Binomial tests
Chi square tests
P-values
Confidence Intervals
Null Hypothesis testing

Practical Statistics for Experimental Biologists
Lecture 2
Tuesday, 28 July 2020
10:00am - 12:00pm

Example 1: Human Sex Ratio

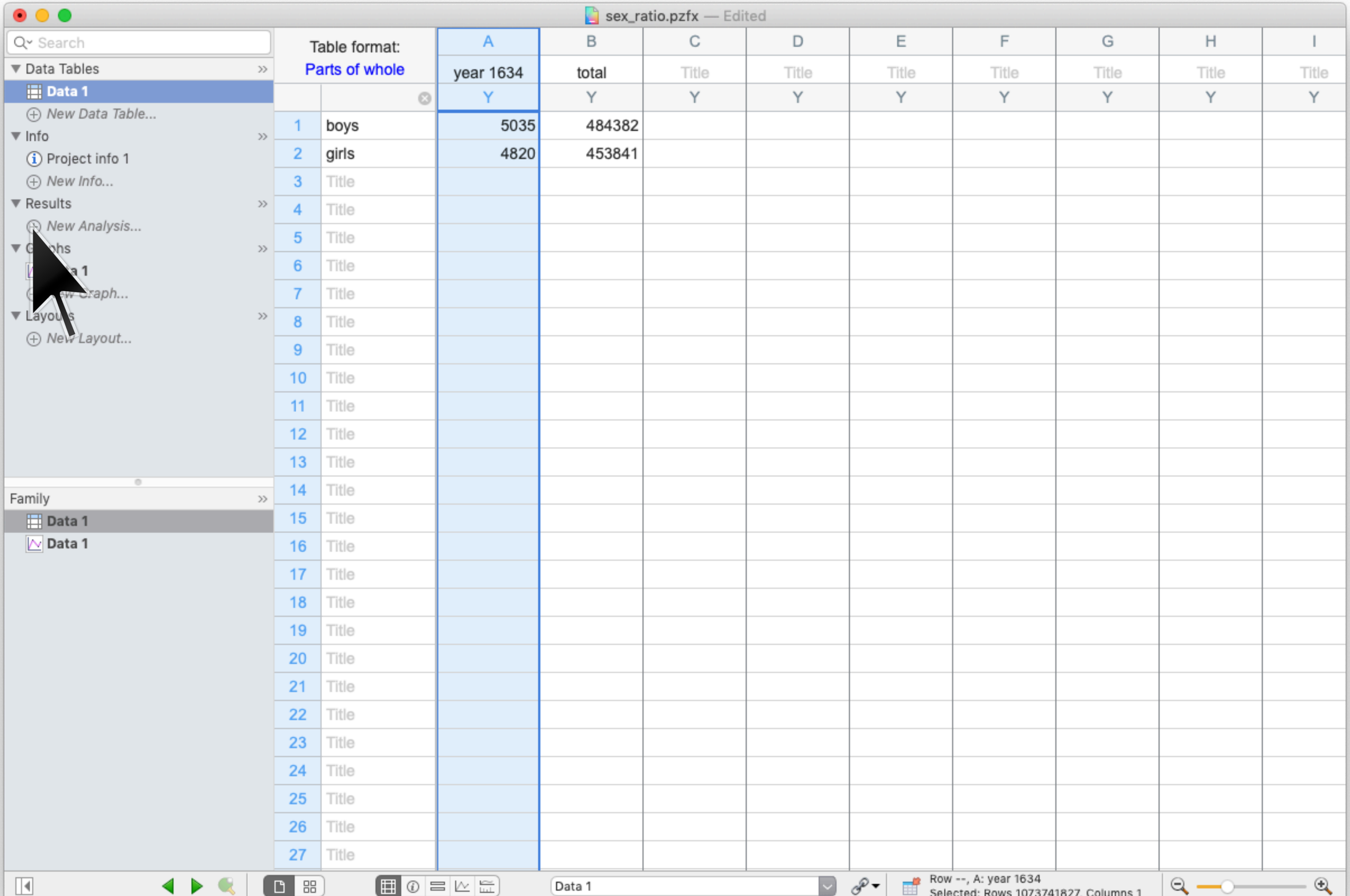
Computing sex ratio of humans is one of the oldest applications of statistics

year	male	female
1629	5218	4683
1630	4858	4457
1631	4422	4102
1632	4994	4590
1633	5158	4839
1634	5035	4820
1635	5106	4928
1636	4917	4605
1637	4703	4457

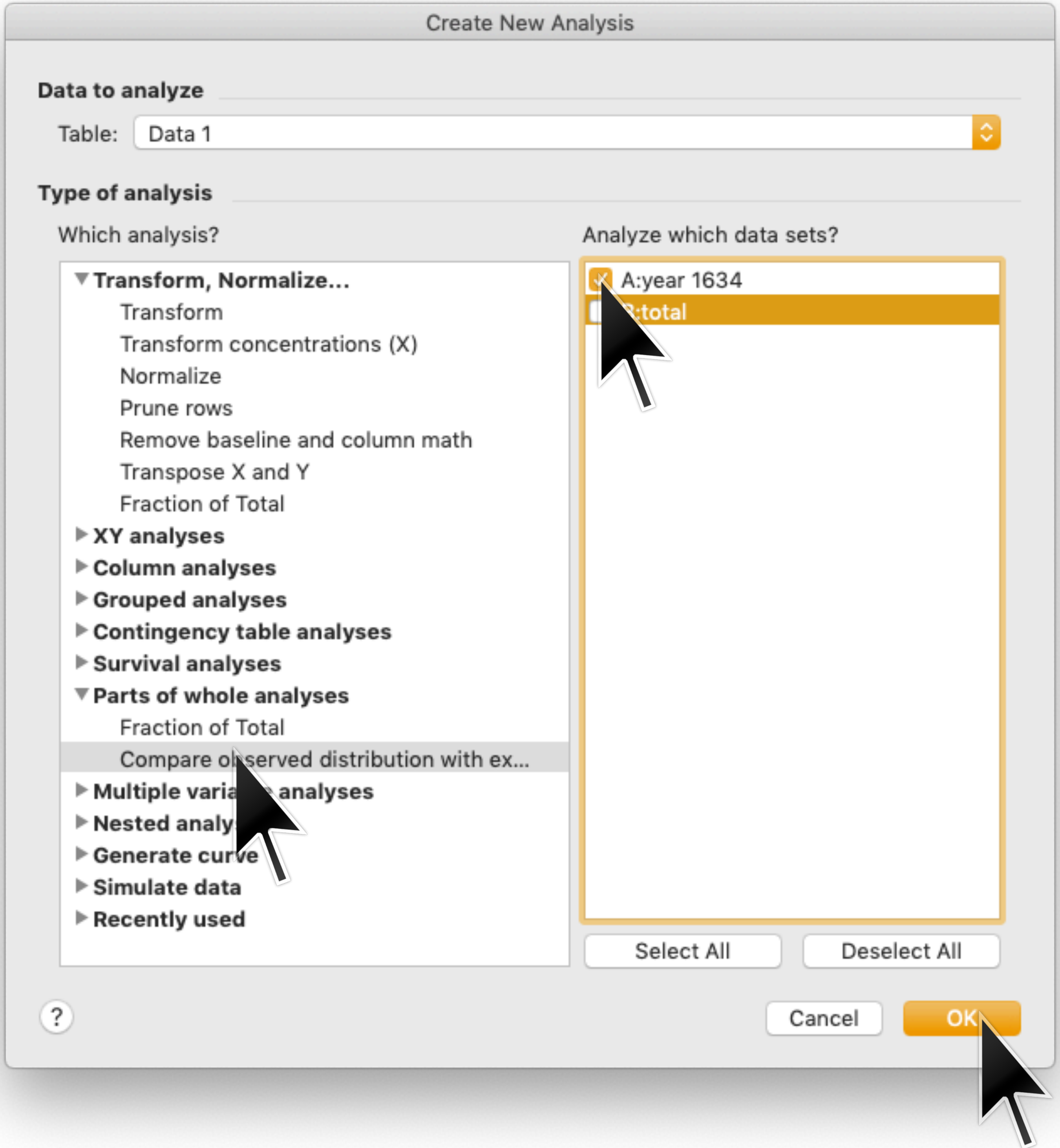
⋮

Arbuthnott J (1711). An Argument for Divine Providence, taken from the Constant Regularity observed in the Births of both Sexes.

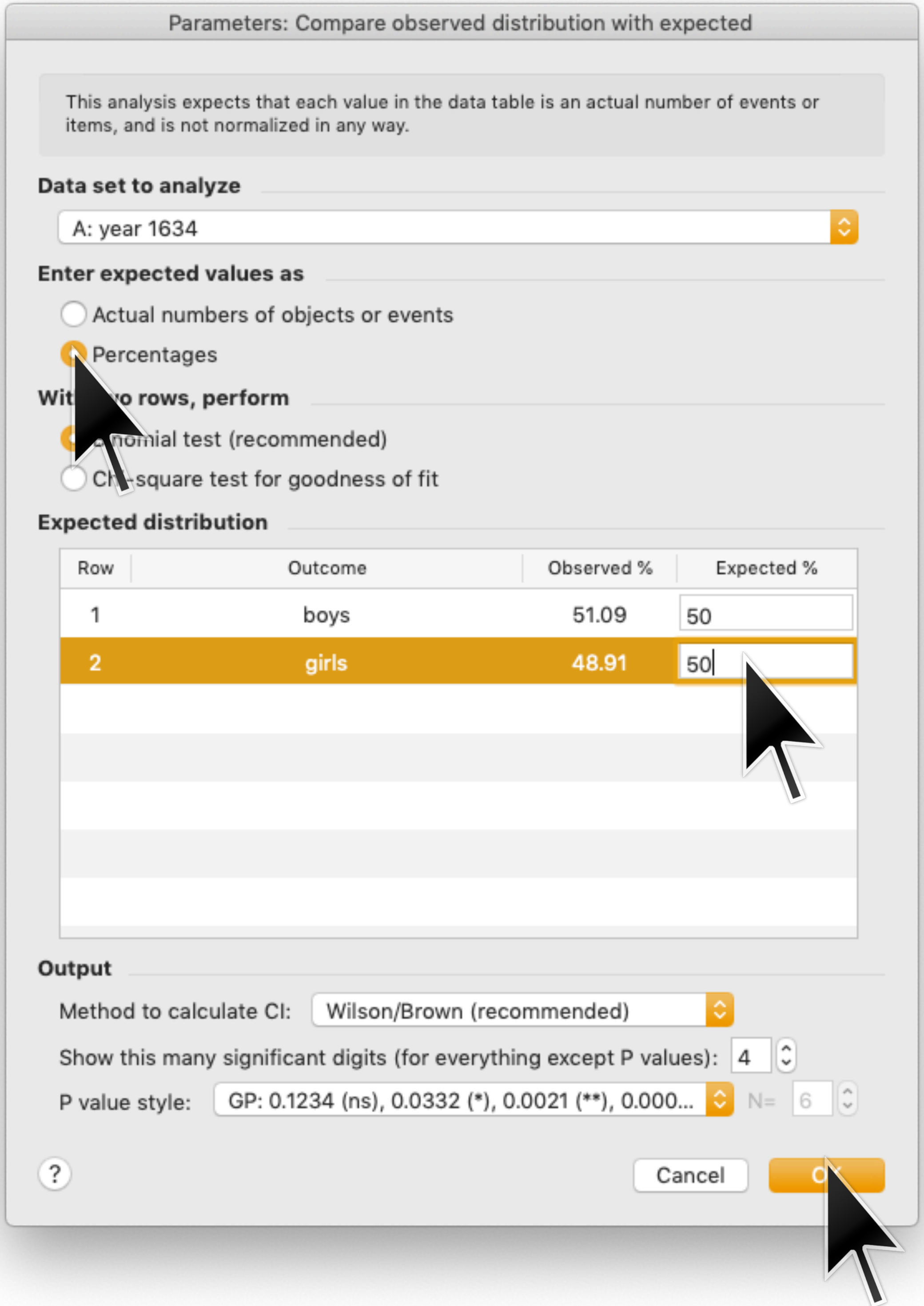
https://www.openintro.org/stat/data/?data=arbuthnot

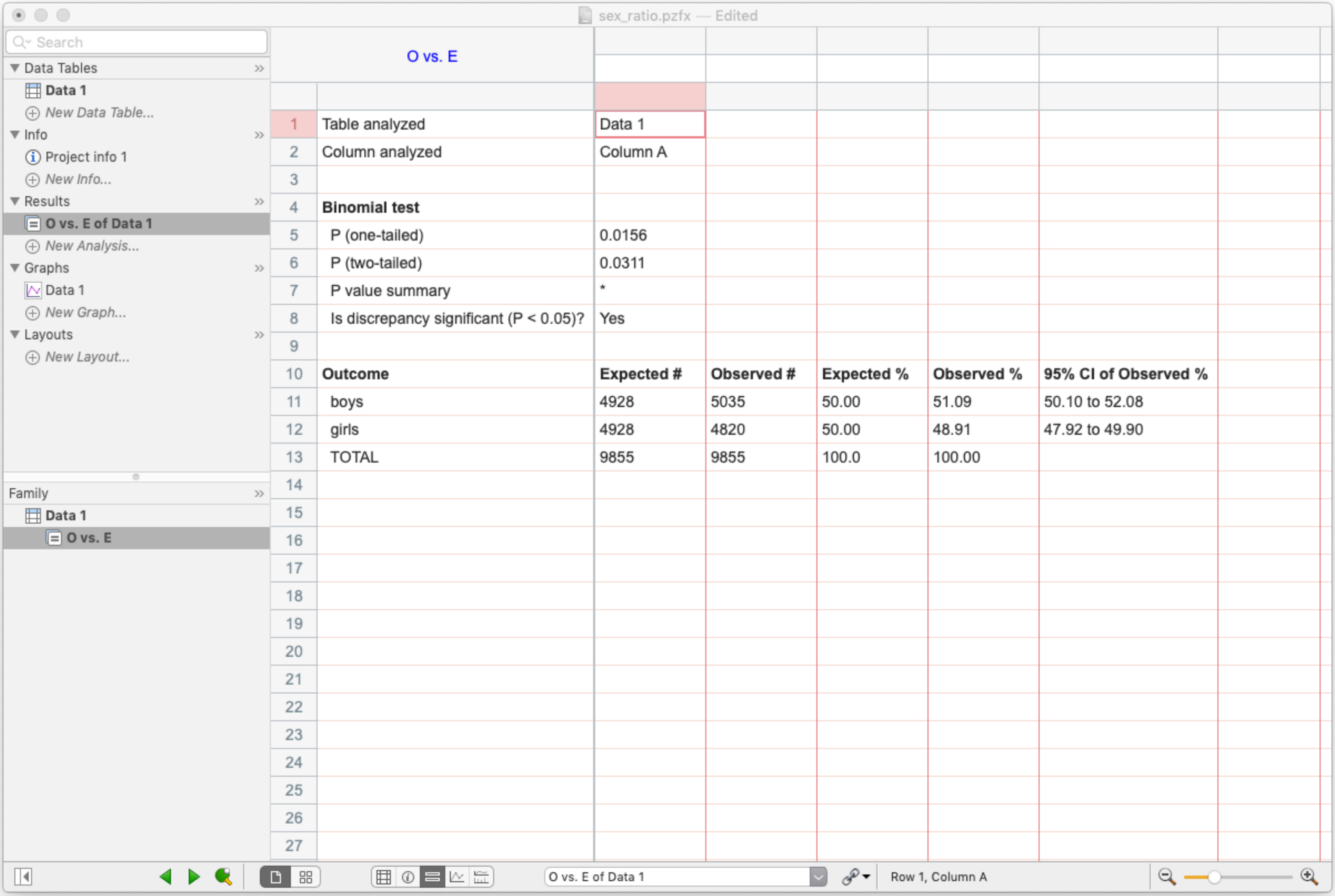


sex_ratio.pzfx — Edited
Search
Data Tables
Data 1
New Data Table...
Info
Project info 1
New Info...
Results
New Analysis...
Data 1
Layouts
New Layout...
Family
Data 1
Data 1
Table format:
Parts of whole
A
B
C
D
E
F
G
H
I
year 1634
total
Title
Y
1 boys 5035 484382
2 girls 4820 453841
Data 1
Row --, A: year 1634
Selected: Rows 1073741827, Columns 1

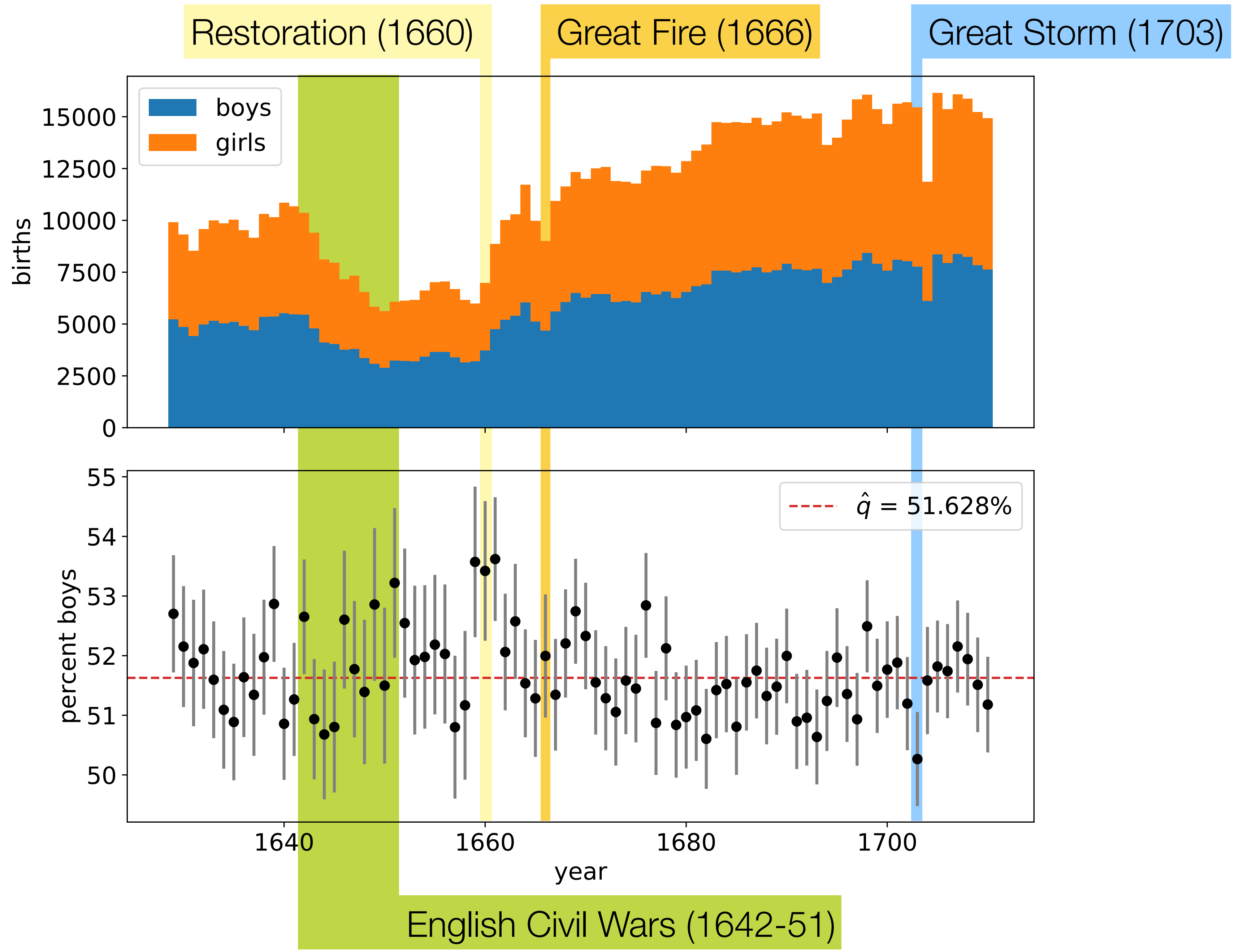


Create New Analysis
Data to analyze
Table: Data 1
Type of analysis
Which analysis?
Transform, Normalize...
Transform
Transform concentrations (X)
Normalize
Prune rows
Remove baseline and column math
Transpose X and Y
Fraction of Total
XY analyses
Column analyses
Grouped analyses
Contingency table analyses
Survival analyses
Parts of whole analyses
Fraction of Total
Compare observed distribution with ex...
Multiple variable analyses
Nested analyses
Generate curve
Simulate data
Recently used
Analyze which data sets?
A:year 1634
B:total
Select All
Deselect All
?
Cancel
OK


Parameters: Compare observed distribution with expected
This analysis expects that each value in the data table is an actual number of events or items, and is not normalized in any way.
Data set to analyze
A: year 1634
Enter expected values as
Actual numbers of objects or events
Percentages
Expected distribution
Row
Outcome
Observed %
Expected %
1
boys
51.09
50
2
girls
48.91
50
Output
Method to calculate CI:
Wilson/Brown (recommended)
Show this many significant digits (for everything except P values):
4
P value style:
GP: 0.1234 (ns), 0.0332 (*), 0.0021 (**), 0.000...
N=
6
?
Cancel

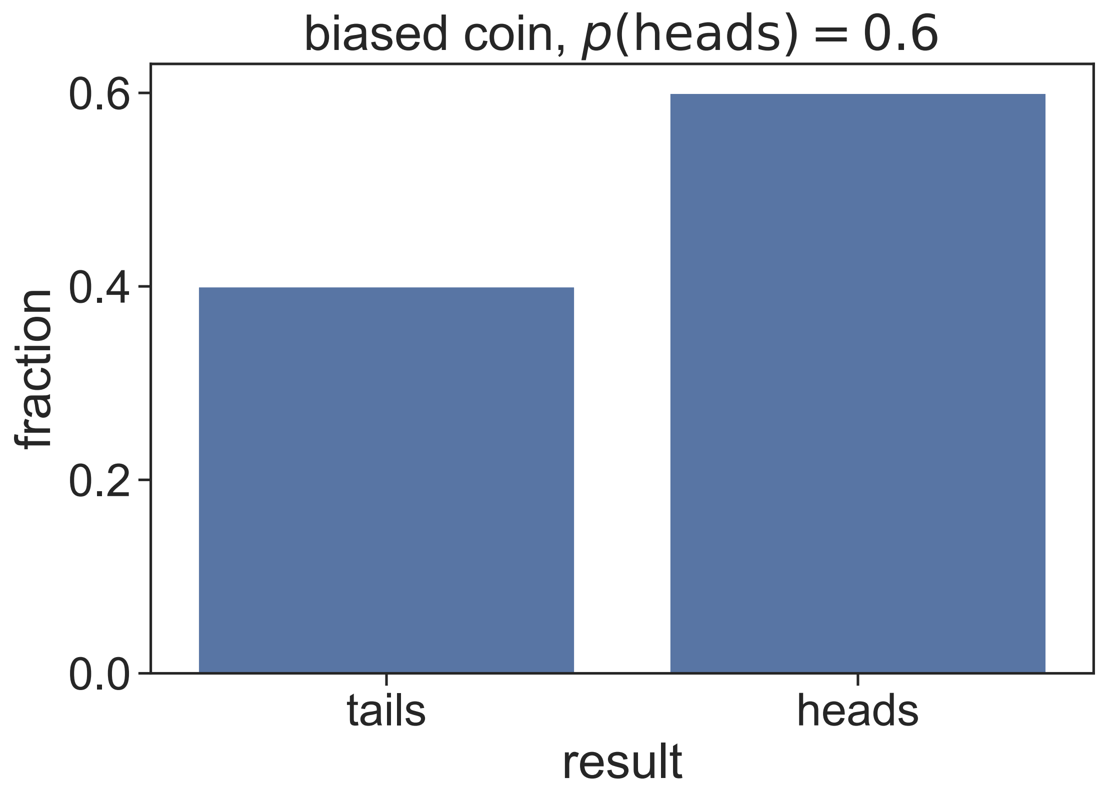

sex_ratio.pzfx — Edited
Search
Data Tables
Data 1
New Data Table...
Info
Project info 1
New Info...
Results
O vs. E of Data 1
New Analysis...
Graphs
Data 1
New Graph...
Layouts
New Layout...
Family
Data 1
O vs. E
O vs. E
1 Table analyzed | Data 1
2 Column analyzed | Column A
3
4 Binomial test
5 P (one-tailed) | 0.0156
6 P (two-tailed) | 0.0311
7 P value summary | *
8 Is discrepancy significant (P < 0.05)? | Yes
9
10 Outcome | Expected # | Observed # | Expected % | Observed % | 95% CI of Observed %
11 boys | 4928 | 5035 | 50.00 | 51.09 | 50.10 to 52.08
12 girls | 4928 | 4820 | 50.00 | 48.91 | 47.92 to 49.90
13 TOTAL | 9855 | 9855 | 100.0 | 100.00
O vs. E of Data 1
Row 1, Column A

Births in London, 1629-1710



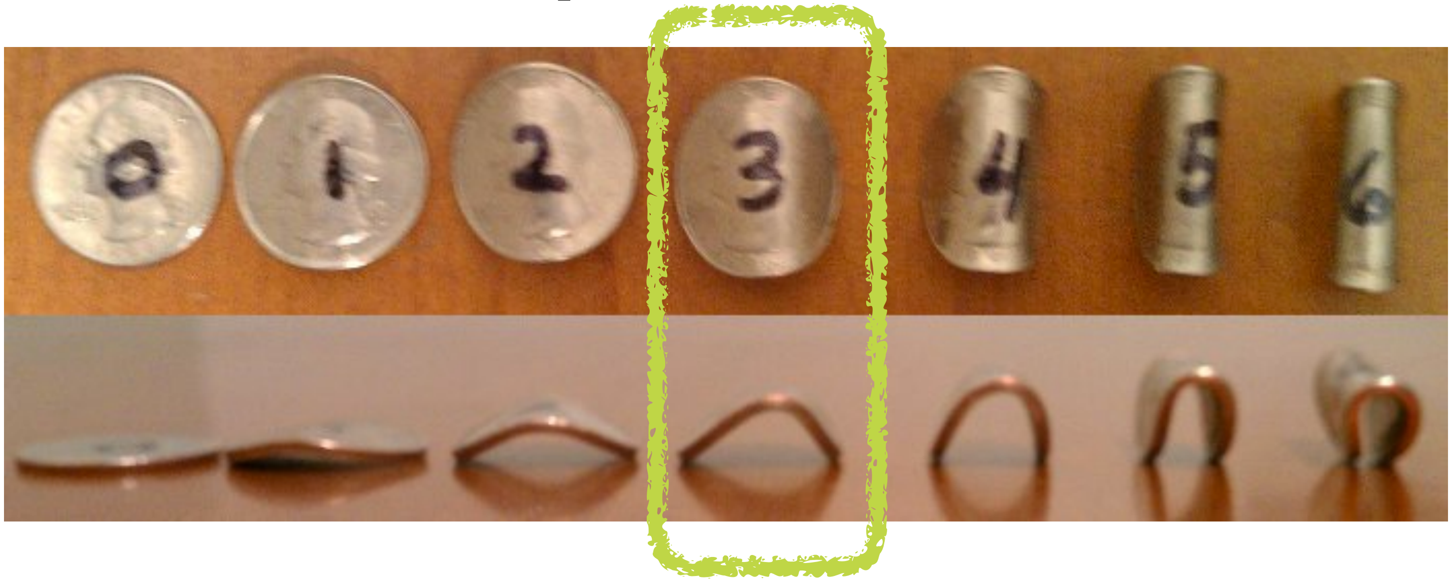
Example 2: A biased coin

Biased coins are modeled using a Bernoulli distribution, which describes probabilities for a binary variable



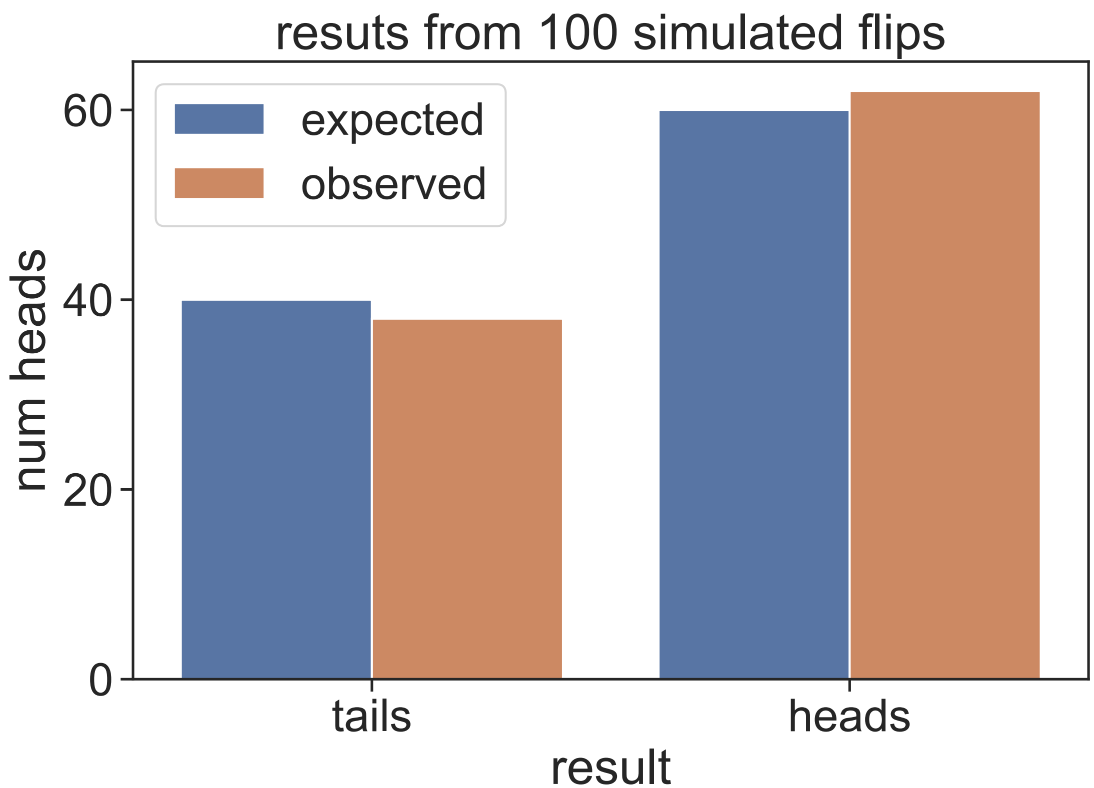
Making a biased coin

$$p(\text{heads}) \approx 60\,\%$$



Mike Izbicki (Claremont McKenna College)
https://izbicki.me/blog/how-to-create-an-unfair-coin-and-prove-it-with-math.html

The number of heads after 100 flips of the biased coin will resemble the underlying probabilities, but will not match exactly



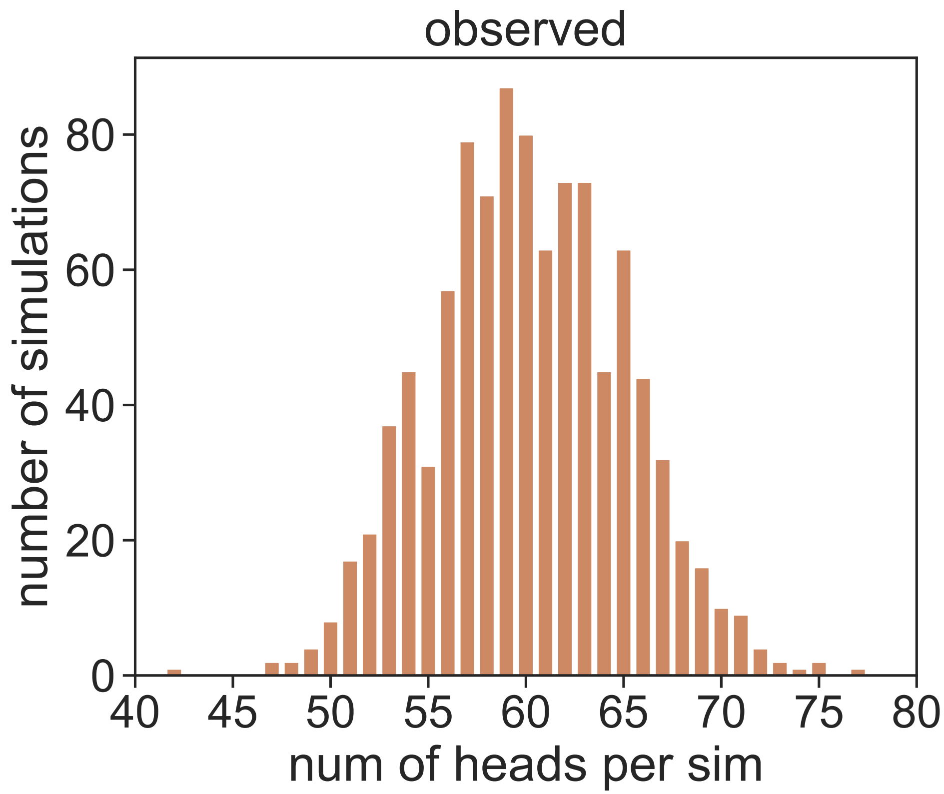
expected: 60 heads, 40 tails

observed: <u>62 heads</u>, <u>38 tails</u>

How much deviation from the expected values do we expect?

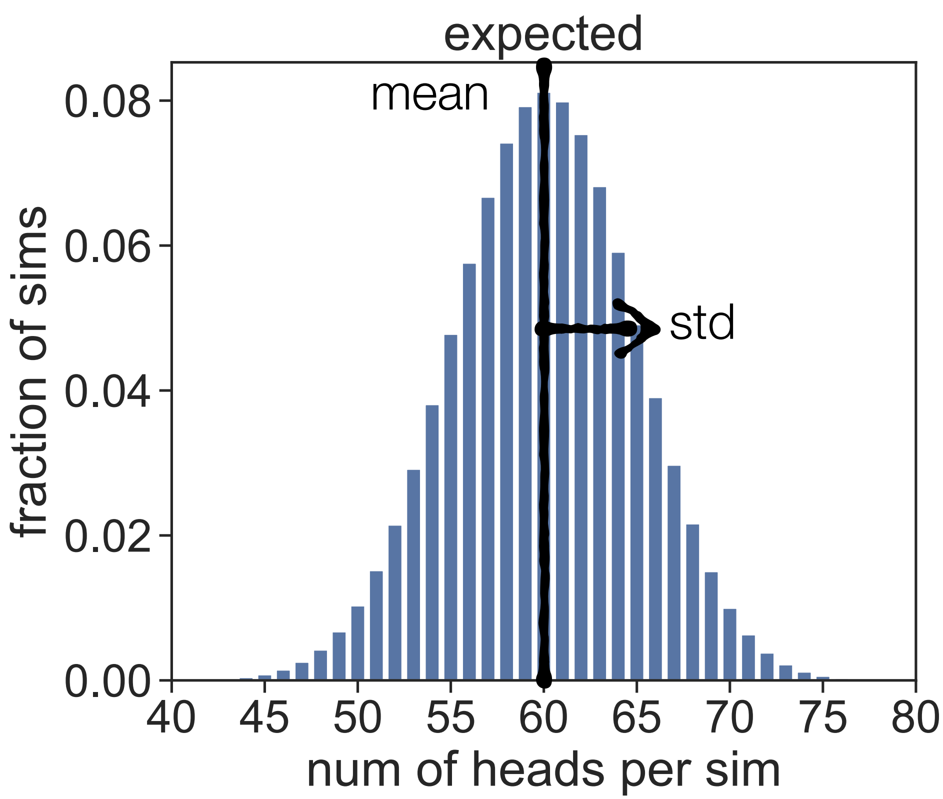
There is substantial variation across replicates. This is to be expected.

Results from 1000 simulations, 100 flips per simulation



The variation in the number of heads from replicate to replicate is described by a binomial distribution

Results from 1000 simulations, 100 flips per simulation



mean = 60

standard deviation (std) = 4.9

Can we determine whether or not a coin is biased by flipping it 100 times?

Suppose we flip a coin 100 times and observe **62 heads** (and 38 tails).

Null hypothesis: heads and tails are equally likely, i.e.

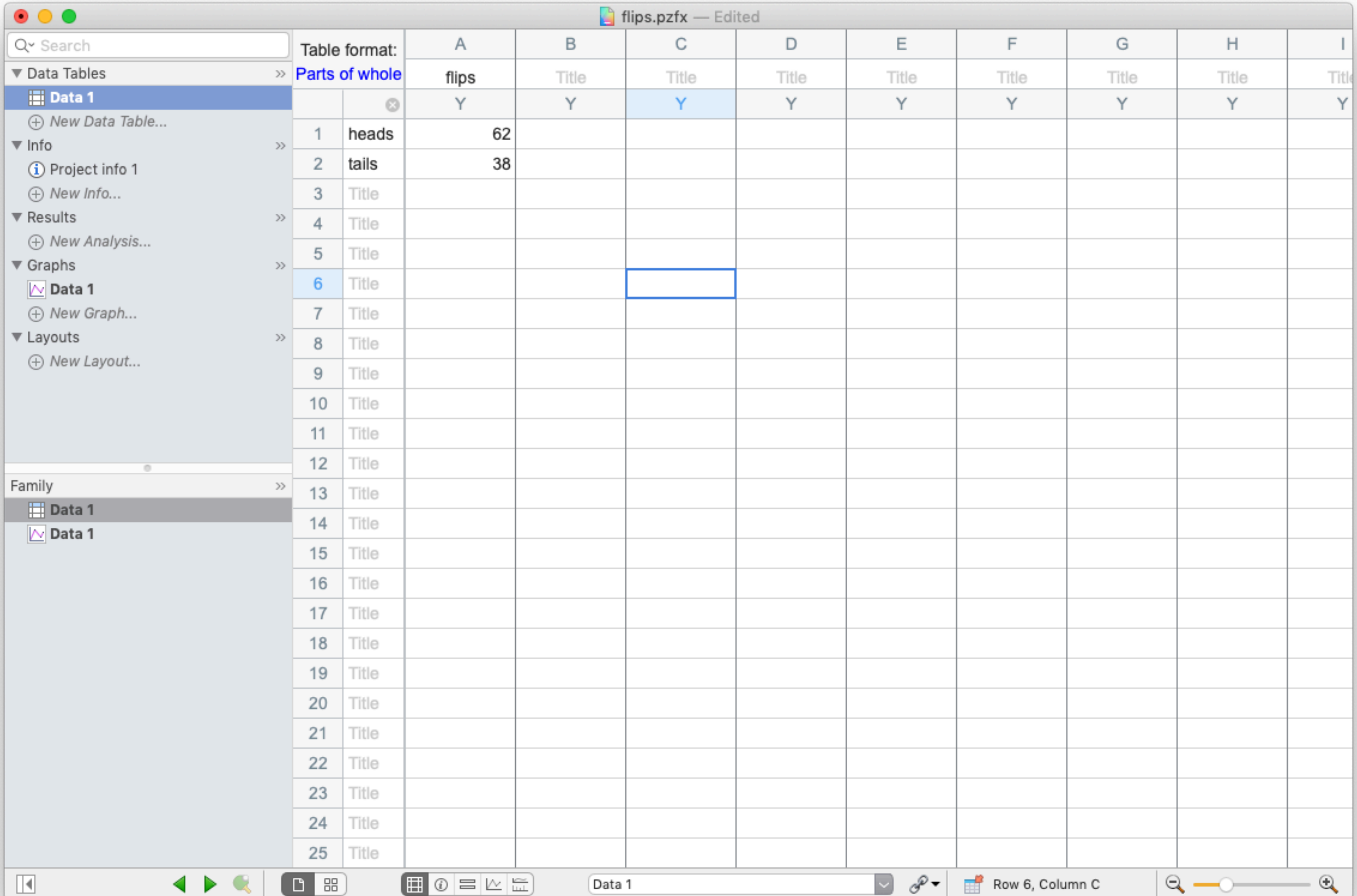
$$p(\text{heads}) = 50\,\%$$

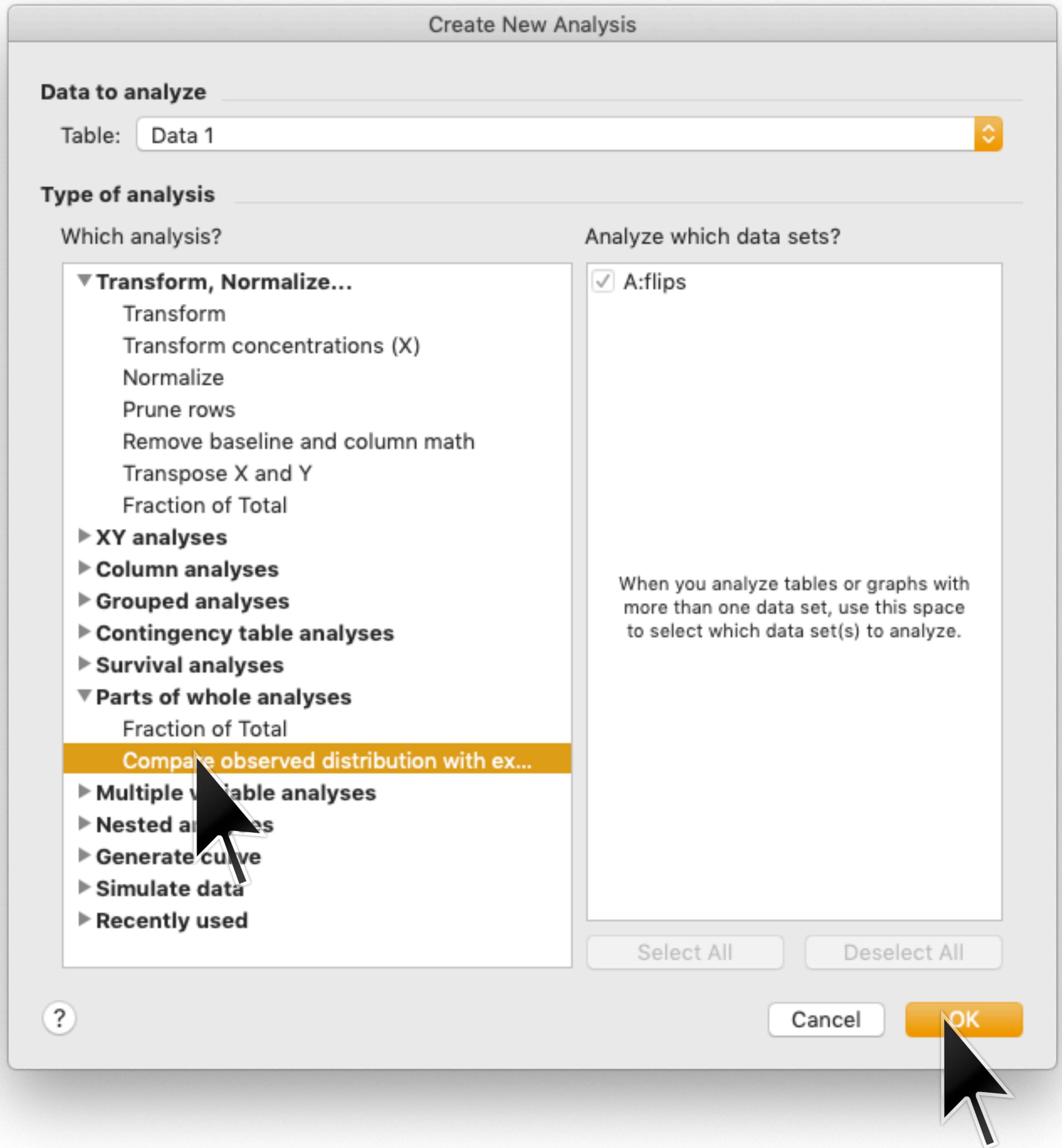
Alternative hypothesis: heads and tails are not equally likely, i.e.

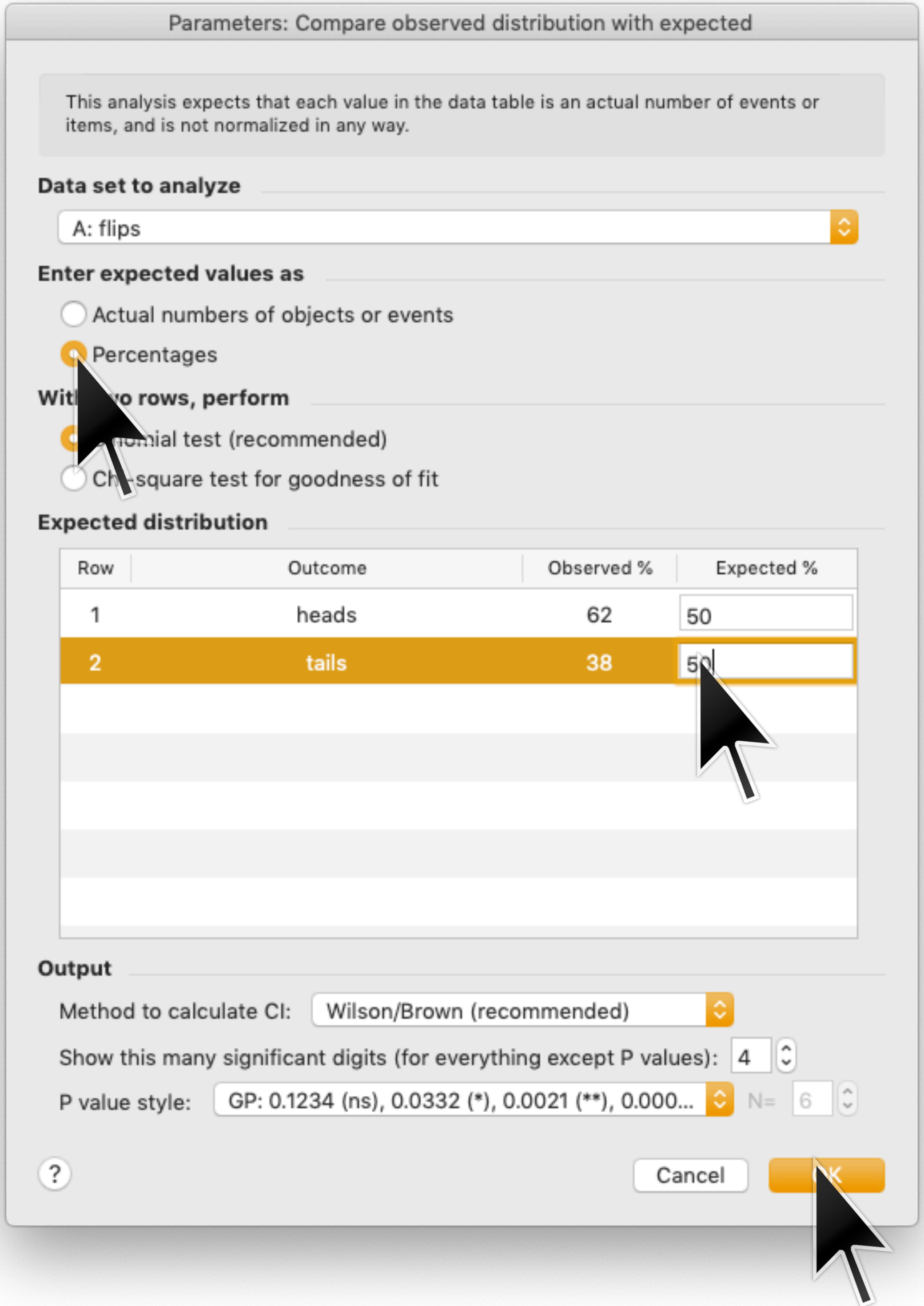
$$p(\text{heads}) \neq 50\,\%$$

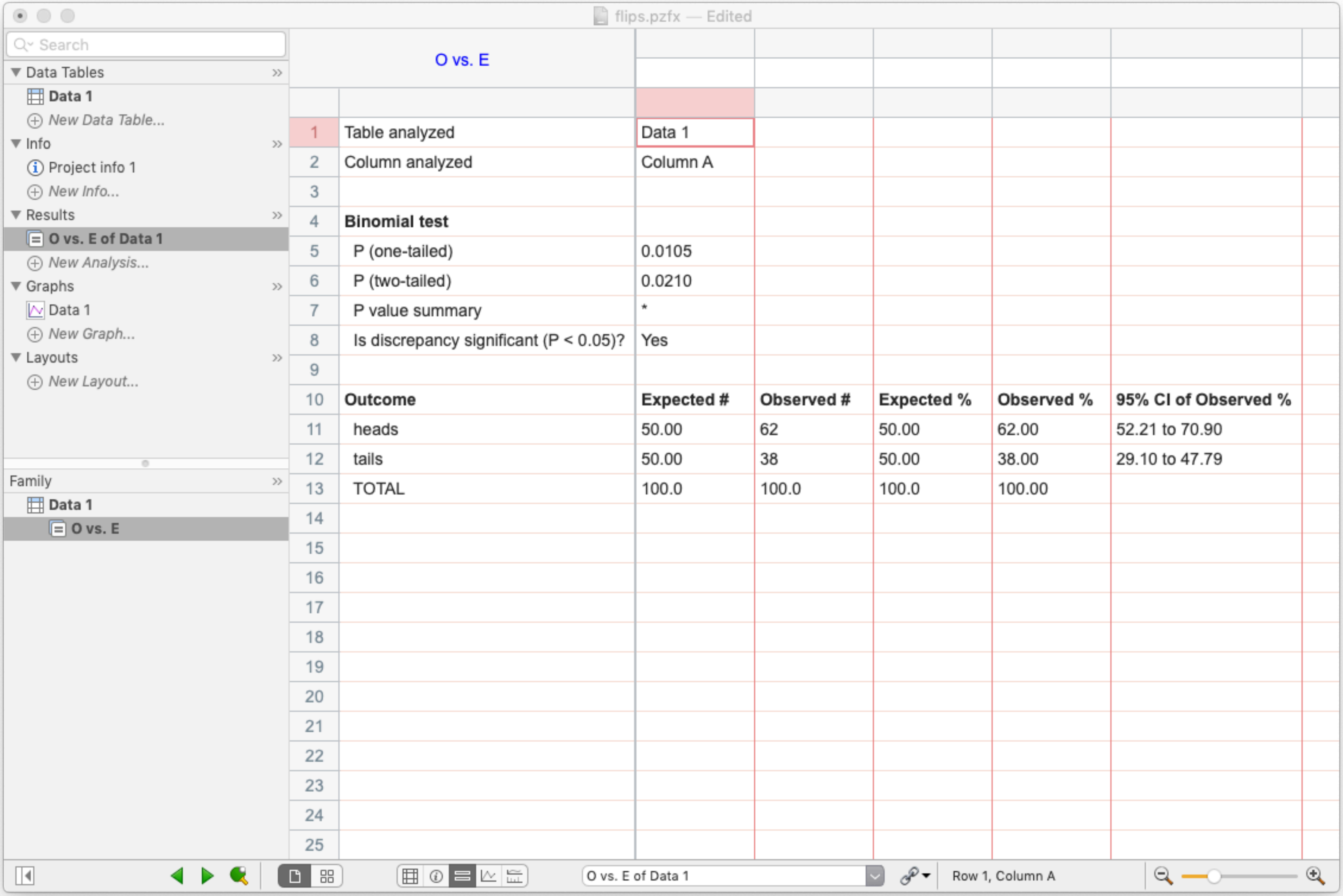
Our observation (62 heads) may or may not allow us to reject the null hypothesis and thus accept the alternative hypothesis.

No amount of data, however, can cause us to accept the null hypothesis.


flips.pzfx — Edited
Search
Data Tables
Data 1
New Data Table...
Info
Project info 1
New Info...
Results
New Analysis...
Graphs
Data 1
New Graph...
Layouts
New Layout...
Family
Data 1
Data 1
Table format:
Parts of whole
A B C D E F G H I
flips
Title
Y
1 heads 62
2 tails 38
Data 1
Row 6, Column C

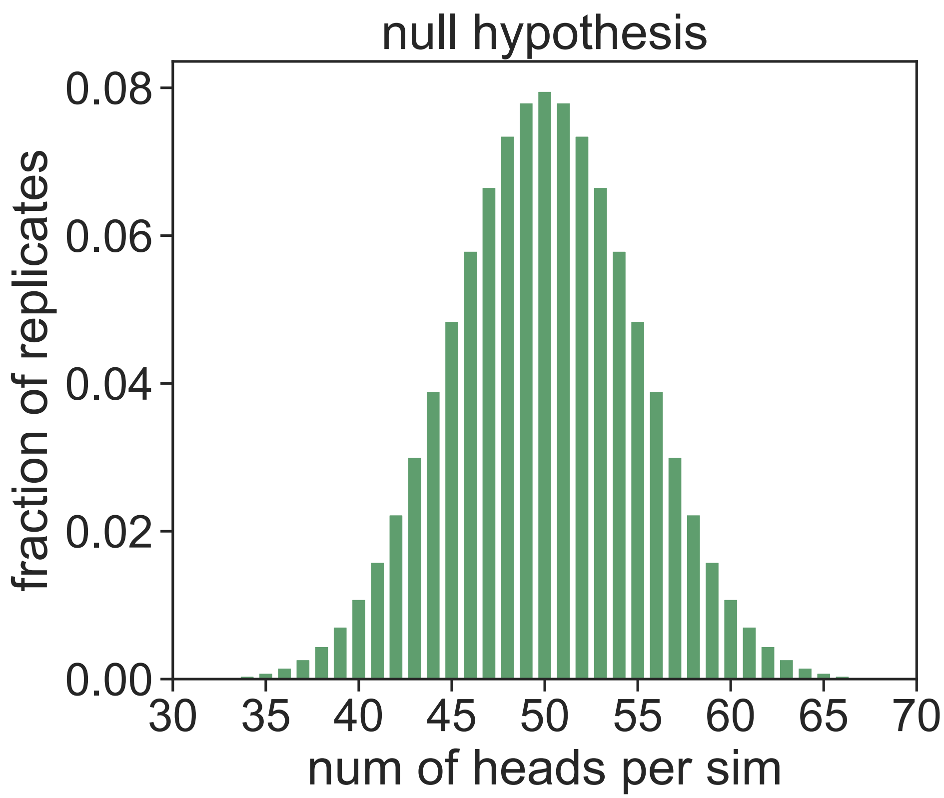

Create New Analysis
Data to analyze
Table:
Data 1
Type of analysis
Which analysis?
Transform, Normalize...
Transform
Transform concentrations (X)
Normalize
Prune rows
Remove baseline and column math
Transpose X and Y
Fraction of Total
XY analyses
Column analyses
Grouped analyses
Contingency table analyses
Survival analyses
Parts of whole analyses
Fraction of Total
Compare observed distribution with ex...
Recently used
Analyze which data sets?
A:flips
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


Parameters: Compare observed distribution with expected
This analysis expects that each value in the data table is an actual number of events or items, and is not normalized in any way.
Data set to analyze
A: flips
Enter expected values as
Actual numbers of objects or events
Percentages
With two rows, perform
Binomial test (recommended)
Chi-square test for goodness of fit
Expected distribution
Row
Outcome
Observed %
Expected %
1
heads
62
50
2
tails
38
50
Output
Method to calculate CI:
Wilson/Brown (recommended)
Show this many significant digits (for everything except P values):
4
P value style:
GP: 0.1234 (ns), 0.0332 (*), 0.0021 (**), 0.000...
N=
6
?
Cancel
OK



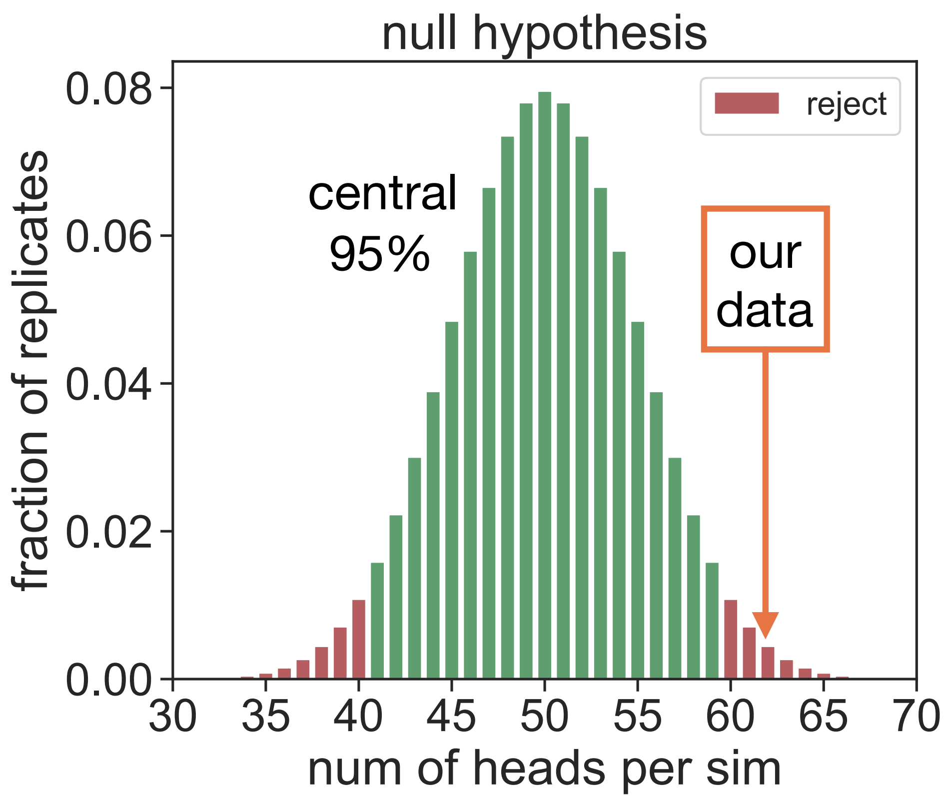
	O vs. E					
1	Table analyzed	Data 1				
2	Column analyzed	Column A				
3						
4	**Binomial test**					
5	P (one-tailed)	0.0105				
6	P (two-tailed)	0.0210				
7	P value summary	*				
8	Is discrepancy significant (P < 0.05)?	Yes				
9						
10	**Outcome**	**Expected #**	**Observed #**	**Expected %**	**Observed %**	**95% CI of Observed %**
11	heads	50.00	62	50.00	62.00	52.21 to 70.90
12	tails	50.00	38	50.00	38.00	29.10 to 47.79
13	TOTAL	100.0	100.0	100.0	100.00	
14						
15						
16						
17						
18						
19						
20						
21						
22						
23						
24						
25						

The null hypothesis is assessed by where the data fall within the null distribution



We reject the null hypothesis of the data fall too far away from the bulk of the distribution

If the null hypothesis is true, data should fall within the green region 95% of the time, and within the red "reject" region 5% of the time.

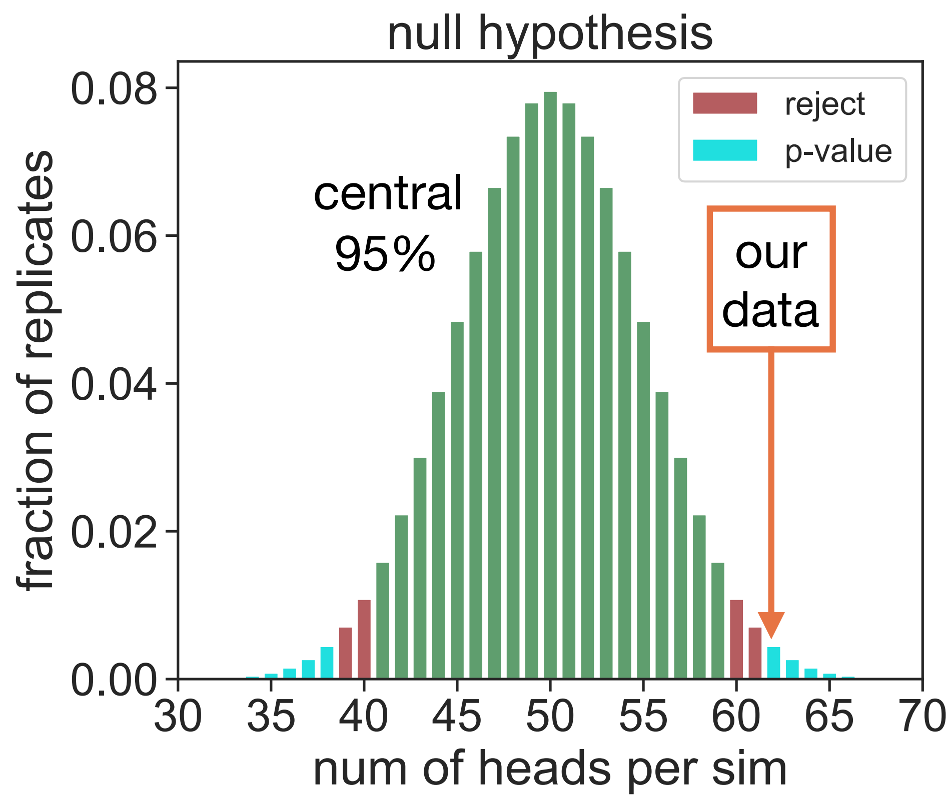


Our assumed dataset (62 heads) lies outside the central 95%.

We can therefore reject the null hypothesis with 95% confidence.

P-values quantify the probability of data being as or more extreme than the data in hand were the null hypothesis true.

The P-value threshold of 0.05 comes from adopting a confidence threshold of 95%.

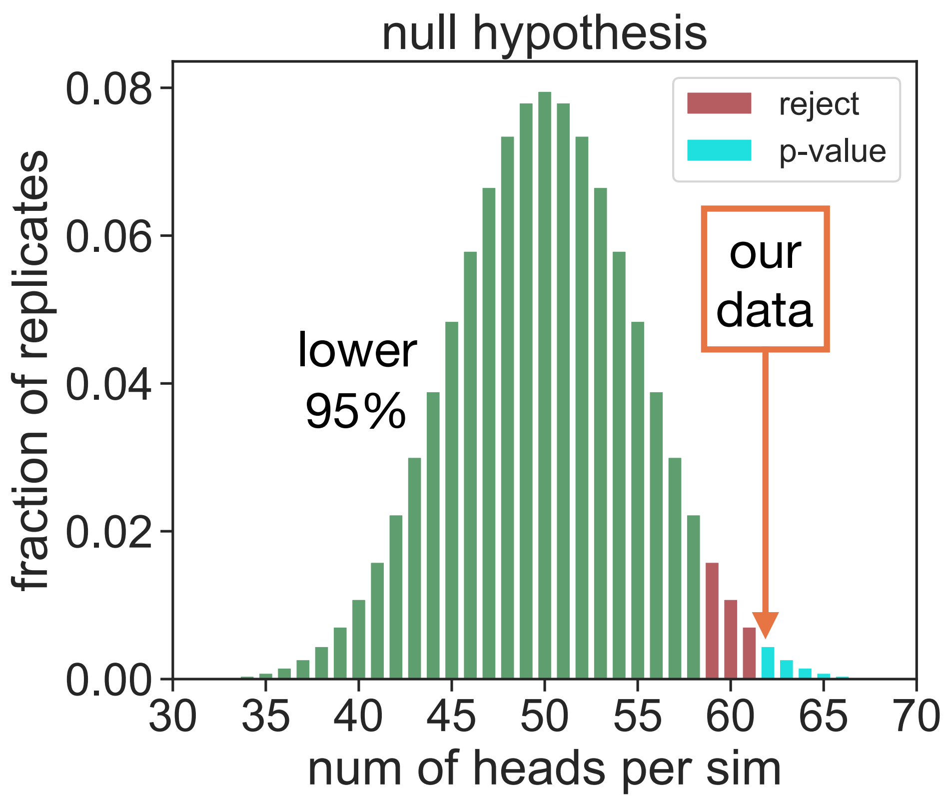


We find that **p=0.0210** for the two-sided test.

We therefore say that our result is "statistically significant"

P-values quantify the probability of data being as or more extreme than the data in hand were the null hypothesis true.

A one-sided hypothesis test only considers one side of the distribution.

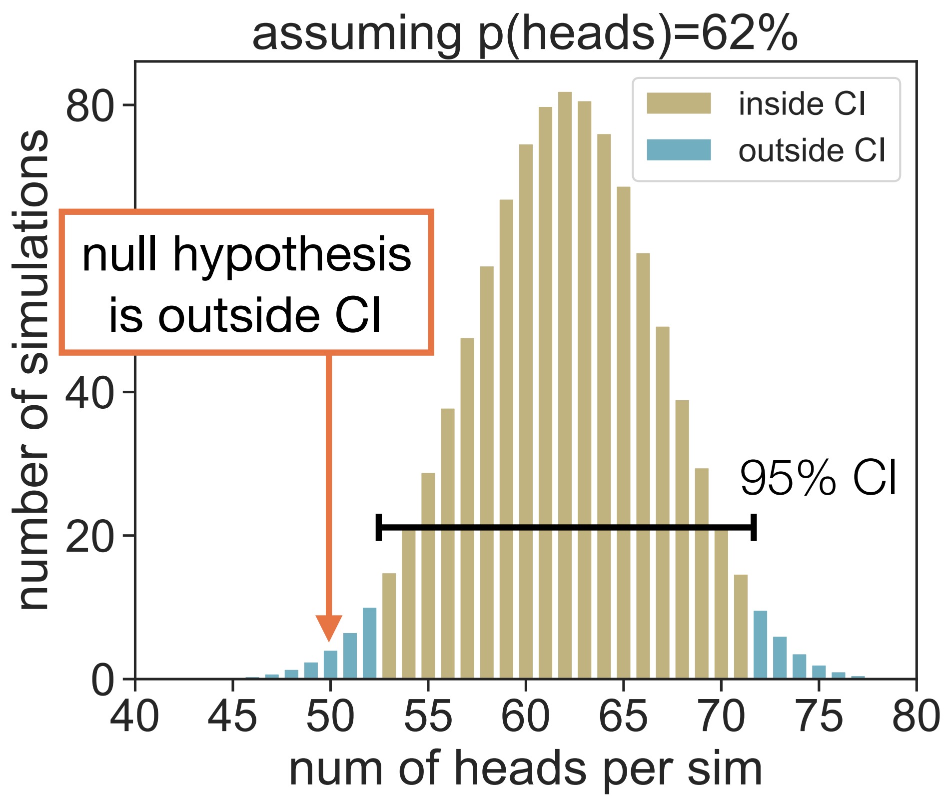


We find that **p=0.0105** for the one-sided test.

In general, two-sided tests are more conservative than one-sided tests.

Unless you have good reason to do otherwise, use two-sided tests.

Confidence intervals (CIs) are more informative than P-values



We conclude that $p(\text{heads})$ lies within [52.5%, 71.5%] with 95% confidence.

We can reject the null hypothesis because it lies outside this confidence interval.

P-values have multiple pitfalls

- "Statistically significant" does not actually mean "significant" in the normal sense. At best, it means "detectable".
- P-values do not say how big an observed effect is.
- P-values do not say how important that observed effect is.
- P-values calculations rely on assumptions, and violation of any of those assumptions can render P-values meaningless.
- Perhaps most severe is the fact that <u>P-values do not actually quantify you how likely or unlikely your null hypothesis is</u>!

Why are Confidence Intervals better than P-values?

- Like a P-value, a CI communicates statistical significance (i.e. detectability).
- A CI also communicates effect size, as well as the uncertainty in that effect size.
- A 95% CI does not actually mean that the true value of a parameter lies within that interval with 95% probability. Still, this (extremely common) misinterpretation is largely benign compared to the misinterpretation of P-values.
- However, P-values are more commonly reported than confidence intervals.

The perils of null hypothesis testing

Summary of null hypothesis testing

Step 1: Specify a null hypothesis.

Step 2: Specify a confidence level (usually 95%)

Step 3: Identify the appropriate statistical test

Then:
evaluate on data

Result: P-value summarizing how unlikely the data is compared to null hypothesis expectations.

Perhaps most problematic is how easily P-values are misinterpreted.

Roughly speaking, P-values quantify how likely our data would be if the null hypothesis were true.

$$p(\text{data} \,|\, \text{null hypothesis})$$

P-values do not quantify the probability of the null hypothesis given our data. Unfortunately, this is the quantity that we actually care about.

$$p(\text{null hypothesis} \,|\, \text{data})$$

My opinion: the use of P-values to reject hypotheses is predicated on the base rate fallacy

By convention $P < 0.05$, then one rejects null hypothesis, supposedly because $p(\text{null hypothesis} \,|\, \text{data})$ is small.

For this to make sense, one has to accept the base rate fallacy, i.e.,

$$p(\text{data} \,|\, \text{null hypothesis}) \approx p(\text{null hypothesis} \,|\, \text{data})$$

Whether or not this is true in a specific case depends on the prior odds,

$$p(\text{null hypothesis}),$$

which Frequentist statistics refuses to consider.

The misinterpretation of P-values reflects the Frequentist / Bayesian divide

Frequentist statistics (a.k.a. classical statistics) focuses on likelihood:

$$p(\text{data} \mid \text{hypothesis}).$$

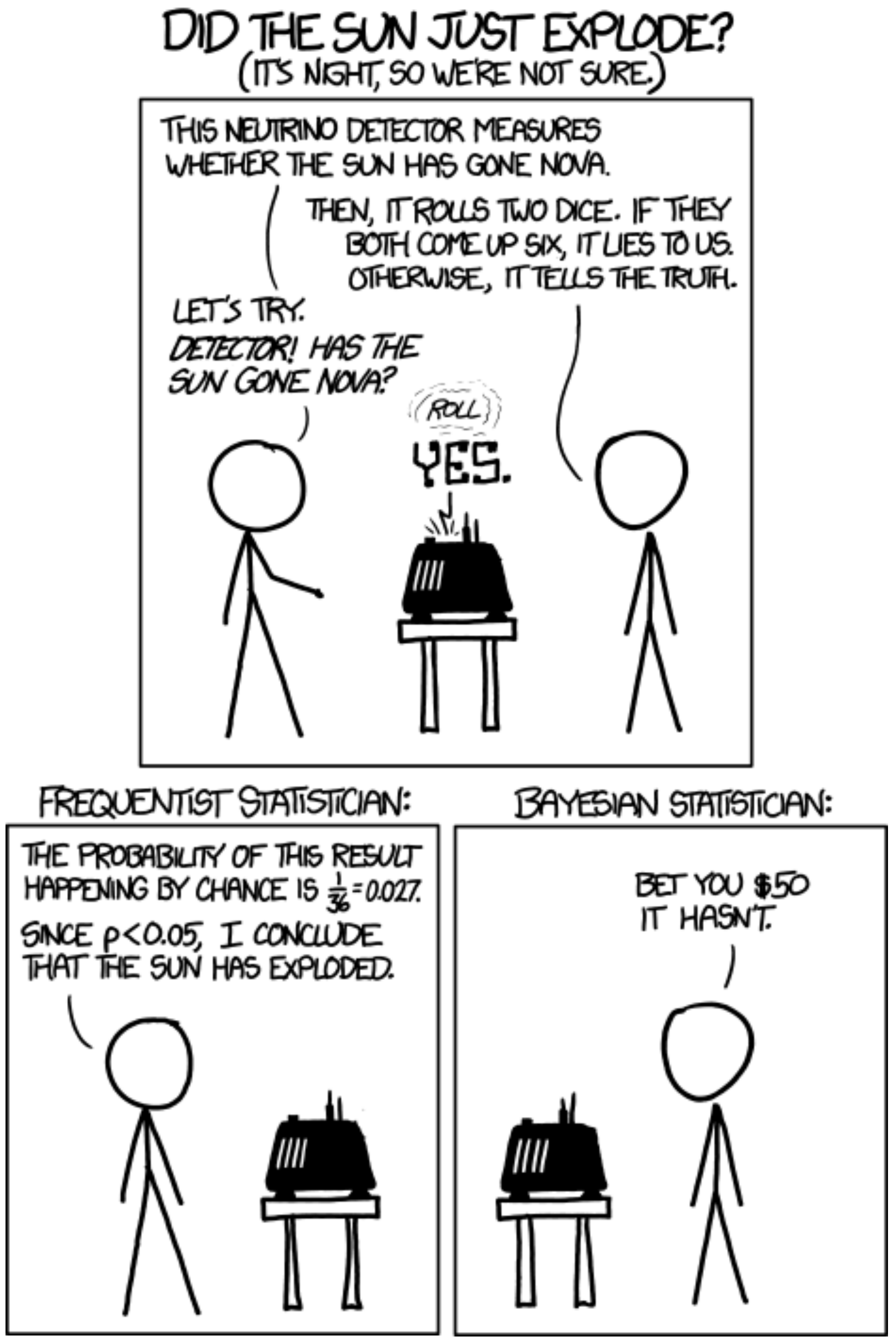
Iron Law of Frequentist Statistics:
Never compute the probability of a hypothesis.

Bayesian statistics focuses on computing posterior probabilities:

$$p(\text{hypothesis} \mid \text{data}).$$

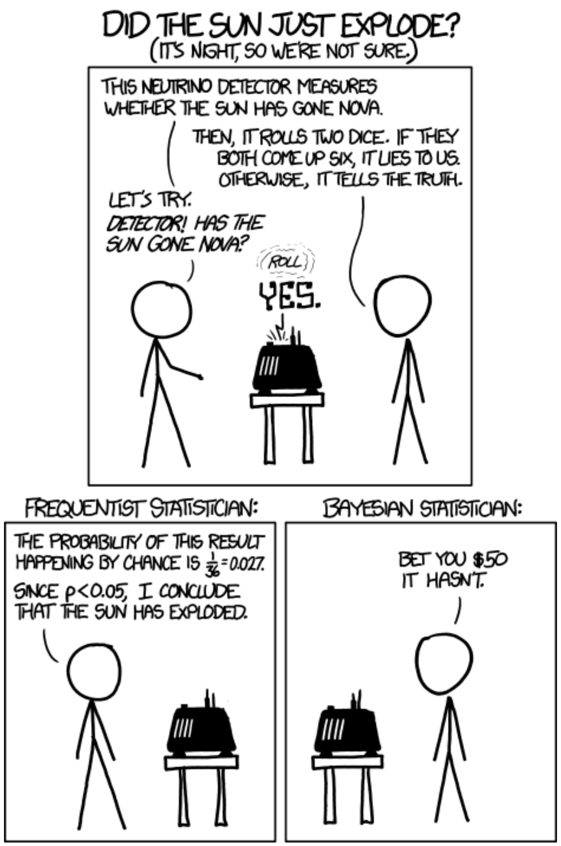
Example 3: Supernova detection machine

Exercise: supernova



https://xkcd.com/1132/

Exercise: supernova



https://xkcd.com/1132/

Bayes's theorem (from yesterday):

$$\frac{p(\text{nova}^+\,|\,\text{detector}^+)}{p(\text{nova}^-\,|\,\text{detector}^+)} = \frac{p(\text{detector}^+\,|\,\text{nova}^+)}{p(\text{detector}^+\,|\,\text{nova}^-)} \times \frac{p(\text{nova}^+)}{p(\text{nova}^-)}$$

$$\left[\frac{35/36}{1/36} = 35\right]$$

If our prior belief is that a supernova is very unlikely, i.e.

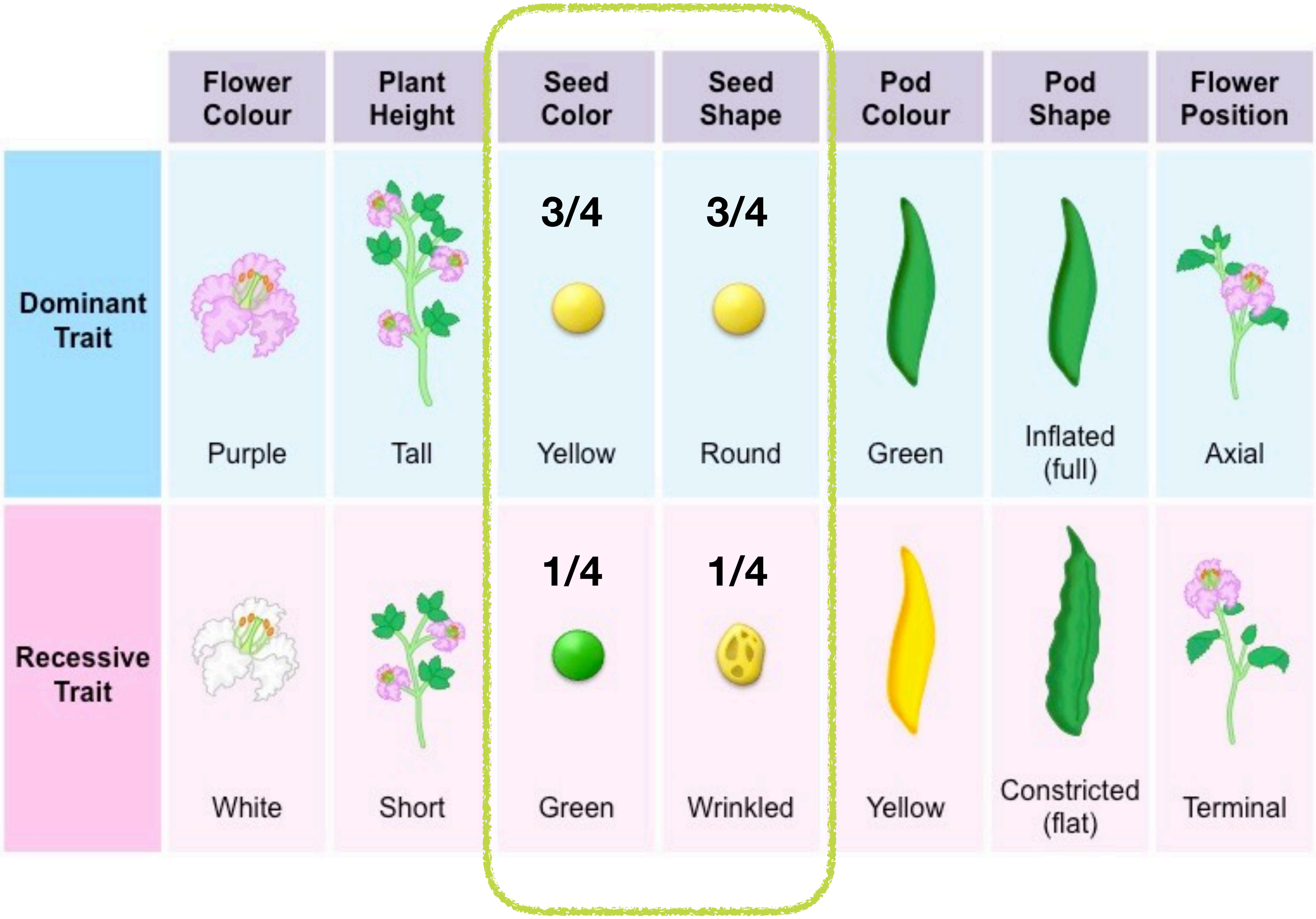
$$\frac{p(\text{nova}^+)}{p(\text{nova}^-)} \ll \frac{1}{35},$$

then we still shouldn't believe the sun has gone nova.

Even though, with a null hypothesis of nova^-,

$$\text{P value} = p(\text{detector}^+\,|\,\text{nova}^-) = \frac{1}{36} = 0.028 < 0.05$$

Example 4: Mendel’s Peas



https://ib.bioninja.com.au/standard-level/topic-3-genetics/34-inheritance/mendels-laws.html

Chi square test (known proportions)

Example: Mendel's peas

	observed	expected proportion	expected counts
Round & yellow	315	9/16	312.75
Round & green	108	3/16	104.25
Angular & yellow	101	3/16	104.25
Angular & green	32	1/16	34.75
Total	556	16/16	556.00

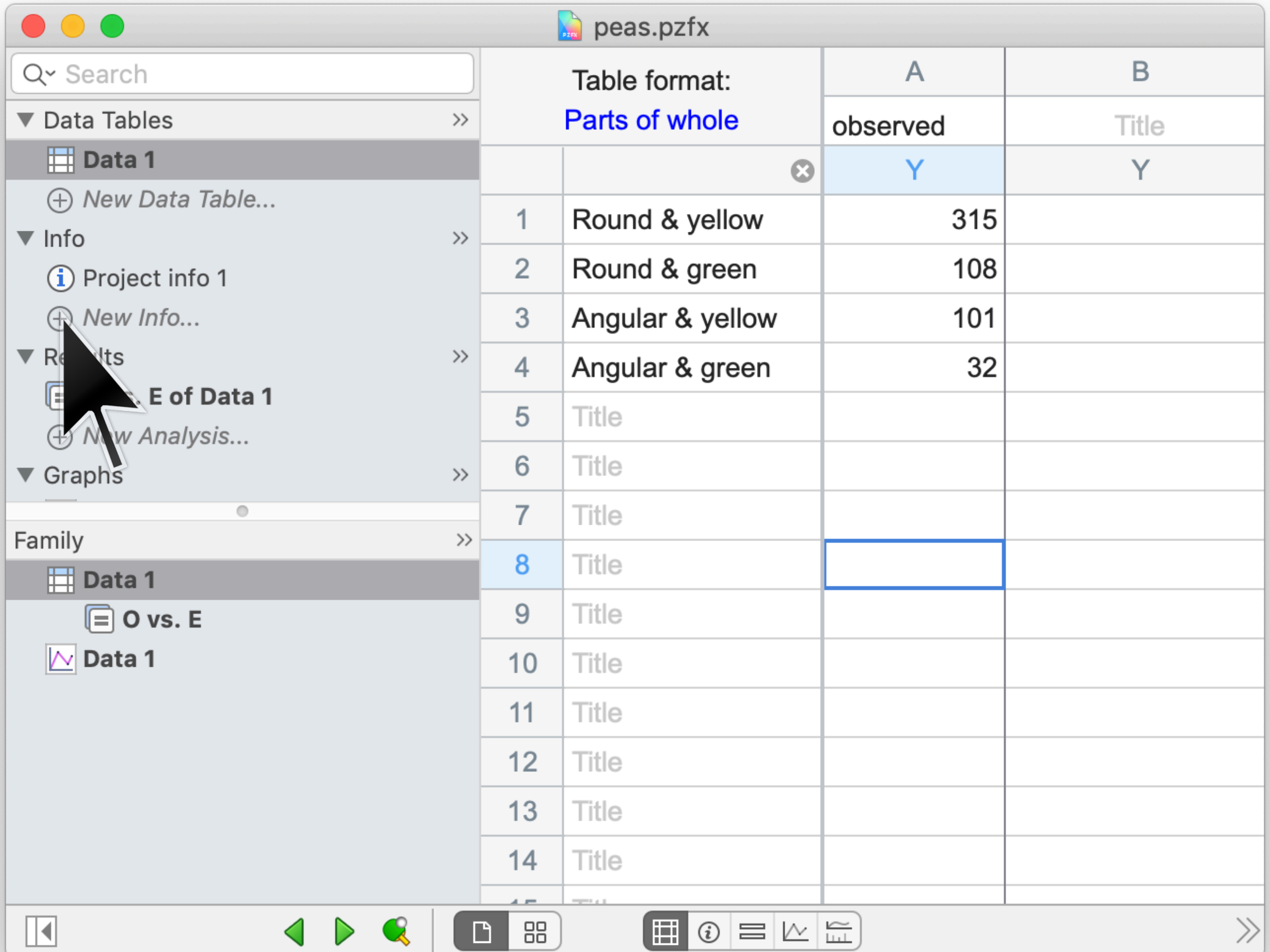
Null Hypothesis:

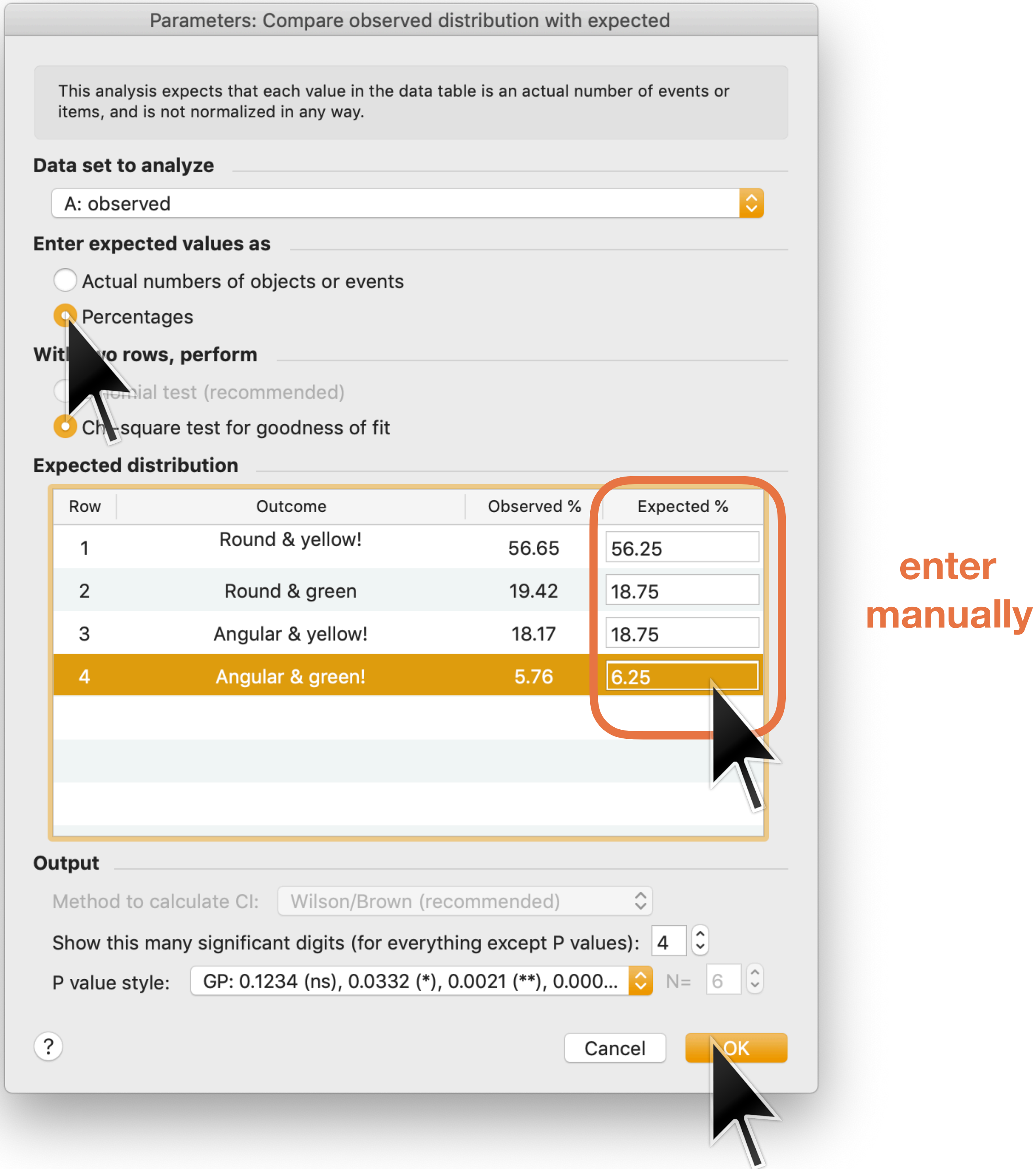
observations in $K = 4$ different categories occur in the expected proportions

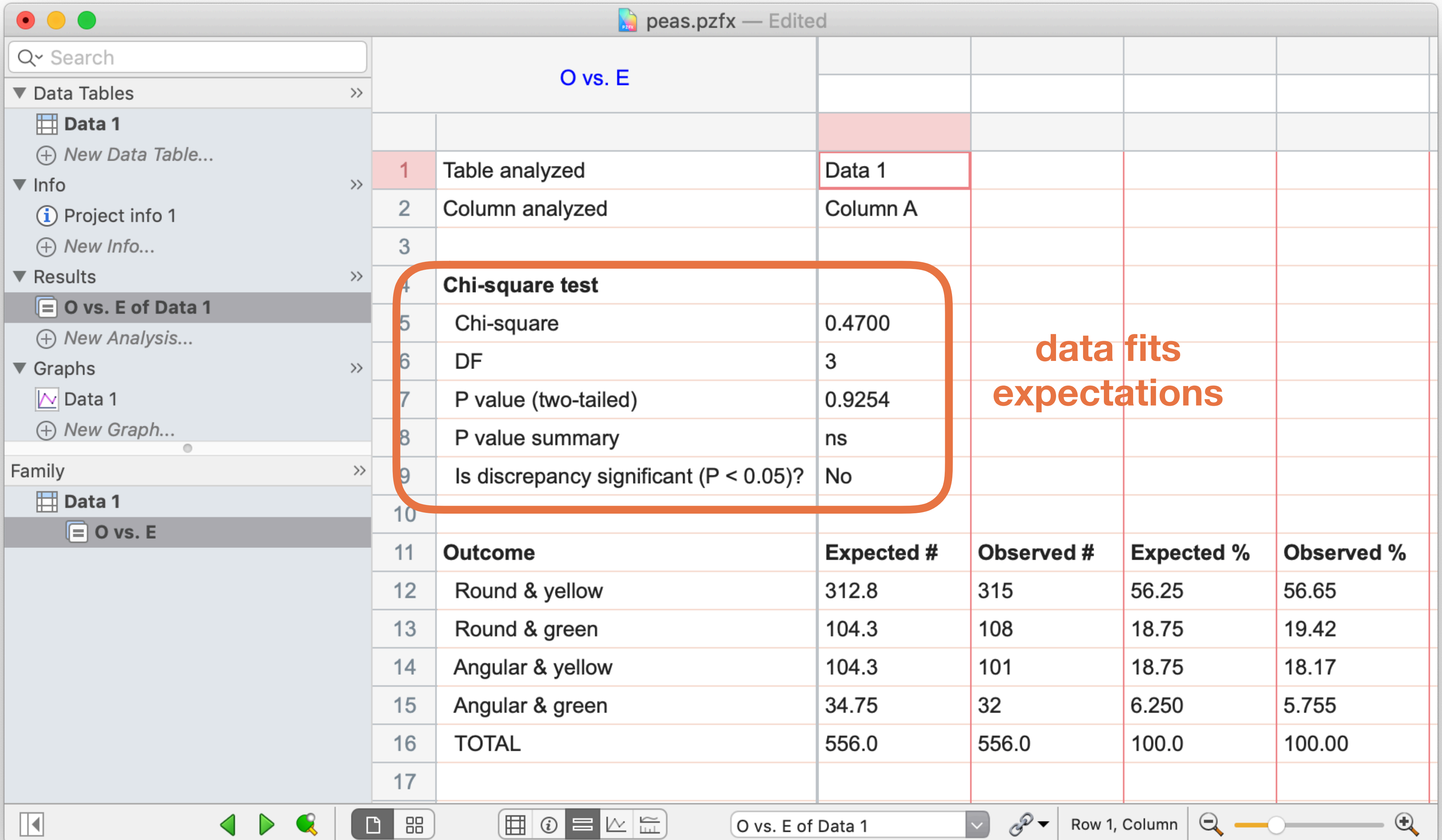
Data: number of observations in each category

Statistic: $\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$

Null distribution: Chi square distribution with $K - 1 = 3$ degrees of freedom (DOF)


peas.pzfx
Search
Data Tables
Data 1
New Data Table...
Info
Project info 1
New Info...
Results
E of Data 1
New Analysis...
Graphs
Family
Data 1
O vs. E
Data 1
Table format:
Parts of whole
A
B
observed
Title
Y
Y
1 Round & yellow 315
2 Round & green 108
3 Angular & yellow 101
4 Angular & green 32
5 Title
6 Title
7 Title
8 Title
9 Title
10 Title
11 Title
12 Title
13 Title
14 Title

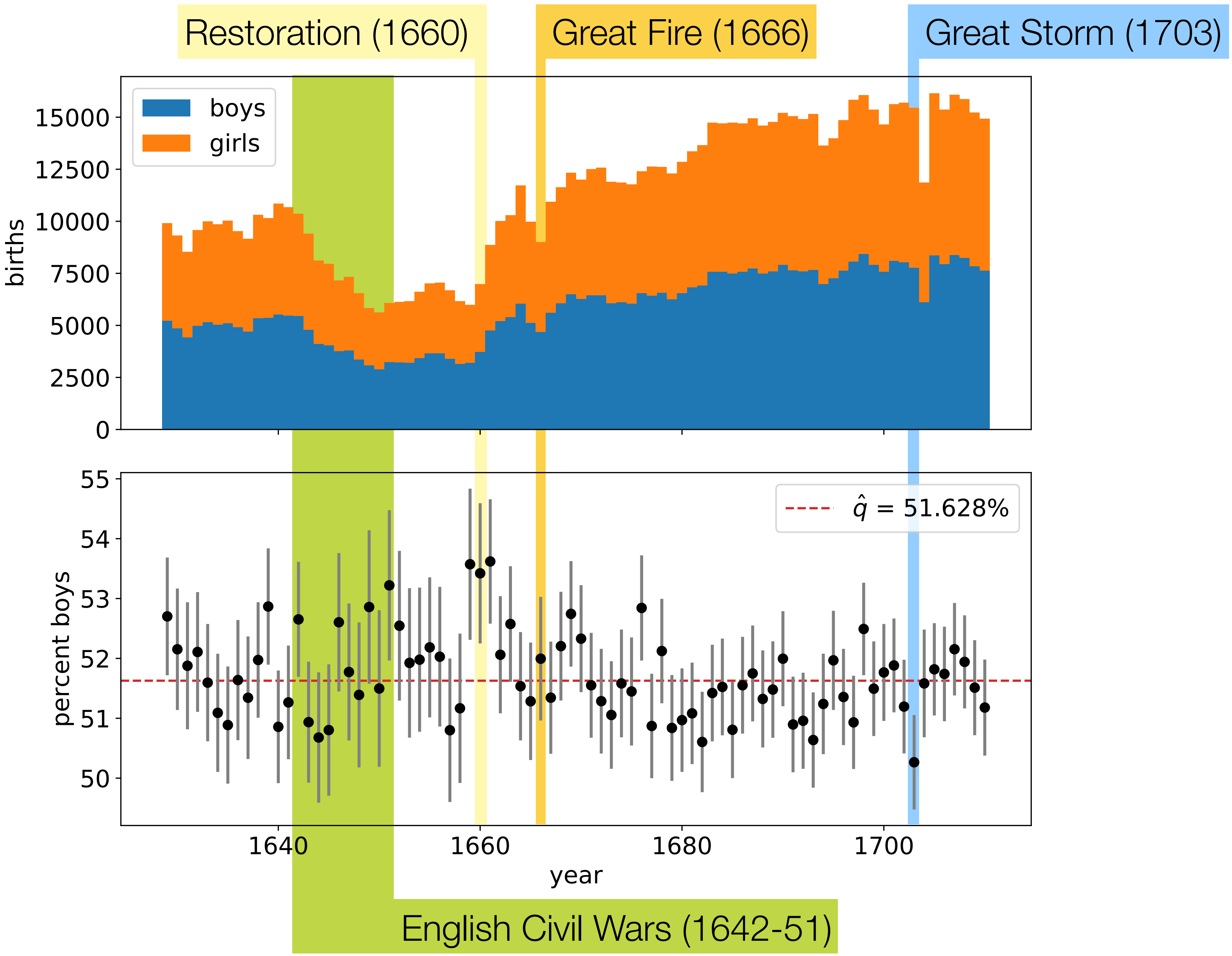

Parameters: Compare observed distribution with expected
This analysis expects that each value in the data table is an actual number of events or items, and is not normalized in any way.
Data set to analyze
A: observed
Enter expected values as
Actual numbers of objects or events
Percentages
With two rows, perform
Binomial test (recommended)
Chi-square test for goodness of fit
Expected distribution
Row
Outcome
Observed %
Expected %
1
Round & yellow!
56.65
56.25
2
Round & green
19.42
18.75
3
Angular & yellow!
18.17
18.75
4
Angular & green!
5.76
6.25
Output
Method to calculate CI:
Wilson/Brown (recommended)
Show this many significant digits (for everything except P values):
4
P value style:
GP: 0.1234 (ns), 0.0332 (*), 0.0021 (**), 0.000...
N=
6
?
Cancel
OK
enter manually



1	Table analyzed	Data 1			
2	Column analyzed	Column A			
3					
4	**Chi-square test**				
5	Chi-square	0.4700			
6	DF	3			
7	P value (two-tailed)	0.9254			
8	P value summary	ns			
9	Is discrepancy significant (P < 0.05)?	No			
10					
11	**Outcome**	**Expected #**	**Observed #**	**Expected %**	**Observed %**
12	Round & yellow	312.8	315	56.25	56.65
13	Round & green	104.3	108	18.75	19.42
14	Angular & yellow	104.3	101	18.75	18.17
15	Angular & green	34.75	32	6.250	5.755
16	TOTAL	556.0	556.0	100.0	100.00
17					

Example 4: Human sex ratio in London over time

Is it possible that the boy/girl ratio changes from year to year?



Chi square test (unknown proportions)

sex / **year**

	male	female
1629	5218	4683
1630	4858	4457
1631	4422	4102
1632	4994	4590
1633	5158	4839
1634	5035	4820
1635	5106	4928
1636	4917	4605
1637	4703	4457
1638	5359	4952

Null Hypothesis:

Two multi-category variables A and B are independent, i.e.,

$$p(A, B) = p(A) \cdot p(B)$$

Statistic:

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

Null distribution:

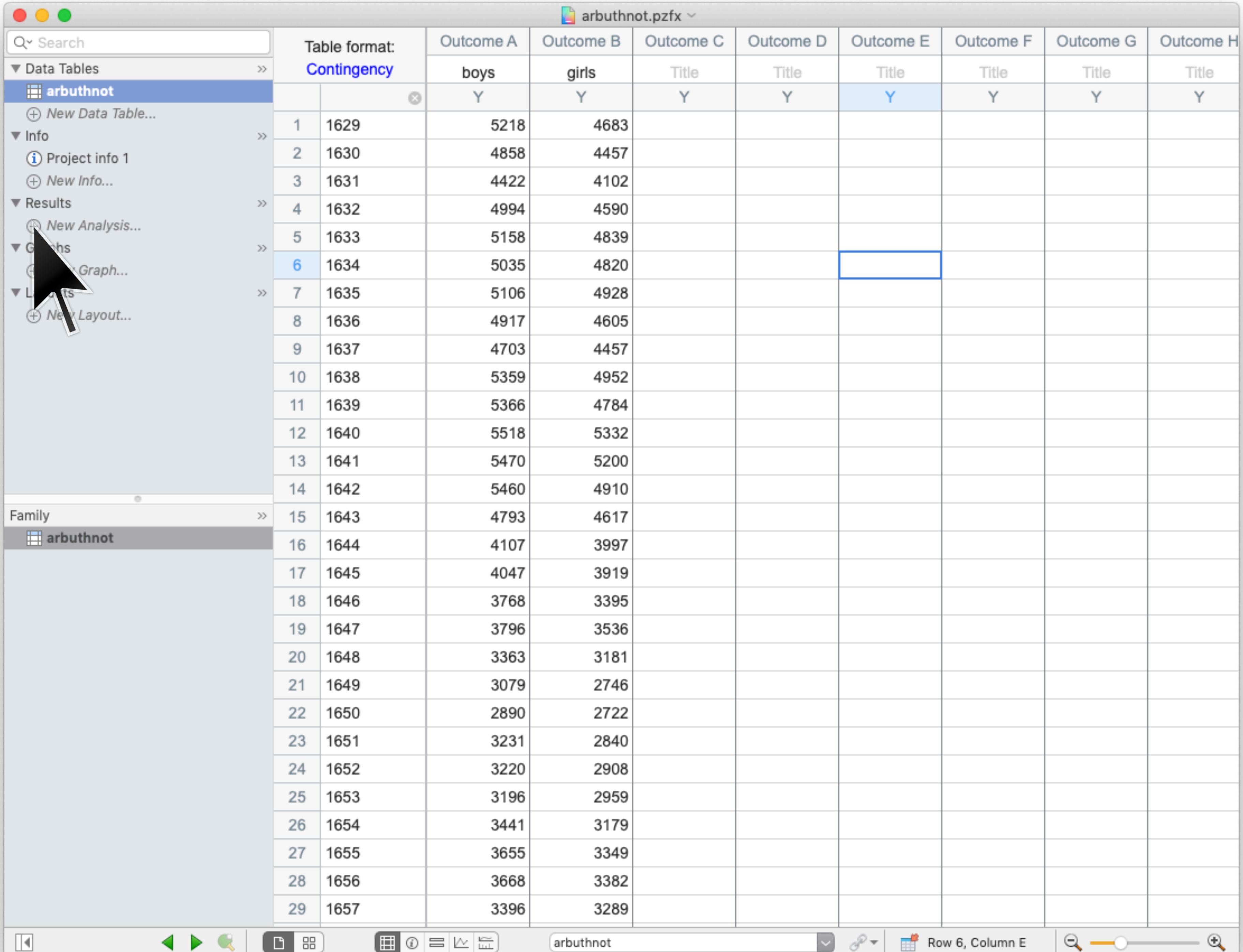
Chi square distribution with

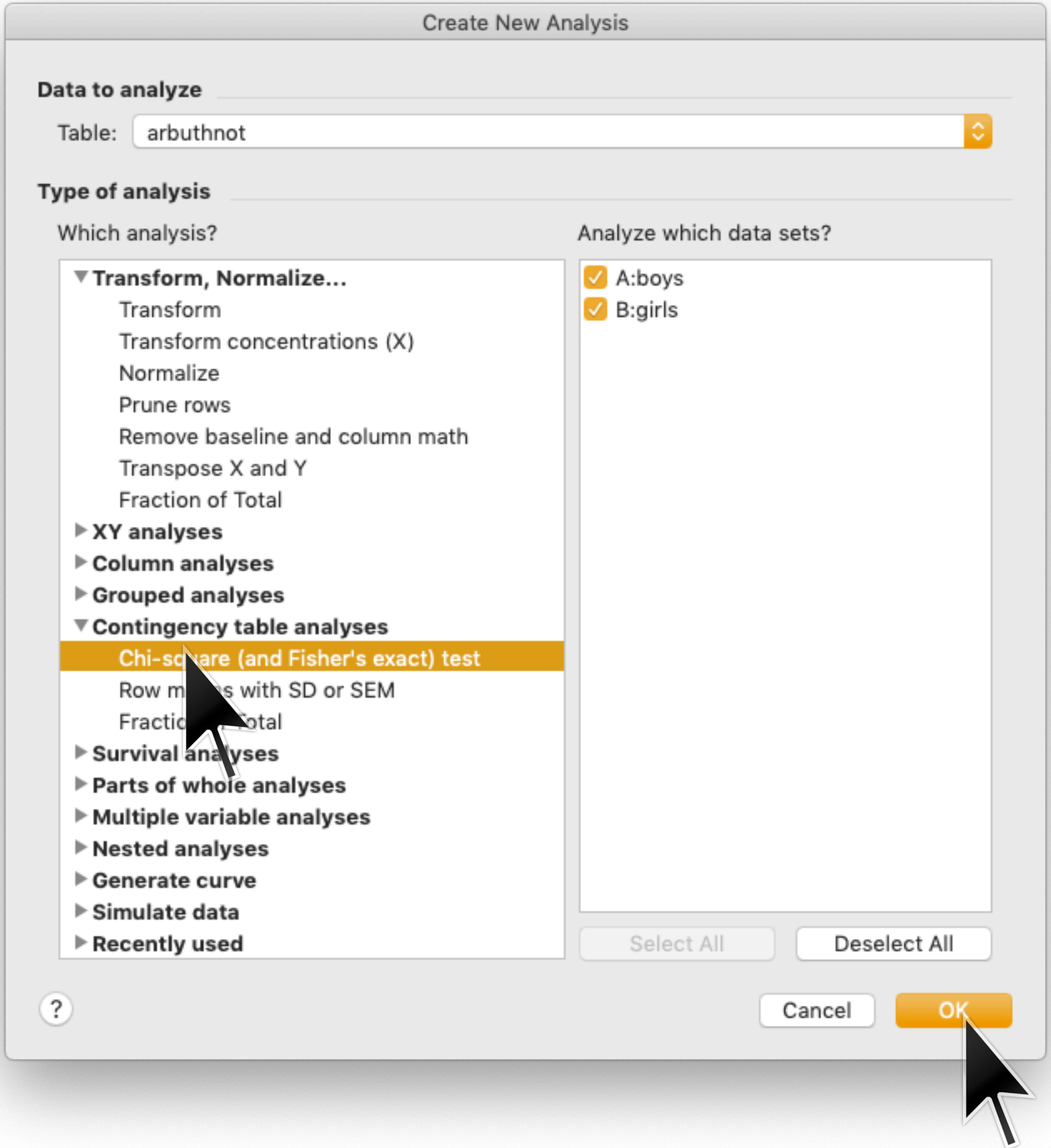
$$\text{DOF} = nm - m - n + 1$$

where

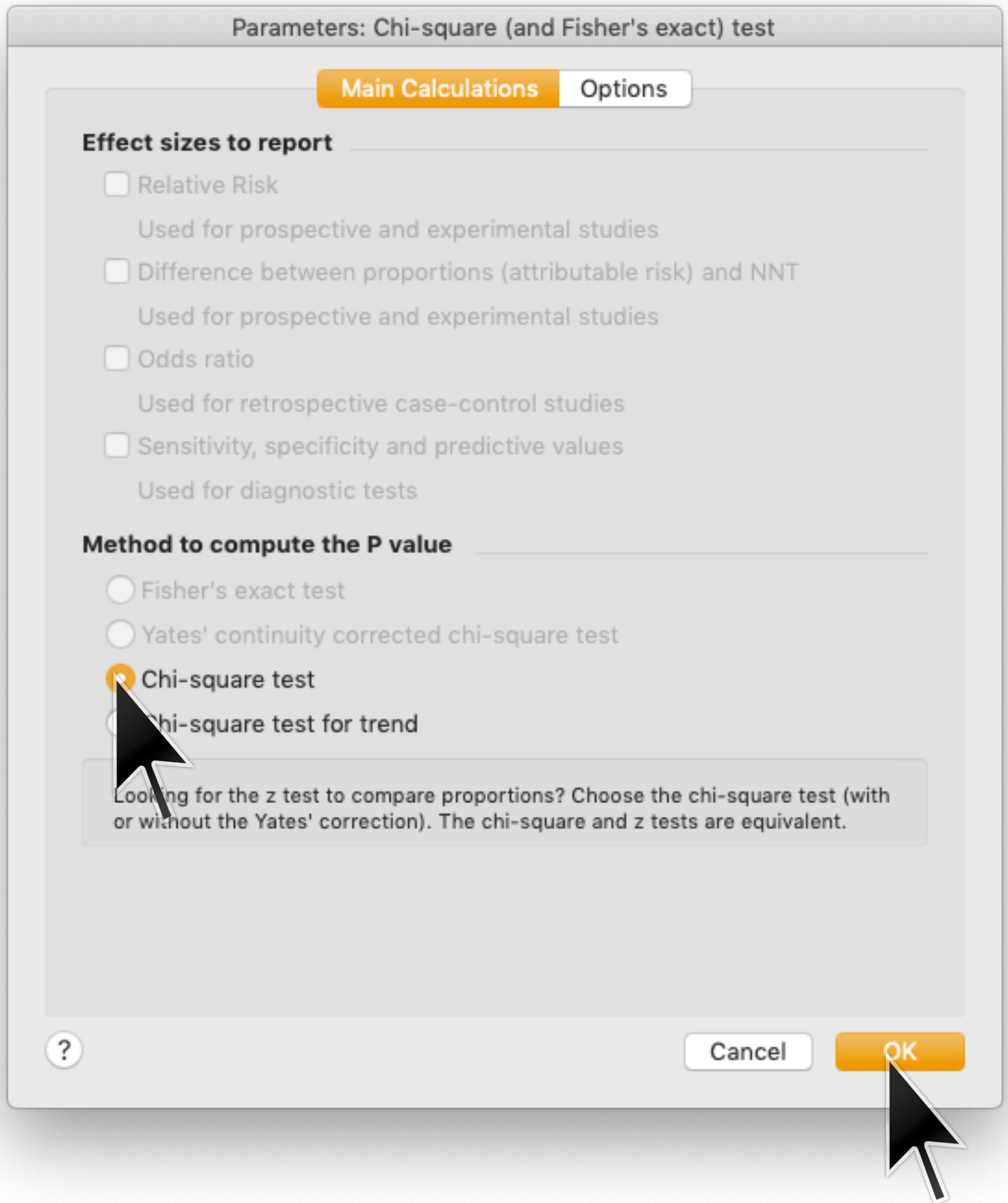
m = number of possible values for A

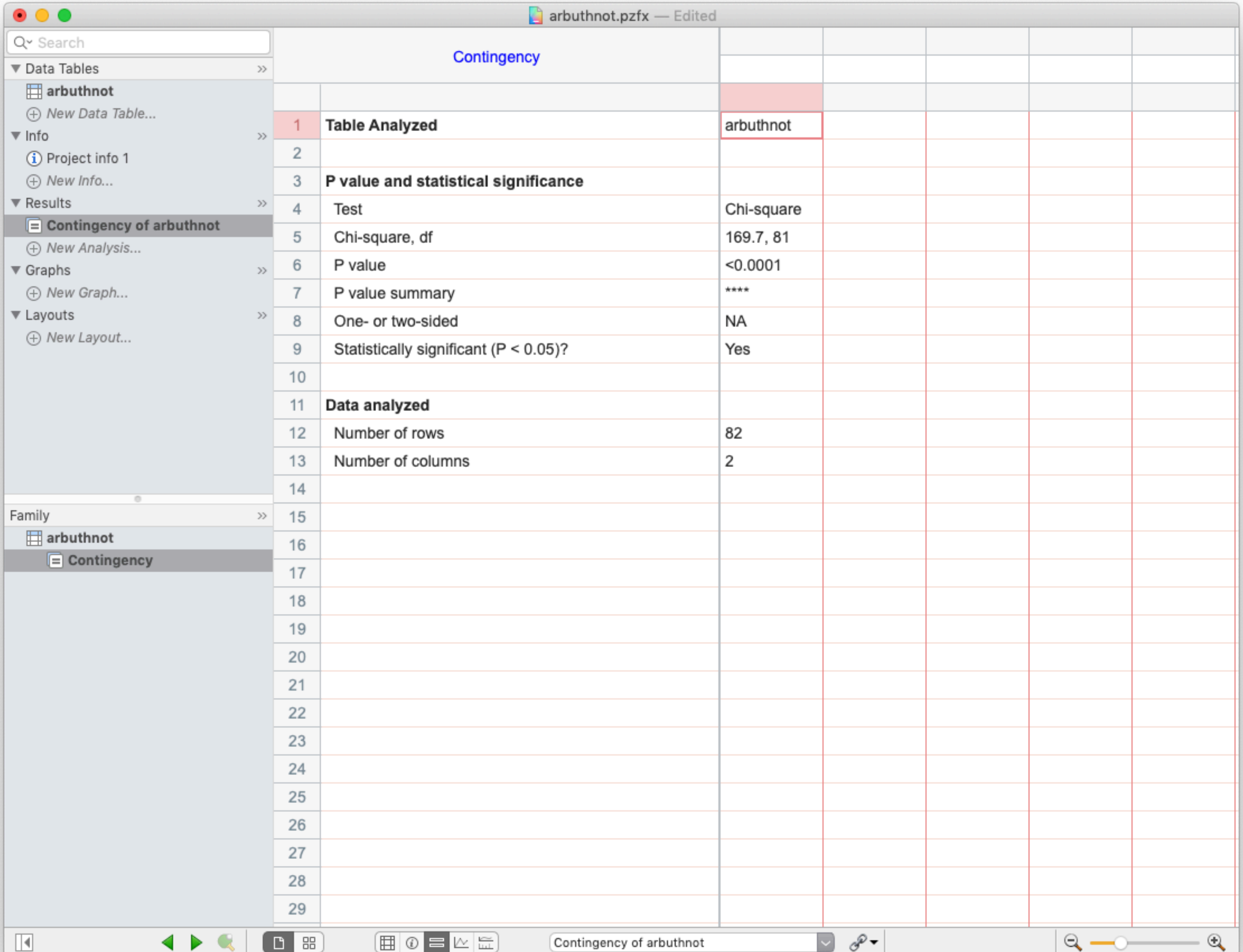
n = number of possible values or B


arbuthnot.pzfx
Search
Data Tables
arbuthnot
New Data Table...
Info
Project info 1
New Info...
Results
New Analysis...
Graph...
New Layout...
Family
arbuthnot
Table format:
Contingency
Outcome A
Outcome B
Outcome C
Outcome D
Outcome E
Outcome F
Outcome G
Outcome H
boys
girls
Title
Y
1 1629 5218 4683
2 1630 4858 4457
3 1631 4422 4102
4 1632 4994 4590
5 1633 5158 4839
6 1634 5035 4820
7 1635 5106 4928
8 1636 4917 4605
9 1637 4703 4457
10 1638 5359 4952
11 1639 5366 4784
12 1640 5518 5332
13 1641 5470 5200
14 1642 5460 4910
15 1643 4793 4617
16 1644 4107 3997
17 1645 4047 3919
18 1646 3768 3395
19 1647 3796 3536
20 1648 3363 3181
21 1649 3079 2746
22 1650 2890 2722
23 1651 3231 2840
24 1652 3220 2908
25 1653 3196 2959
26 1654 3441 3179
27 1655 3655 3349
28 1656 3668 3382
29 1657 3396 3289
arbuthnot
Row 6, Column E



Create New Analysis
Data to analyze
Table:
arbuthnot
Type of 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
(and Fisher's exact) test
with SD or SEM
Survival
Parts of
analyses
Multiple variable analyses
Nested analyses
Generate curve
Simulate data
Recently used
A:boys
B:girls
Select All
Deselect All
?
Cancel
OK


Parameters: Chi-square (and Fisher's exact) test
Main Calculations
Options
Effect sizes to report
Relative Risk
Used for prospective and experimental studies
Difference between proportions (attributable risk) and NNT
Used for prospective and experimental studies
Odds ratio
Used for retrospective case-control studies
Sensitivity, specificity and predictive values
Used for diagnostic tests
Method to compute the P value
Fisher's exact test
Yates' continuity corrected chi-square test
Chi-square test
hi-square test for trend
Loo ng for the z test to compare proportions? Choose the chi-square test (with or wi out the Yates' correction). The chi-square and z tests are equivalent.
?
Cancel
OK


arbuthnot.pzfx — Edited
Search
Data Tables
arbuthnot
New Data Table...
Info
Project info 1
New Info...
Results
Contingency of arbuthnot
New Analysis...
Graphs
New Graph...
Layouts
New Layout...
Family
arbuthnot
Contingency
Contingency
1 Table Analyzed arbuthnot
2
3 P value and statistical significance
4 Test Chi-square
5 Chi-square, df 169.7, 81
6 P value <0.0001
7 P value summary ****
8 One- or two-sided NA
9 Statistically significant (P < 0.05)? Yes
10
11 Data analyzed
12 Number of rows 82
13 Number of columns 2
Contingency of arbuthnot