# A5Q3

**Summary**

Often, we can use bootstrap to estimate the sampling distribution of a pivotal quantity. For example, if we have the population $\mathcal{P}$ and samples $S$, we can see that a quantity $Z = \frac{\tilde{a}(S)-a(\mathcal{P})}{\hat{SD}[\tilde{a}(S)]}$ is approximately pivotal and the distribution can be approached by t-density.

When we use bootstrap to approximate the sampling distribution of $Z$, we will follow the following bootstrap procedure:

- Obtain a sample $S$ from the population $\mathcal{P}$.
- generate bootstrap samples $S_1^*, ...S_B^*$ from the sample $S$ and calculate $Z_B^* = \frac{\tilde{a}(S_b)-a(\mathcal{S})}{\hat{SD}[\tilde{a}(S_b^*)]}$, we get the bootstrap estimate of the distribution $\{Z_1^*, ...Z_B^*\}$. (We need to use an estimate of standard deviation $\hat{SE}[a(S_i)] = \hat{SD}[\tilde{a}(S_i)] = \frac{\hat{\sigma}}{\sqrt{n}} \times \sqrt{\frac{N-n}{N-1}}$).
- With a $p \in (0, 1)$ we are able to obtain the upper and lower constants $Z_{upper}^*$ and $Z_{lower}^*$ so that $1 - p = Pr(Z*_{lower} \leq Z^* \leq Z_{upper}^*) \approx Pr(Z*_{lower} \leq Z \leq Z_{upper}^*)$.
- Then we can get a confidence interval from the bootstrap estimate: $[a(s) - Z_{upper}^* \hat{SD}[\tilde{a}(S)], a(S) - Z_{lower}^* \hat{SD}[\tilde{a}(S)]]$.

**Example**

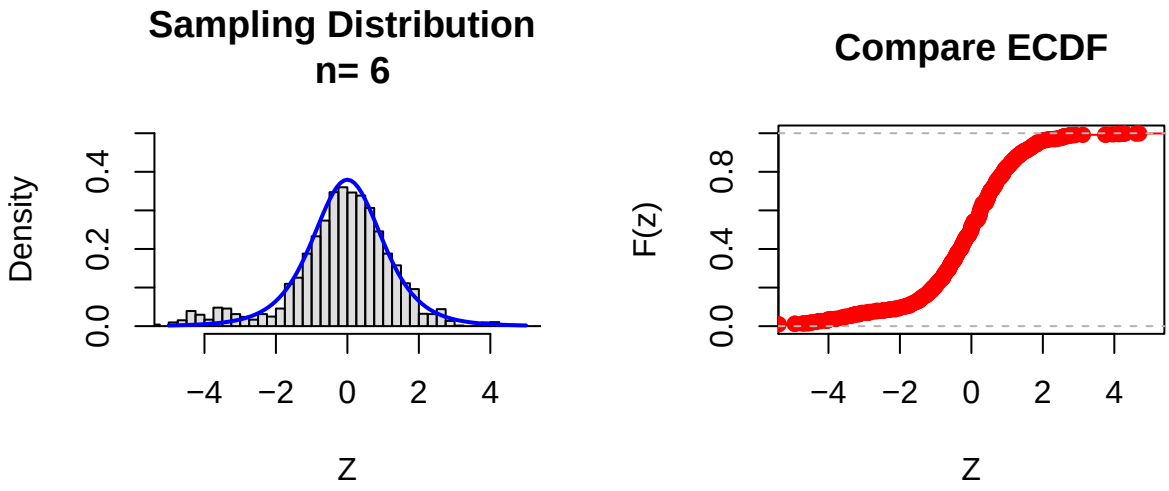
'cars.csv' contains information of used cars with different conditions. We will use the first 100 rows from 'cars.csv' as the population. We will use a sample with n=5 and bootstraps with M=5000. The target column is the price of cars (price_usd).

```
car <- read.csv('cars.csv',header = T)[1:100, ]
popCars <- rownames(cars)
```

```
M <- 5000
n = 6
set.seed(341)
samCars <- sample(popCars, n, replace = FALSE)
samStar <- sapply(1:M, FUN = function(m) sample(samCars, n, replace = TRUE))
aveSam <- mean(car[samCars, "price_usd"])

avesBoot <- apply(samStar, MARGIN = 2, FUN = function(s) {
    mean(car[s, "price_usd"])
})

SEaveBoot <- apply(samStar, MARGIN = 2, FUN = function(s) {
    se.avg(car[s, "price_usd"])
})
```

```r
ZBoot <- (avesBoot - aveSam)/SEaveBoot
```

```r
par(mfrow = c(1, 2))
brk = seq(-50, 50, by = 0.25)
hist(ZBoot, freq = FALSE, breaks = brk, col = adjustcolor("grey", 0.5),
     main = paste("Sampling Distribution \n n=",
    n), xlab = "Z", xlim = c(-5, 5), ylim = c(0, 0.5))
lines(x = seq(-5, 5, 0.1), y = dt(x = seq(-5, 5, 0.1), df = n - 1),
      col = "blue",
    lwd = 2)

plot(ecdf(ZBoot), xlim = c(-5, 5), col = "red", main = "Compare ECDF",
     xlab = "Z",
    ylab = "F(z)")
```




A Bootstrap-$t$ confidence interval using $\hat{SD}[\tilde{a}(S)]$ is:

```r
samCarPrice = car[samCars, "price_usd"]

zStar.lower = quantile(ZBoot, 0.025)
zStar.upper = quantile(ZBoot, 0.975)

round(mean(samCarPrice) - c(zStar.upper, zStar.lower) * se.avg(samCarPrice),
    2)
```

```
##    97.5%     2.5%
##  3351.46 16842.96
```

Conclusion:

Both graphs show a roughly normal distribution. We see that the bootstrap distribution nearly matches the sampling distribution of the pivotal quantity except there are some data at two tails. This procedure works. And we get the confidence interval (3351.46, 16842.96) based on the standard deviation of the estimator.