7.2.3 Bootstrap Confidence Interval

Sometimes, it is difficult to known distribution of statistic of interest and/or difficult to compute its variance, e.g., a trimmed mean, a ratio of two random variables. How to construct the confidence intervals?

Suppose we have infinite resources are are interested in the distribution of a statistics, μ, We might take n samples of the size m from the population and for each sample calculate $\hat{\mu}$. We could determine the distribution of μ based on these samples.

But resources are always limiting! We can only afford to sample one from the population, e.g., clinical trial each patient might cost $1000 or more to enroll into the study.

In a bootstrap sample, we treat the sample as it were the population. We sample with replacement. For each bootstrap sample, calculate $\hat{\mu}_j^B$, $j = 1, \cdots, B$ where the B indicates that the estimate was computed for a bootstrap sample. The distribution of μ is approximated by the distribution of $\hat{\mu}_j^B$. Intuitively it makes sense that the approximation is better in large sampels. Here we consider how to use the distribution of $\hat{\mu}_j^B$'s to compute the bootstrap percentile interval. Consider the following algorithm:

1. Generate B bootstrap samples;
2. Compute $\hat{\mu}_j^B$ for each $j = 1, \cdots, B$ samples;
3. Compute the variance $V^B(\hat{\mu})$ of the $\hat{\mu}_j^B$ and use this to estimate the variance of $\hat{\mu}$;

4a Construct approximately normal-based confidence intervals:

$$\hat{\mu} \pm Z_{\alpha/2}\sqrt{V^B(\hat{\mu})}.$$

4b Order $\hat{\mu}_j^B$ and choose $(\hat{\mu}_{(\alpha/2*B)}, \hat{\mu}_{((1-\alpha/2)B)})$ as $1-\alpha$ confidence interval based on the percentile.

Example 7.2.5 (Law School Data)**.** *They tried to study the correlation between LSAT and GPA. They want to estimate the 95% CI of the correlation.*

The standard procedure (Fisher Transformation) for generating a pivot is to use

$$\xi = 0.5 \log \frac{1+\rho}{1-\rho}, \hat{\xi} = 0.5 \log \frac{1+\hat{\rho}}{1-\hat{\rho}},$$

where

$$\hat{\xi} \sim N(\xi, Var(\hat{\xi})).$$

This interval is asymptotically valid. But there is some issues with sample samples. There exists some bias issues.

Consider a statistic that has a skewed distribution. To create a "standard" confidence interval we might transform the statistic to give it a symmetric distribution. For example, we might take the logs of the data and create the endpoints of the confidence interval and then use the inverse transformation (exponential) to go back to the original scale. The endpoints based ont the transformed version are usually different than the endpoints that would result if you used the original scale without the transformation.

7.3 Confidence intervals for two populations

7.3.1 Difference of two means

There are two cases for deriving the confidence intervals for comparing the difference of two means.

(i) **Equal Variances:** If the variances is assumed to be equal, their common value is estimated as a weighted average of the two individual sample variances just like doing the hypothesis testing. This is referred to as *pooling*.

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}.$$

The pooled variance has more degrees of freeedom as n_1+n_2-2. Then the $1-\alpha$ CI is

$$\bar{y}_1 - \bar{y}_2 \pm t_{\alpha/2,n_1+n_2-2} s_p \sqrt{\frac{1}{n_1}+\frac{1}{n_2}}.$$

(ii) **Unequal variance case:** if the variances are not the same, then the $1-\alpha$ CI is

$$\bar{y}_1 - \bar{y}_2 \pm t_{\alpha/2,df} \sqrt{\frac{s_1^2}{n_1}+\frac{s_2^2}{n_2}},$$

where df is calculated based on the Satterthwaite approximation.

Example 7.3.1 (Microarray dataset)**.** *Construct 95% confidence interval.*

7.3.2 Confidence intervals for two variances

Assume that we have two independent random samples from two populations. Let σ_1^2 and σ_2^2 be the variances of two populations. s_1^2 and s_2^2 be the sample variances. Then $1-\alpha$ confidence interval for a ratio of variances of two normal populations $\frac{\sigma_1^2}{\sigma_2^2}$ is

$$(\frac{s_1^2}{s_2^2}\frac{1}{F_{Low}}, \frac{s_1^2}{s_2^2}F_{High}),$$

where $F_{Low} = F_{1-\alpha/2,n_1-1,n_2-1}$ and $F_{High} = F_{1-\alpha/2,n_2-1,n_1-1}$.

Example 7.3.2 (Microarray dataset)**.** *Construct 95% for comparing two variances for the 1-st gene.*

7.3.3 Two population proportions

THe need for confidence intervals on the difference of two proportions is frequently encountered. For instance, we might wish to estimate the difference in the proportions of voters in two populations who favor a particular candidate.

Labeling the populations as 1 and 2, the traditional confidence interval is

$$\hat{p}_1 - \hat{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Agresti and Caffo also provided an improved interval. Let $\tilde{p}_i = \frac{Y_i+1}{n_i+2}$. Then the improved interval is

$$\tilde{p}_1 - \tilde{p}_2 \pm z_{\alpha/2}\sqrt{\frac{\tilde{p}_1(1-\tilde{p}_1)}{n_1+2} + \frac{\tilde{p}_2(1-\tilde{p}_2)}{n_2+2}}.$$

Example 7.3.3 (Exit Poll)**.**

7.4 Sample Size Determination

When considering the confidence interval, one needs to find a balance between the coverage probability and the length of the interval or the margin of error which is defined as half of the length.

For a fixed sample,

1. Increase the coverage probabilites $\rightarrow$ larger margin of error;

2. Decrease the margin of error $\rightarrow$ decrease the coverage probabilities.

What if we want to have a "short" interval without sacrificing the coverage probabilities?

For a CI on a single mean, assuming a known variance σ^2. If we want to have a $1-\alpha$ confidence interval with a certain MOE, then the required sample size is

$$n \geq \frac{\sigma^2 (z_{1-\alpha/2})^2}{MOE^2}.$$

If considering the population proportion, then the sample size for the Agresti and Caffo CI on a single population is

$$n \geq \frac{(z_{1-\alpha/2})^2}{4MOE^2} - 4.$$