Lab 6: Introduction to the Bootstrap
STAT 630, Fall 2020

The bootstrap is a data based simulation method for statistical inference that can be used to compute standard errors and construct confidence intervals. The term "bootstrap" derives from the phrase "to pull oneself up by one's bootstrap." It is a useful technique for constructing confidence intervals when there is no analytic (mathematical) formula to work with, or if the assumption for using a formula are not well satisfied.

Bootstrap Algorithm:

1. Let $x_1, x_2, \cdots, x_n$ be a sample of size n; and $\hat{\theta}$ a statistic computed using this sample (e.g., mean, median, trimmed mean, standard deviation, etc.)

2. Take a sample with replacement of size n from the original sample. Call this the **bootstrap sample**.

3. Recompute the statistic of interest using the bootstrap sample. Call this the **bootstrap replicate** of the statistic, denoted by θ^*.

4. Repeat steps 2 and 3 B times to generate B bootstrap replicates of the statistic: $\theta_1^*, \theta_2^*, \cdots, \theta_B^*$. The distribution of the bootstrap replicates is called the **bootstrap distribution**.

Sampling with replacement means that an observation can occur more than once in the bootstrap sample. For example, if x_1, x_2, x_3, x_4, x_5 is a sample of size $n = 5$, then x_1, x_2, x_3, x_3, x_5 is a possible bootstrap sample. Or think of it this way: write the numbers 1 though 100 on tickets and place in a hat, then each time you draw a random number from that hat throw it back in.

Bootstrap Standard Error: A bootstrap estimate of the standard error of a statistic can be computed as the the standard deviation of the bootstrap replicates of the statistic, i.e., the standard deviation of $\theta_1^*, \theta_2^*, \cdots, \theta_B^*$.

Bootstrap Percentile Confidence Interval: A 95% bootstrap confidence interval for a parameter θ can be computed as the 0.025 and 0.975 quantiles of the bootstrap replicates of the statistic, i.e., the 0.025 and 0.975 quantiles of $\theta_1^*, \theta_2^*, \cdots, \theta_B^*$. We can use the `quantile()` function in R to do this.

Remarks

- The main idea behind the bootstrap procedure is that if the sample is representative of the population, then the bootstrap distribution should approximate the shape and spread of the sampling distribution of a statistic. So we can use the bootstrap method to learn something about the sampling distribution of a statistic, and how close that statistic is to the true value (population parameter).
- The bootstrap distribution should be centered around the statistic $\hat{\theta}$ (not the parameter θ). So the bootstrap is not useful for getting a more accurate estimate of θ, but rather for characterizing the variability (standard error) of that statistic.

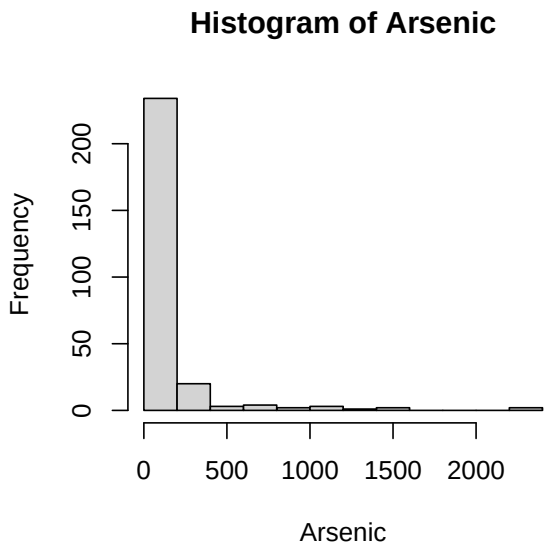
Example

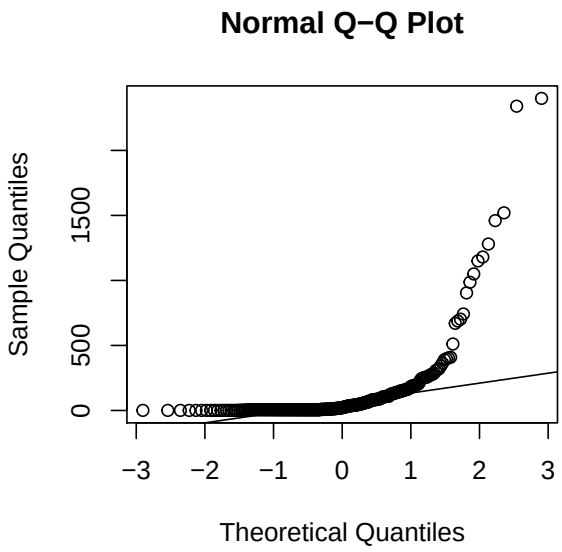
For this example, we use a data set containing arsenic concentration in 271 wells in Bangladesh. Arsenic is a naturally occurring element in the groundwater in Bangladesh; since much of this water is used for drinking in rural areas, arsenic poisoning is a major health problem. To access the data set first install the `resampledata` package, and then use `library()` to load the package into your workspace.

```
library(resampledata)
```

First, let's look at some descriptive statistics for Arsenic. Note that the arsenic measurements are in micrograms per liter (μg/L).

```
Arsenic <- Bangladesh$Arsenic
summary(Arsenic)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##     0.5     6.0    22.0   125.3   109.0  2400.0

sd(Arsenic)

## [1] 297.9755

par(mfrow=c(1,2))
hist(Arsenic)
qqnorm(Arsenic)
qqline(Arsenic)
```

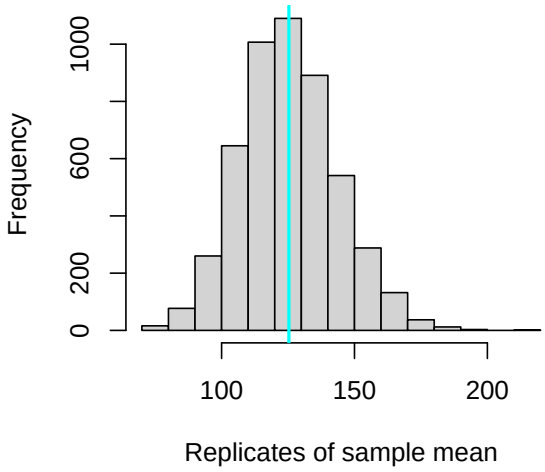


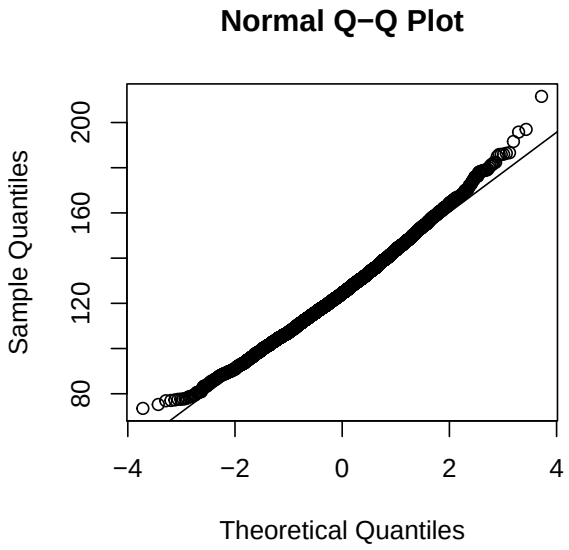


We see that the shape of the histogram of the sample is heavily skewed to the right.

Next, draw 5000 bootstrap samples of arsenic. Compute the mean of each bootstrap sample to get 5000 bootstrap replicates. Then use the 5000 bootstrap replicates to compute the bootstrap standard error and 95% confidence interval.

```
set.seed(9999)
n <- length(Arsenic); n

## [1] 271

replicates <- rep(0, 5000)
for(i in 1:5000) {
  boot_samp <- sample(Arsenic, size = n, replace = TRUE)
  replicates[i] <- mean(boot_samp)
}

# bootstrap distribution
par(mfrow=c(1,2))
hist(replicates, xlab="Replicates of sample mean", main='')
abline(v=mean(Arsenic), col="cyan", lwd=2)
qqnorm(replicates)
qqline(replicates)
```






```
# bootstrap standard error
sd(replicates)

## [1] 18.14656

# 95% bootstrap CI
quantile(replicates, c(0.025, 0.975))

##      2.5%     97.5%
##  91.89089 162.95601
```

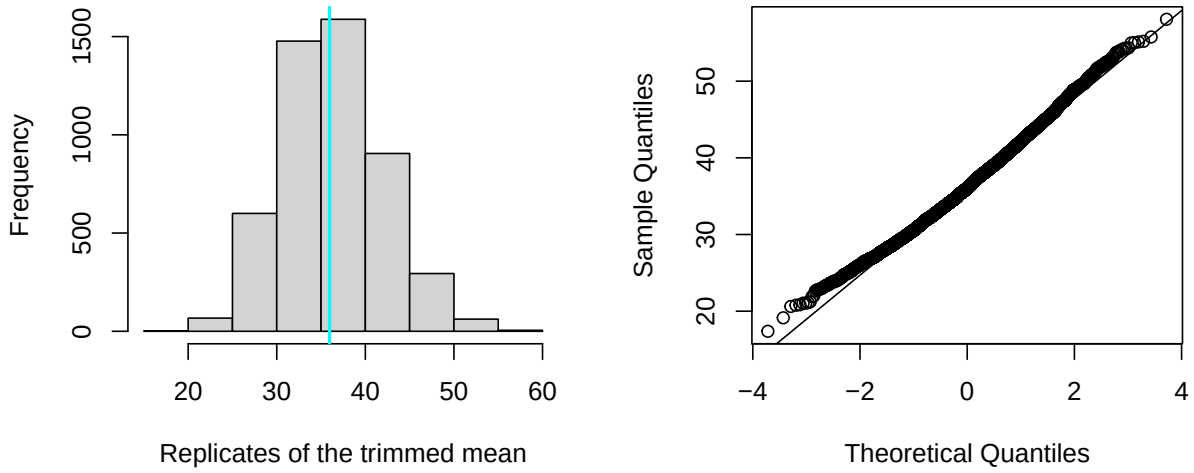
We are 95% confident that the true mean arsenic level is between 91.89 and 162.96 micrograms per liter. We can compare this to a traditional confidence interval for the mean calculated with a z-critical value. Even though the population distribution is heavily skewed, the sample size $n = 271$ is perhaps large enough so that the CLT provides justification.

```
ci_lower <- mean(Arsenic) - 1.96 * sd(Arsenic) / sqrt(n)
ci_upper <- mean(Arsenic) + 1.96 * sd(Arsenic) / sqrt(n)
round(c(ci_lower, ci_upper), 2)

## [1]  89.84 160.80
```

Indeed, we see that the endpoints of 95% bootstrap and z-confidence intervals are close.

A major motivation for the bootstrap is that it can be used with wide variety of statistics (means, medians, trimmed means, standard deviations, correlation coefficients). The bootstrap is especially useful when there is no easy formula that we can work with to get a confidence interval. For example, let's use the bootstrap to construct a confidence interval for the 25% trimmed mean, also called the midmean; that is, the mean of the middle 50% of observations.

```
mean(Arsenic, trim=0.25) # 25% trimmed mean

## [1] 35.95985

replicates <- rep(0, 5000)
for(i in 1:5000) {
  boot_samp <- sample(Arsenic, size = n, replace = TRUE)
  replicates[i] <- mean(boot_samp, trim = 0.25)
}

# boostrap distribution
par(mfrow=c(1,2))
hist(replicates, xlab="Replicates of the trimmed mean", main='')
abline(v=mean(Arsenic, trim=0.25), col="cyan", lwd=2)
qqnorm(replicates)
qqline(replicates)
```



```r
# bootstrap standard error
sd(replicates)

## [1] 5.717114

# 95% bootstrap CI
quantile(replicates, c(0.025, 0.975))

##     2.5%    97.5%
## 26.26692 48.59909
```

We are 95% confident that the true midmean for arsenic is between 26.27 and 48.6 micrograms per liter. Notice that the bootstrap distribution for the trimmed mean has much smaller spread than the bootstrap distribution for the mean. This is because the trimmed mean is more robust, and less sensitive to extreme values.

Reference: Chihara, L., and Hesterberg T. Mathematical statistics with resampling and R, 2nd edition, Chapter 5. [Electronic version: `http://library.csueastbay.edu/home`]

Practice Problem. Using the arsenic data:

(a) Generate 5000 bootstrap replicates of the median. Make a histogram of the 5000 replicates that you generated.

(b) Compute the bootstrap standard error for the median.

(c) Compute a 95% bootstrap percentile confidence interval for the median. Interpret the interval.