Bootstrap Sampling: Two-Sample Example

BIOS 6611

CU Anschutz

Week 6

Two-Sample Bootstrap

Two-Sample Bootstrap

Bootstrap sampling *mimics how the data were obtained*. For an experiment designed to compare two populations, we randomly take a sample *from each*. Hence, the bootstrap sample will mimic this process:

Given independent samples of sizes m and n from two populations,

1. Draw a resample of size m with replacement from the first sample and a separate resample of size n with replacement from the second sample. Compute a statistic that compares the two groups, such as the difference between the two sample means.
2. Repeat this resampling process many times, say 10,000.
3. Construct the bootstrap distribution of the statistic. Inspect its spread, bias, and shape.

Two-Sample Example

Example 5.4 from our *Chihara and Hesterberg* textbook is a comparison of commercial length between basic and extended cable during random half-hour periods from 7am-11pm:

Basic	7.0	10.0	10.6	10.2	8.6	7.6	8.2	10.4	11.0	8.5
Extended	3.4	7.8	9.4	4.7	5.4	7.6	5.0	8.0	7.8	9.6

We will compare the difference in means between the two cable types and calculate the corresponding bootstrap.

Describe the Distribution

First we will create two vectors with our data:

```
times.Basic <- c(7,10,10.6,10.2,8.6,7.6,8.2,10.4,11.0,8.5)
times.Ext <- c(3.4,7.8,9.4,4.7,5.4,7.6,5.0,8.0,7.8,9.6)
```

The mean (SD) for each group is

```
mean(times.Basic); sd(times.Basic)
```

```
## [1] 9.21
```

```
## [1] 1.395588
```

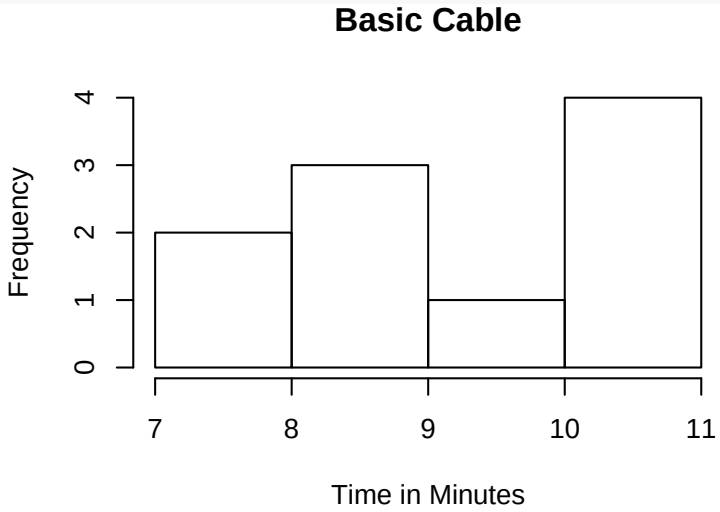
```
mean(times.Ext); sd(times.Ext)
```

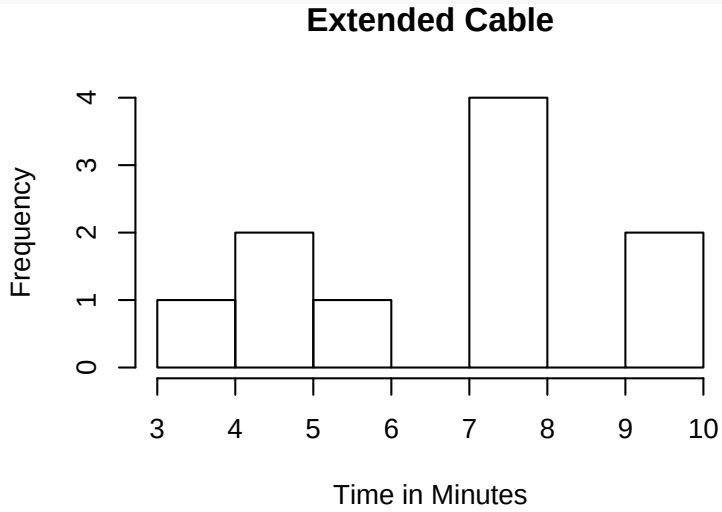
```
## [1] 6.87
```

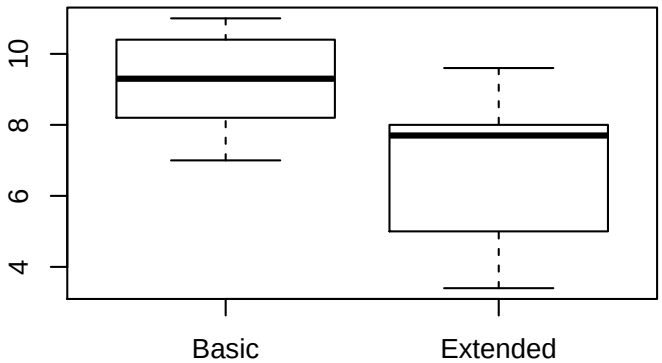
```
## [1] 2.102934
```

Plots of the Sample

```
par(mfrow=c(2,2))
hist(times.Basic, main='Basic Cable', xlab='Time in Minutes')
hist(times.Ext, main='Extended Cable', xlab='Time in Minutes')
boxplot(times.Basic, times.Ext, names=c('Basic','Extended'))
```







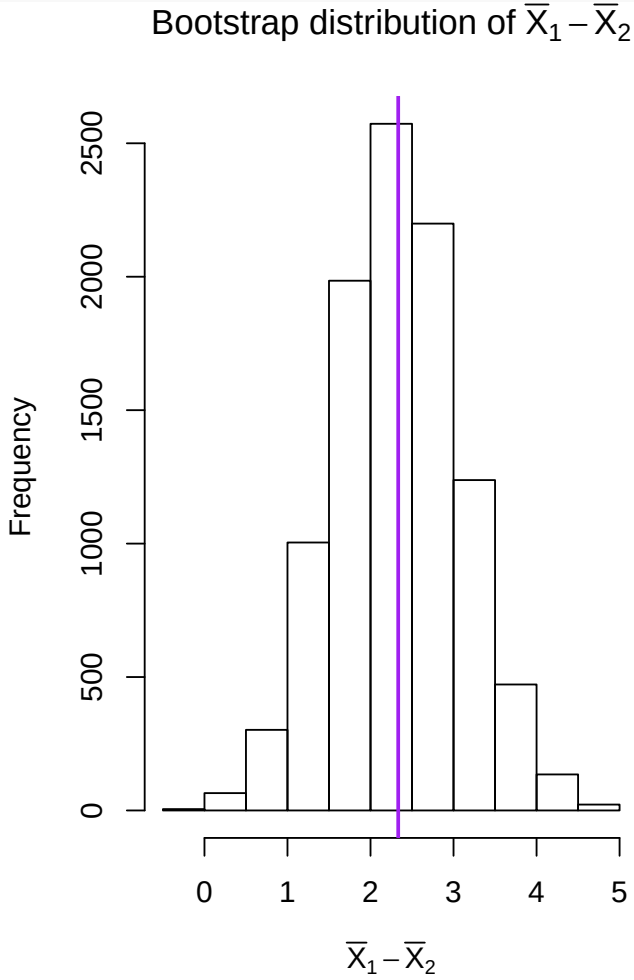
Bootstrap Sample of the Mean

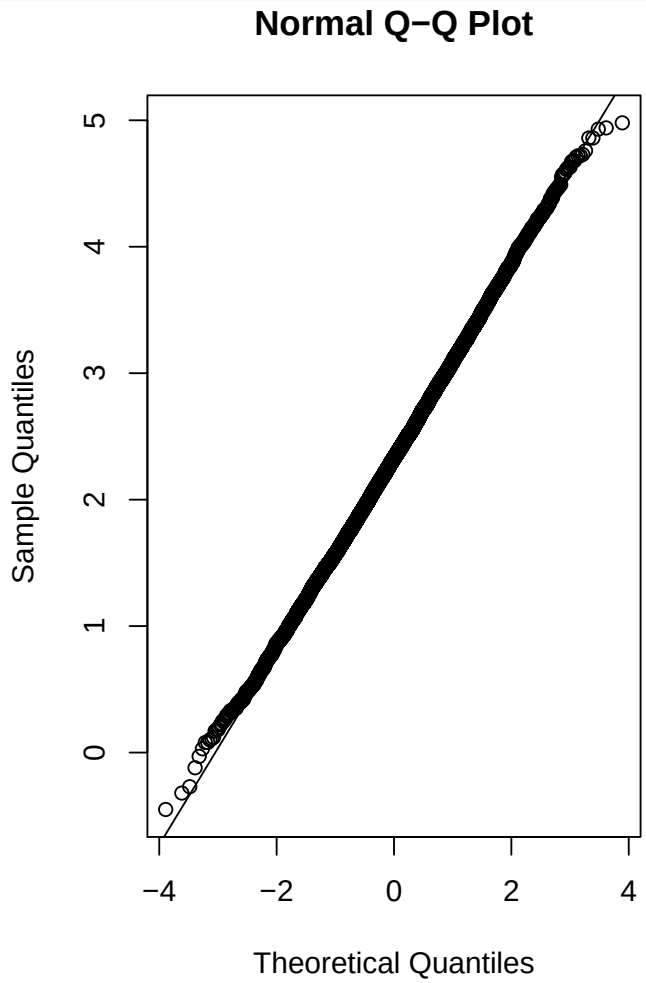
Now let's implement our bootstrap with 10,000 samples:

```
set.seed(54)
n.Basic <- length(times.Basic)
n.Ext <- length(times.Ext)

B <- 10^4
times.diff.mean <- numeric(B)

for (i in 1:B){
  # resample basic cable:
  Basic.boot <- sample(times.Basic, n.Basic, replace=TRUE)

  # resample extended cable
  Ext.boot <- sample(times.Ext, n.Ext, replace=TRUE)

  # calculate difference in means
  times.diff.mean[i] <- mean(Basic.boot)-mean(Ext.boot)
}
```

Plots of the Bootstrap Distribution

```
hist(times.diff.mean, main=expression(paste('Bootstrap distribution of ',
  bar(X)[1] - bar(X)[2])), xlab=expression(bar(X)[1] - bar(X)[2]) )
abline(v=mean(times.diff.mean), col='purple', lwd=2)
qqnorm(times.diff.mean); qqline(times.diff.mean)
```





Bootstrap Mean, Bias, and SE

From our bootstrap distribution we can calculate our mean, bias, and standard error:

```
mean(times.Basic)-mean(times.Ext) #sample difference
```

```
## [1] 2.34
```

```
mean(times.diff.mean) #bootstrap estimated difference
```

```
## [1] 2.333784
```

```
mean(times.diff.mean)-(mean(times.Basic)-mean(times.Ext))   #bias
```

```
## [1] -0.006216
```

```
sd(times.diff.mean) #bootstrap SE
```

```
## [1] 0.7561167
```

We see there is minimal bias in our bootstrap estimator for the difference in mean commercial length.

Bootstrap Percentile CI

Let's calculate the 95% bootstrap percentile CI:

```
quantile(times.diff.mean,c(0.025,0.975)) #bootstrap CI
```

```
##     2.5%    97.5%
## 0.88975 3.84000
```

```
(mean(times.diff.mean)-(mean(times.Basic)-mean(times.Ext)))/
    sd(times.diff.mean)
```

```
## [1] -0.008220953
```

Interpretation: The 95% bootstrap percentile CI is (0.89, 3.84) minutes. We are 95% confident that the true mean difference is in this interval. The ratio of the bias/SE is -0.008, which does not exceed ± 0.10 so we should have good accuracy.

Since the 95% CI excludes 0, we can conclude it is unlikely that the duration of commercials between basic and extended are equal and that basic cable has more commercials.

Matched Data

In the case where we have matched pairs for our cable data, we could simply conduct a one-sample bootstrap on the distribution of the matched differences:

```
set.seed(54)
n <- 10
B <- 10^4
times.diffpair <- times.Basic - times.Ext
times.diffpair.mean <- numeric(B)

for (i in 1:B){
    diff.boot <- sample(times.diffpair, n, replace=T)
    times.diffpair.mean[i] <- mean(diff.boot)
}

quantile(times.diffpair.mean,c(0.025,0.975))
```

```
##  2.5% 97.5%
##  1.18  3.40
```

Since the data is now matched, we have a tighter confidence interval than our unmatched example.

We are 95% confident that the true mean difference is between (1.18, 3.40) minutes for basic and extended cable at the same time of day.

Bootstrap Final Notes

Advantage of Bootstrap Sampling

- Ideal for understanding the sampling distribution of a sample(s) and/or statistics from a sample(s) without assuming anything about the distribution.
- Better for CI than parametric methods (e.g. normal approximation or t-distribution based intervals) when population has moderately (or more) skewed distribution and sample sizes are inadequate.
- Bootstrap diagnostics are easy to apply to determine amount of skewness in underlying population and its impact on coverage from asymptotic confidence intervals.
- Many different "schemes" of bootstrapping exist. We discussed what most consider "case resampling" where the entire case is sampled. There are variations including smooth, Bayesian, wild, parametric, block, etc. bootstraps.

Bootstrap Reminder

Remember, bootstrap sampling will NOT improve our estimators (we can't do any better than the sample we have).

However, we can use them to estimate the variability (SE or confidence interval) of our sampling distribution. Even though we do not have a p-value accompanying this, we can compare our CIs to an expected null value to evaluate potential significance.