Bootstrapping

- Bootstrapping is a technique that allows us to estimate the sampling distribution of almost any statistic using resampling methods. We can estimate the distribution of a mean, a median, a variance, a standard deviation, or anything else like quartiles or percentiles.

- We are always interested in the large group, the Population, denoted by N. We collect data from a subset of the population, call it our sample, and use this data to estimate population parameters by calculating sample statistics.

- In bootstrapping, we treat our actual sample as a population from which we take random samples with replacement. In other words, if our sample contains say 20 observations, take multiple samples of size 20 from the sample of size 20 (with replacement).

- Bootstrap many resample (3000 should do the job), and then from the distribution of resamples, you can estimate population parameters.

- The idea is such: The population is unknown, so the true error in a sample statistic against its population is unknown. In bootstrap resamples, the 'population' is in fact known, so the quality of inference from the resample is measureable.

When to use a Bootstrapping Methods

1. When the theoretical distribution of a statistic of interest is complicated or unknown.
2. When the sample size is insufficient for straightforward statistical inference.

Advantages

1. Simplicity.
2. Straightforward way to derive estimates of standard errors and confidence intervals for complex parameters.

Steps To Do Bootstrapping on StatCrunch

1. Open your dataset and select Stat → Resample → Statistic

2. You will need to do a tiny bit of coding. First select your column to resample. Then, in the Statistic window:

   To analyze a mean, type mean("variable name") where "variable name" is the column label.

   To analyze a median, type median("variable name")

   To analyze a standard deviation, type std("variable name")

   To analyze the first quartile, type q1("variable name")

   To analyze the third quartile, type q3("variable name")

3. Change the number of resamples to 3000.

4. Hit Next. In the Percentiles Box, the default is the 2.5th percentile, the 5th, the 95th, and the 97.5th. If you are interested in bootstrapping a 99% confidence interval for your statistics, you must add the 0.5th and 99.5th to that box. Type 0.5 and 99.5, separated by commas.

5. Hit Resample Statistic.

Here's what happens:

- StatCrunch takes 3000 samples of size *n* with replacement from your data with sample size *n*.

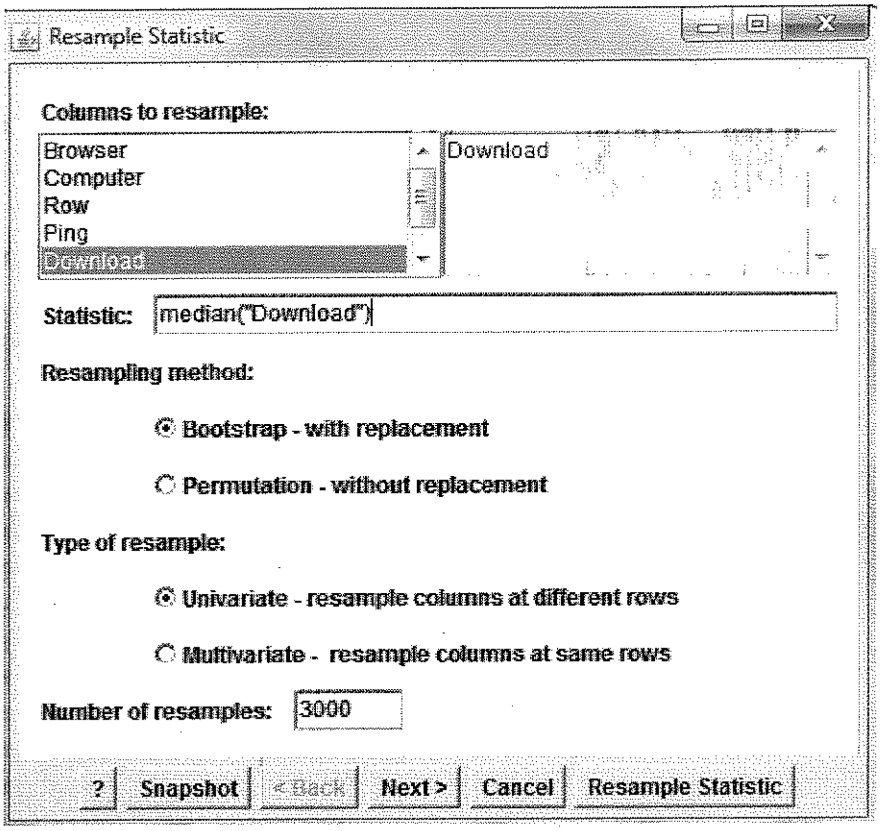
- StatCrunch calculates the statistic of your choosing from 2. above for each resample.

- A common way to use the data from the resamples is to create percentile based confidence intervals.

- To bootstrap a 95% confidence interval for any statistic we choose, look at the 2.5th percentile as your lower bound and the 97.5th percentile as your upper bound.

- A 90% bootstrap confidence interval uses the 5th and 95th percentiles, a 99% bootstrap confidence interval uses the 0.5th and 99.5th percentiles.

Example: Load up the "***Internet Speed***" dataset that we collected in class back in March 2013 and analyze the median download speed with a 95% bootstrap confidence interval.



Columns to resample:

Browser
Computer
Row
Ping
Download

Download

Statistic: median("Download")

Resampling method:

● Bootstrap - with replacement

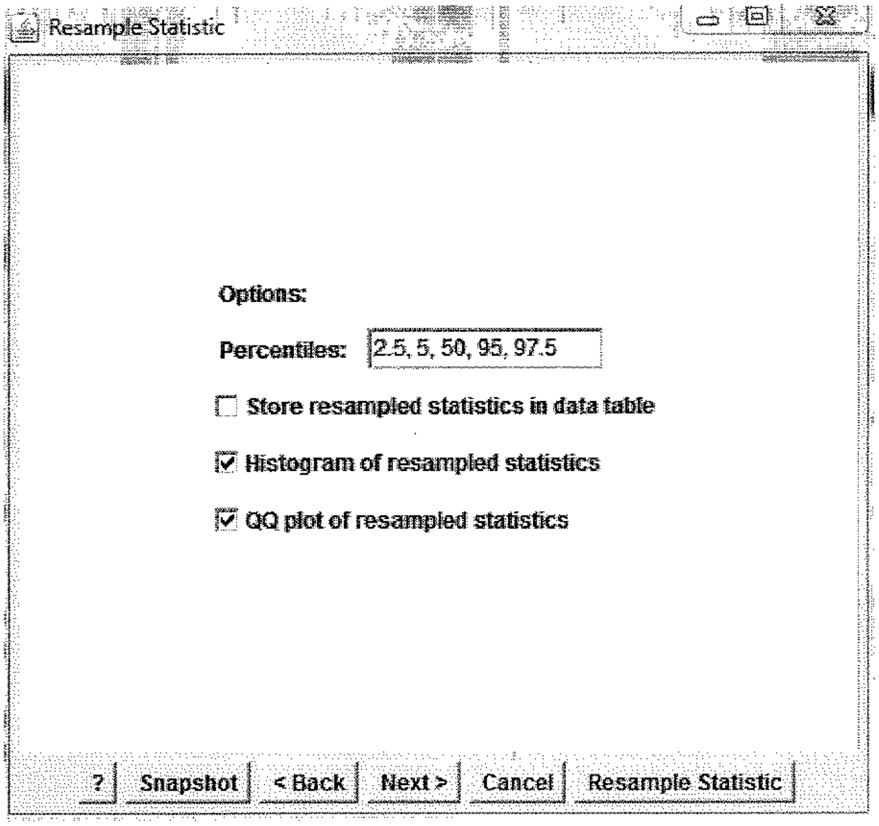
○ Permutation - without replacement

Type of resample:

● Univariate - resample columns at different rows

○ Multivariate - resample columns at same rows

Number of resamples: 3000



Options:

Percentiles: 2.5, 5, 50, 95, 97.5

□ Store resampled statistics in data table

☑ Histogram of resampled statistics

☑ QQ plot of resampled statistics

Observed Median
1.775

2.5th Percentile
1.48

97.5th Percentile
2

95% Bootstrap
CI
for Median
is
(148 to 2)

95% conf....

Example: Load up "***Kupe Bowling***" and calculate a 99% bootstrap confidence interval for the mean difference in his first game and his third game.

1. Need to get all Game 1 in a column and all of game 3 in a column (and in order!).

Data → Split Column → Column: Scores
Group Column: Match

2. Take Differences: Put in new column.

Data → Compute Expression
"1 Scores" – "3 Scores"

3. Now ~~Bootstrap~~ → Stat → Resample
Add 0.5, 99.5 Percentiles.

4. Observed Mean is 17.58 pins
(could change as Kupe updates data)

99% CI for mean difference is

(–5.27 pins to 38.48 pins)

With 99% confidence, the true mean difference is captured by this interval.
Zero is in it, indicating no difference, Game 1 versus Game 3.

Example: How much variation is there in the "**Albuquerque NM Real Estate**" market for the size of homes? Bootstrap the square footage of the homes in the dataset and run a 1% significance level test to determine the standard deviation in home sizes exceeds 400 square feet.

Test $H_0 : \sigma \leq 400$ SQ FT
$H_A : \sigma > 400$ SQ FT
Run at $\alpha = 0.01$

Bootstrap and add 1st, 99th percentiles

std("SQFT")

Observed $S = 523.7228$

98% CI for σ is $(420.39, 623.95)$

Has $\alpha = 0.01$ in each tail.

Lower bound of 420.39

exceeds 400 in HA.

Reject Ho.

At $\alpha = 0.01$, conclude the standard deviation of SQFT in Alb. NM exceeds 400.