MODULE 2: July 31st 2015

Main Discussion Points:

- Sampling Distributions
- Bootstrap for Uncertainty Quantification

Two Main Questions:

1. How sure can the answer be?
   - Sample Distribution
2. Are these two "things" associated with one another?
   - Permutation Test See Scribe Notes Module 3: (July 31st 2015):
   - Answering old question with new technological power (timeless question that always come up)

Sampling Distribution -

Statistic –> any function of the data
Sample = $x_1, ..., x_n$
Compute sample statistic: $t = T(X_1, ..., X_n)$
T = (random variable Tn) –> a function of all observations
t_n = observed statistic for sample size n
Key notion: A different sample will return a different statistic

Posiblity of a different observed statistic is dependent upon the sample taken from the population.
If the sample is random and the statistic is random: how different are they going to be? $P(Tn)$: sampling distribution –> all possible T_n results for all samples of tn:
1. Determine Meaningfulness by equating confidence with stability
2. If Tn varies dramatically across samples, wildly unstable, the answer for each independent sample is untrustworthy.
3. If T_n has minimal variance –> the stability of the estimate improves credibility
4. What are some features of $P(T_n)$?

Evaluate the Standard Error –> Standard Deviation of P(Tn)
Example:
Sample 1: $x_1^1, x_2^1, ..., x_n^1$ –> t_n^1
Sample 2: $x_1^2, x_2^2, ..., x_n^2$ –> t_n^2
...
...
...
Sample 1000: $x_1^{1000}, x_2^{1000}, ..., x_n^{1000}$ –> t_n^{1000}
...
...
...
Sampling distribution = collection of $tn(1)$, $tn(2)$, ... $tn(1000)$, ...
Convert into histogram –> This provides transparency to the sampling distribution.
Determine the spread of distribution –> Evaluate the standard error by constructing error bars on the histogram.
Quote the standard deviation as the distribution standard error

Typically a lot of assumptions are made that the errors are randomly and normally centered around zero. However, these assumptions often times do not prove to be true in more complex models. In order to account for this, a more robust method is necessary.

The Bootstrap method provides an all purpose tool that uses a sample to approximate a collection of statistics across the entire population.

BOOTSTRAP tool:

Ideally to approximate $P(T_n)$ we would sample the entire population repeatedly.
- $x_1^1, x_2^1, \ldots.., x_n^1$ –>
However, given the nature of population data and the extensive effort required to obtain samples, the act of sampling the entire population repeatedly is far to expensive to both collect and evaluate.

One Alternative:

Objectively evaluate a single surrogate sampling of the entire population. With a surrogate sample of size n, a random sample of size n is needed 1 major flaw:
When n = n, the results are going to be the same every time given the entire sample is being used repeatedly. Using n = n on the sample data would produce a static result no different than sampling the entire population. With n = n, there will only be a single observable statistic (tn).

Modifying the sampling process (implementing with replacement):

Sampling with replacement: The idea is that as an x is randomly selected from the sample data, it is accounted for than then returned back to the sample. Once returned it becomes a candidate for selection with the probability of being reselected unchanged from the initial probability of selection. This leads $n = n$ sampling with a unique observable statistic (tn) for each sample.
Example: Catch and Release Fishing
Once a fish is caught it is released. The probability of catching that fish again remains the same.

Variability associated with Bootstrapping:

1. Monte Carlo Variability - The variability associated with random sampling through Monte Carlo Simulation. This variance is reducible in Bootstrapping by increasing the number of random samples used.

2. Single Sample Nature - With a single sample, the chances that the sample does not truly represent the entire population leads to a level of irreducible variability relative to the population.

With Bootstrapping there will always be unique standard errors for each and every sample run due to the with replacement nature of sampling. This eliminates the n = n problem that is present when sampling without replacement.

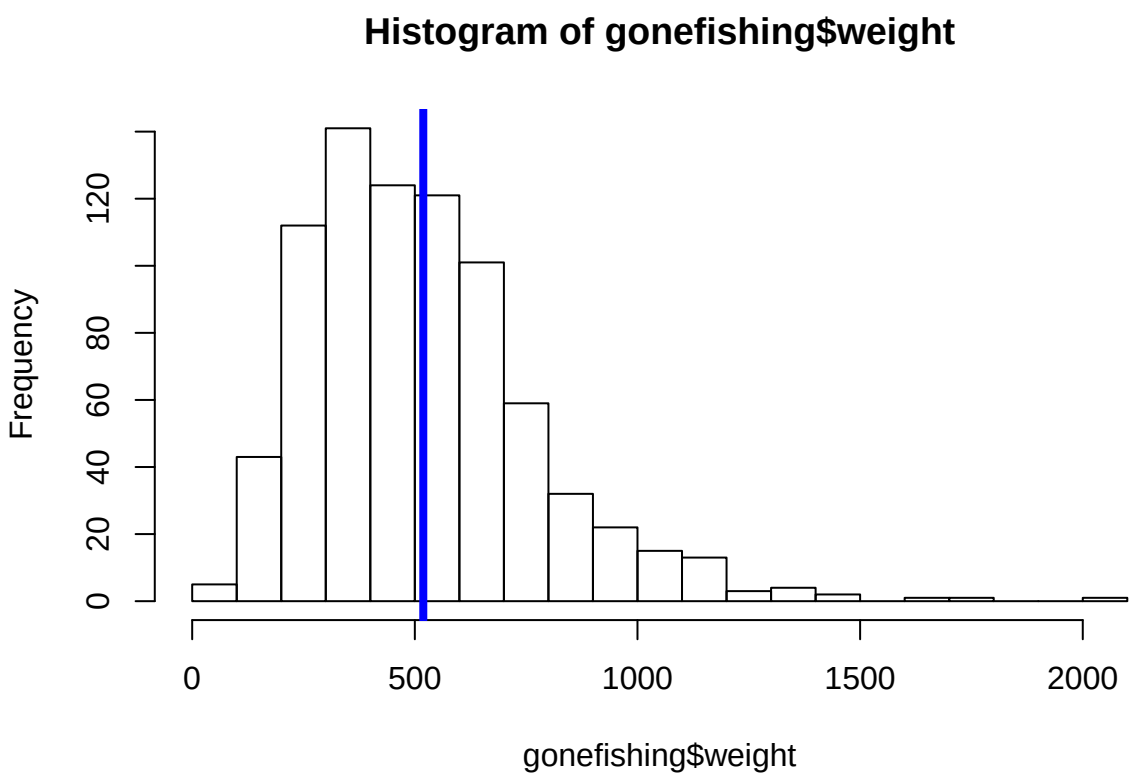
We will use the example “Gone Fishing” to illustrate bootstrapping.

The data used in Gone Fishing is a csv with the length, width, and height of a fictious set of caught fish.

First we take a look at the histogram of the weights of the entire population:

```
library(mosaic)
library(foreach)
gonefishing = read.csv('../data/gonefishing.csv', header=TRUE)
```

```r
# Histogram of weights
hist(gonefishing$weight, breaks=20)
mean_weight_pop = mean(gonefishing$weight)
abline(v=mean_weight_pop, lwd=4, col='blue')
```



Next we look at sampling the entire population 25 times (without replacement)

```r
n_fish = 30

foreach(i = 1:25, .combine='c') %do% {
  fishing_trip = sample(gonefishing, n_fish)
  mean_weight_sample = mean(fishing_trip$weight)
  mean_weight_sample
}
```

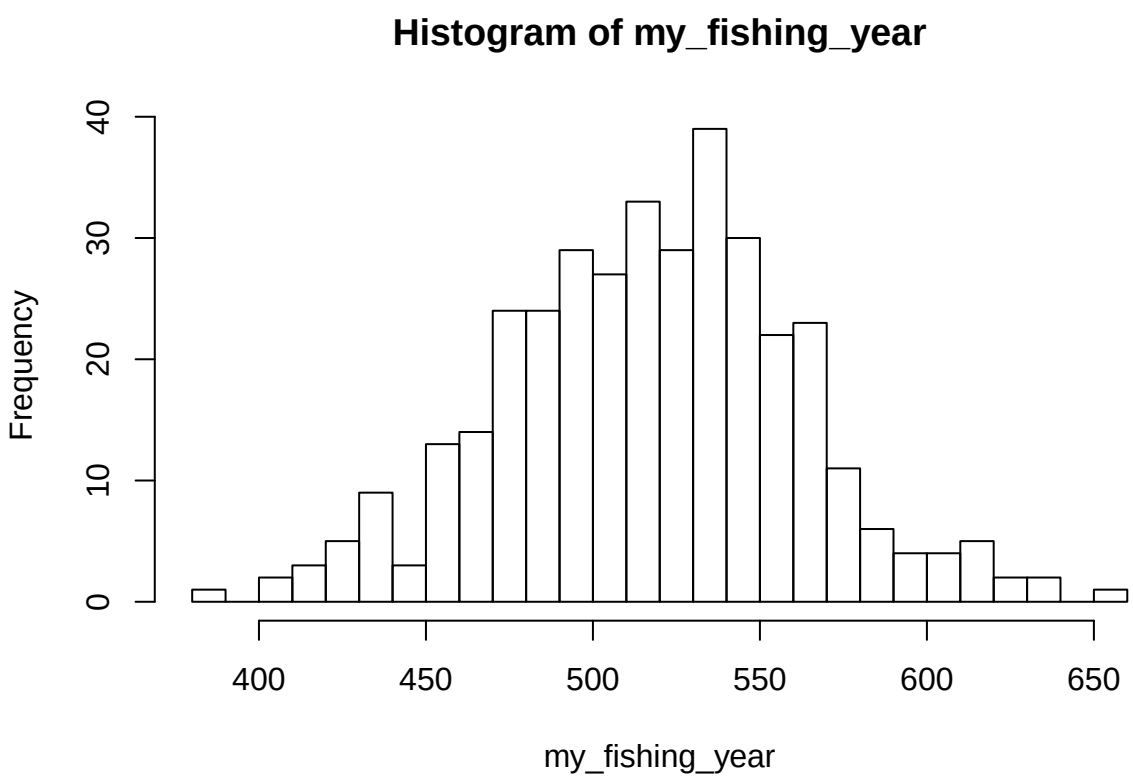
```
##  [1] 502.2667 463.7333 604.6333 587.5667 585.3000 617.2667 533.7000
##  [8] 588.9667 515.6000 558.9000 560.4000 508.4000 579.7333 494.8333
## [15] 537.7333 483.1333 515.2333 462.9667 523.5000 557.3667 464.6000
## [22] 443.8333 546.6667 515.3333 471.4333
```

Next we look at an entire year of catching 30 fish per day:

```r
my_fishing_year = foreach(i = 1:365, .combine='c') %do% {
  fishing_trip = sample(gonefishing, n_fish)
  mean_weight_sample = mean(fishing_trip$weight)
  mean_weight_sample
}
```

Which has a historgram and standard deviation of weights that looks like:

```
hist(my_fishing_year, 25)
```




```
sd(my_fishing_year)
```

```
## [1] 44.54067
```

This takes a lot of work- catching 30 fish per day for an entire year? Let's try something else.

Bootstrapping:

So for a single sample of the total population

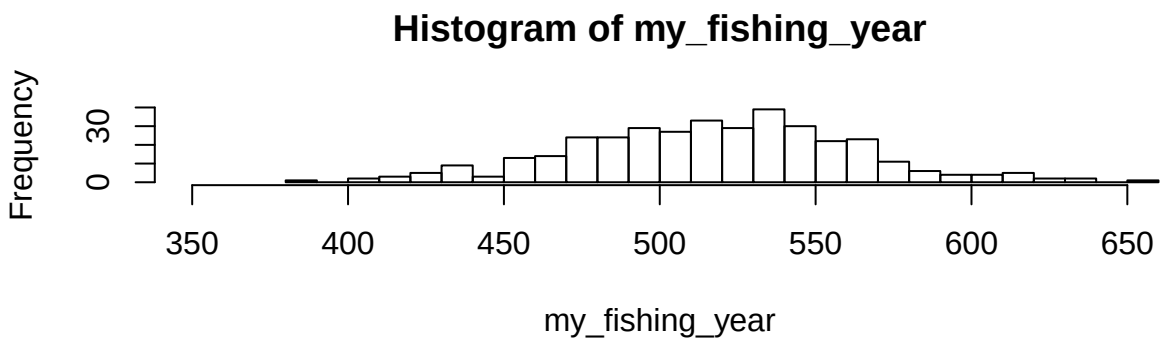
```
fishing_trip = sample(gonefishing, n_fish)
mean_weight_sample = mean(fishing_trip$weight)
mean_weight_sample
```

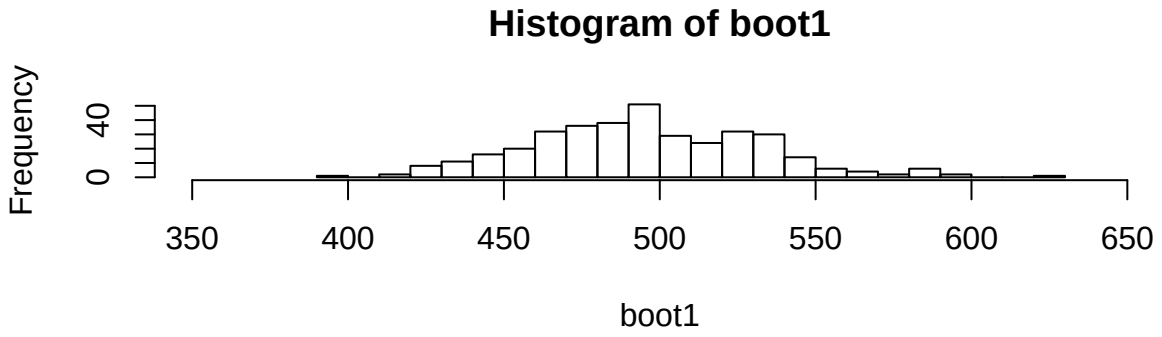
```
## [1] 496.8
```

We can now bootstrap that single sample

```
boot1 = foreach(i = 1:365, .combine='c') %do% {
  fishing_trip_bootstrap = resample(fishing_trip, n_fish)
  mean_weight_bootstrap = mean(fishing_trip_bootstrap$weight)
  mean_weight_bootstrap
}
```

Compare the two histograms:





And let's compare the two standard errors

```
sd(my_fishing_year)
```

```
## [1] 44.54067
```

```
sd(boot1)
```

```
## [1] 36.50103
```

A second example: GDP Growth- Quantifying uncertainity between two estimations.

Our goal is to Bootstrap ordianry correlation coefficienat (a pearson non-robust correlation coefficeient)

And compare to the Spearman Correlation Coefficient.

```
gdpgrowth = read.csv('../data/gdpgrowth.csv', header=TRUE)
head(gdpgrowth)
```

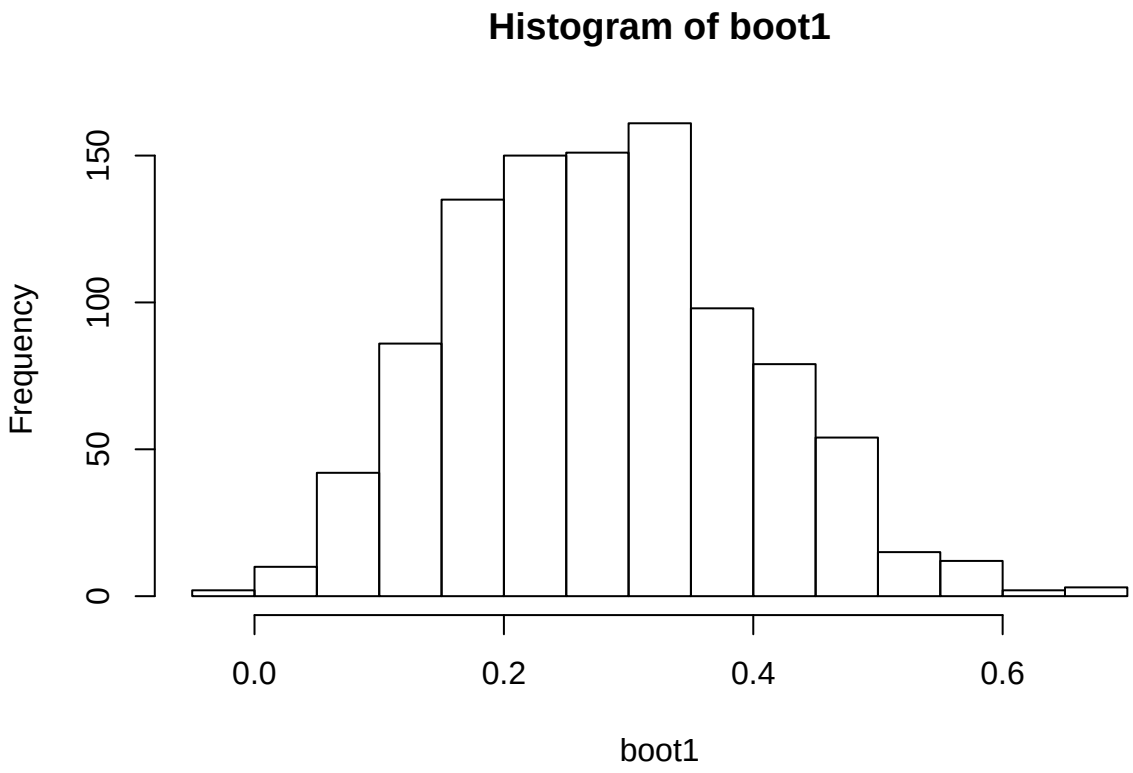
```
##   CODE          COUNTRY  GR6096   DENS60  COAST65  POPGR6090 EAST DEF60
## 1  DZA          Algeria  0.0110 5.396041 4.327307 0.02841708    0 0.030
## 2  BEN            Benin  0.0011 3.900966 4.607945 0.02396531    0 0.018
## 3  BDI          Burundi  0.0046 2.164587 0.000000 0.02027949    0 0.014
## 4  CMR         Cameroon  0.0024 4.475757 3.024604 0.02634325    0 0.024
## 5  CAF Cent'l Afr. Rep. -0.0252 6.006636 0.000000 0.02255507    0 0.009
## 6  COG            Congo  0.0151 5.845420 2.595199 0.02747309    0 0.025
##     LGDP60 EDUC60 LIFE60
## 1 7.451822 0.0297   47.3
## 2 7.003065 0.0248   38.9
```

```
## 3 6.461468 0.0183   41.8
## 4 6.463029 0.0221   43.4
## 5 6.556778 0.0231   39.3
## 6 7.023759 0.0311   47.3
```

```
NMC = 1000
boot1 = foreach(i=1:NMC, .combine='c') %do% {
    gdp_boot = resample(gdpgrowth)
    rho = cor(gdp_boot$DEF60, gdp_boot$GR6096)
}
```

And here is the histogram

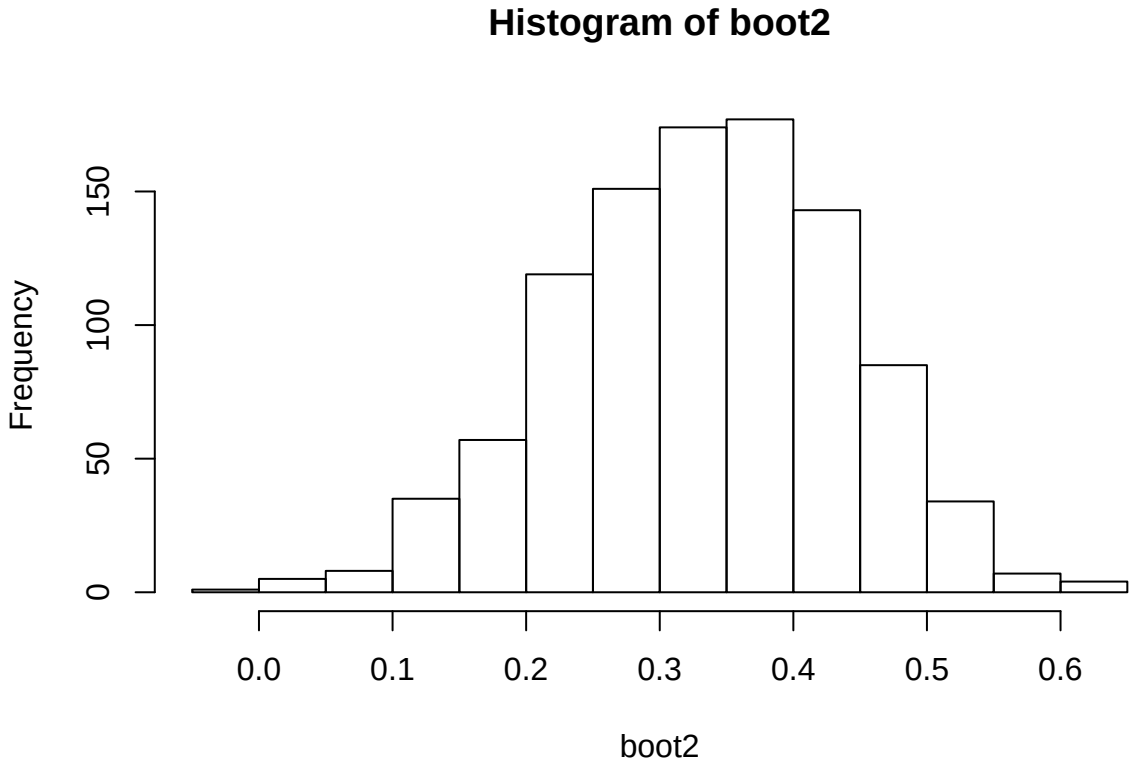
```
par(mfrow=c(1,1))
hist(boot1)
```



Now the Spearman Correlation

```
boot2 = foreach(i=1:NMC, .combine='c') %do% {
    gdp_boot = resample(gdpgrowth)
    rho = cor(gdp_boot$DEF60, gdp_boot$GR6096, method='spearman')
}

hist(boot2)
```



The Bootstrap procedure is an all purpose tool for quantifying estimates. It works for any statistic, both common and unorthodox.