Some Considerations for Using the Bootstrap Feature in JMP® Pro

Jason Brinkley, PhD
Senior Researcher
American Institutes for Research

AIR®
AMERICAN INSTITUTES FOR RESEARCH®

Introduction

- The bootstrap has become a popular technique for statistical analysis for a wide variety of metrics.
- Bootstrapping is the process of repeated sampling with replacement from a given dataset.
- The technique relies more on the observed data and computational acumen rather than assumptions about the underlying structure or statistical model for the data.

Naïve Bootstrap

- The simple or naïve bootstrap for the mean is a relatively simple procedure.
- Starting with an original set of observations, denoted here as $X_1, X_2, \dots, X_n$, create a new sample of observations, denoted here as $X_{11}, X_{12}, \dots, X_{1n}$by sampling the original dataset.
- Note that the naïve bootstrap creates a resampled version of the data whose size is the same as the original sample (*n*). To keep the samples from being exactly the same, the bootstrapped sample has been created with replacement, which means that one X_i in the original data may appear many times in the bootstrapped sample.
- The general idea is that the behavior of the bootstrapped sample mimics features of the original sample but is potentially different.

Example

- Suppose your original data is: 2,4,6,12,14,16
- An example bootstrapped dataset is: 2,2,6,12,14,14
- The key is to do this many times, the idea is that these resampled datasets offer some insights into the variability of the original data.

Bootstrapping for Analysis

- Traditionally the bootstrap has been utilized as an alternative technique for providing estimates of variation and interval estimates for non-standard metrics.
- However, there is a growing base that has started considering the bootstrap as a data analytic tool. The end goal may not always be direct inference.

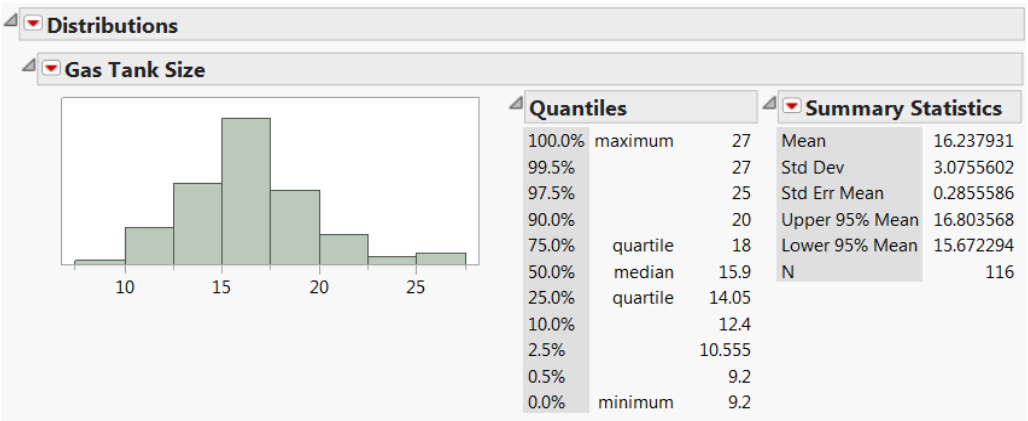
Bootstrapping in JMP Pro

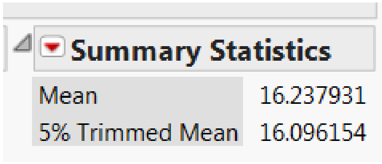
- JMP introduced bootstrapping as a standard option in many different analyses in JMP Pro 10.
- There will be no deviation from the standard options that JMP uses to bootstrap and examine the data.
- The goal here is to provide some examples and ideas to motivate the reader into using this feature in their day–to-day work.

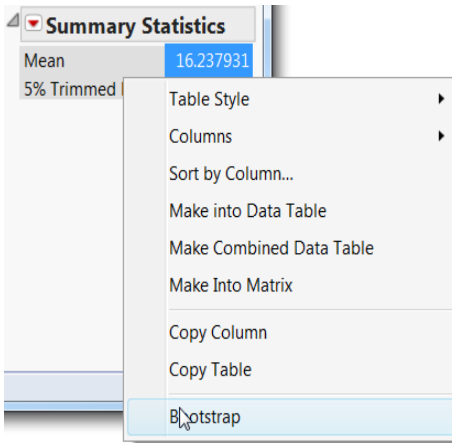
Sample Dataset

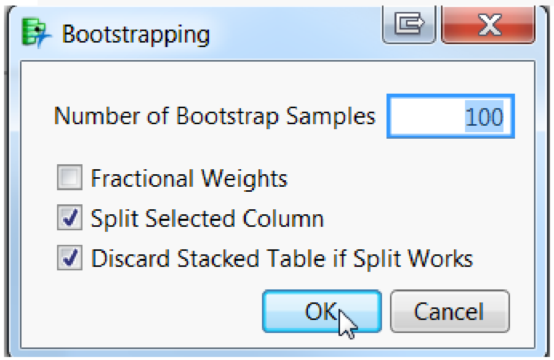
- The Car Physical Data sample dataset was collected in 1990 and consist of 116 different car models from manufacturer's, which are grouped into three geographic regions (USA, Japan, Other). The data also list vehicle type (Large, Medium, Compact, Small, Sport) and vehicle metrics for weight, turning circle displacement, horsepower and gas tank size.

Trimmed Mean









The bootstrapping option creates the bootstrap samples in the background and recalculates the selected metric.

Untitled 4 - JMP Pro

File Edit Tables Rows Cols DOE Analyze Graph SAS Tools View Window Help

Untitled 4
Source

Columns (4/0)
Y
BootID•
5% Trimmed Mean
Mean

	Y	BootID•	5% Trimmed Mean	Mean
1	Gas Tank Size	0	16.096153846	16.237931034
2	Gas Tank Size	1	15.8125	15.951724138
3	Gas Tank Size	2	16.399038462	16.56637931
4	Gas Tank Size	3	16.819230769	16.982758621
5	Gas Tank Size	4	15.990384615	16.121551724
6	Gas Tank Size	5	16.140384615	16.289655172
7	Gas Tank Size	6	15.972115385	16.156896552
8	Gas Tank Size	7	16.370192308	16.490517241
9	Gas Tank Size	8	16.276923077	16.453448276
10	Gas Tank Size	9	15.861538462	16.068965517

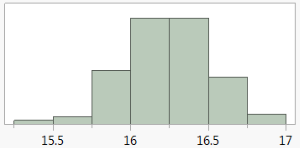
Original

Summary Statistics

Mean	16.237931
Std Dev	3.0755602
Std Err Mean	0.2855586
Upper 95% Mean	16.803568
Lower 95% Mean	15.672294
N	116

Bootstrapped

Mean



Quantiles

100.0%	maximum	16.9828
99.5%		16.9828
97.5%		16.9134
90.0%		16.566
75.0%	quartile	16.4455
50.0%	median	16.2392
25.0%	quartile	16.0563
10.0%		15.8802
2.5%		15.6299
0.5%		15.4759
0.0%	minimum	15.4759

Summary Statistics

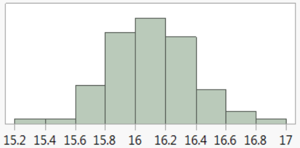
Mean	16.235319
Std Dev	0.2783723
Std Err Mean	0.0278372
Upper 95% Mean	16.290554
Lower 95% Mean	16.180084
N	100

Bootstrap Confidence Limits

Coverage	Pct Lower	Pct Upper
0.95	15.6299	16.9134
0.90	15.7674	16.6591
0.80	15.8802	16.566
0.50	16.0563	16.4455

Original Estimate 16.23793

5% Trimmed Mean



Quantiles

100.0%	maximum	16.8192
99.5%		16.8192
97.5%		16.7818
90.0%		16.4716
75.0%	quartile	16.2921
50.0%	median	16.1144
25.0%	quartile	15.951
10.0%		15.7786
2.5%		15.5409
0.5%		15.3798
0.0%	minimum	15.3798

Summary Statistics

Mean	16.115029
Std Dev	0.2784228
Std Err Mean	0.0278423
Upper 95% Mean	16.170274
Lower 95% Mean	16.059784
N	100

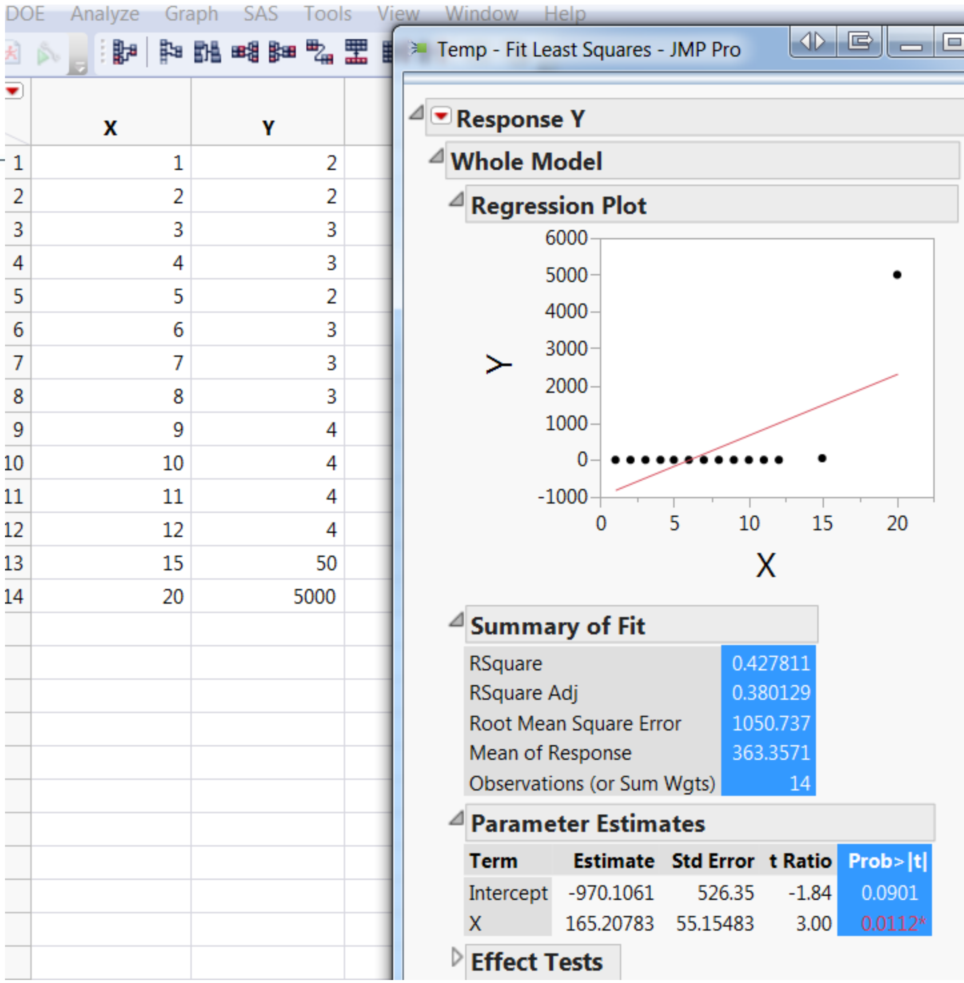
Bootstrap Confidence Limits

Coverage	Pct Lower	Pct Upper
0.95	15.5409	16.7818
0.90	15.6849	16.5722
0.80	15.7786	16.4716
0.50	15.951	16.2921

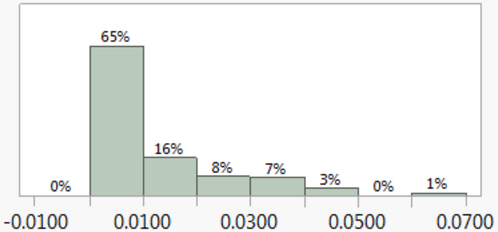
Original Estimate 16.09615

Another Example

Obviously from the plot we see that this is not a very good model. I illustrate it here because it gets a lot harder to see in many dimensions where plots aren't always handy.



X



Quantiles

100.0%	maximum	0.0694
99.5%		0.0694
97.5%		0.04651
90.0%		0.03032
75.0%	quartile	0.01433
50.0%	median	0.00359
25.0%	quartile	0.00053
10.0%		1.1e-5
2.5%		1.48e-7
0.5%		1.21e-8
0.0%	minimum	1.21e-8

Summary Statistics

Mean	0.0098357
Std Dev	0.0133298
Std Err Mean	0.001333
Upper 95% Mean	0.0124807
Lower 95% Mean	0.0071908
N	100

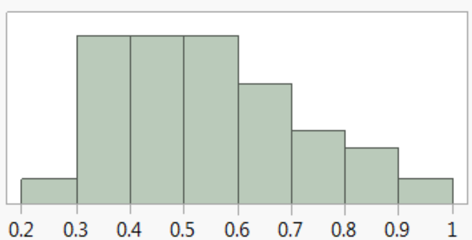
Bootstrap Confidence Limits

Coverage	Pct Lower	Pct Upper
0.95	1.48e-7	0.04651
0.90	1.59e-6	0.03822
0.80	1.1e-5	0.03032
0.50	0.00053	0.01433

Original Estimate 0.011163

99% of the resamplings have significant p-values for the X variable in the regression. But look at the distribution of R-square. Whether it has “true” coverage for R-square is immaterial. It is large and weird.

RSquare



Quantiles

100.0%	maximum	0.93889
99.5%		0.93889
97.5%		0.90997
90.0%		0.81381
75.0%	quartile	0.64638
50.0%	median	0.52046
25.0%	quartile	0.40552
10.0%		0.33431
2.5%		0.29117
0.5%		0.24885
0.0%	minimum	0.24885

Summary Statistics

Mean	0.5402756
Std Dev	0.1685247
Std Err Mean	0.0168525
Upper 95% Mean	0.5737145
Lower 95% Mean	0.5068366
N	100

Bootstrap Confidence Limits

Coverage	Pct Lower	Pct Upper
0.95	0.29117	0.90997
0.90	0.31109	0.86348
0.80	0.33431	0.81381
0.50	0.40552	0.64638

Original Estimate 0.427811

Example (Chase and Drummer)

- Turning to the last example, a different sample dataset is needed. The sample dataset “Children’s Popularity” contains 480 observations from a study by Chase and Dummer (1992). JMP notes showing the following description:

“*Subjects were students in grades 4-6 from three school districts in Ingham and Clinton Counties, Michigan. Chase and Dummer stratified their sample, selecting students from urban, suburban, and rural school districts with approximately 1/3 of their sample coming from each district. Students indicated whether good grades, athletic ability, or popularity was most important to them. They also ranked four factors: grades, sports, looks, and money, in order of their importance for popularity. The questionnaire also asked for gender, grade level, and other demographic information.*”

- The ranked factors are the values of primary interest.

Spearman Correlations

Multivariate Gender=boy

Nonparametric: Spearman's ρ

Variable	by Variable	Spearman ρ	Prob>|ρ|
Sports	Grades	-0.1594	0.0162*
Looks	Grades	-0.5424	<.0001*
Looks	Sports	-0.3283	<.0001*
Money	Grades	-0.4103	<.0001*
Money	Sports	-0.2448	0.0002*
Money	Looks	-0.2198	0.0009*

Multivariate Gender=girl

Nonparametric: Spearman's ρ

Variable	by Variable	Spearman ρ	Prob>|ρ|
Sports	Grades	-0.1592	0.0115*
Looks	Grades	-0.5612	<.0001*
Looks	Sports	-0.3144	<.0001*
Money	Grades	-0.4770	<.0001*
Money	Sports	-0.4642	<.0001*
Money	Looks	0.0694	0.2734

Here we see something of a different constellation of correlations between the genders. With the measured association between Money and Looks to be -0.2198 for boys and 0.0694 for girls, the question that arises is whether the correlation is significantly higher for boys than girls. We have an estimate of that difference to be -0.2892, but can one find a 95% bootstrap confidence interval for that difference to determine if it contains zero?

Note – You have to split the data

File Edit Tables Rows Cols DOE Analyze Graph SAS Tools View Window Help

Untitled 19
Source
Columns (2/0)
BootID•
Girl Money Versus L

	BootID•	Girl Money Versus Looks
1	0	0.0694
2	1	0.0153
3	2	-0.0016
4	3	0.1050
5	4	0.0736
6	5	0.0830
7	6	0.0574
8	7	0.2436
9	8	0.0920
10	9	0.1438
11	10	-0.0297

File Edit Tables Rows Cols DOE Analyze Graph SAS Tools View Window Help

Untitled 22
Source
Columns (2/0)
BootID•
Boy Money Versus L

	BootID•	Boy Money Versus Looks
1	0	-0.2198
2	1	-0.1817
3	2	-0.1324
4	3	-0.1586
5	4	-0.3101
6	5	-0.2357
7	6	-0.3626
8	7	-0.2380
9	8	-0.1699
10	9	-0.2009
11	10	-0.3174

File Edit Tables Rows Cols DOE Analyze Graph SAS Tools View Window Help

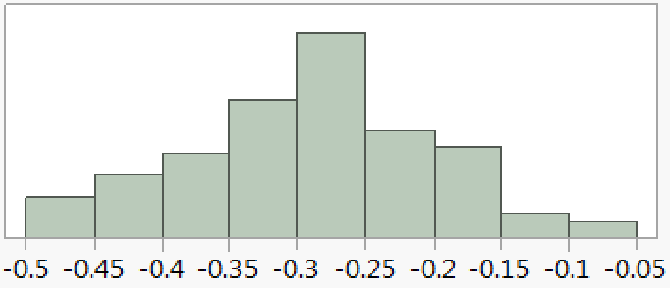
Untitled 23
Source
Columns (5/1)
BootID• of Untitled 22
Boy Money Versus Looks
BootID• of Untitled 19
Girl Money Versus Looks
Difference

	BootID• of Untitled 22	Boy Money Versus Looks	BootID• of Untitled 19	Girl Money Versus Looks	Difference
1	0	-0.2198	0	0.0694	-0.289145973
2	1	-0.1817	1	0.0153	-0.196976911
3	2	-0.1324	2	-0.0016	-0.13082771
4	3	-0.1586	3	0.1050	-0.263555292
5	4	-0.3101	4	0.0736	-0.383650716
6	5	-0.2357	5	0.0830	-0.318778821
7	6	-0.3626	6	0.0574	-0.4[illegible]2939
8	7	-0.2380	7	0.2436	-0.481645146
9	8	-0.1699	8	0.0920	-0.261871441
10	9	-0.2009	9	0.1438	-0.344619027
11	10	-0.3174	10	-0.0297	-0.287668367

Difference

Distributions

Difference



Quantiles

100.0%	maximum	-0.0879
99.5%		-0.0879
97.5%		-0.105
90.0%		-0.1617
75.0%	quartile	-0.231
50.0%	median	-0.2848
25.0%	quartile	-0.3467
10.0%		-0.4189
2.5%		-0.473
0.5%		-0.4947
0.0%	minimum	-0.4947

Be careful with Inference

- There are some well known limitations of the bootstrap. In particular, we know that there are times when a bootstrap interval does not have good coverage of the truth.
- It is unclear as to whether every option in JMP that produces a bootstrap interval is of a type that has good coverage. The software implements the technique with a wide brush. It is likely that some measures that are bootstrapped here produce biased intervals.

Don’t be careful with analysis and descriptives

- From Mammen and Nandi:
  “In a data analysis the statistician wants to get a basic understanding of he stochastic nature of the data…We will argue that the bootstrap and other resampling methods offer a simple way to get a basic understanding for the stochastic nature of plots that depend on random data.”
- While their examples were overlaying the results of many bootstrap samples with the original data in a concentrated visualization, the same ideas extend to other areas of analysis.
- Looking at what happens in a particular analysis across many resamplings can offer many insights into the stability of not just the data but the techniques being implemented.
- Plus it’s easy to implement, so if you have JMP Pro then one can implement the procedure with relative ease.

How to use it?

- These are my own personal reflections and not necessarily grounded in any statistical theory. More of a direct applied approach.
- Most of the time, when I utilize the bootstrap in this way, I am looking for something ‘weird’.
- Is the mean of the resampled values ‘close’ to the original data? How wide are the interval estimates? Looking at the range of the interval estimates, how would I have felt if I had received the upper or lower bounds in my actual data?
- Sometimes what you find will surprise you.

Acknowledgements and Contact Info

Special Thanks to Jennifer Mann (OSU) for advice on earlier work on this.

Contact Information:

Jason Brinkley, PhD
Senior Researcher
American Institutes for Research
email: jbrinkley@air.org

100 Europa Drive Suite 315, Chapel Hill, NC 27517
TEL:919.918.2318|WEBSITE WWW.AIR.ORG