The Bootstrap

Peter Ralph

15 October – Advanced Biological Statistics

Estimating sampling error

Standard error

From a single set of numbers

$$x_1, x_2, \ldots, x_n$$

we can get both a *mean*:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Standard error

From a single set of numbers

$$x_1, x_2, \ldots, x_n$$

we can get both a *mean*:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

and an estimate of the *variability* of the mean, the *standard error*:

$$\frac{s}{\sqrt{n}} = \sqrt{\frac{1}{n(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2}.$$

This is amazing!

This is amazing!

Sadly, most other types of estimates don't have this amazing property.

This is amazing!

Sadly, most other types of estimates don't have this amazing property.

What to do?

Enter the bootstrap

Idea:

Idea:

- We'd *like* to get a whole new dataset, and repeat the estimation, to see how different the answer is.

Idea:

- We'd *like* to get a whole new dataset, and repeat the estimation, to see how different the answer is.
- And, well, our best guess at what the data look like is **our dataset itself**,

Idea:

- We'd *like* to get a whole new dataset, and repeat the estimation, to see how different the answer is.
- And, well, our best guess at what the data look like is **our dataset itself**,
- *sooooo*, let's just *resample from the dataset*, with replacement, to make a "new" dataset!

Idea:

- We'd *like* to get a whole new dataset, and repeat the estimation, to see how different the answer is.
- And, well, our best guess at what the data look like is **our dataset itself**,
- *sooooo*, let's just *resample from the dataset*, with replacement, to make a "new" dataset!
- If we resample and re-estimate lots of times, this should give us a good idea of the variability of the estimate.

The bootstrap resampling algorithm

To estimate the uncertainty of an estimate:

1. Use the computer to take a random sample of observations from the original data, with replacement.
2. Calculate the estimate from the resampled data set.
3. Repeat 1-2 many times.
4. The standard deviation of these esimates is the **bootstrap standard error**.

Advantages

- Applies to most any statistic
- Works when there's no simple formula for the standard error (e.g., median, trimmed mean, eigenvalue, etc)
- Is *nonparametric*, so doesn't make specific assumptions about the distribution of the data.
- Applies to even complicated sampling procedures.

Exercise

```
x <- c(0.6, 1, 3.1, 3.7, 4.8, 6.2, 12.5, 12.5, 13.4, 24.1)
```

- Use R to make 1000 “pseudo-samples” of size 10 (with replacement),
- and store the mean of each in a vector.
- Plot the histogram of the resampled means, and calculate their standard deviation (with `sd()`).
- How does this compare to the usual standard error of the mean, `sd(x) / sqrt(length(x))`?

Confidence intervals?

The 2.5% and 97.5% percentiles of the bootstrap samples estimate a 95% confidence interval. (use the `quantile( )` function)

Confidence intervals?

The 2.5% and 97.5% percentiles of the bootstrap samples estimate a 95% confidence interval. (use the `quantile( )` function)

Exercise: get a 95% CI and compare it to that given by `t.test( )`.

// reveal.js plugins

/