An Elaboration of the Motivation and Theory Underlying the Bootstrap-t Interval

Imagine we have a sample $\mathcal{S}$ of size n that we assume was drawn from the normal distribution $N(\mu, \sigma^2)$, and the attribute that we care about is the population average (μ). Then we know that $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$ and also that

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

and also that

$$Z \equiv \frac{\overline{X} - \mu}{\widetilde{\sigma}/\sqrt{n}} \sim t_{(n-1)}.$$

Thus we can write

$$Pr(c_{lower} \leq Z \leq c_{upper}) = 1 - p$$

where c_{lower} is the $p/2$ quantile of the $t_{(n-1)}$ distribution and c_{upper} is the $1 - p/2$ quantile of the $t_{(n-1)}$ distribution.

The probability statement above can be arranged to yield a *random interval* for μ as follows:

$$\begin{aligned}
Pr\left(c_{lower} \leq Z \leq c_{upper}\right) &= 1 - p \\
Pr\left(c_{lower} \leq \frac{\overline{X}-\mu}{\widetilde{\sigma}/\sqrt{n}} \leq c_{upper}\right) &= 1 - p \\
Pr\left(\overline{X} - c_{upper} \times \frac{\widetilde{\sigma}}{\sqrt{n}} \leq \mu \leq \overline{X} - c_{lower} \times \frac{\widetilde{\sigma}}{\sqrt{n}}\right) &= 1 - p
\end{aligned}$$

and so a $(1 - p) \times 100\%$ random interval for μ is

$$\left[\overline{X} - c_{upper} \times \frac{\widetilde{\sigma}}{\sqrt{n}}, \overline{X} - c_{lower} \times \frac{\widetilde{\sigma}}{\sqrt{n}}\right].$$

The corresponding $(1 - p) \times 100\%$ *observed interval* is found by substituting *random* quantities above by their corresponding *observed* sample estimates:

$$\left[\overline{x} - c_{upper} \times \frac{\widehat{\sigma}}{\sqrt{n}}, \overline{x} - c_{lower} \times \frac{\widehat{\sigma}}{\sqrt{n}}\right].$$

Now, the development above assumed that the attribute we care above is the population average, and with that assumption came two convenient consequences:

- We know how to calculate $SD[\overline{X}]$
- The distribution of Z is known

So, how would we proceed if we wanted to calculate a confidence interval for an attribute $a(\mathcal{P})$ other than the average?

Well, analogously the estimator $\widetilde{a}(\mathcal{S})$ has some distribution, just like $\overline{X}$ did above, but now we don't know what it is. Similarly, the standardized ratio

$$Z = \frac{\widetilde{a}(\mathcal{S}) - a(\mathcal{P})}{\widetilde{SD}[\widetilde{a}(\mathcal{S})]}$$

has some distribution – no longer the $t_{(n-1)}$ distribution – but some distribution nonetheless. From this distribution (whatever it is) we can in theory determine c_{lower} and c_{upper} which are respectively its $p/2$ and $1 - p/2$ quantiles and write:

$$Pr\left(c_{lower} \leq Z \leq c_{upper}\right) = 1 - p$$

As before this statement can be pivoted, (socially) isolating for $a(\mathcal{P})$ in the middle:

$$\begin{aligned} Pr\left(c_{lower} \leq Z \leq c_{upper}\right) &= 1 - p \\ Pr\left(c_{lower} \leq \frac{\widetilde{a}(\mathcal{S}) - a(\mathcal{P})}{\widetilde{SD}[\widetilde{a}(\mathcal{S})]} \leq c_{upper}\right) &= 1 - p \\ Pr\left(\widetilde{a}(\mathcal{S}) - c_{upper} \times \widetilde{SD}[\widetilde{a}(\mathcal{S})] \leq a(\mathcal{P}) \leq \widetilde{a}(\mathcal{S}) - c_{lower} \times \widetilde{SD}[\widetilde{a}(\mathcal{S})]\right) &= 1 - p \end{aligned}$$

This yields the $(1 - p) \times 100\%$ *random interval*

$$\left[\widetilde{a}(\mathcal{S}) - c_{upper} \times \widetilde{SD}[\widetilde{a}(\mathcal{S})], \widetilde{a}(\mathcal{S}) - c_{lower} \times \widetilde{SD}[\widetilde{a}(\mathcal{S})]\right].$$

As above, we want to convert this into an *observed interval.* To do so, we could replace $\widetilde{a}(\mathcal{S})$ with the sample estimate $a(\mathcal{S})$. However, we also need

(1) to be able to estimate $SD[\widetilde{a}(\mathcal{S})]$, and
(2) to know the distribution of Z in order to determine c_{lower} and c_{upper}

To resolve each of these issues, we'll employ the bootstrap. In particular, to address (1), we will approximate the distribution of $\widetilde{a}(\mathcal{S})$ using the boostrap distribution determined by $a(S_1^\star), a(S_2^\star), \ldots a(S_B^\star)$, and thus estimate $SD[\widetilde{a}(\mathcal{S})]$ by the bootstrap standard deviation

$$\widehat{SD}_\star[\widetilde{a}(\mathcal{S})] = \sqrt{\frac{\sum_{b=1}^{B}\left(a(S_b^\star) - \overline{a}^\star\right)^2}{B - 1}}$$

where $\overline{a}^\star = \frac{1}{B}\sum_{b=1}^{B} a(S_b^\star)$ is the average attribute value over all of the bootstrap samples.

To address (2), we will approximate the distribution of Z by the bootstrap distribution of

$$Z^\star = \frac{\widetilde{a}(\mathcal{S}^\star) - a(\mathcal{P}^\star)}{\widetilde{SD}[\widetilde{a}(\mathcal{S}^\star)]} = \frac{\widetilde{a}(\mathcal{S}^\star) - a(\mathcal{S})}{\widetilde{SD}[\widetilde{a}(\mathcal{S}^\star)]}.$$

Thus c_{lower} and c_{upper} will be *estimated* as the $p/2$ and $1 - p/2$ quantiles of $z_1^\star, z_2^\star, ..., z_B^\star$ where

$$z_b^\star = \frac{a(\mathcal{S}_b^\star) - a(\mathcal{S})}{\widehat{SD}\left[a(\mathcal{S}_b^\star)\right]}$$

(Note that a double bootstrap will likely be necessary to determine $\widehat{SD}\,[a(\mathcal{S}_b^\star)]$).

Thus the $(1-p)\times 100\%$ *observed interval* is given by

$$\left[a(\mathcal{S}) - c_{upper} \times \widehat{SD}_\star[\widetilde{a}(\mathcal{S})], a(\mathcal{S}) - c_{lower} \times \widehat{SD}_\star[\widetilde{a}(\mathcal{S})]\right].$$

This is what is referred to as the bootstrap-t confidence interval for $a(\mathcal{P})$.