# Stat 322: Statistical Theory
# Bootstrap Activity

Based on Section 4.5 in *Investigating Statistical Concepts, Applications, and Methods* (2006) by Beth Chance and Allan Rossman. Relevant R code can be found in bootstrap.R.

The population of all 268 words in the Gettysburg address is stored in gettysburgpop.csv along with the length of each word. Suppose we take a random sample of n=10 words from this population, and let $X_i$ be the length of the i$^{th}$ word from our sample

1. What can we say about the sampling distribution of $\overline{X}$ in terms of center, spread, and shape?

2. Is the distribution of t-statistics (i.e. $(\overline{X} - \mu)/(s/\sqrt{n})$) symmetric?

Consider the specific sample of n=10 words contained in gettysburgsample.csv.

3. Find the sample mean, sample standard deviation, and 95% frequentist confidence interval for the population mean based on this sample.

4. Take 1000 bootstrap samples from our original sample of n=10. Create a plot showing the distribution of bootstrap means.

5. Do the mean and standard deviation of the bootstrap means agree with your predictions from (1)?

6. Create a t-interval with bootstrap SE for the population mean by taking:

Original sample mean $\pm$ $t^*_{n-1}$ x bootstrap estimate of SD of sample mean

7. Create a bootstrap percentile interval for the population mean by picking off the 2.5$^{th}$ and 97.5$^{th}$ percentiles. How does this interval compare to the interval in (6)?

When the bootstrap distribution is skewed, we should consider alternatives to the "t-interval with bootstrap SE" in (6) and the "bootstrap percentile interval" in (7). One such alternative is the "basic bootstrap" (Davison and Hinkley 1997), also called the "reverse bootstrap percentile interval" (Hesterberg 2014).

Let $\theta$ be the population parameter of interest, $\hat{\theta}$ be the sample estimate of $\theta$, and $\hat{\theta}^*$ be a bootstrap sample estimate of $\theta$. If we call $\hat{\theta}^*_{.975}$ the 97.5th percentile and $\hat{\theta}^*_{.025}$ the 2.5th percentile of bootstrap sample estimates, then:

$$P(\hat{\theta}^*_{.025} \le \hat{\theta}^* \le \hat{\theta}^*_{.975}) = .95$$

$$\Rightarrow \quad P(\hat{\theta}^*_{.025} - \hat{\theta} \le \hat{\theta}^* - \hat{\theta} \le \hat{\theta}^*_{.975} - \hat{\theta}) = .95$$

$$\Rightarrow \quad P(\hat{\theta}^*_{.025} - \hat{\theta} \le \hat{\theta} - \theta \le \hat{\theta}^*_{.975} - \hat{\theta}) = .95$$ assuming the distribution of $\hat{\theta} - \theta$ is closely approximated by the distribution of $\hat{\theta}^* - \hat{\theta}$

$$\Rightarrow \quad P(2\hat{\theta} - \hat{\theta}^*_{.975} \le \theta \le 2\hat{\theta} - \hat{\theta}^*_{.025}) = .95$$

8. Find the reverse bootstrap percentile interval for the population mean using gettysburgsample.csv. How does this compare with your intervals in (6) and (7)?

Hesterberg (2014) contends the interval in (8) performs poorly in practice and is asymmetrical in the wrong direction for skewed data. He (and other authors) recommend the "bootstrap-t" or "percentile-t" interval. In this case, bootstrapping is used to find the appropriate multipliers for a confidence interval, since the t-statistic does not have a t-distribution with skewed data. This confidence interval is of the form:

$$(\hat{\theta} - q_{1-\alpha/2}\hat{S}, \hat{\theta} - q_{\alpha/2}\hat{S})$$

where $\hat{S}$ is the estimate of the standard error from the original sample, and the *q* terms are quantiles from the bootstrapped distribution of t-statistics – i.e. from the distribution of $t^* = \dfrac{\hat{\theta}^* - \hat{\theta}}{\hat{S}^*}$ values obtained from the bootstrap samples.

9. Find the bootstrap-t interval for the population mean using gettysburgsample.csv and compare this interval with your 3 previous intervals.

10. Find 95% bootstrapped intervals for mu using the boot() function in R applied to gettysburgsample.csv.

Hesketh and Everitt (2000) report on a study by Caplehorn and Bell (1991) that investigated the times that heroin addicts remained in a clinic for methadone maintenance treatment. The data in heroin.csv include the amount of time that the subjects stayed in the facility until treatment was terminated.

11. Using both a bootstrap percentile interval and a bootstrap-t interval, find a 95% confidence interval for:
    a) mean survival time
    b) median survival time
    c) 25% trimmed mean survival time