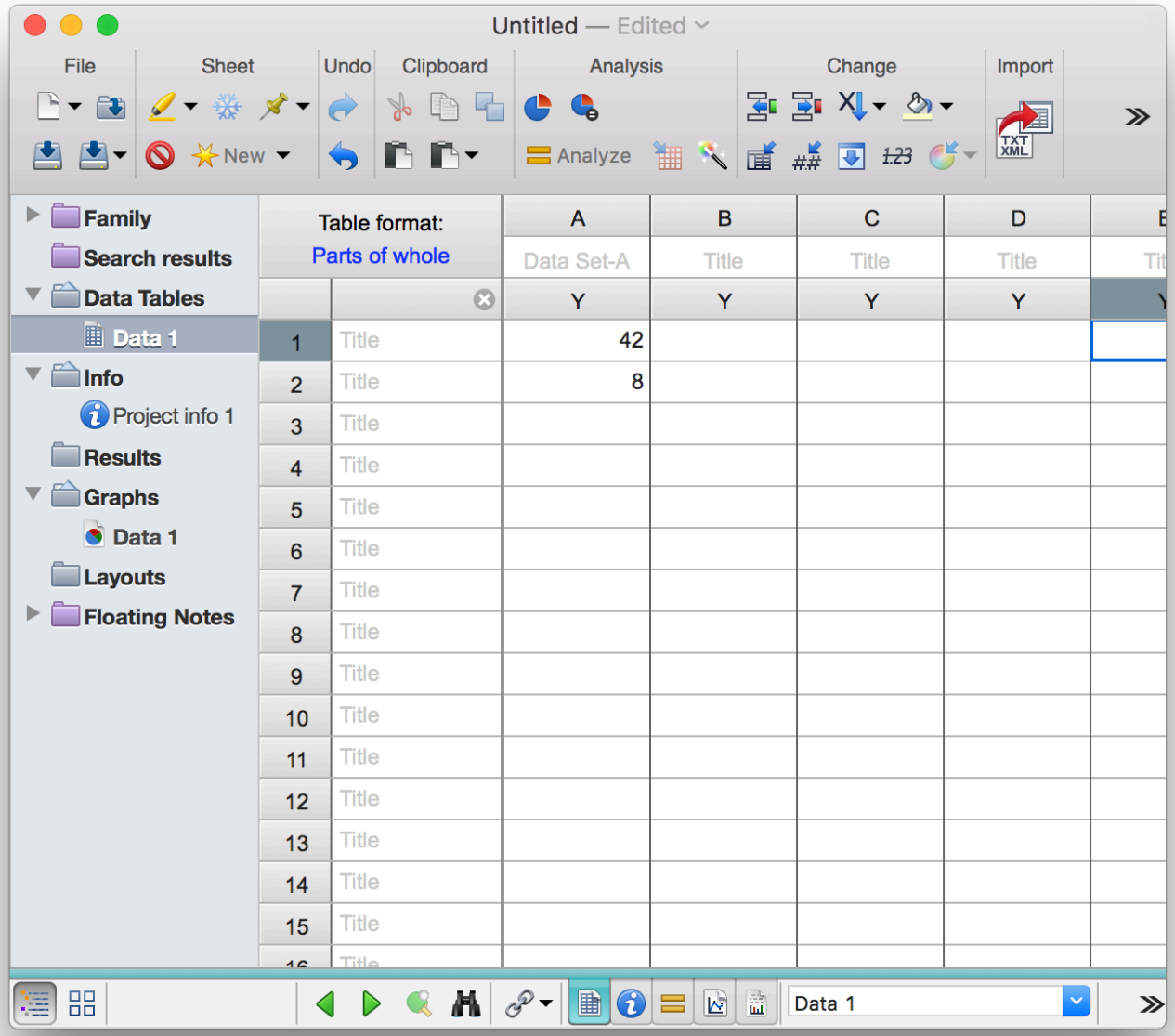
Proportion CI in Prism

1.Create a new table using ‘Parts of Whole’ format

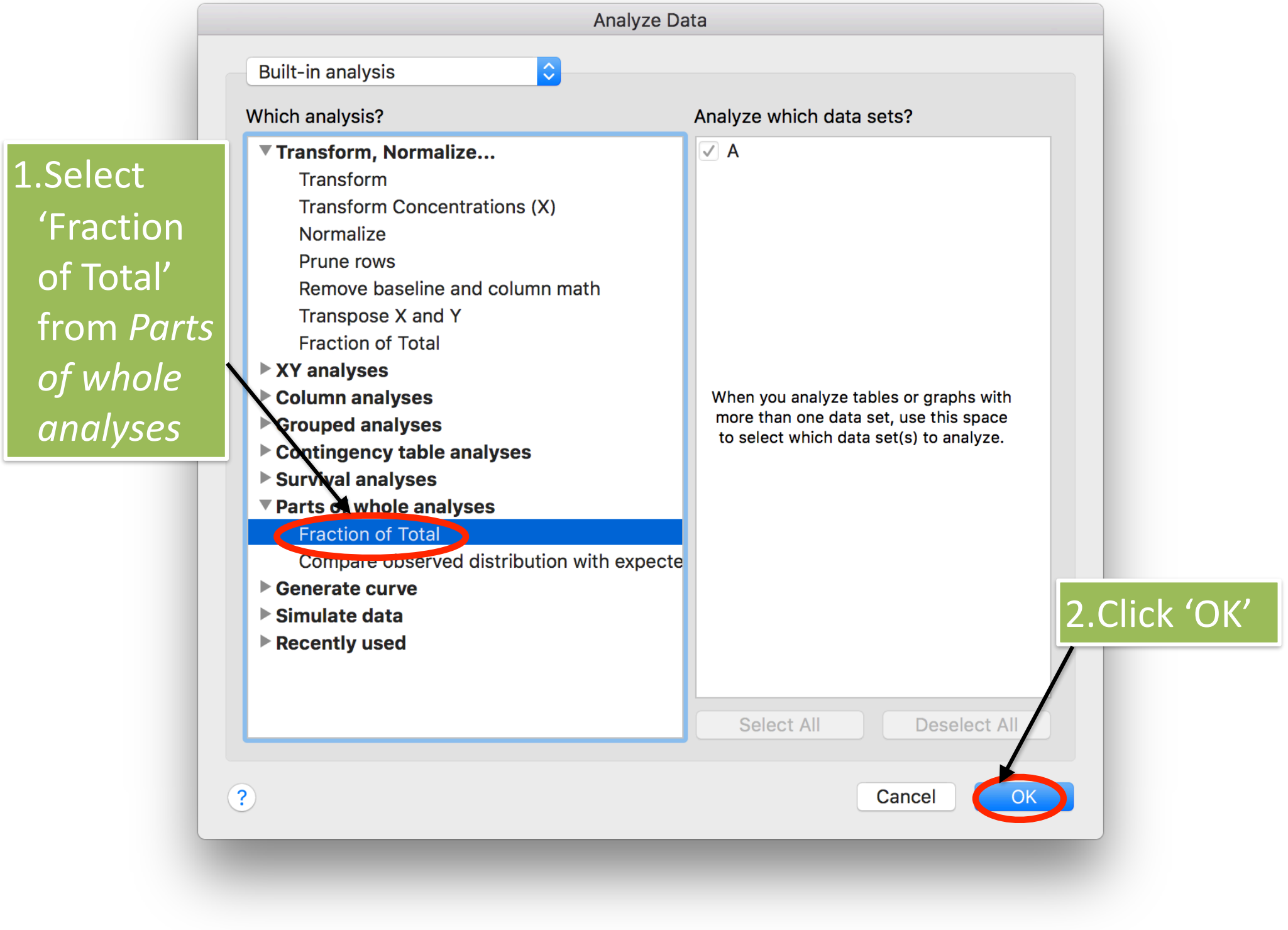


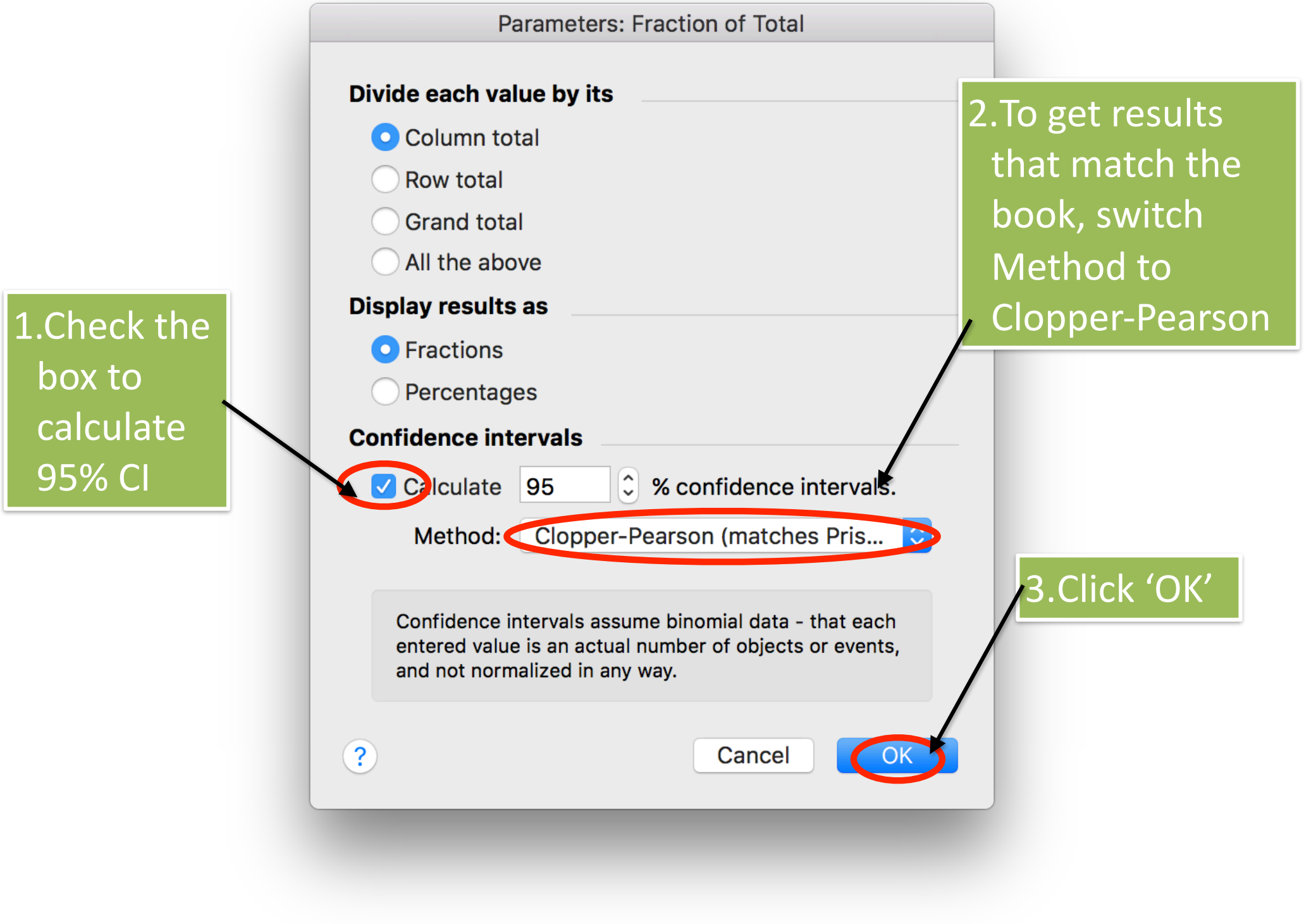
2.Start with an empty table

3.Click ‘Create’



Enter the number of observation in each category in a single column.


1.Select ‘Fraction of Total’ from *Parts of whole analyses*
Analyze Data
Built-in analysis
Which analysis?
Analyze which data sets?
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
Grouped analyses
Contingency table analyses
Survival analyses
Parts of whole analyses
Fraction of Total
Compare observed distribution with expecte
Generate curve
Simulate data
Recently used
A
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
2.Click ‘OK’
Cancel
OK


Parameters: Fraction of Total
Divide each value by its
Column total
Row total
Grand total
All the above
Display results as
Fractions
Percentages
Confidence intervals
Calculate
95
% confidence intervals.
Method:
Clopper-Pearson (matches Pris...
Confidence intervals assume binomial data - that each entered value is an actual number of objects or events, and not normalized in any way.
?
Cancel
OK
1.Check the box to calculate 95% CI
2.To get results that match the book, switch Method to Clopper-Pearson
3.Click 'OK'

Fraction of Total		A		
		Data Set-A		
		Mean	Upper Limit	Lower Limit
1		0.840	0.928	0.709
2		0.160	0.291	0.072
3				

95% C. I. for proportions

Number examined:	50
Number detected (Successful):	42
Point estimate (Successful):	84.0%
95% CI for proportion (Successful):	70.9 - 92.8%

Nominal data is not very efficient

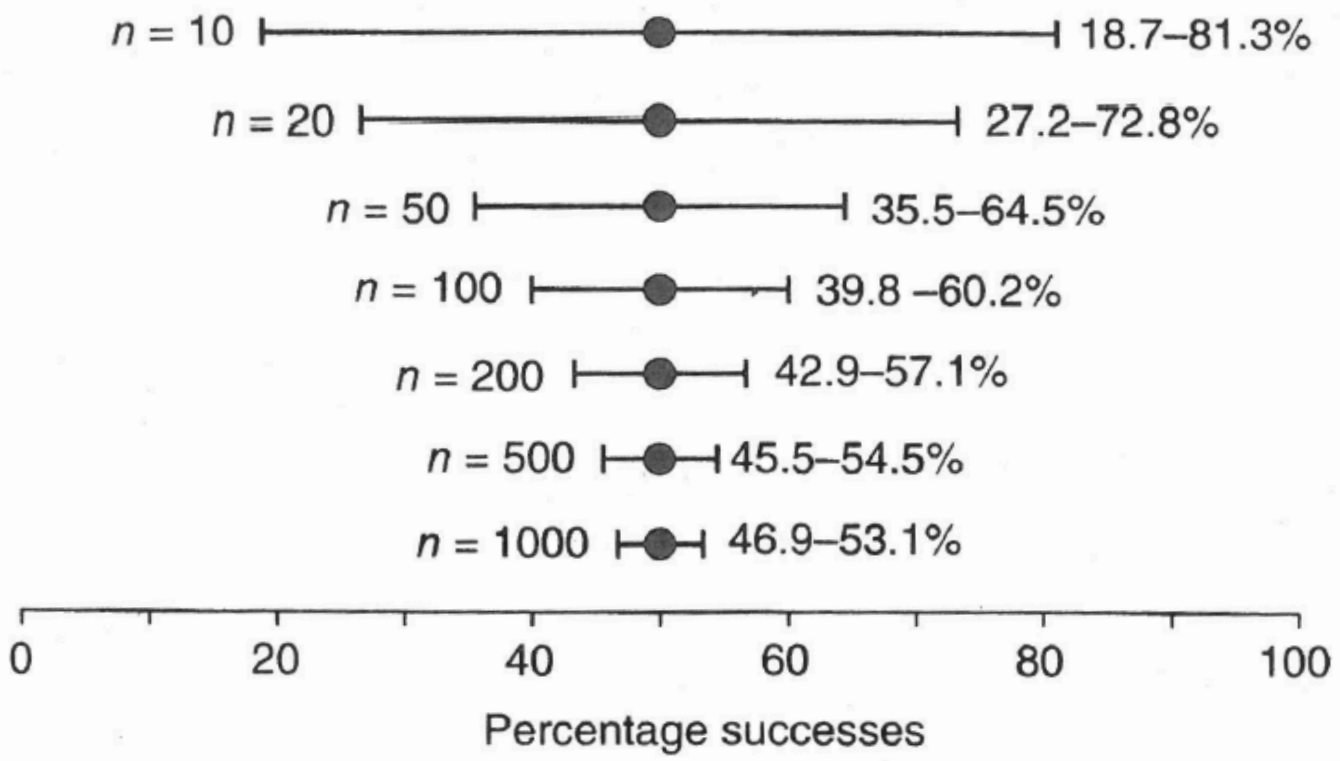


Figure 17.3 95% C.I.s for the proportion of successful outcomes with varying sample sizes. Point estimate equals 50% in all cases

Alternative category

If you need the CI for the proportion in the opposite category (failures in the case below), subtract values from 100%.

95% C. I. for proportions	Successes	Failures
Number examined:	50	50
Number detected:	42	8
Point estimate:	84.0%	12.0%
95% CI for proportion:	70.9 - 92.8%	7.2% - 29.1%

Asymmetric confidence intervals

Confidence intervals for proportions are always asymmetrical, unless the point estimate happens to be exactly 50%. Asymmetry is due to upper and lower constraints.

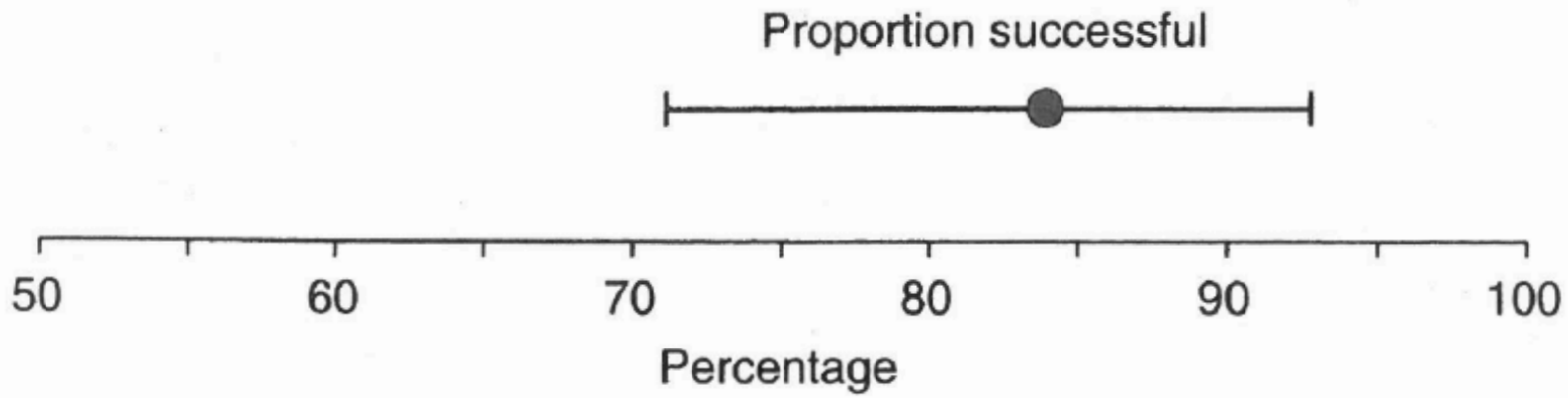


Figure 17.4 95% C.I. for the proportion of successful outcomes (42/50 in sample)

COMPARING OBSERVED PROPORTION TO SOME PRE-DETERMINED FIGURE

Using the 95% CI for the proportion

- Examine the 95% CI to determine if it includes the pre-determined figure, e.g., 50%.

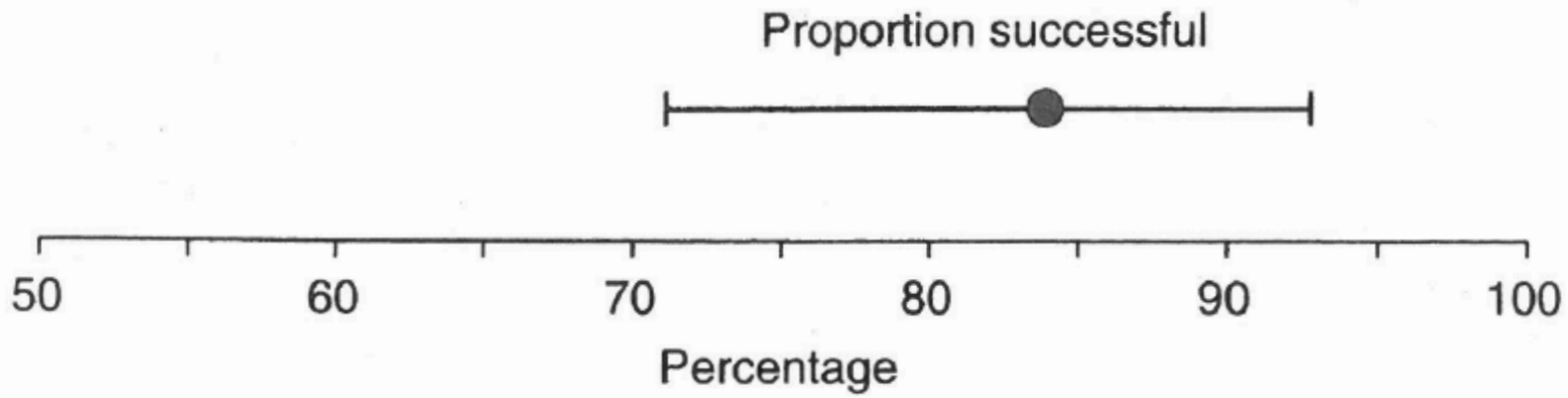


Figure 17.4 95% C.I. for the proportion of successful outcomes (42/50 in sample)

Goodness of fit chi-square test

Compares the proportion within a sample with some hypothesized proportion for the population.

Is the sample data consistent with the specified proportion?

Chi-square example

A company produces pressurized canisters of drug for asthma inhalers.

- 2 machines that produce the inhalers
  - Allegro - produces 61% of the canisters
  - Andante - produces 39% of the canisters
- 126 canisters were returned as faulty
  - 52 from the Allegro machine
  - 74 from the Andante machine

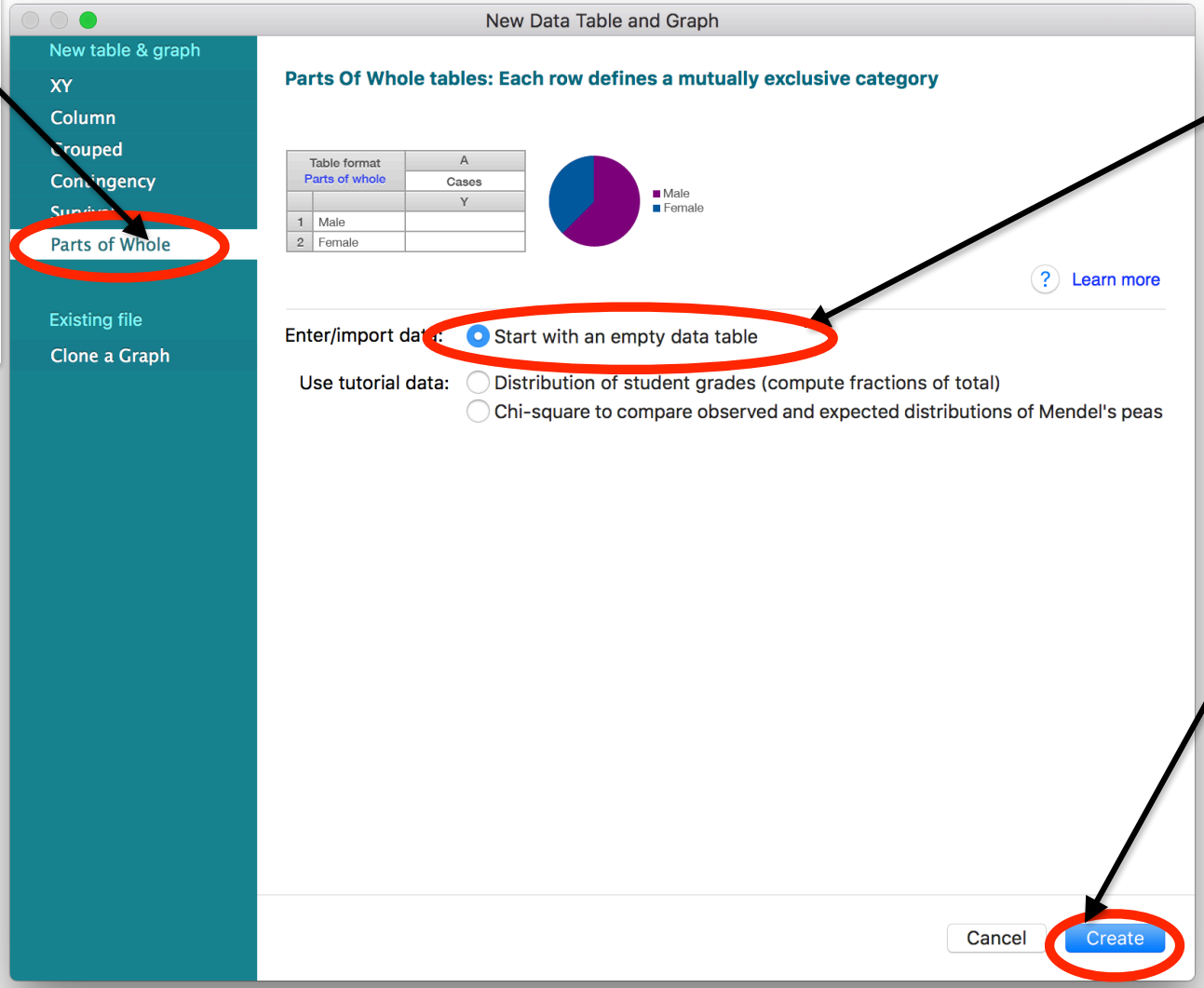
Are the machines equally reliable?

	Allegro	Andante
Observed	52	74
Expected	126 ✖ 0.61 = 76.86	126 ✖ 0.39 = 49.14
Obs - Exp	52 - 76.86 = -24.86	74 - 49.14 = 24.86
$(\text{Obs - Exp})^2$	618	618
$(\text{Obs - Exp})^2/\text{Exp}$	8.04	12.58

$$\chi^2 = 8.04 + 12.58 = 20.62 \qquad \text{df} = \text{number of categories - 1}$$

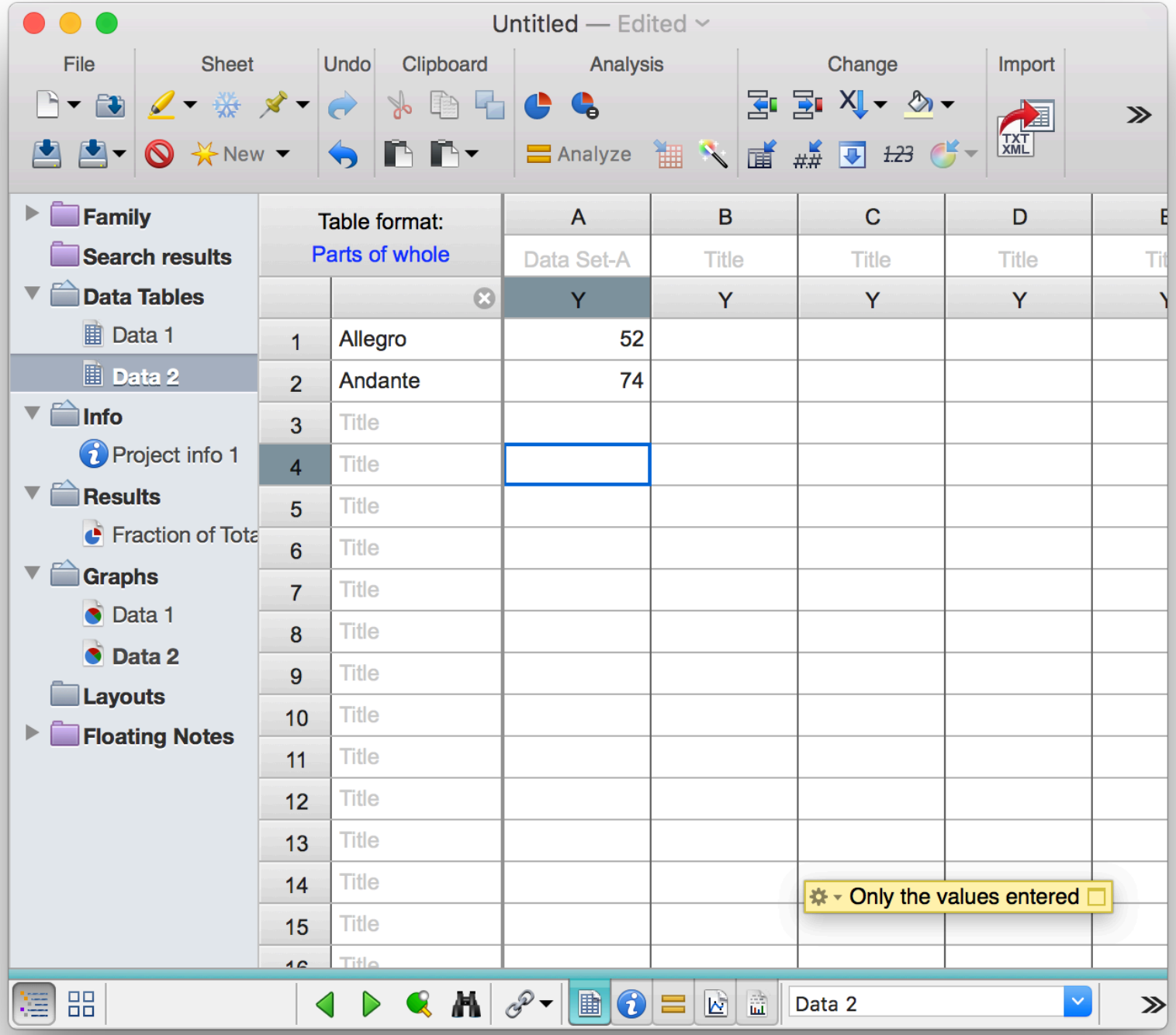
Chi-square goodness of fit in Prism

1.Create a new table using ‘Parts of Whole’ format

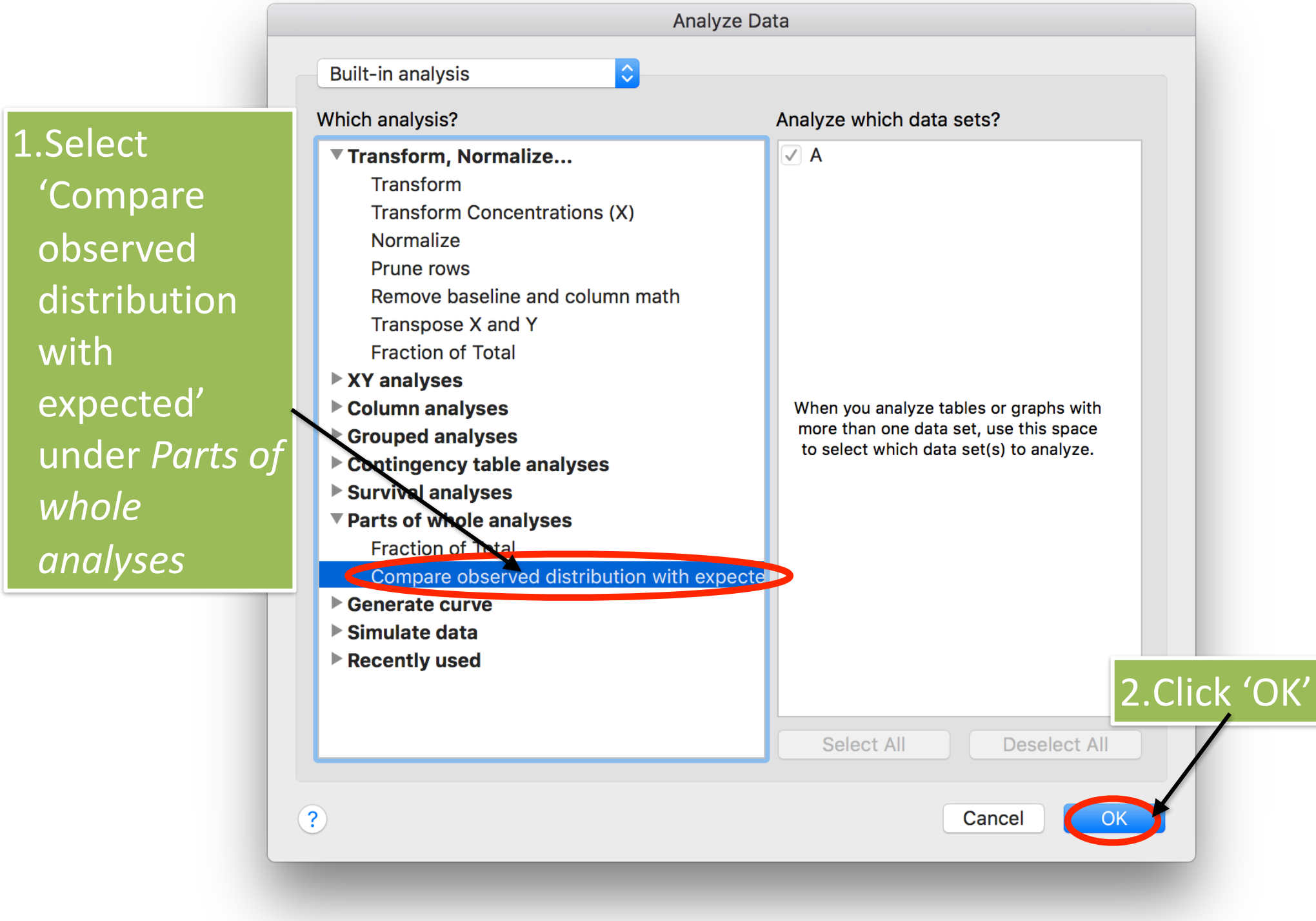


2.Start with an empty table

3.Click ‘Create’



Input the number of observed for each category in a single column

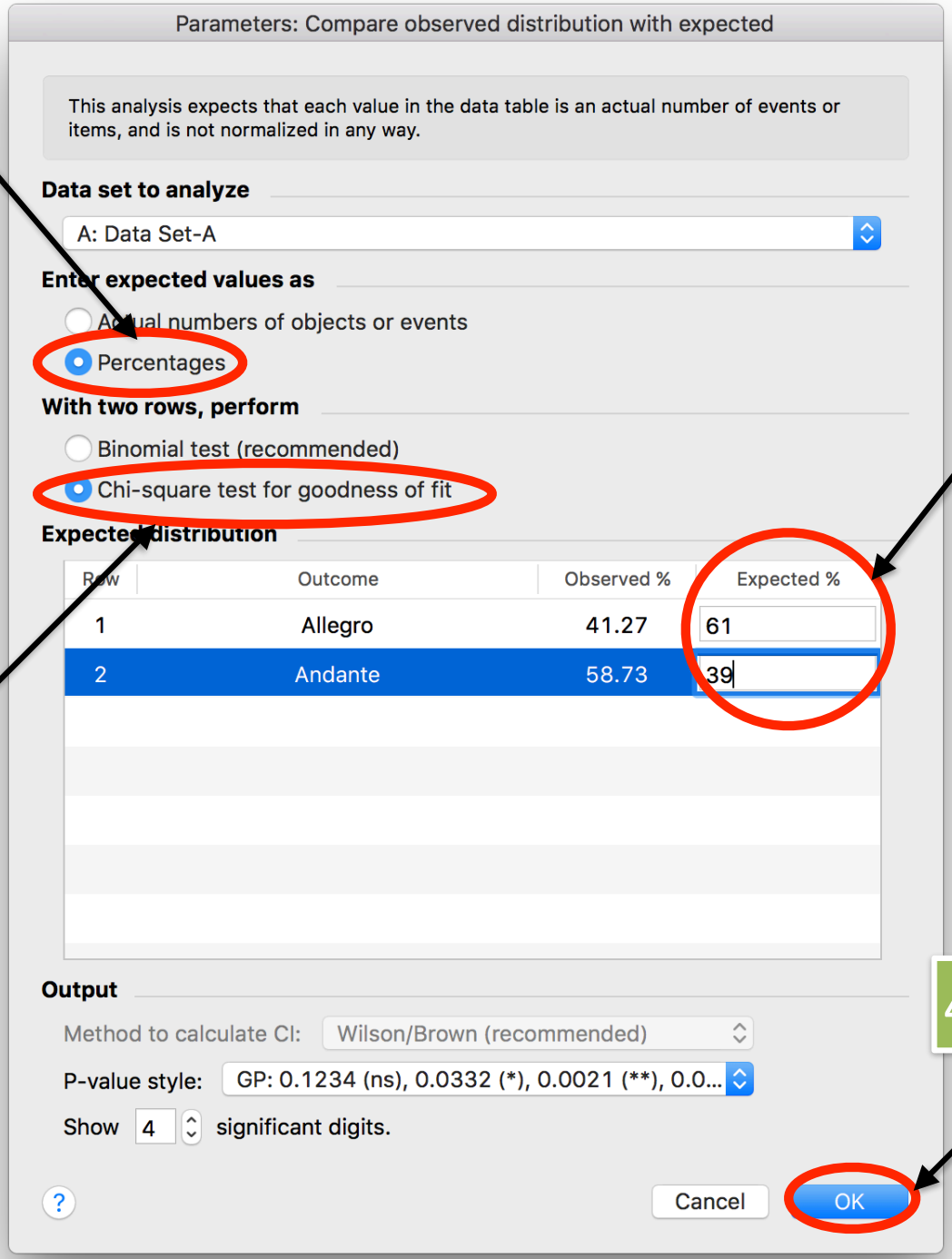

1.Select ‘Compare observed distribution with expected’ under *Parts of whole analyses*
Analyze Data
Built-in analysis
Which analysis?
Analyze which data sets?
Transform, Normalize...
Transform
Transform Concentrations (X)
Normalize
Prune rows
Remove baseline and column math
Transpose X and Y
Fraction of Total
XY analyses
Column analyses
Grouped analyses
Contingency table analyses
Survival analyses
Parts of whole analyses
Fraction of Total
Compare observed distribution with expecte
Generate curve
Simulate data
Recently used
A
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
2.Click ‘OK’
Cancel
OK

1.Select ‘Percentages’ for *Enter expected values as*

2.Select ‘Chi-square test for goodness of fit’ for *With two rows, perform*

3.Enter expected percentages

4.Click ‘OK’



O vs. E					
1	Table analyzed	Data 2			
2	Column analyzed	Column A			
3					
4	Chi-square test				
5	Chi-square	20.62			
6	DF	1			
7	P value (two-tailed)	<0.0001			
8	P value summary	****			
9	Is discrepancy significant (P < 0.05)?	Yes			
10					
11	Outcome	Expected #	Observed #	Expected %	Observed %
12	Allegro	76.86	52	61	41.27
13	Andante	49.14	74	39	58.73
14	TOTAL	126	126	100	100.00

The ‘Continuity’ problem

- Mathematical basis of test includes the assumption that chi-squared values are ‘continuous’.
- BUT, since we are examining discrete outcomes, this isn’t completely true

Yates continuity correction

- Adjust the discrepancies (observed count minus expected count) by 0.5 towards zero.
- Only noticeable when counts are small
- Only use when there are just two categories
- NOT available in Prism for Chi-square goodness of fit test

Cases with more than two categories

Same procedure, more ambiguous result

O vs. E					
1	Table analyzed	Data 3			
2	Column analyzed	Column A			
3					
4	Chi-square test				
5	Chi-square	4.467			
6	DF	2			
7	P value (two-tailed)	0.1072			
8	P value summary	ns			
9	Is discrepancy significant (P < 0.05)?	No			
10					
11	Outcome	Expected #	Observed #	Expected %	Observed %
12	Leaflet A	30	23	33.33	25.56
13	Leaflet B	30	39	33.33	43.33
14	Leaflet C	30	28	33.33	31.11
15	TOTAL	90	90	100	100.00

What did we learn?

- Nominal outcome data is treated differently than continuous (interval) outcome data
- Proportion of subjects within each category is used to summarize the data.
- Sampling error of a proportion depends on sample size and the presence of a rare category.
- 95% CI and a chi-square goodness of fit test can be used to compare an observed proportion to an expected proportion.