STAT 135 Lab 5
Bootstrapping and Hypothesis Testing

Rebecca Barter

March 2, 2015

The Bootstrap

Bootstrap

Suppose that we are interested in estimating a parameter θ from some population with members $x_1, \ldots, x_M$, but that this population is so large that we cannot possibly observe every x_i.

We have seen that we can take a sample $X_1, \ldots, X_n$ and generate an estimate, $\hat{\theta}$ (e.g. MLE, MOM).

What if we want to know properties of this estimator, but we don't know any of the theory?

Bootstrap

Suppose that we can feasibly take another N samples from our population. Then from each sample we can generate another estimator:

$$x_1, \dots, x_M \quad \overset{sample}{\longrightarrow} \quad \begin{cases} X_1^{(1)}, \dots, X_n^{(1)} & \rightarrow \hat{\theta}^{(1)} \\ & \vdots \\ X_1^{(N)}, \dots, X_n^{(N)} & \rightarrow \hat{\theta}^{(N)} \end{cases}$$

we can estimate properties of $\hat{\theta}$ using the estimates $\hat{\theta}^{(1)}, \dots, \hat{\theta}^{(N)}$. For example:

$$\widehat{E(\hat{\theta})} = \frac{1}{N} \sum_{i=1}^{N} \hat{\theta}^{(i)}$$

In most cases, however, it is not feasible to take new samples from our population. How else can we generate samples to estimate properties of $\hat{\theta}$?

Bootstrap

Suppose that we **know the distribution** of the population (e.g. $F_\theta = Exponential(4)$). Then we can simulate N IID samples from F_θ and calculate the estimators from each sample

$$F_\theta \quad \overset{simulate}{\rightarrow} \quad \begin{cases} X_1^{(1)}, ..., X_n^{(1)} & \rightarrow \hat{\theta}^{(1)} \\ & \vdots \\ X_1^{(N)}, ..., X_n^{(N)} & \rightarrow \hat{\theta}^{(N)} \end{cases}$$

we can estimate properties of $\hat{\theta}$ using the estimates $\hat{\theta}^{(1)}, ..., \hat{\theta}^{(N)}$. For example:

$$\widehat{Var(\hat{\theta})} = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{\theta}^{(i)} - \frac{1}{N} \sum_{j=1}^{N} \hat{\theta}^{(j)} \right)^2$$

In most cases, however, we don't know what F_θ is.

Bootstrap

Bootstrapping is a method which uses random sampling techniques to estimate properties (such as bias, variance, confidence intervals, etc) of an estimator, $\hat{\theta}$, **when we don't know the true distribution, F_θ, of our data (and we cannot feasibly draw new samples from our population)**.

Bootstrap

We will focus on two approaches to generating bootstrap samples:

- **Parametric bootstrap**: Estimate $\hat{\theta}$ from our original sample, $X_1, ..., X_n$ and generate samples $X_1^*, ..., X_n^*$ from $F_{\hat{\theta}}$, which approximates F_θ.
- **Non-parametric bootstrap**: Take samples $X_1^*, ..., X_n^*$ with replacement from our original sample $X_1, ..., X_n$.

Once we have B bootstrap samples, we can generate B estimates of $\hat{\theta}$:

$$X_1^{*(1)}, ..., X_n^{*(1)} \quad \rightarrow \quad \hat{\theta}^{*(1)}$$

$$\vdots$$

$$X_1^{*(B)}, ..., X_n^{*(B)} \quad \rightarrow \quad \hat{\theta}^{*(B)}$$

We can use these bootstrapped estimators, $\hat{\theta}^{*(1)}, ..., \hat{\theta}^{*(B)}$, to estimate properties of $\hat{\theta}$

Parametric Bootstrap

Parametric Bootstrap

Recall that we are using bootstrapping because we don't know the true parameter θ, but we want to identify the distribution of our estimator $\hat{\theta}$ (e.g. the MLE).

Suppose that we have observed a sample $X_1, ..., X_n$, $X_i \sim F_\theta$. Suppose further that we know F, but θ is unknown (For example, we know that our sample is $N(\theta, 2)$ distributed, but we don't know θ).

We can calculate $\hat{\theta} = T(X_1, ..., X_n)$ for our sample.

- To identify the distribution of our estimator, $\hat{\theta}$, we want to obtain more observations of $\hat{\theta}$, but we only have one sample!

How can we get more (approximated) values of $\hat{\theta}$?

Parametric Bootstrap

Why not approximate the distribution, F_θ, using our estimate, $\hat{\theta}$? We can then draw samples from the approximated distribution $F_{\hat{\theta}}$:

$$F_{\hat{\theta}} \quad \overset{simulate}{\rightarrow} \quad \begin{cases} X_1^{*(1)}, \dots, X_n^{*(1)} & \rightarrow \hat{\theta}^{*(1)} \\ \qquad\qquad \vdots & \\ X_1^{*(N)}, \dots, X_n^{*(N)} & \rightarrow \hat{\theta}^{*(N)} \end{cases}$$

We can now estimate properties of $\hat{\theta}$ using these bootstrapped estimates, for example:

$$Var_{boot}(\hat{\theta}) = \frac{1}{B} \sum_{i=1}^{B} \left(\hat{\theta}_i^* - \frac{1}{B} \sum_{j=1}^{B} \hat{\theta}_j^* \right)$$

Non-parametric Bootstrap

Non-parametric Bootstrap

Recall for the parametric bootstrap, we try to approximate the distribution unknown F_θ using our estimate $\hat{\theta}$ as the parameter.

But what if we don't even know the form F?

Non-parametric Bootstrap

For non-parametric bootstrap, we don't assume any distributional form F. In fact, we draw our bootstrap samples from our original sample $X_1, ..., X_n$ (taken from the population $x_1, ..., x_M$):

$$X_1, ..., X_n \quad \overset{sample}{\rightarrow} \quad \begin{cases} X_1^{*(1)}, ..., X_n^{*(1)} & \rightarrow \hat{\theta}^{*(1)} \\ & \vdots \\ X_1^{*(B)}, ..., X_n^{*(B)} & \rightarrow \hat{\theta}^{*(B)} \end{cases}$$

Bootstrap Confidence Intervals

Bootstrap Confidence Intervals

Recall that the goal of the bootstrapping was to learn about our estimator $\hat{\theta}$ (e.g. to estimate variance of $\hat{\theta}$ by calculating the variance of the bootstrapped estimators $\hat{\theta}^*_1, ..., \hat{\theta}^*_B$)

We can also use our bootstrapped estimators, $\hat{\theta}^*_1, ..., \hat{\theta}^*_B$, to calculate approximate **confidence intervals** for θ.

Bootstrap Confidence Intervals

Three types of confidence intervals:

- **Normal interval**:

  $$\hat{\theta} \pm q_{(1-\alpha/2)} \sqrt{Var_{boot}(\hat{\theta})}$$

  where q is the $1 - \alpha/2$ quantile of the standard normal distribution.

- **Percentile interval**:

  $$(\hat{\theta}^*_{(\alpha/2)}, \hat{\theta}^*_{(1-\alpha/2)})$$

- **Pivotal interval**:

  $$(2\hat{\theta} - \hat{\theta}^*_{(1-\alpha/2)}, 2\hat{\theta} - \hat{\theta}^*_{(\alpha/2)})$$

Exercise

Exercise: Bootstrapping

Suppose that a sample consists of the following observations:

78, 86, 97, 91, 83, 89, 92, 88, 79, 68

Then the α-trimmed mean is the average of the inner $(1 - 2\alpha)$ values in the sample. For example, if $\alpha = 0.2$, then the trimmed mean is the average of the inner 60% of the observations

68, 78, **79**, **83**, **86**, **88**, **89**, **91**, 92, 97

Why might we prefer to use the trimmed mean over the usual sample mean in this case?

Exercise: Bootstrapping

We can compute the $\alpha = 0.2$ trimmed mean in R for the above sample using the command `mean(x, trim = 0.2)`. Suppose that we don't know a general formula for the standard error of the trimmed mean. Our goal is to estimate it using bootstrapping.

1. Suppose that we are fairly confident that the data comes from a $\mathit{Normal}(\mu, 8^2)$ distribution (but we don't know μ). Use parametric bootstrap to estimate the standard error of the trimmed mean.
2. Suppose we have absolutely no idea what the underlying distribution of the data is. Use non-parametric bootstrap to estimate the standard error of the trimmed mean.
3. Calculate 95% pivot confidence intervals for each of these estimates.

Hypothesis Testing

Hypothesis Testing

- **Hypothesis testing** is a method of using statistical inference for testing a hypothesis.
- For example, suppose that the DMV claims that the average waiting time is 20 minutes. We might want to test whether the average waiting time at the DMV is in fact more than 20 minutes.
- To test this hypothesis, we examine the DMV waiting times for a random sample of people and determine whether or not we have enough evidence to show that the average waiting time for the population is more than 20 minutes.

In general, we are interested in testing the null hypothesis

$$H_0 : \mu = 20$$

against the alternative hypothesis

$$H_1 : \mu > 20$$

Hypothesis Testing

The terminology we use when conducting a hypothesis test is that we either:

- Have enough evidence (based on our statistic, $T(X_1, ..., X_n)$) to reject the null hypothesis H_0 in favor of the alternative hypothesis H_1, or
- We do not have enough evidence to reject the null hypothesis H_0, and so our data is consistent with the null hypothesis.
    - This doesn't mean that the alternative hypothesis, H_1, is false, just that we don't have enough evidence to show that it is true!

Hypothesis Testing

Suppose we are testing a hypothesis about the mean, μ, of a distribution, for example:

$$H_0 : N(\mu = 5, 4)$$

against the **composite** alternative hypothesis that

$$H_1 : N(\mu > 5, 4)$$

Given that we have observed a sample $X_1, ..., X_{20}$, what statistic, $T(X_1, ..., X_{20})$, could we consider in order to determine whether we have enough evidence to suggest that $\mu > 5$ rather than $\mu = 5$?

$$T(X_1, ..., X_n) = \frac{\text{observed/estimated value} \; - \; \text{null value}}{\text{SD of estimator}} = \frac{\bar{X}_n - \mu_0}{SD(\bar{X}_n)}$$

Hypothesis Testing

Recall that we are testing the null hypothesis that the true distribution is

$$H_0 : N(\mu = 5, 4)$$

versus the alternative hypothesis that the true distribution is

$$H_1 : N(\mu > 5, 4)$$

Suppose that from our sample, $X_1, ..., X_{20}$, we observed

$$\bar{X}_n = 5.4 \;\text{ and }\; SD(\bar{X}_n) = SD(X)/\sqrt{n} = 0.45$$

$$T(X_1, ..., X_n) = \frac{\bar{X}_n - \mu_0}{SD(\bar{X}_n)} = \frac{5.4 - 5}{0.45} = 0.89$$

This is not a big number – our data seems fairly consistent with the null hypothesis (our observed sample mean is within 1 SD of $\mu_0 = 5$).

Hypothesis Testing

Recall that we are testing the null hypothesis that the true distribution is

$$H_0 : N(\mu = 5, 4)$$

versus the alternative hypothesis that the true distribution is

$$H_1 : N(\mu > 5, 4)$$

If, on the other hand, we observed

$$\bar{X}_n = 7 \text{ and } SD(\bar{X}_n) = 0.45$$

$$T(X_1, ..., X_n) = \frac{\bar{X}_n - \mu_0}{SD(\bar{X}_n)} = \frac{7-5}{0.45} = 4.44$$

which is much larger. In this case, we have enough evidence to reject H_0, (that the true mean is 5) in favor of the alternative, H_1 (that the true mean is > 5).

Hypothesis Testing

How do we determine what values of $T(X_1, ..., X_n)$ are "big" enough such that we can reasonably conclude that our null hypothesis is false?

1. First, we identify the distribution of the test statistic $T(X_1, ..., X_n)$ assuming the null hypothesis is true.
2. We then calculate the probability of observing a test statistic that is as or "more extreme" (in terms of the alternate hypothesis) than $T(X_1, ..., X_n)$. This probability is called the **p-value**.

Hypothesis Testing

In our example, we had

$$T(X_1, ..., X_n) = \frac{\bar{X}_n - \mu_0}{SD(\bar{X}_n)}$$

which, if we assume the null hypothesis is true (i.e. $X_i \sim N(\mu_0, 4)$, where $\mu_0 = 5$), then

$$T(X_1, ..., X_n) \sim N(0, 1)$$

Next, we want to calculate the probability of observing a test statistic that is as or "more extreme" than $T(X_1, ..., X_n)$.

- Since our alternate hypothesis is $H_1 : \mu > 5$, a test statistic that is as or "more extreme" corresponds to a value **greater** than our observed $T(X_1, ..., X_n)$.
- If our alternate hypothesis was $H_1 : \mu < 5$, a test statistic that is as or "more extreme" corresponds to a value **less** than $T(X_1, ..., X_n)$.

Hypothesis Testing

In our example, we had $\bar{X}_n = 7$ and $SD(\bar{X}_n) = SD(X)/\sqrt{n} = 2/\sqrt{20} = 0.45$, so that out test statistic is

$$T(X_1, ..., X_n) = \frac{\bar{X}_n - \mu_0}{SD(\bar{X}_n)} = 4.44$$

We know that if the H_0 is true then $T(X_1, ..., X_n) \sim N(0, 1)$, so our p-value for the hypothesis test with $H_1 : \mu > 5$ is given by

$$P(Z > T(X_1, ..., X_n)) = P(Z > 4.44) = 0.0000045$$

This tells us that, assuming H_0 is true, the probability of seeing a test statistic as extreme or more extreme than what we have observed is so extremely small, that we can reject $H_0 = 5$ in favor of $H_1 < 5$

Exercise

Exercise: Hypothesis Testing

Suppose that you had read an article which claimed that the average weight of red pandas was 6.3kg. However, suppose that you had recently weighed $n = 121$ red pandas and found that these pandas had an average weight of 5.1kg.

Assume that it is well-known that the standard deviation of the weights of red pandas was 1.4.

Do you have enough evidence to claim with 95% probability that the true average weight of red pandas is in fact less than 6.3kg?

Hypothesis Testing

Recall that we are trying to make a decision on whether to reject H_0 in favor of H_1 based on a statistic $T(X_1, ..., X_n)$.

- **Rejection region**: the range of values of $T(X_1, ..., X_n)$ for which we reject the null hypothesis, H_0.
- **Acceptance region**: the range of values of $T(X_1, ..., X_n)$ for which we do not reject the null hypothesis, H_0.

Hypothesis Testing

There are two primary types of error associated with hypothesis testing

1. **Type I error**: the error arising from rejecting H_0 when it is actually true
   - A type I error occurs with probability α
   - The **significance level** of the test is equal to α
2. **Type II error**: the error arising from accepting H_0 when it is actually false
   - A type II error occurs with probability β
   - The **power** of a test is equal to $1 - \beta$ (the probability of rejecting H_0 when it is actually false)
     - The power can be thought of as the ability of the test to detect an effect if the effect actually exists

Hypothesis Testing: The Neyman-Pearson Approach

Hypothesis Testing

We will now examine hypothesis testing from the Neyman-Pearson (NP) approach:

- Suppose that we observe a sample of DMV waiting times, $X_1, ..., X_n$, and we are interested identifying the distribution, F_θ, of waiting times.
- Suppose that we know that $F_\theta = \mathit{Poisson}(\theta)$, but we don't know θ.
- Then based on our observed data $X_1, ..., X_n$, we might want to test the null hypothesis that

  $$H_0 : \text{ the true dist of waiting times is } \mathit{Poisson}(\theta_0 = 20)$$

  against the alternative hypothesis that

  $$H_1 : \text{ the true dist of waiting times is } \mathit{Poisson}(\theta_1 = 30)$$

Hypothesis Testing

Note that the hypotheses

$$H_0 : \mathit{Poisson}(\theta_0 = 20)$$

and

$$H_1 : \mathit{Poisson}(\theta_1 = 30)$$

are both examples of **simple hypothesis** (each hypothesis assumes only *one value* for the parameter).

All hypotheses in the Neyman-Pearson framework are of this type.

Hypothesis Testing

The Neyman-Pearson approach to hypothesis testing:

- Fix α (the probability of a Type I error) at some small pre-defined value
- Us the likelihood ratio to find the corresponding test with the highest power (we want to maximize the probability of rejecting the H_0 when it is false)

Hypothesis Testing

Suppose we are testing the *simple* null hypothesis

$$H_0 : \theta = \theta_0$$

against the *simple* alternative

$$H_1 : \theta = \theta_1$$

The **likelihood ratio** is the ratio of the likelihood function, $f_0(X)$, of the data assuming that $\theta = \theta_0$ (H_0 is true) to the likelihood function, $f_1(X)$, of the data assuming that $\theta = \theta_1$ (H_1 is true):

$$\frac{f_0(X_1, ..., X_n)}{f_1(X_1, ..., X_n)}$$

Hypothesis Testing

Assuming H_0 and H_1 are both simple, the **Neyman-Pearson lemma**, tells us that:

If the likelihood ratio test (LRT) that rejects H_0 when

$$\frac{f_0(X)}{f_1(X)} < c(\alpha)$$

has significance level α (probability of incorrectly rejecting H_0 when it is true), then any other test with significance level $\alpha' \leq \alpha$ has power less than or equal to that of the LRT.

Exercise

Exercise: Neyman-Pearson hypothesis testing

Suppose we have a sample $X_1, ..., X_n$ such that $X_i \sim Normal(\mu, 16)$ where μ is unknown. We want to test the null hypothesis

$$H_0 : \mu = 10$$

against the alternative hypothesis

$$H_1 : \mu = 15$$

Find the best test with sample size $n = 16$ and significance level $\alpha = 0.05$. What is the power of this test?