36-700: Probability and Mathematical Statistics I **Fall 2016**

Lecture 14: September 30

Lecturer: Siva Balakrishnan

14.1 Review and Outline

Last class we discussed:

- The Empirical CDF
- Estimating functionals via the plug-in method

Today we will talk about the bootstrap. The bootstrap is a method to assess variability of a point estimate, that was introduced by Brad Efron in the late 70s. It is a simple method (although its validity often relies on technical conditions) and is widely used. It is one of the methods that is often credited with pushing the discipline of statistics towards its modern computer intensive state.

14.2 The Setup

Suppose we have data $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$, where μ is unknown but σ^2 is known. The MLE is:

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

In this case, one can easily verify that:

$$C_n = \left[\hat{\mu} - \frac{\sigma z_{\alpha/2}}{\sqrt{n}}, \hat{\mu} + \frac{\sigma z_{\alpha/2}}{\sqrt{n}}\right],$$

is an exact $1 - \alpha$ confidence interval for the unknown μ.

In the non-parametric case we discussed in the last lecture, the estimator above for the mean is still the plug-in estimator, even if we do not know anything about the underlying distribution. A natural question is if there is a similar agnostic confidence interval.

The bootstrap is surprisingly more general. It works for other (non-linear) statistics just as well, and can be useful even in parametric cases because understanding the distribution of an arbitrary statistic analytically can be quite challenging.

14.3 Resampling

Before we get to the main ideas let us get used to the idea of resampling a dataset. If I have a dataset $X_1, \ldots, X_n$, then I could resample the dataset, i.e. sample from the dataset uniformly with replacement.

Suppose we sample m times from the dataset with replacement. Following standard bootstrap terminology we will denote the resampled dataset: $X_1^*, \ldots, X_m^*$.

Notice that we m could be smaller than n, equal to n or even larger than n. Further, we could repeat this process many times to generate many different datasets from the original dataset.

14.4 Bootstrap Variance Estimation

Suppose we have a the plug-in estimate of some functional:

$$\widehat{T}(F) = T(\widehat{F}_n) = T(X_1, \ldots, X_n),$$

and would like to estimate the variance $\text{Var}_F(T(X_1, \ldots, X_n))$. The notation is explicit to remind you that this depends on the unknown distribution F.

The idea of the bootstrap is to approximate:

$$\text{Var}_F(T(X_1, \ldots, X_n)) \approx \text{Var}_{\widehat{F}_n}(T(X_1^*, \ldots, X_n^*)),$$

and then to approximate the RHS by sampling from $\widehat{F}_n$ many times and appealing to the LLN.

Computationally, the procedure is quite straightforward:

1. We select a large number B.
2. We repeat the following B times:
   - Create a new dataset by resampling the original one n times, i.e. create $X_1^*, \ldots, X_n^*$.
   - Estimate the functional on this dataset as $T(X_1^*, \ldots, X_n^*)$. Denote the i^{th} estimate as T_i^*.
3. Estimate the variance of our point estimate of the functional as:

$$\widehat{V} = \frac{1}{B} \sum_{i=1}^{B} \left(T_i^* - \frac{1}{B} \sum_{i=1}^{B} T_i^* \right)^2.$$

To summarize, the bootstrap principle says we can approximate the distribution F by $\widehat{F}_n$ in order to assess the variability of a point estimate. Further, we can approximate this variability via a Monte Carlo approximation, i.e., by resampling the dataset several times. There are two approximation steps when using the bootstrap.

14.5 Bootstrap Confidence Intervals

The next logical step is to try to obtain confidence intervals. There are several ways to do this:

1. **The Normal Interval:** If we assumed that the distribution of the statistic is (approximately) Gaussian then all we need is an estimate of the variance to construct an interval. We calculated this in the previous section. So we could use:

$$[\widehat{\theta} - \sqrt{\widehat{V}} z_{\alpha/2}, \widehat{\theta} + \sqrt{\widehat{V}} z_{\alpha/2}].$$

   The Gaussian assumption typically does not hold in non-parametric cases so most often in practice people use the third method.

2. **Percentile Intervals:** Suppose that we have the B bootstrap estimates: $T_1^*, \ldots, T_B^*$. We could use the quantiles of this distribution to construct a confidence set, i.e. find two values $t_{\alpha/2}$ and $t_{1-\alpha/2}$ such that:

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(T_i \leq t_{\alpha/2}) \leq \alpha/2$$
$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(T_i \geq t_{1-\alpha/2}) \leq \alpha/2,$$

   and then use the confidence interval $[t_{\alpha/2}, t_{1-\alpha/2}]$.

3. **Pivotal Intervals:** Here the idea is to approximate the distribution of $T - \mu$ by the distribution of $T^* - T$, where T is the value of the statistic computed on the full data, μ is the true parameter and T^* is what we compute on the bootstrap samples. Let

$$R_i^* = T_i^* - T,$$

   then we would compute the $\alpha/2$ and $1 - \alpha/2$ quantiles of the distribution of R^*, i.e. we would find $r^*_{1-\alpha/2}$ and $r^*_{\alpha/2}$ such that:

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(R_i^* \leq r^*_{\alpha/2}) \leq \alpha/2$$
$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(R_i^* \geq r^*_{1-\alpha/2}) \leq \alpha/2,$$

and use the interval:

$$[T - r^*_{1-\alpha/2}, T - r^*_{\alpha/2}].$$

Notice that both the signs are negative. This is because the r^* are quantiles of the difference distribution not of the statistic itself. As an exercise show that this is correct.

14.6 The Parametric Bootstrap

In parametric settings, there is an alternative popular way to bootstrap. Here rather than re-sampling the empirical distribution, we resample the estimated distribution $f_{\widehat{\theta}}$.

Here is a description of the procedure for variance estimation:

1. Compute a point estimate $\widehat{\theta}$ and select a large number B.
2. We repeat the following B times:
   - Create a new dataset by drawing n samples from $f_{\widehat{\theta}}$, i.e. generate $X_1^*, \ldots, X_n^* \sim f_{\widehat{\theta}}$.
   - Estimate the functional on this dataset as $T(X_1^*, \ldots, X_n^*)$. Denote the i^{th} estimate as T_i^*.
3. Estimate the variance of our point estimate of the functional as:

$$\widehat{V} = \frac{1}{B} \sum_{i=1}^{B} \left(T_i^* - \frac{1}{B} \sum_{i=1}^{B} T_i^* \right)^2.$$

14.7 Why/when does the Bootstrap work?

This turns out to be a somewhat delicate question. The name bootstrap is supposed to remind you of the phrase "pull oneself up be their own bootstraps". It might be a bit absurd that re-using the same data in this fashion can give "new information".

In point estimation for instance, why can't I generate many re-samples and then average them to obtain a "better" (say lower MSE) point estimate? The answer is that doing this will result in a better estimate of our point estimate (not of the true parameter).

It is worth pointing out that in general T_i^* (the statistic computed on the i^{th} bootstrap sample) will be far from T (the statistic computed on the original data). The bootstrap is computing the distribution of: $T_i^* - T$ instead of what we want $T - \mu$ (where μ is the true parameter). The subtle point is that although these distributions are "centered" at different points, they can still have the same variance or tail behaviour.