Question 3

Bootstrap-t Confidence Intervals

The t-distribution approximation of a sampling distribution works for certain attributes usually when $\tilde{a}(S)$ is approximately normal over all possible samples. If $a(P)$ is the median or a measure of skewness, the t-distribution cannot be a good approximation

In the Bootstrap-t Confidence Interval approach we use bootstrap to approximate the sampling distribution of a pivotal quantity and then construct a confidence interval based on it. This approach is different from that of t-approximation since we also require an estimate of the standard error of an attribute.

How to calculate confidence interval of the pivotal quantity using bootstrap estimate?

To use the bootstrap to approximate the sampling distribution of Z, we first estimate the population P with the estimate $P^* = S$ (the sample). Then we estimate the sample S with the bootstrap sample S^* and generate $S_1^*, ..., S_B^*$. We then calculate

$$Z_b^* = \frac{\tilde{a}(S_b) - a(S)}{\hat{SD}[\tilde{a}(S_b^*)]}$$

From above the bootstrap estimate of the sampling distribution is $\{z_1^*, ..., z_b^*\}$. Using a p $\in$ (0,1) the bootstrap estimate we can find Z_{lower}^* and Z_{upper}^* such that

$$1 - p = Pr(Z_{lower}^* \leq Z^* \leq Z_{upper}^*) \approx Pr(Z_{lower}^* \leq Z \leq Z_{upper}^*)$$

And a confidence interval of the pivotal quantity using the above bootstrap estimate is

$$(a(S) - Z_{upper}^* \times \hat{SD}[\tilde{a}(S)], a(S) - Z_{lower}^* \times \hat{SD}[\tilde{a}(S)])$$

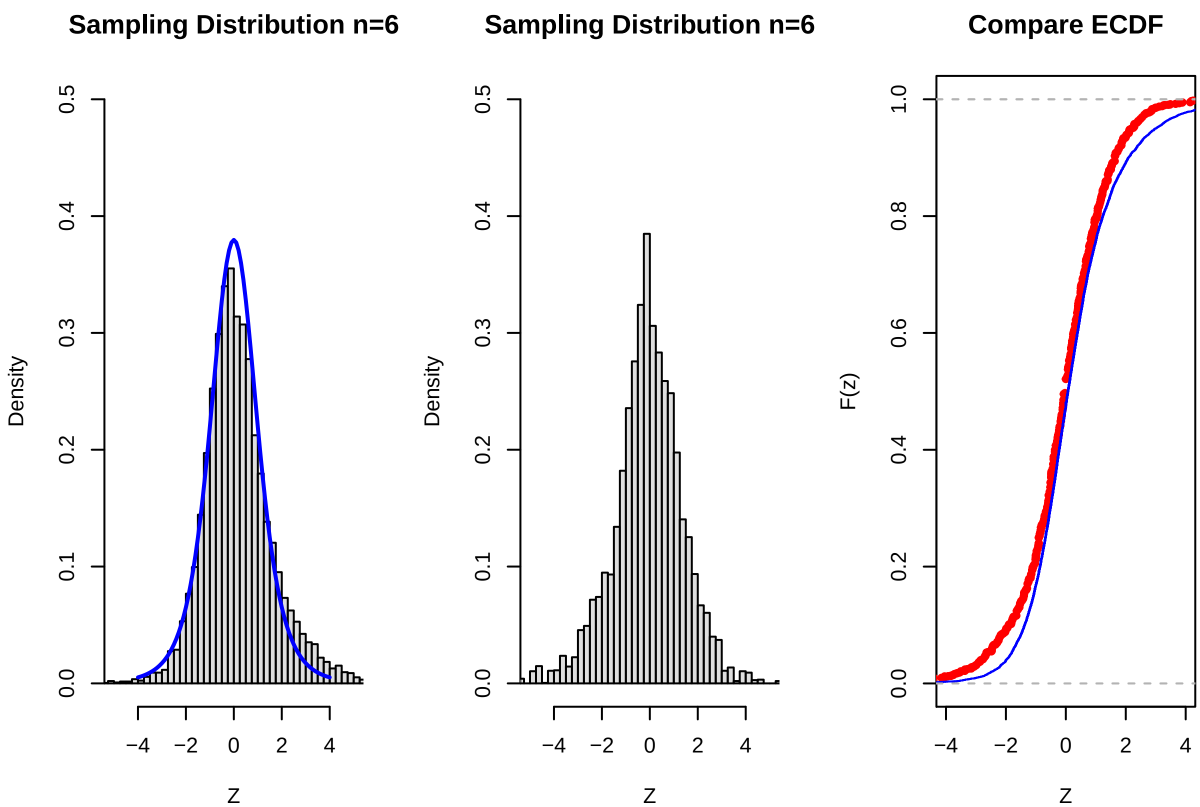
Example 1

We will now look at the spotify data and compare the bootstrap estimate to the sampling distribution for the average energy of each song.

```
ZPop <- (avesSamp - mean(spotify[,"energy"]))/SEaveSamp # Zi
ZBoot <- (avesBoot - aveSam)/SEaveBoot
```

The plots below compares both the approaches and we can see that the histogram and the t_5 line (in blue) are quite close

```
savePar <- par(mfrow = c(1, 3))
brk = seq(-60, 60, by = 0.25)
hist(ZPop, freq = FALSE, breaks = brk, col = adjustcolor("grey", 0.5), main = paste("Sampling Distributi
lines(x = seq(-4, 4, 0.1), y = dt(x = seq(-4, 4, 0.1), df = n - 1), col = "blue",
lwd = 2) # t5 line
hist(ZBoot, freq = FALSE, breaks = brk, col = adjustcolor("grey", 0.5), main = paste("Sampling Distribu
plot(ecdf(ZBoot), xlim = c(-4, 4), col = "red", main = "Compare ECDF", xlab = "Z",
ylab = "F(z)")
lines(ecdf(ZPop), col = "blue")
```



The General Approach to calculating Bootstrap-t confidence interval:

For a given sample S, attribute $a(S)$ and standard error $\hat{SD}[\tilde{a}(S)]$.

- We first calculate $a(S)$ and $\hat{SD}[\tilde{a}(S)]$ based on the sample. We then generate B bootstrap samples $S_1^*, ..., S_B^*$ from S *with replacement*.
- For each of the B bootstrap samples from above, we then calculate $a(S_b^*)$ and $\hat{SD}[\tilde{a}(S_b^*)]$ such that

$$Z_b^* = \frac{a(S_b^*) - a(S)}{\hat{SD}[\tilde{a}(S_b^*)]}$$

- From the values $z_1^*, ..., z_B^*$ we find $c_{lower} = Q_z(p/2)$ and $c_{upper} = Q_z(1 - p/2)$ The pair $\{c_{lower}, c_{upper}\}$ is nothing but quantiles from $\{z_1^*, ..., z_B^*\}$ which are estimates of the sampling distribution.
- We can now find a $100(1 - p)\%$ bootstrap-t confidence interval, given by:

$$(a(S) - c_{upper} \times \hat{SD}[\tilde{a}(S)], a(S) - c_{lower} \times \hat{SD}[\tilde{a}(S)])$$

Example 2

Let's calculate the bootstrap-t confidence interval using the standard error $\hat{SD}[\tilde{a}(S)]$

```
samSpotifyEnergy = spotify[samSpotify, "energy"]
zStar.lower = quantile(ZBoot, 0.025)
zStar.upper = quantile(ZBoot, 0.975)
round(mean(samSpotifyEnergy) - c(zStar.upper, zStar.lower) * se.avg(samSpotifyEnergy), 2)
```

```
## 97.5%  2.5%
## 60.90 88.68
```