Hypothesis Testing Details

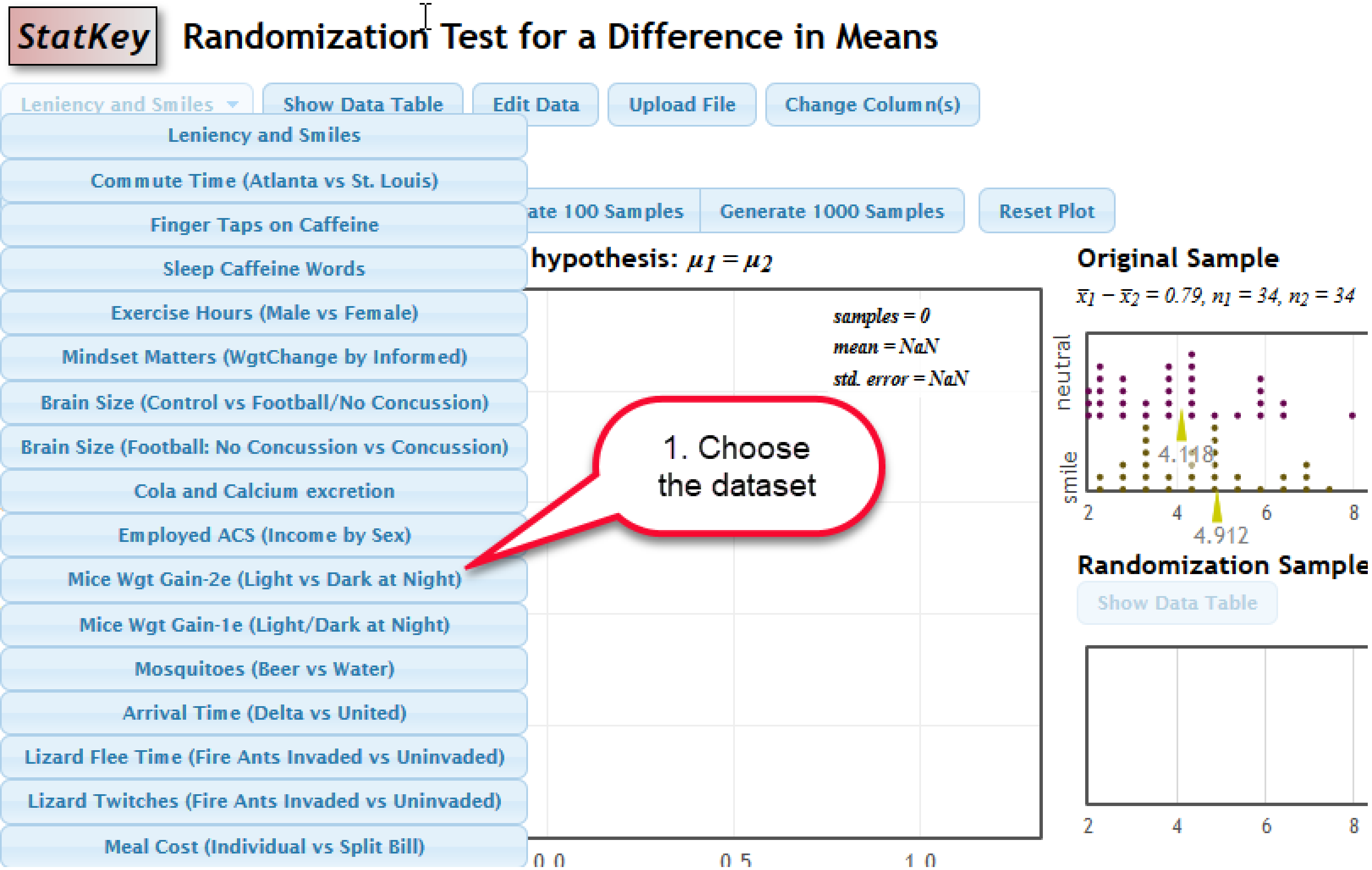
As we go through this example of comparing two means, we'll also address how to interpret the results.

Compare Two Means

Why is this particular type of test used?

- Illustrate a two-parameter test.
- Show / discuss different options for the randomization method.

Does light at night affect mice’s weight gain?


StatKey
Leniency and Smiles
Show Data Table
Edit Data
Upload File
Change Column(s)
Leniency and Smiles
Commute Time (Atlanta vs St. Louis)
Finger Taps on Caffeine
Sleep Caffeine Words
Exercise Hours (Male vs Female)
Mindset Matters (WgtChange by Informed)
Brain Size (Control vs Football/No Concussion)
Brain Size (Football: No Concussion vs Concussion)
Cola and Calcium excretion
Employed ACS (Income by Sex)
Mice Wgt Gain-2e (Light vs Dark at Night)
Mice Wgt Gain-1e (Light/Dark at Night)
Mosquitoes (Beer vs Water)
Arrival Time (Delta vs United)
Lizard Flee Time (Fire Ants Invaded vs Uninvaded)
Lizard Twitches (Fire Ants Invaded vs Uninvaded)
Meal Cost (Individual vs Split Bill)
ate 100 Samples
Generate 1000 Samples
Reset Plot
Original Sample
neutral
smile
4.118
4.912
2
4
6
8
Randomization Sample
Show Data Table
2
4
6
8
0.0
0.5
1.0

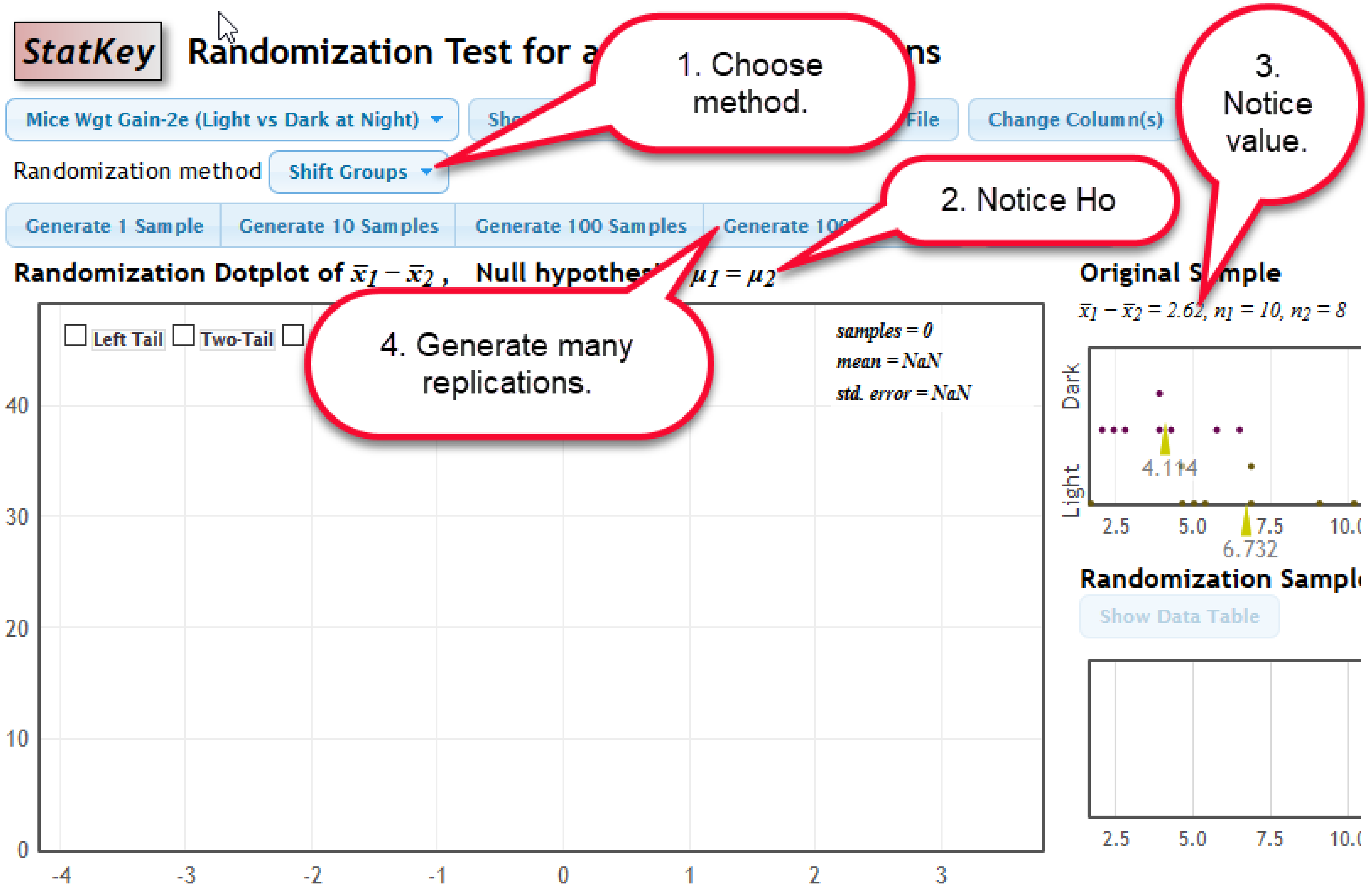
Randomization Test for a Difference in Means

hypothesis: $\mu_1 = \mu_2$

samples = 0
mean = NaN
std. error = NaN

$\bar{x}_1 - \bar{x}_2 = 0.79$, $n_1 = 34$, $n_2 = 34$

1. Choose the dataset



Randomization method: Shift Groups

Randomization Dotplot of $\bar{x}_1 - \bar{x}_2$, Null hypothes $\mu_1 = \mu_2$

samples = 0
mean = NaN
std. error = NaN

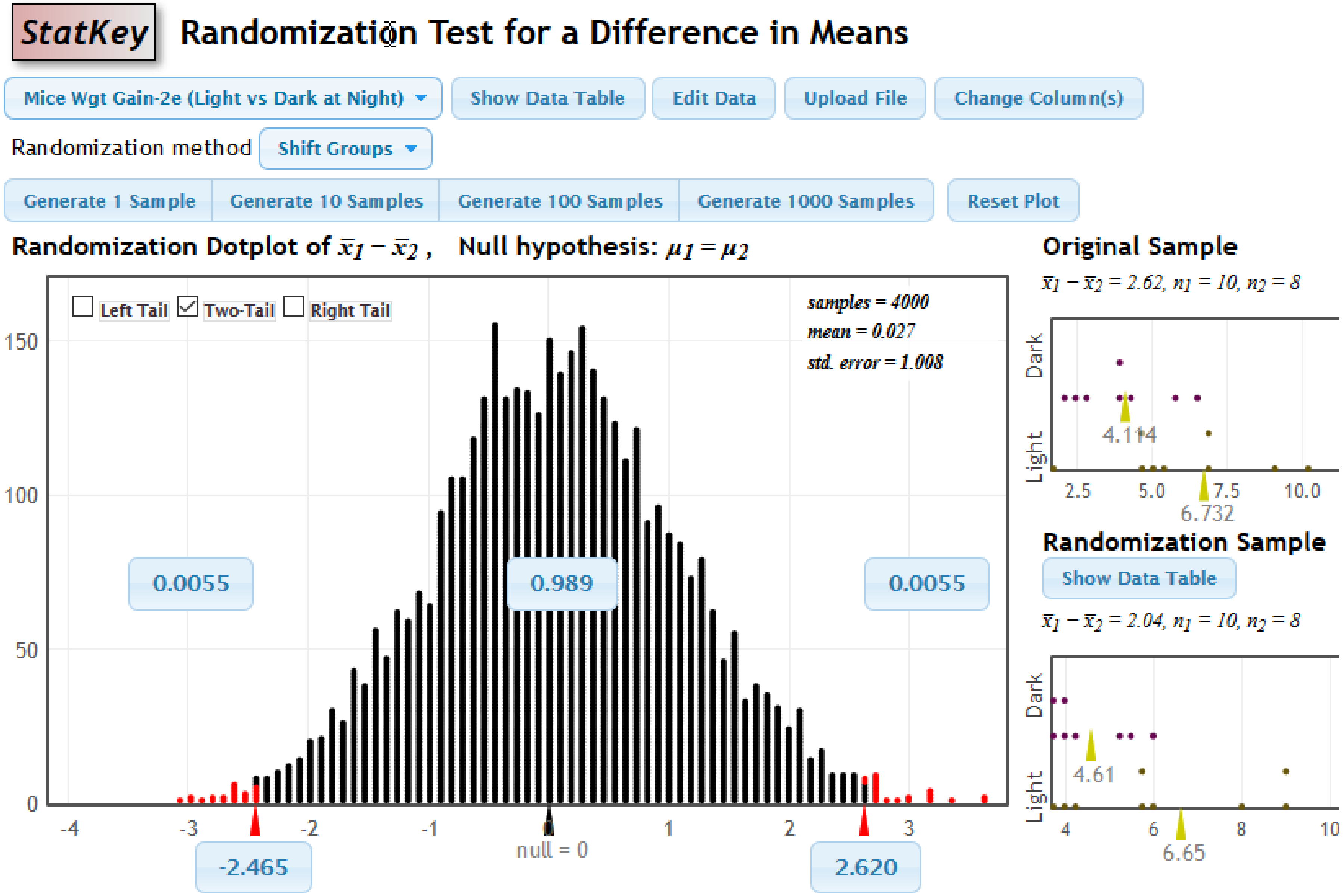
$\bar{x}_1 - \bar{x}_2 = 2.62$, $n_1 = 10$, $n_2 = 8$

1. Choose method.

2. Notice Ho

3. Notice value.

4. Generate many replications.



StatKey
Randomization Test for a Difference in Means
Mice Wgt Gain-2e (Light vs Dark at Night)
Show Data Table
Edit Data
Upload File
Change Column(s)
Randomization method
Shift Groups
Generate 1 Sample
Generate 10 Samples
Generate 100 Samples
Generate 1000 Samples
Reset Plot
Randomization Dotplot of $\bar{x}_1 - \bar{x}_2$, Null hypothesis: $\mu_1 = \mu_2$
□ Left Tail ☑ Two-Tail □ Right Tail
samples = 4000
mean = 0.027
std. error = 1.008
150
100
50
0
0.0055
0.989
0.0055
-4
-3
-2
-1
1
2
3
null = 0
-2.465
2.620
Original Sample
$\bar{x}_1 - \bar{x}_2 = 2.62, n_1 = 10, n_2 = 8$
Dark
Light
4.114
6.732
2.5
5.0
7.5
10.0
Randomization Sample
Show Data Table
$\bar{x}_1 - \bar{x}_2 = 2.04, n_1 = 10, n_2 = 8$
Dark
Light
4.61
6.65
4
6
8
10

Light at Night: Steps

1. Chose a randomization method (Shift groups)
2. Generated MANY replications (4000).
3. Chose to do a two-tailed test.
4. Put the value of the test statistic (difference of the two sample means) on the horizontal axis.
5. Read the p-value to be 2(0.0055)=0.011

Light at Night: Interpretation of result

The randomization dist’n here shows the values of the sample statistic IF the null hypothesis (Ho) is true.

The actual data we observed is quite far out in the tails of this dist’n, so it is very unlikely to happen if Ho is true.

We summarize that comment as …

Light at Night: Conclusion

The data give quite strong evidence (p=0.011) that there is a difference in average weight gain between mice who are exposed to light at night and those who were not.

Light at Night:Conclusion (Statistical Significance)

In some situations, a decision must be made, such as
Quit using light at night or do not quit.

1. Choose a cut-off value for how small the p-value has to be to decide to act on the alternative hypothesis.
2. Decide: Reject Ho and act on Ha if the p-value is smaller than that cut-off.

Several issues can be discussed here

1. Why was this a two-sided test?
2. How would we choose a cut-off value on which to act?
3. Why choose this method of randomization?

1. Why was this a two-sided test?

- I approached this dataset without an opinion about whether or how a change in the light conditions for mice would result in a change in their weight gain. (Possibly the researcher had an opinion and did a one-sided test.)
- Without a reason to believe one direction or another, it was most reasonable to simply test whether there is difference.

2. How would one choose a cut-off value? 1

- When making a decision to DO something, there are always two possible errors that can be made, with different consequences.
    - Error: Reject Ho when Ho true. (Called a Type I error.)
    - Error: Fail to reject Ho when Ha is true. (Called a Type II error.)

2. How would one choose a cut-off value? 2

- There is a “trade-off” between making the two possible types of error.

- The appropriate way to choose a cut-off value is for the person who will eventually make the decision
to make a judgment about how to balance the negative impact of the consequences of each type of error.

2. How would one choose a cut-off value? 3

- In a given field, there may be a customary cut-off level used, based on judgment of many people about the relative “cost” of the consequences.
- This “cut-off” value for the p-value is called the “significance level.”

Why did I choose this randomization method? 1

An appropriate randomization sample

- Must use the data in the original sample (to preserve the variability.)
- Must be consistent with the null hypothesis.
- Should reflect the way the original data was collected.

Why did I use this randomization method? 2

- First, in most situations any of the three methods will give similar results.
- Having a choice mainly illustrates that there is not just one way to do it.
- Here, I chose this method because, if the "treatment" (light exposure) changes not only the mean but the variability in weight gain, this method preserves the different variability in the different groups.

Student Smokers

The dataset is “Student Survey”

Consider the proportions of each sex who smoke?

H_o:p_m= p_f
H_A:p_m ≠ p_f

We want to use

Hypothesis Test > Difference of two proportions,
but there is no option to open a data file in that tool.

Student Smokers: Finding the data

1. Find the description of the data file in www.lock5stat.com/Datasets (pdf file at the end)
2. Find the datafile itself in the alphabetical list of files at that same link.

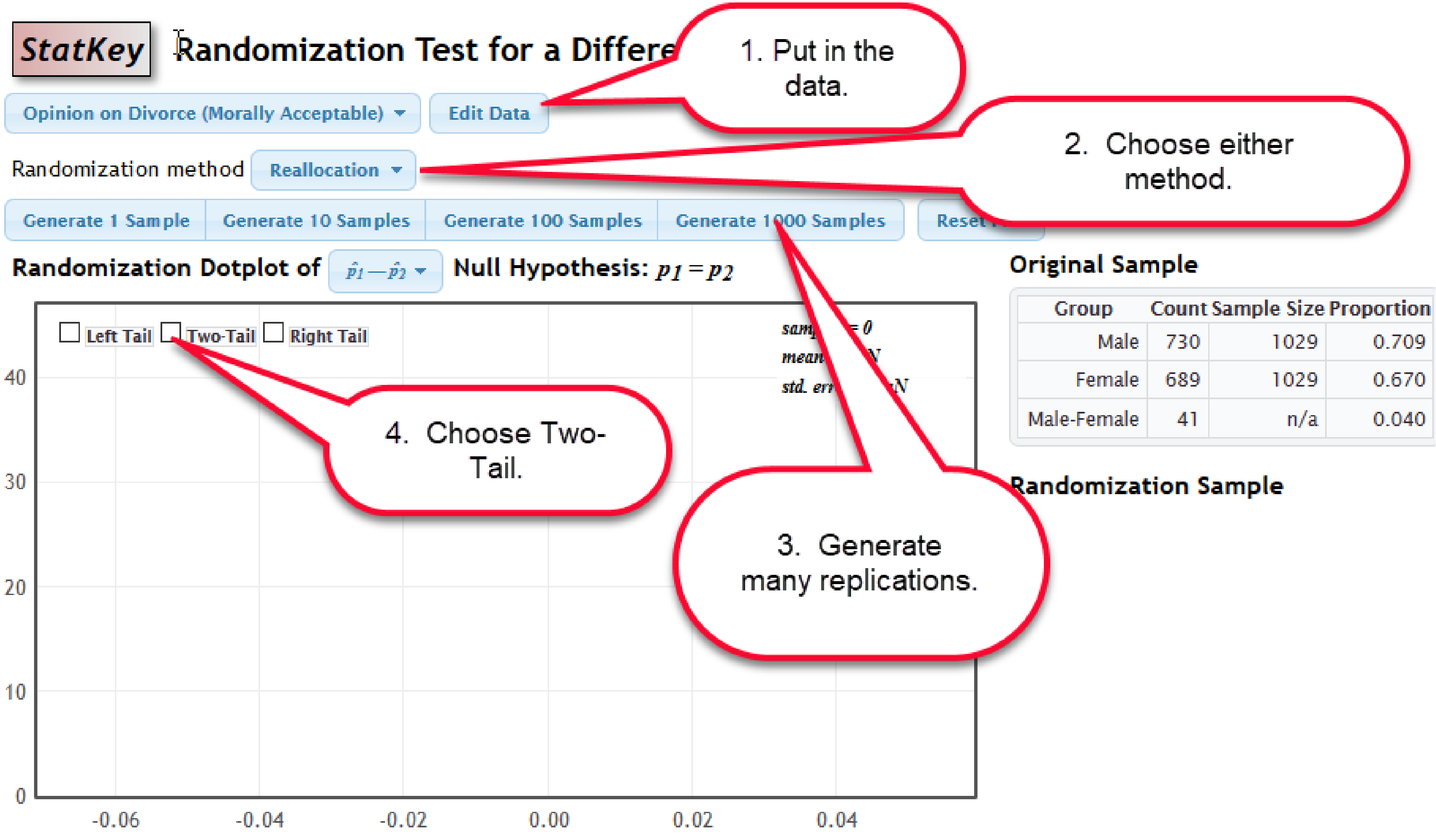
Instead use

- Descriptive Statistics and Graphs > Two Categorical Variables
- and open the data file Student Survey to obtain the counts and sample sizes for the males and females.

- Data: 27 out of 193 males are smokers and 16 out of 169 females are smokers.

Student Smokers: Find the p-value

1. Chose either randomization method
2. Generated MANY replications (5000).
3. Chose to do a two-tailed test.
4. Put the value of the test statistic (difference of the two sample proportions) on the horizontal axis.
5. Read the p-value to be 2(0.123)=0.246



Randomization Test for a Differe

Randomization method Reallocation

Randomization Dotplot of $\hat{p}_1 - \hat{p}_2$ Null Hypothesis: $p_1 = p_2$

1. Put in the data.

2. Choose either method.

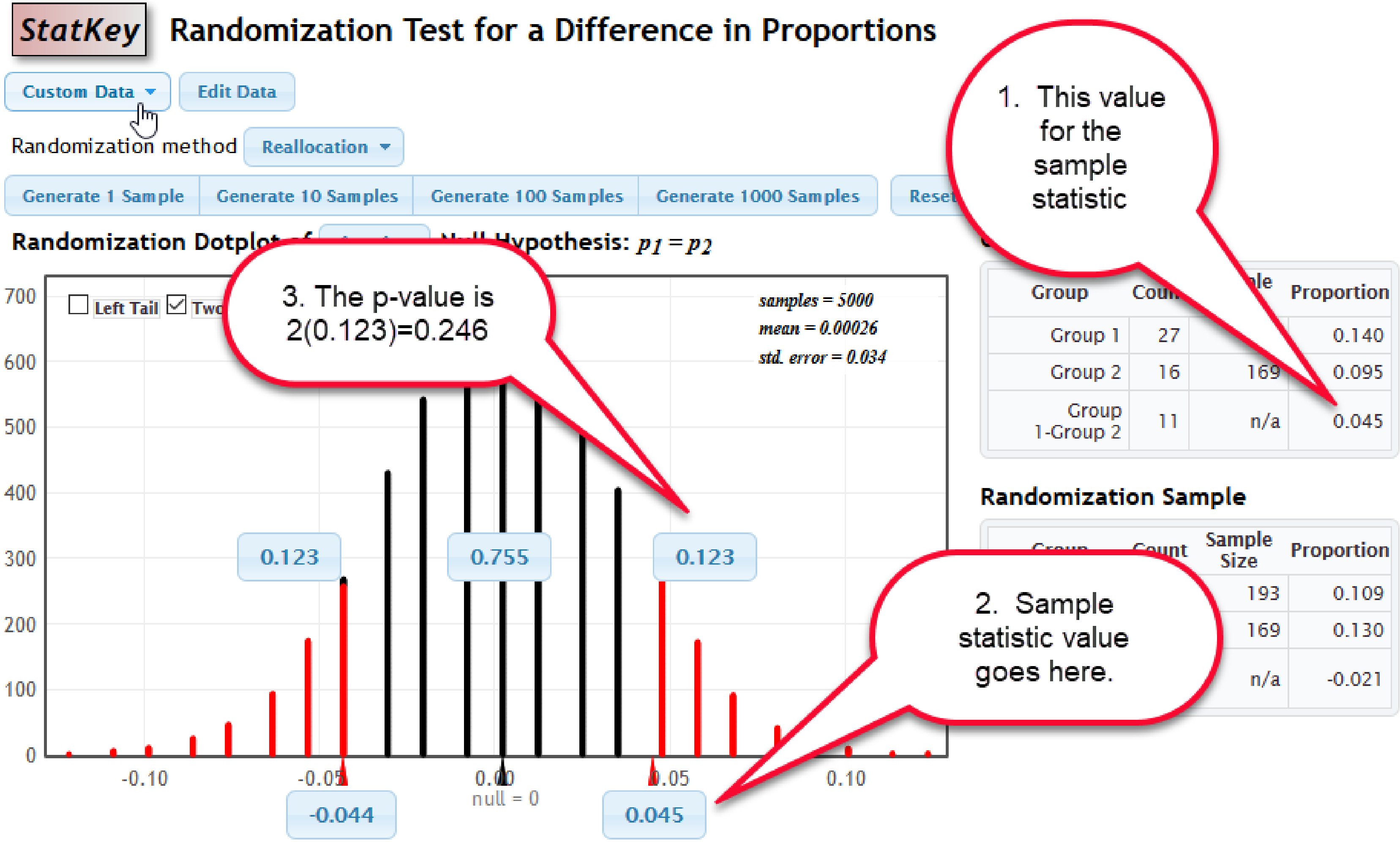
3. Generate many replications.

4. Choose Two-Tail.

Original Sample

Group	Count	Sample Size	Proportion
Male	730	1029	0.709
Female	689	1029	0.670
Male-Female	41	n/a	0.040

Randomization Sample



StatKey Randomization Test for a Difference in Proportions

Randomization method Reallocation

Randomization Dotplot of Null Hypothesis: $p_1 = p_2$

□ Left Tail ☑ Two

samples = 5000
mean = 0.00026
std. error = 0.034

Group	Coun	le	Proportion
Group 1	27		0.140
Group 2	16	169	0.095
Group 1-Group 2	11	n/a	0.045

Randomization Sample

Group	Count	Sample Size	Proportion
		193	0.109
		169	0.130
		n/a	-0.021

1. This value for the sample statistic

2. Sample statistic value goes here.

3. The p-value is 2(0.123)=0.246

Student Smokers: Interpretation of result

The randomization dist’n here shows the values of the sample statistic IF the null hypothesis (Ho) is true.

The actual data we observed is not at all far out in the tails of this dist’n, so it is reasonably likely to happen if Ho is true.

We summarize that comment as …

Student Smokers: Conclusion

The p-value for this hypothesis test is 0.246, which is not at all small.

Data values this extreme could easily happen by chance alone if there is no real difference in the population proportions of smokers.

Why did I choose to include this example?

- Putting the appropriate proportion data into the right places is usually somewhat confusing.
  - The software calls for “summary” data for proportions,
where we simply put in a column of the individual values for quantitative variables.

Datasets that “don’t fit usual conditions”

We’ll look at several datasets that do not fit the conditions to use our usual theoretical dist’ns to form Confidence Intervals.

We’ll see how, in some cases, it is reasonable to form a Bootstrap interval.

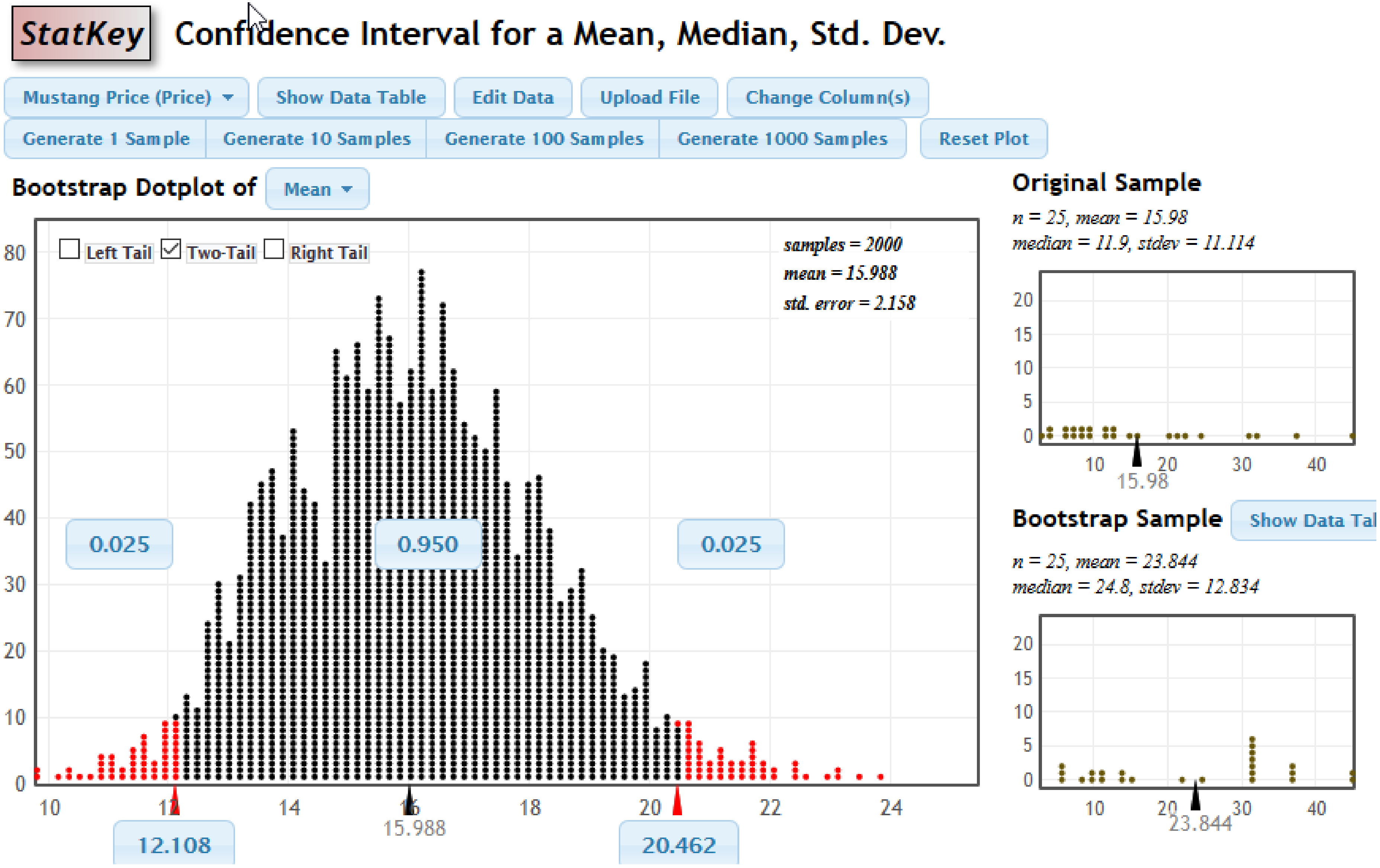
The theoretical models assume symmetry, but the bootstrap method of forming Confidence Intervals is just as useful for distributions with some deviations from symmetry.

Example 1.   Mustang Price dataset

Find a bootstrap CI for the mean of the Mustang Price data.

(The data is already in the drop-down menu.)

Note that the sample size is under 30 and the data distribution is strongly right-skewed.



StatKey Confidence Interval for a Mean, Median, Std. Dev.

Bootstrap Dotplot of Mean

samples = 2000
mean = 15.988
std. error = 2.158

Original Sample

n = 25, mean = 15.98
median = 11.9, stdev = 11.114

Bootstrap Sample

n = 25, mean = 23.844
median = 24.8, stdev = 12.834

Mustang Price CI

It is reasonable to summarize this bootstrap dist’n by a confidence interval with the middle 95% of the dist’n.

It is just not symmetric.

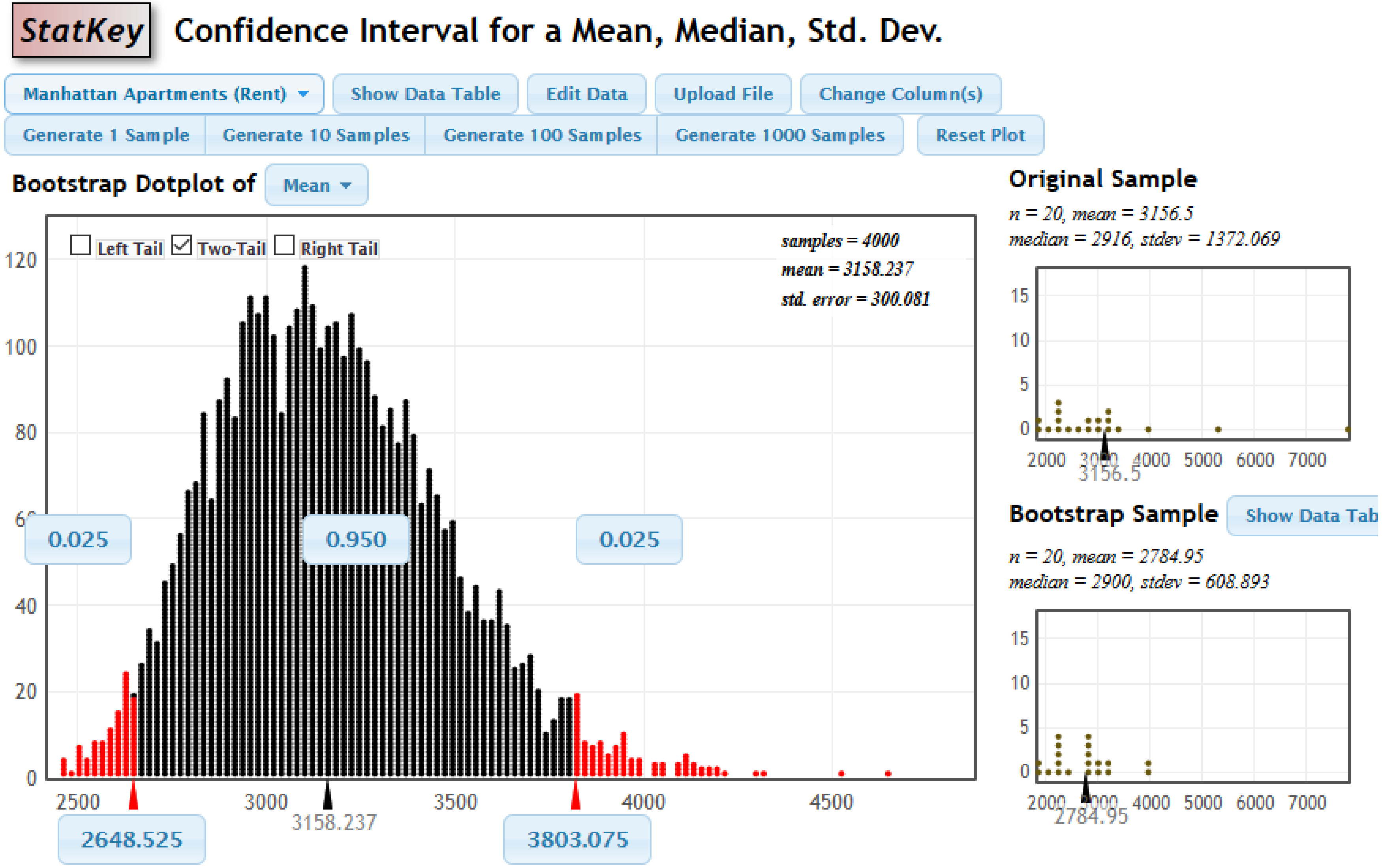
> (Notice the bootstrap sample shown. I made that one show by clicking on the dot with the largest value in the bootstrap dist’n. That is the bootstrap sample that produced the dot.)

Example 2. Manhattan Apts. Rent dataset

Find a bootstrap CI for the mean of the Manhattan Apts. Rent dataset.

(The data is already in the drop-down menu.)

Note that the sample size is under 30 and the data distribution is even more strongly right-skewed than the Mustang Price data.



Confidence Interval for a Mean, Median, Std. Dev.

Bootstrap Dotplot of Mean

samples = 4000
mean = 3158.237
std. error = 300.081

Original Sample

n = 20, mean = 3156.5
median = 2916, stdev = 1372.069

Bootstrap Sample

n = 20, mean = 2784.95
median = 2900, stdev = 608.893

Manhattan Apts. Rent CI for mean

Even though this dataset has several much more extreme outliers than the Mustang Price dataset, and is a fairly small sample size, the bootstrap dist'n still looks reasonable for summarizing with a confidence interval.

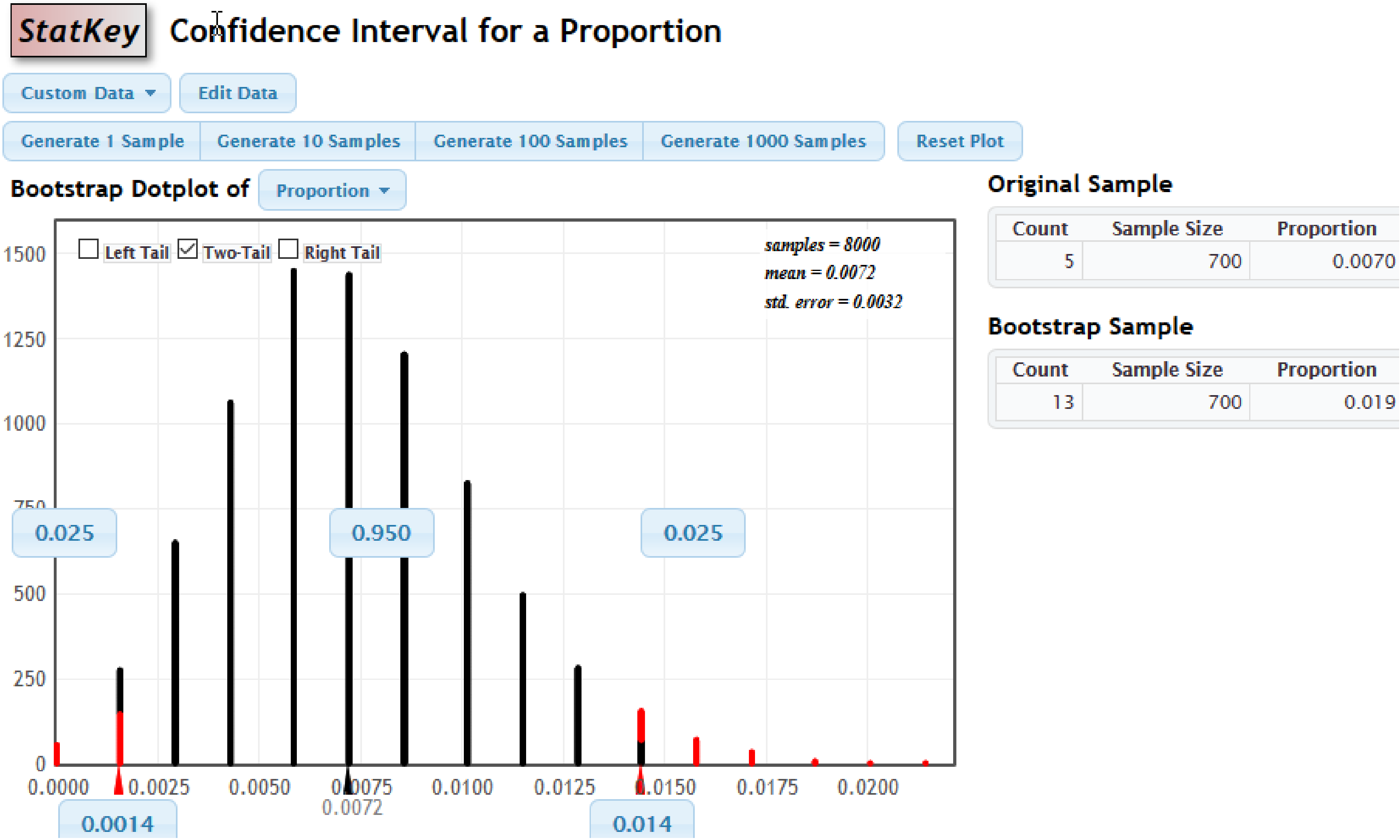
This time, I put my cursor on a dot at the far left of the bootstrap distribution to see the sample it came from.

Explore!!

Example 3. Very Small Proportion dataset

Our dataset has 5 successes in 700 trials.

This does not meet the conditions for using our usual normal approximation to the binomial dist'n to from a confidence interval for the population proportion.



samples = 8000
mean = 0.0072
std. error = 0.0032

Original Sample

Count	Sample Size	Proportion
5	700	0.0070

Bootstrap Sample

Count	Sample Size	Proportion
13	700	0.019

Example 3. CI for Very Small Proportion

The dist’n is clearly skewed, but it is reasonable enough to think of using it to form a confidence interval for p.

Other Statistics

Median

We use this as a descriptive statistic. Why don't we find confidence intervals for pop'n medians very much?

Standard Deviation

Have you thought about the actual sampling dist'n of the sample standard deviation?

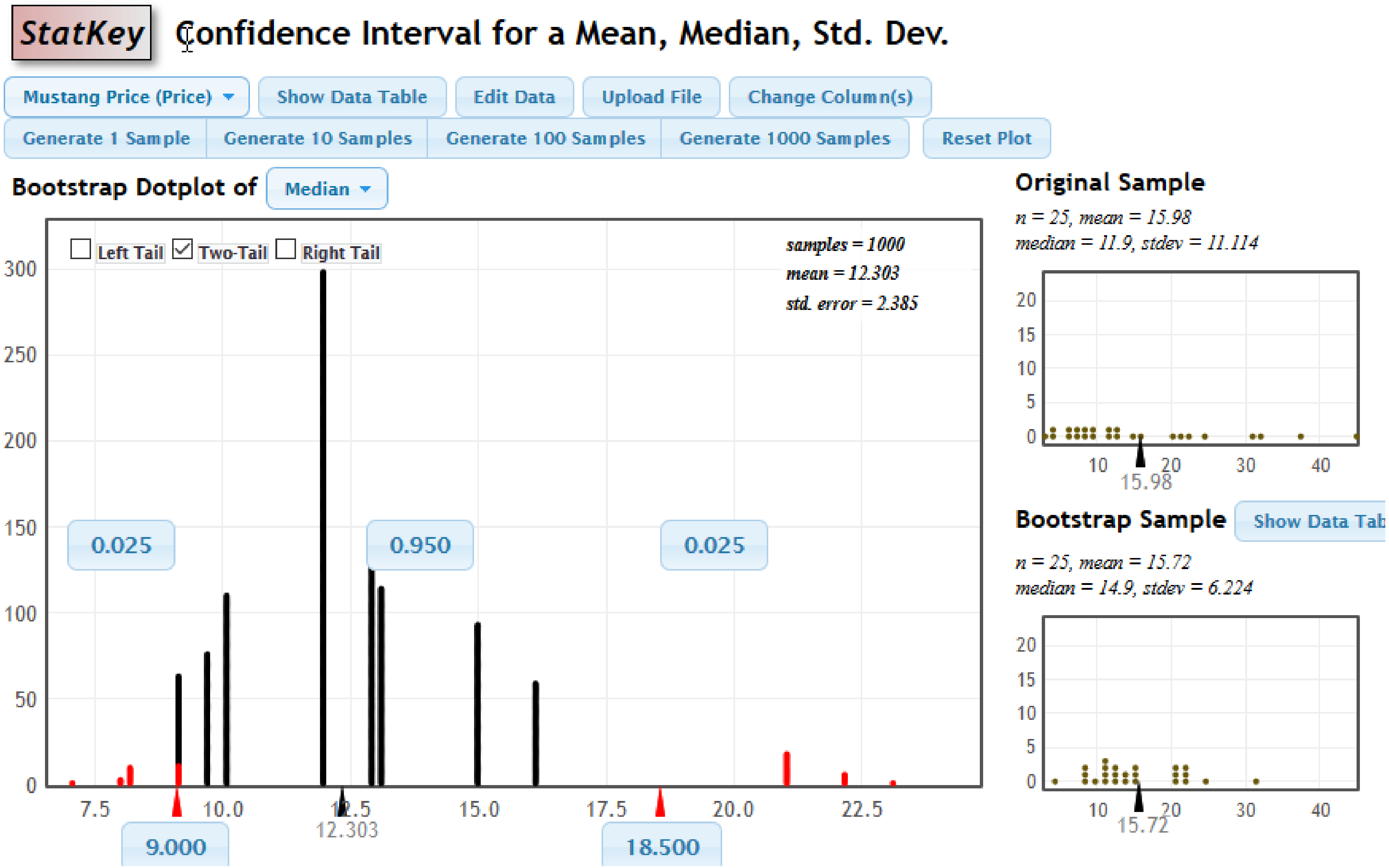
Let's look at some examples.

Bootstrap CI Median of Mustang Price

This doesn’t look like a bootstrap dist’n that we should summarize with an interval.

Look at the upper end of the interval. It is in a very wide gap between two values in the bootstrap dist’n.

How much attention to people pay to the endpoints of a CI?
Is this worth that attention?



Confidence Interval for a Mean, Median, Std. Dev.

Bootstrap Dotplot of Median

samples = 1000
mean = 12.303
std. error = 2.385

0.025

0.950

0.025

9.000

18.500

Original Sample

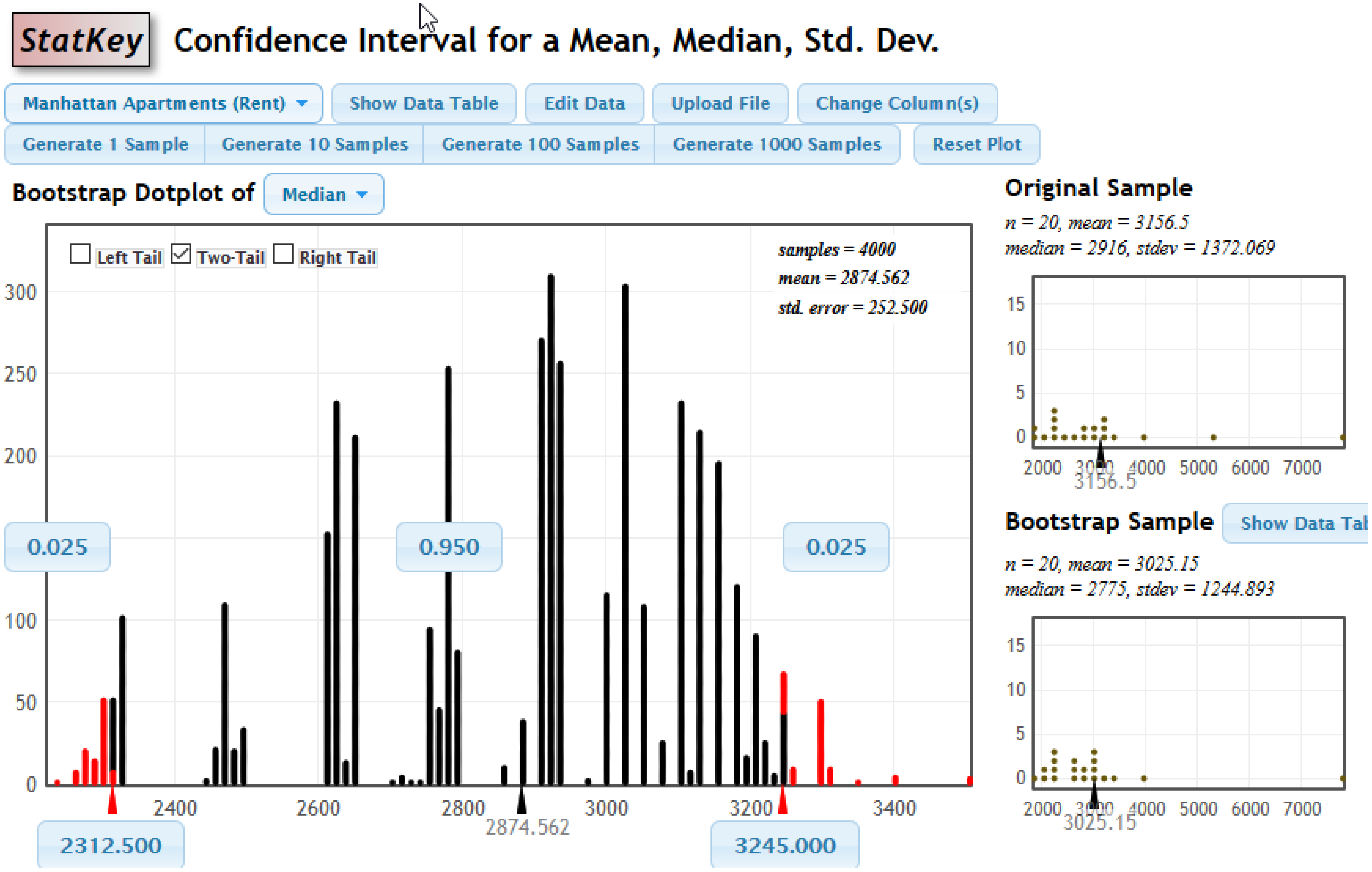
n = 25, mean = 15.98
median = 11.9, stdev = 11.114

Bootstrap Sample

n = 25, mean = 15.72
median = 14.9, stdev = 6.224

Bootstrap CI for Median of Manhattan rents

This bootstrap dist’n clearly has “clumps” and “gaps.” Those suggest that confidence intervals are not particularly useful for describing the possible values of the population median.



StatKey Confidence Interval for a Mean, Median, Std. Dev.

Bootstrap Dotplot of Median

□ Left Tail ☑ Two-Tail □ Right Tail

samples = 4000
mean = 2874.562
std. error = 252.500

0.025 0.950 0.025

2312.500 3245.000

Original Sample

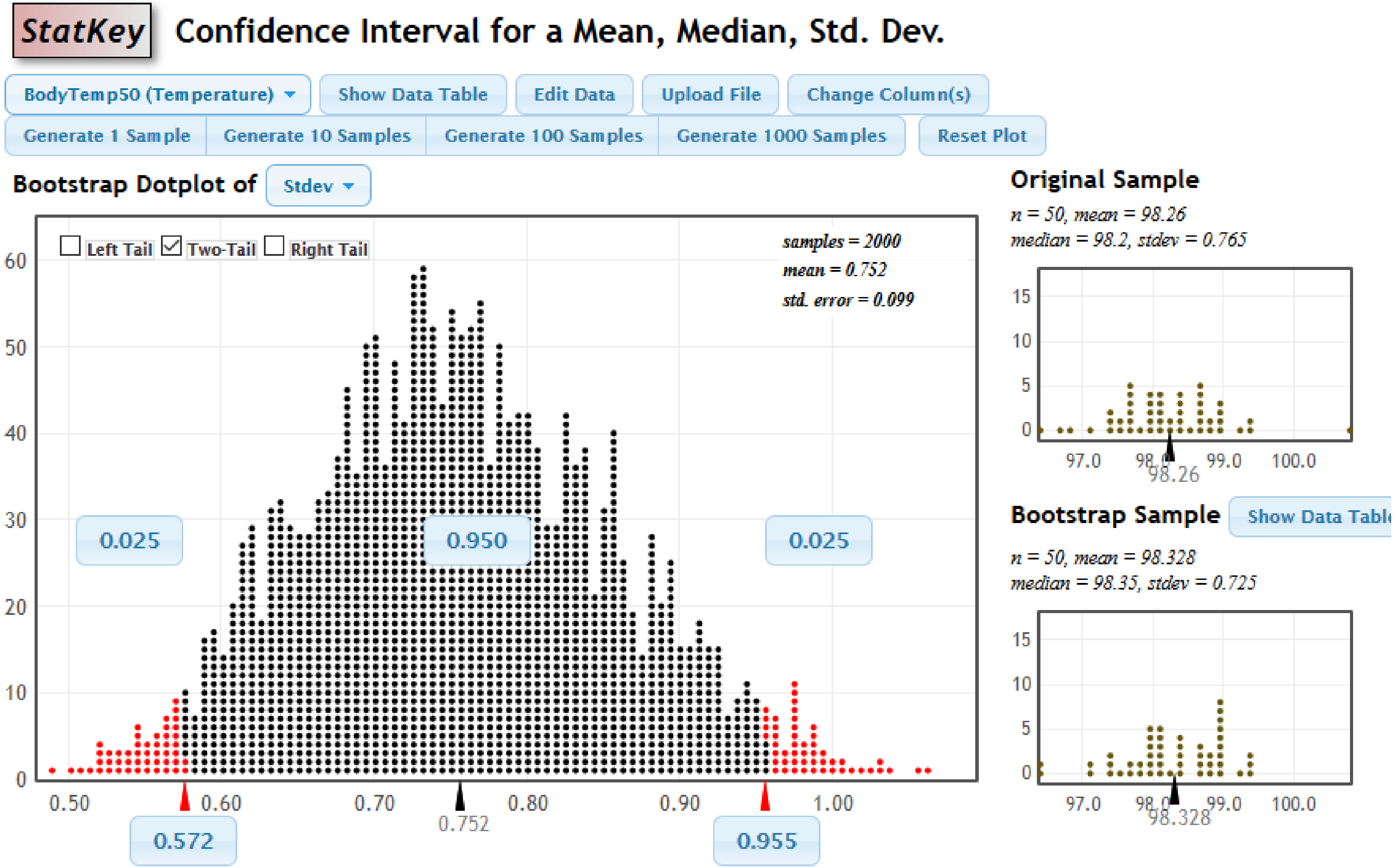
n = 20, mean = 3156.5
median = 2916, stdev = 1372.069

Bootstrap Sample

n = 20, mean = 3025.15
median = 2775, stdev = 1244.893

Standard Deviation: Body Temperature dataset

The BodyTemp50 dataset gives body temperatures of a random sample of healthy adults. It’s not a skewed dist’n and it’s interesting to see that the bootstrap dist’n of the standard deviation looks very reasonable to use to form confidence intervals.



Confidence Interval for a Mean, Median, Std. Dev.

Bootstrap Dotplot of Stdev

samples = 2000
mean = 0.752
std. error = 0.099

Original Sample

n = 50, mean = 98.26
median = 98.2, stdev = 0.765

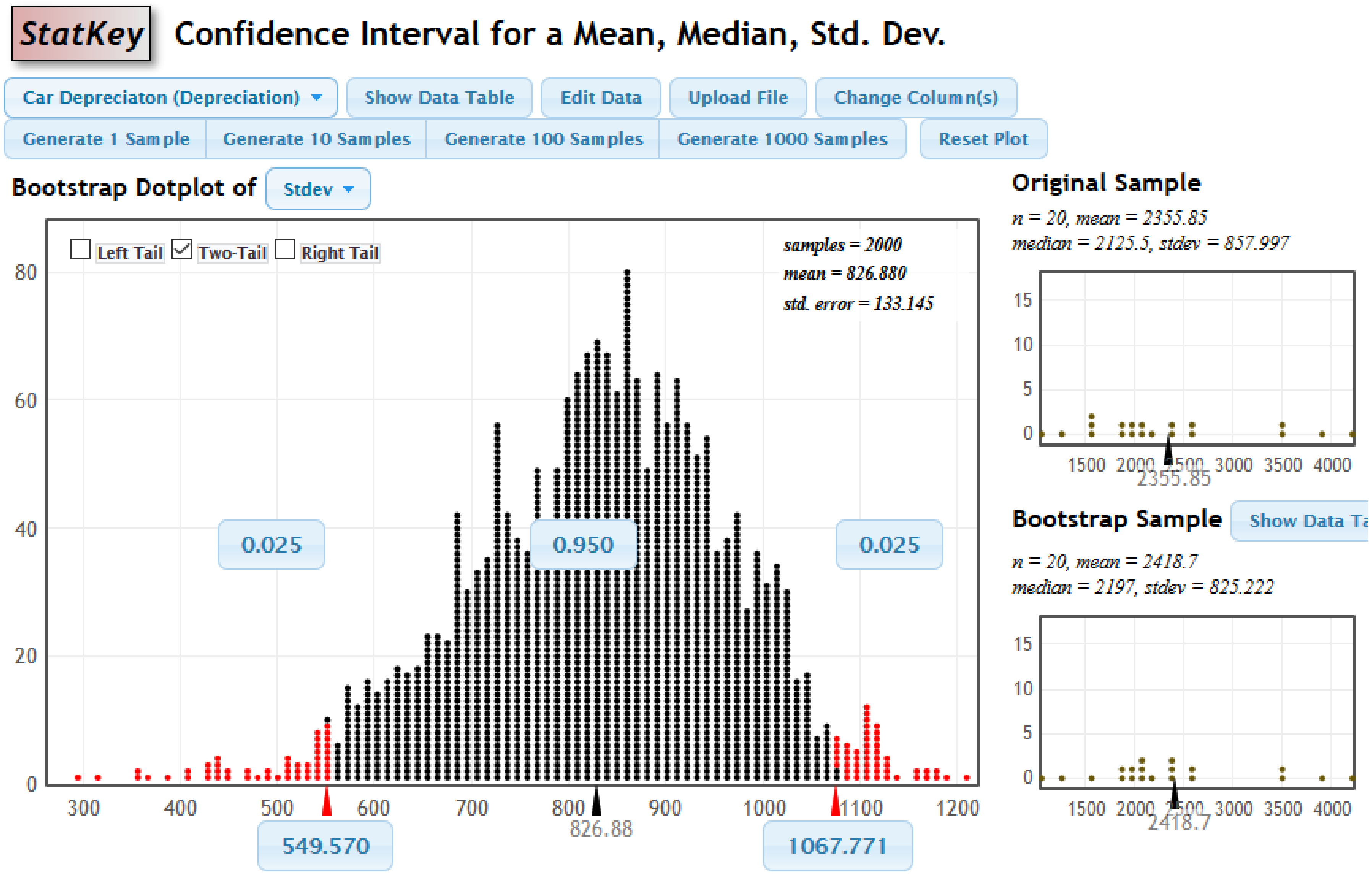
Bootstrap Sample

n = 50, mean = 98.328
median = 98.35, stdev = 0.725

Standard Deviation: Car Depreciation Data

It’s interesting to see what pattern in the data leads to what type of pattern in the sampling dist’n of the standard deviation. What do you see here?

For data that is somewhat right-skewed?



StatKey Confidence Interval for a Mean, Median, Std. Dev.

Car Depreciaton (Depreciation)

Bootstrap Dotplot of Stdev

□ Left Tail ☑ Two-Tail □ Right Tail

samples = 2000
mean = 826.880
std. error = 133.145

0.025 0.950 0.025

549.570 826.88 1067.771

Original Sample

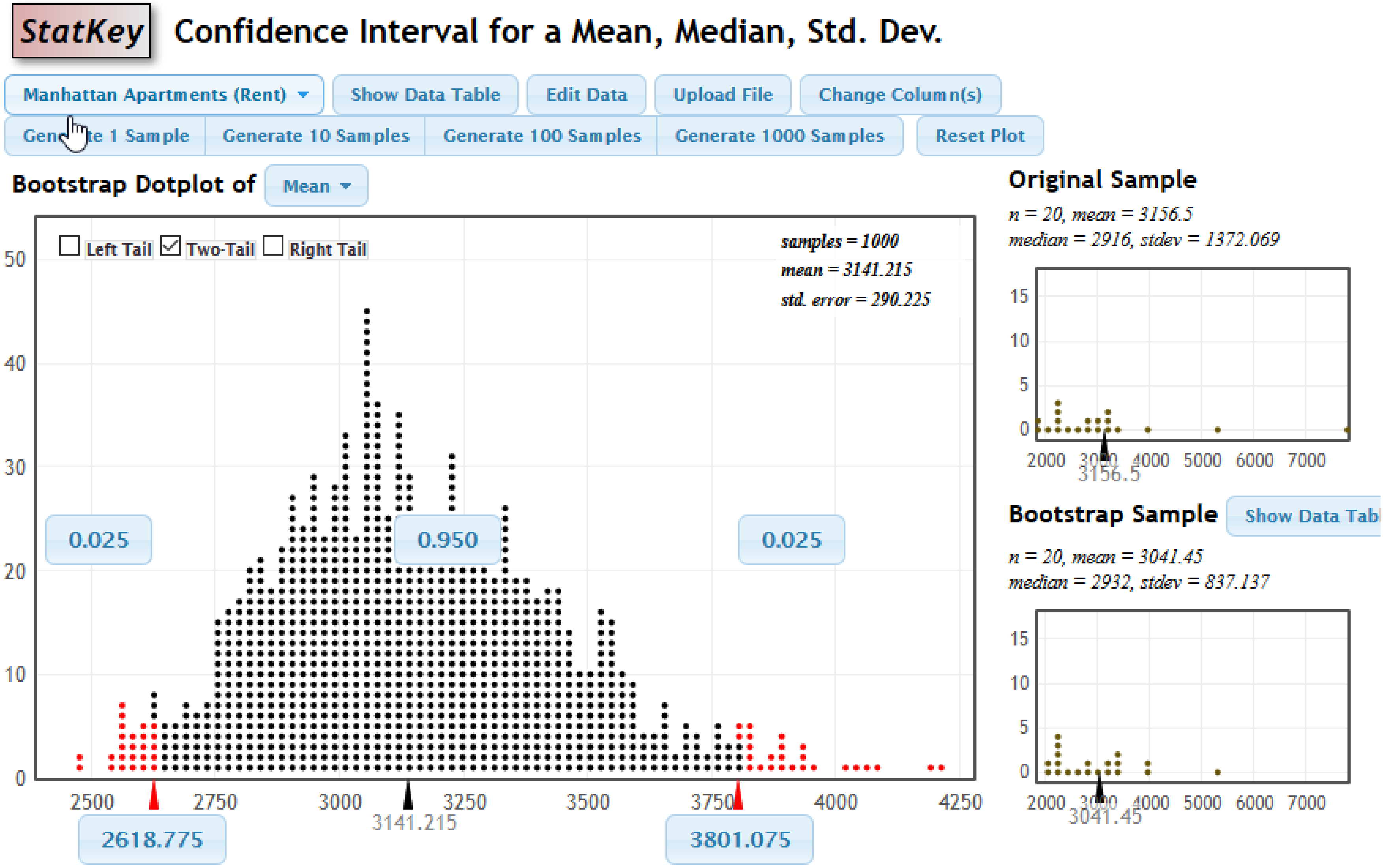
n = 20, mean = 2355.85
median = 2125.5, stdev = 857.997

Bootstrap Sample

n = 20, mean = 2418.7
median = 2197, stdev = 825.222

Standard Deviation: Manhattan Apts. Rent

The Manhattan Apts. Rent dataset is more strongly skewed to the right than the previous dataset.   What shape do we see here for the sampling dist’n of the standard deviation?



StatKey Confidence Interval for a Mean, Median, Std. Dev.

Bootstrap Dotplot of Mean

☐ Left Tail ☑ Two-Tail ☐ Right Tail

samples = 1000
mean = 3141.215
std. error = 290.225

0.025 0.950 0.025

2618.775 3801.075

Original Sample

n = 20, mean = 3156.5
median = 2916, stdev = 1372.069

Bootstrap Sample

n = 20, mean = 3041.45
median = 2932, stdev = 837.137

Types of Simulation

Our work is just an introduction to the ideas of simulation to study the properties of various statistics.

Many modifications are possible to address specific situations where more is known (or can be assumed.)

How useful is it to simulate?

1. What have you learned?
2. Do you think you can use these simulation tools to form the “usual” confidence intervals and find p-values for “usual” hypothesis tests?
3. Can you see that the “conditions” you learned for applying theoretical methods in your applied statistics course were just guidelines and the reality is more complex?
4. Can you use these StatKey simulations and datasets to explore?
5. Are you motivated to learn to make your own simulations to have more control over what you can explore?