13. Bootstrap Sampling

Readings: Chihara and Hesterberg: Ch. 5
Rosner: Ch. 6.11-6.12

R: qqnorm, sample, quantile

Homework: Homework 6 due by midnight on October 15
Homework 7 due by midnight on October 29

Overview

A) Basic ideas behind bootstrap sampling
B) The Plug-in Principle
C) Confidence intervals
D) Two-sample bootstrap, independent vs. paired means; other statistics
E) Accuracy of the bootstrap
F) How many bootstrap samples?

A. Basic Ideas Behind Bootstrap Sampling

In classical statistics, we often assume a distribution for the population, and use this distributional assumption to derive sampling distributions for a statistic.

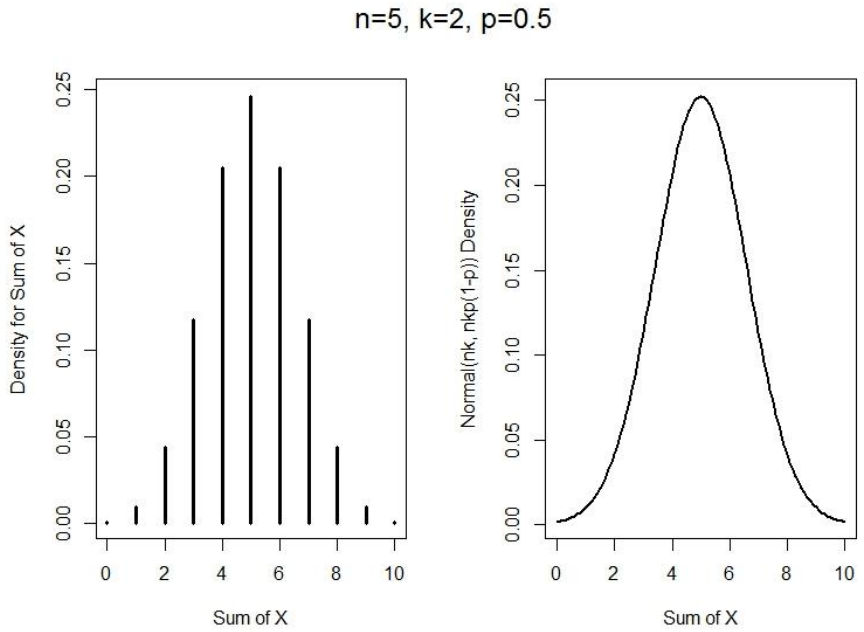
Review:

What is a sampling distribution?

How does the Central Limit Theorem fit into sampling distributions?

Example: Given a random sample $X_1, X_2, \ldots, X_n$ from a $\text{Binomial}(k, 0.5)$ distribution: What is the distribution of $\sum_{i=1}^{n} X_i$?

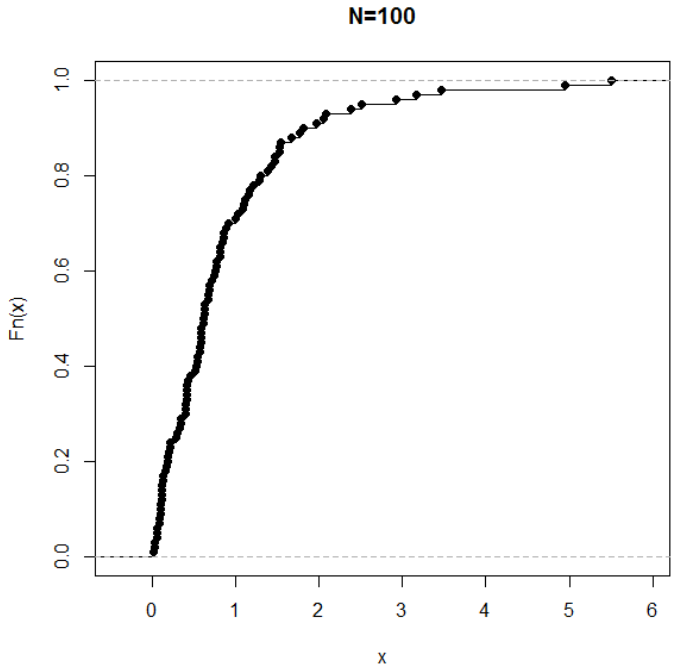
What is a reasonable approximation for the distribution of $\sum_{i=1}^{n} X_i$ if n is large?



What if the underlying distribution is unknown, and we cannot use the CLT? How do we estimate the sampling distribution for a statistic? How do we estimate the distribution of a random variable?

Assuming random sampling, our sample is a representation of the population.

The *empirical cumulative distribution function* is our best "estimate" of the population distribution.



This is the basic idea behind the bootstrap:

The Bootstrap Idea

The original sample approximates the population from which it was drawn. So resamples from this sample approximate what we would get if we took many samples from the population. The bootstrap distribution of a statistic, based on many resamples, approximates the sampling distribution of the statistic, based on many samples.

Source: Chihara and Hesterberg (Pg. 100)

This mimics the use of the empirical distribution. Note that **sampling is done with replacement**.

Side combinatorics question: How many possible bootstrap samples exist for a sample of size n?

Bootstrap for a Single Population

Given a sample of size *n* from a population,

1. Draw a resample of size *n* with replacement from the sample. Compute a statistic that describes the sample, such as the sample mean.
2. Repeat this resampling process many times, say 10,000.
3. Construct the bootstrap distribution of the statistic. Inspect its spread, bias, and shape.

Source: Chihara and Hesterberg (Pg. 101)

Note: It's easier to assume order matters when drawing samples with replacement.

Consider the sample (1, 3, 4, 6). How many bootstrap samples are there?

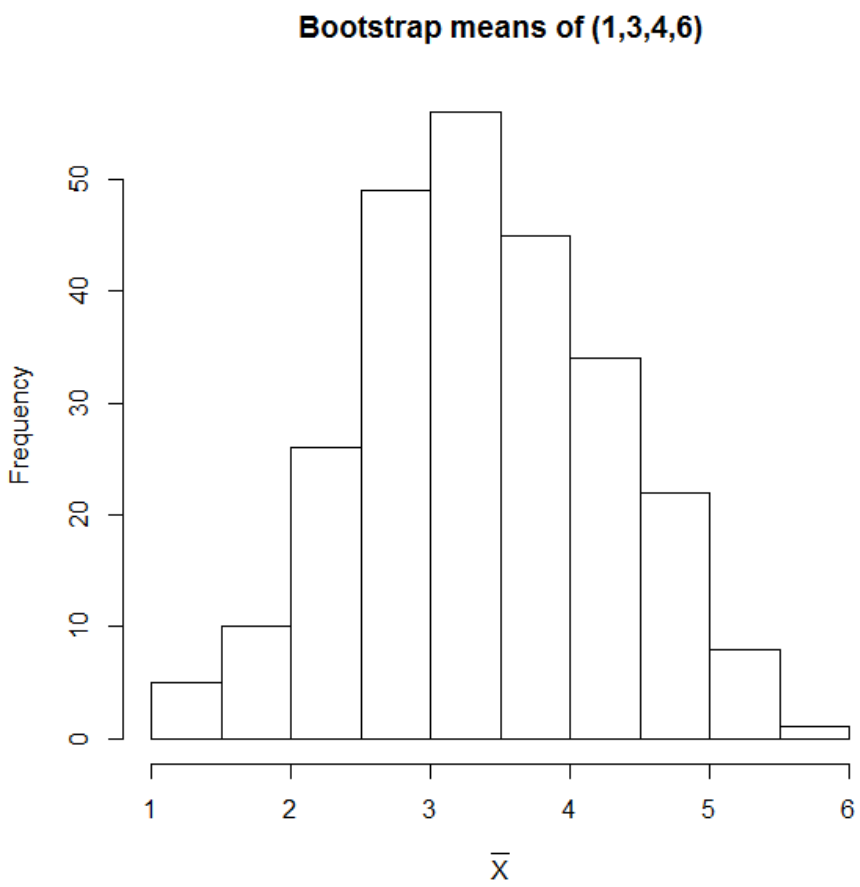
What is the probability the mean is 1?

What is the probability the maximum is 6?

Verify using R calculations:

```
dat <- c(1,3,4,6)

boot <- numeric(0)
for(i in 1:4)
  for(j in 1:4)
    for(k in 1:4)
      for(l in 1:4)
          boot <-
rbind(boot,c(dat[i],dat[j],dat[k],dat[l]))
dim(boots)


boot.mean <- apply(boot,MARGIN=1,mean)
mean(boot.mean==1)
1/256


boot.max <- apply(boot, MARGIN=1,max)
mean(boot.max == 6)
1-(3^4)/(4^4)


hist(boot.mean, main="Bootstrap means of
(1,3,4,6)", xlab=expression(bar(X)))

mean(boot.mean)
sd(boot.mean)


mean(dat)
sd(dat)
```

```
> dim(boot)
[1] 256   4
> boot
      [,1] [,2] [,3] [,4]
  [1,]    1    1    1    1
  [2,]    1    1    1    3
  [3,]    1    1    1    4
.
.
.
[255,]    6    6    6    4
[256,]    6    6    6    6

> boot.mean <- apply(boot,MARGIN=1,mean)
> mean(boot.mean==1)
[1] 0.00390625
> 1/256
[1] 0.00390625
>
> boot.max <- apply(boot, MARGIN=1,max)
> mean(boot.max == 6)
[1] 0.6835938
> 1-(3^4)/(4^4)
[1] 0.6835938
>
> hist(boot.mean, main="Bootstrap means of (1
,3,4,6)", xlab=expression(bar(X)))
>
> mean(boot.mean)
[1] 3.5
> sd(boot.mean)
[1] 0.9031535
>
> mean(dat)
[1] 3.5
> sd(dat)
[1] 2.081666
```



Bootstrap Standard Error

The *bootstrap standard error* of a statistic is the standard deviation of the bootstrap distribution of that statistic.

Source: Chihara and Hesterberg (Pg. 102)

Example 5.1: Consider a sample of size 50 drawn from a $N(23,7^2)$ distribution.

```
# Replicating Figure 5.2 from C & H text
set.seed(515) # note: C&H don't specify a seed
par(mfrow=c(2,2))
u <- 23; sd<-7; n<-50; L<-u-3*sd; U<-u+3*sd
X <- seq(L,U,.1)
dat <- rnorm(n=n,mean=u,sd=sd)

B <- 1000
my.boot <- numeric(B)
for (i in 1:B)
{
  x <- sample(dat, size=50, replace=TRUE)#draw resample
  my.boot[i] <- mean(x) #compute mean, store in my.boot
}


plot(X,dnorm(X,mean=u,sd=sd), type="l", lwd=2, col="red",
   ylab="Density", main="Population, N(23,49)",
   xlim=c(L,U))
plot(X,dnorm(X,mean=u,sd=sd/sqrt(n)), type="l", lwd=2,
   col=1, ylab="Density", main="Population, N(23,49/50)",
   xlim=c(L,U), xlab=expression(bar(X)))
lines(c(u,u), c(0,1), col="blue", lwd=2, lty=2)
hist(dat,xlab="X",ylab="Density", xlim=c(L,U), lwd=2, main="Sample, n=50")
hist(my.boot,xlab=expression(bar(X)),ylab="Density", xlim=c(L,U), lwd=2, main="Bootstrap
distribution")
boot.mean <- mean(my.boot)
lines(c(boot.mean, boot.mean), c(0,250), col="blue", lwd=2, lty=2)
```

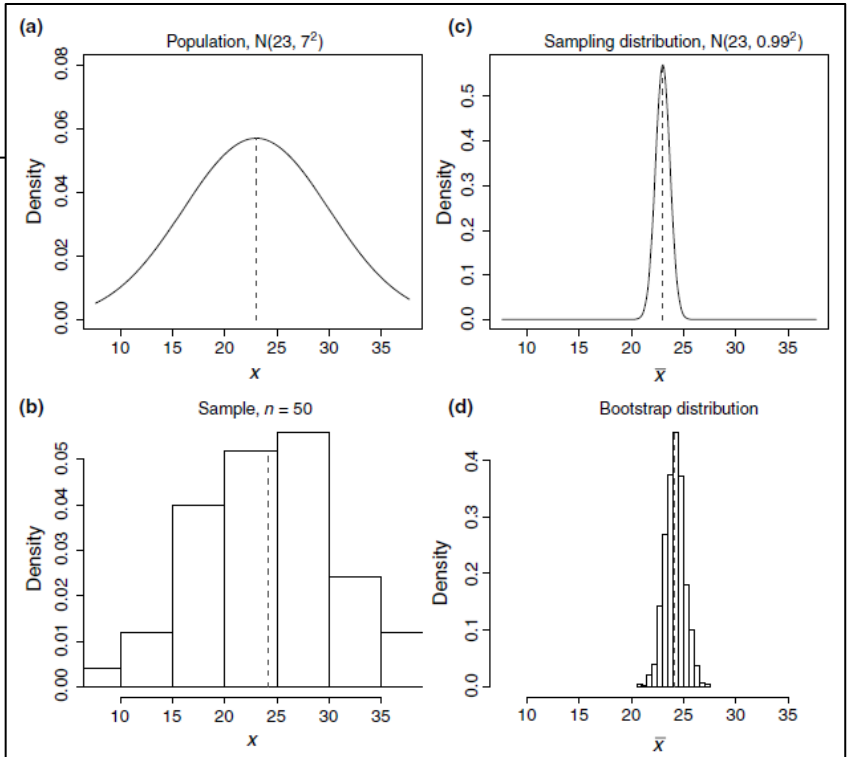
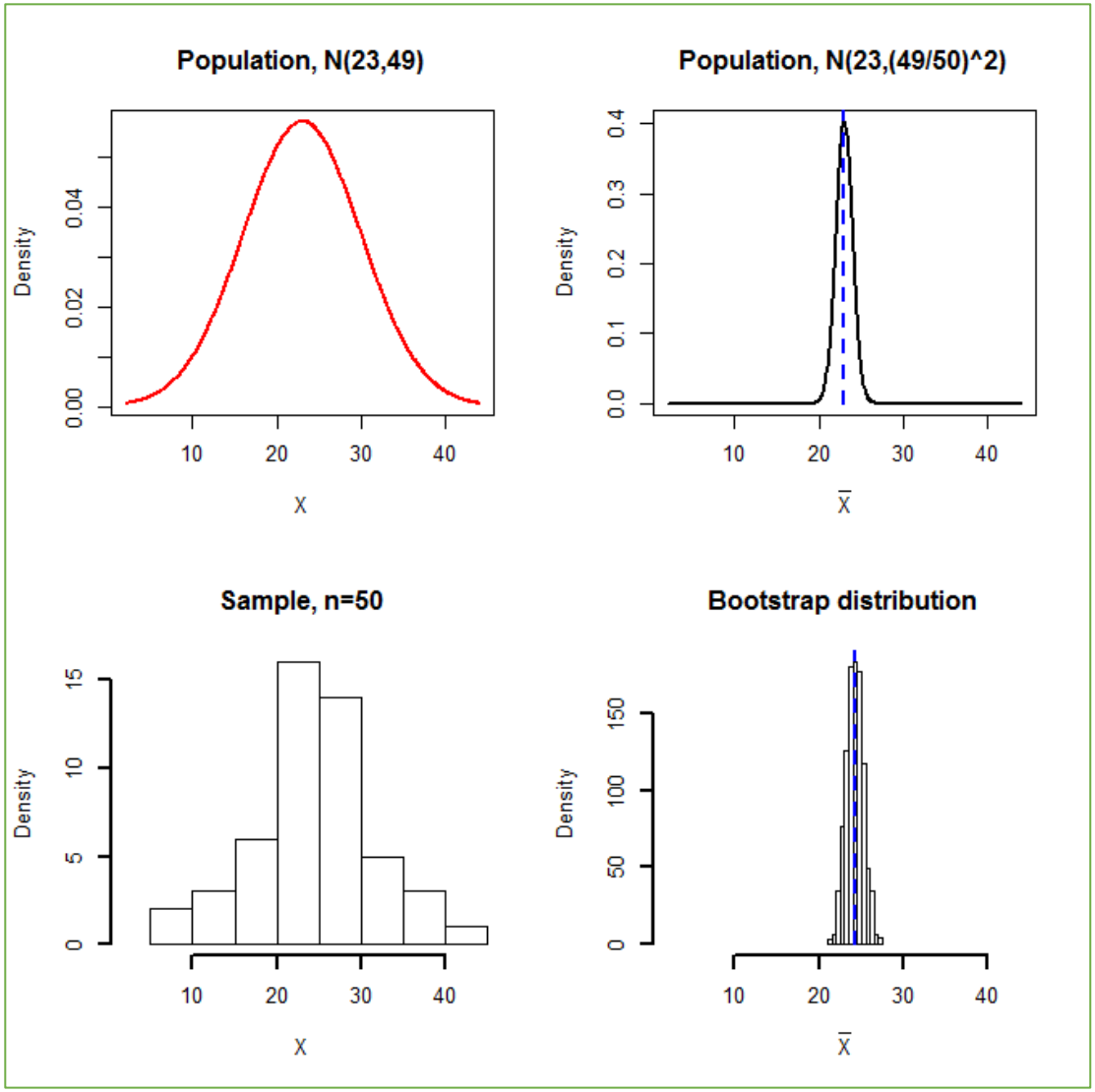


FIGURE 5.2 Sampling and bootstrap distributions of the mean for $N(23, 7^2)$. (a) The population distribution, $N(23, 7^2)$. (b) The distribution of one sample of size 50 from $N(23, 7^2)$. (c) The theoretical sampling distribution of $\bar{X}$, $N(23, 7^2/50)$. (d) The bootstrap distribution. Vertical lines mark the means.

Take home points:

- Shape and spread of the bootstrap distribution are comparable to that of the sampling distribution
- Centers of the bootstrap and sampling distributions are <u>different</u> (23 for population, 24.204 for bootstrap mean)
- Comparing the center of the bootstrap distribution to the observed statistic gives a measure of *bias*



B. The Plug-In Principle

If our goal is to estimate the sampling distribution for some statistic, we need to know:

- The underlying population (which is unknown!)
- The sampling procedure
- The statistic, e.g., $\bar{X}$

The bootstrap uses the *plug-in principle*:

The Plug-In Principle

To estimate a parameter, a quantity that describes the population, use the statistic that is the corresponding quantity for the sample.

Source: Chihara and Hesterberg (Pg. 106)

Using the plug-in principle is a natural and frequent approach in statistics. Think $\bar{X}$ for μ and s^2 for σ^2. This idea can be extended to the sample: it's an estimate of the whole population.

Set-up: Let F and f denote the cdf and pdf for some unknown distribution with $x_1, x_2, \ldots, x_n$ a random sample from this distribution. Without making further assumptions about the distribution, we can use the empirical distribution:

$$\hat{F}(s) = \frac{1}{n}\{\text{number of points} \leq s\}, \text{ note this is a } \underline{\text{discrete}} \text{ function}$$

$$\hat{f}(s) = \frac{1}{n}\{\text{number of points} = s\}$$

Now consider the mean of $X \sim F$:

$$E_F(X) = \mu_F = \begin{cases} \int_{-\infty}^{\infty} xf(x)dx & \text{for continuous random variables} \\ \sum_x xf(x) & \text{for discrete random variables} \end{cases}$$

How do you use the plug-in principal via the bootstrap to estimate the mean (and variance) of X?

$$\begin{aligned} E_{\hat{F}}(X) &= \mu_{\hat{F}} \\ &= \sum_x x\hat{f}(x) \\ &= \sum_{i=1}^{n} x_i\left(\frac{1}{n}\right) = \bar{x} \end{aligned} \qquad \begin{aligned} Var_{\hat{F}}(X) &= \sigma^2_{\hat{F}} \\ &= E_{\hat{F}}[(X - \mu_{\hat{F}})^2] \\ &= \sum_{i=1}^{n} (x_i - \bar{x})^2\left(\frac{1}{n}\right) \end{aligned}$$

C. Confidence Intervals

But, **how well does the bootstrap do?**

Definition 5.1
If $X_1, X_2, \ldots, X_n$ are random variables from a distribution with parameter θ and $g(X_1, X_2, \ldots, X_n)$ an expression used to estimate θ, then we call this function an *estimator.*

Source: Chihara and Hesterberg (Pg. 110)

For most common estimators and under fairly general distribution assumptions:

- **Spread** – the spread does reflect the spread of the sampling distribution
- **Skewness** – the skewness of the bootstrap sample does reflect the skewness of the sampling distribution
- **Center** – the bootstrap distribution is NOT an accurate estimator for the center of the sampling distribution
- **Bias** – the bootstrap can be used to estimate the bias of the sampling distribution (more later)

Thus, bootstrap sampling is useful for studying the sampling behavior of estimators (e.g. s.e., skewness, bias) and obtaining confidence intervals for a parameter. It's not used to improve estimators.

Example 5.3 – Arsenic in Bangladesh groundwater. Let's walk through the code provided by the Chihara & Hesterberg text to:

1. Describe the distribution of arsenic in groundwater along with the sample mean and standard deviation
2. Obtain and describe a bootstrap sample for the mean
3. Obtain a confidence interval for the bootstrap distribution using normal percentiles
4. Obtain a confidence interval for the bootstrap distribution using bootstrap percentiles:

> **Bootstrap Percentile Confidence Intervals**
> The interval between 2.5 and 97.5 percentiles of the bootstrap distribution of a statistic is a 95% *bootstrap percentile confidence interval* for the corresponding parameter.

Source: Chihara and Hesterberg (Pg. 113)

5. Compare the two intervals from parts 3 and 4 and comment.

Note: The bootstrap percentile CI is easy to compute, but *it's not always optimal*. See the Hesterberg (2011) review paper for some proposed modifications.

R code

```
library(resampledata)
set.seed(53)
Arsenic <- resampledata::Bangladesh$Arsenic

# Part 1
hist(Arsenic)
qqnorm(Arsenic); qqline(Arsenic)

mean(Arsenic)
[1] 125.3199
sd(Arsenic)
[1] 297.9755

# Part 2
n <- length(Arsenic)
N <- 10^4
arsenic.mean <- numeric(N)
for (i in 1:N){
    x <- sample(Arsenic, n, replace = TRUE)
    arsenic.mean[i] <- mean(x)
}

hist(arsenic.mean, main = "Bootstrap distribution of means")
abline(v = mean(Arsenic), col = "blue", lty = 2) # observed mean
qqnorm(arsenic.mean); qqline(arsenic.mean)

mean(arsenic.mean) # bootstrap mean
[1] 125.5482
mean(arsenic.mean)-mean(Arsenic) # bias
[1] 0.2282611
sd(arsenic.mean) # bootstrap SE
[1] 18.2324
```

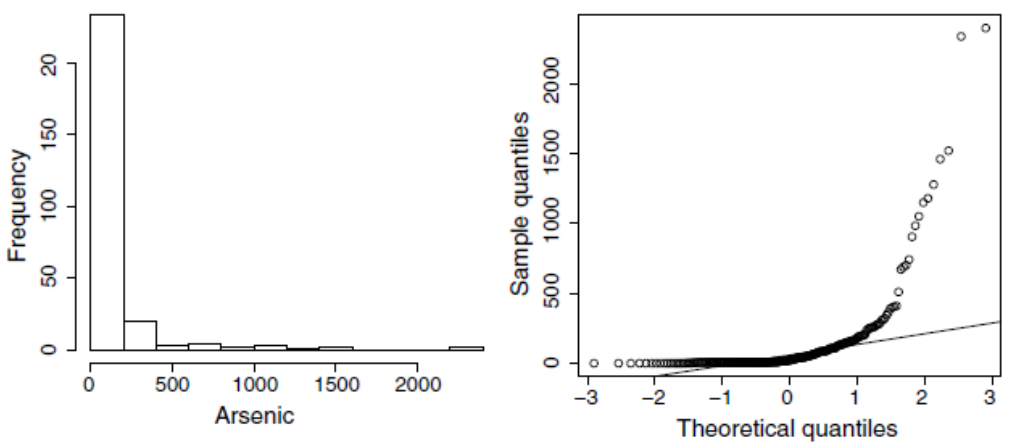


FIGURE 5.6 Arsenic levels in 271 wells in Bangladesh.

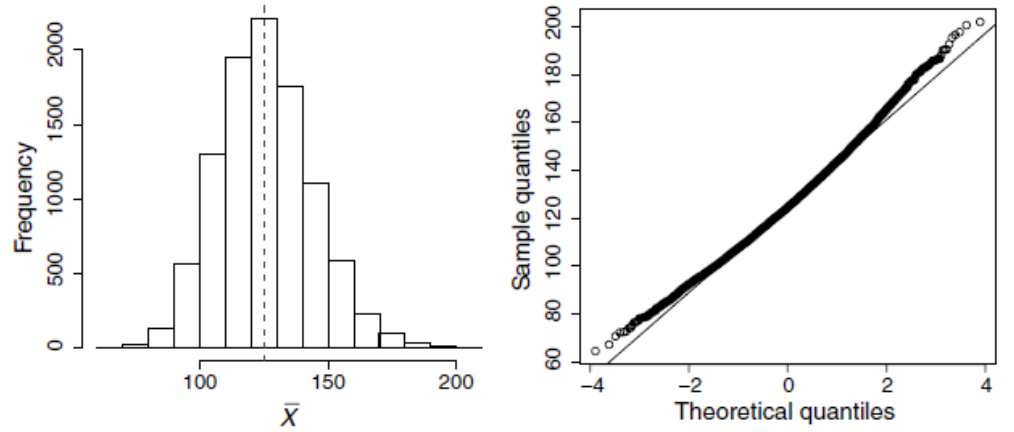
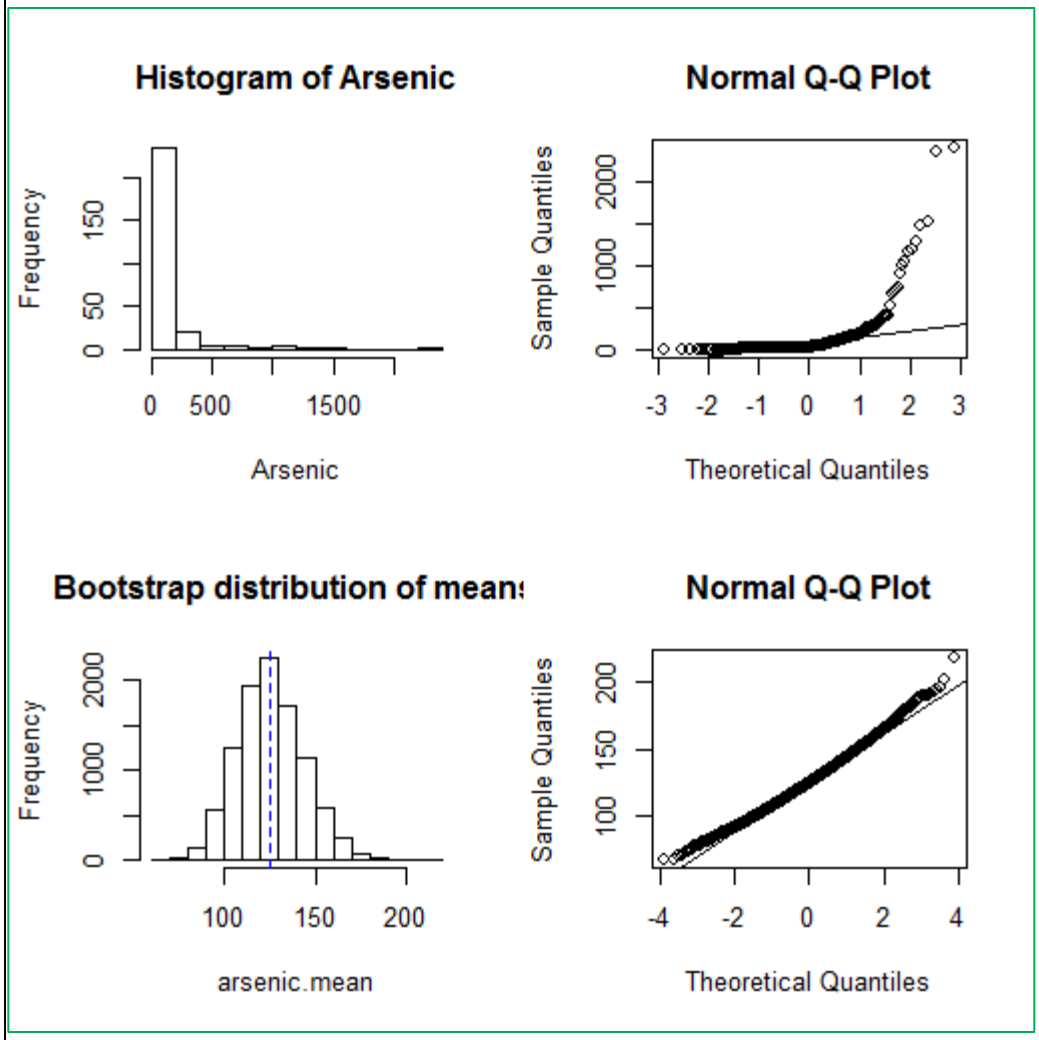


FIGURE 5.7 Histogram and QQ plot of the bootstrap distribution for the arsenic concentrations.



R code cont.

```
# Part 3: Obtain Normal percentile 95% CI
LL <- mean(arsenic.mean)-1.96*sd(arsenic.mean) #Lower limit of 95% Normal CI
LL
[1] 89.81268
UL <- mean(arsenic.mean)+1.96*sd(arsenic.mean) #Upper limit of 95% Normal CI
UL
[1] 161.2837

sum(arsenic.mean < LL)/N  # Coverage of CI at lower end
[1] 0.0158
sum(arsenic.mean > UL)/N  # Coverage of CI at upper end
[1] 0.0327

# Part 4: Obtain bootstrap percentile 95% CI
quantile(arsenic.mean, c(0.025, 0.975))
     2.5%     97.5%
 92.25681 163.77824
```

Part 5 (comparing Part 3 and 4 intervals):

D. Two-Sample Bootstrap, Independent vs. Paired Means, Other Statistics

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:

Bootstrap for a Comparing Two Populations

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* 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.

Source: Chihara and Hesterberg (Pg. 114)

Example 5.4: Comparison of commercial length between basic and extended cable during random half-hour periods from 7am-11pm

TABLE 5.4 Length of Commercials (Minutes) During Random Half-Hour Periods from 7a.m. to 11p.m.

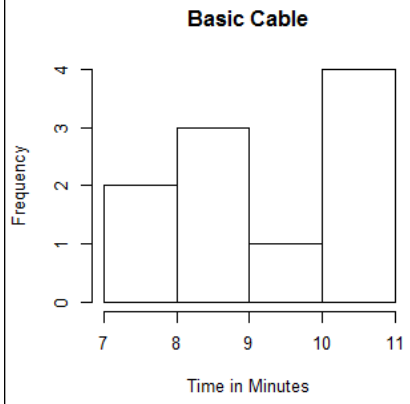
Basic	7.0	10.0	10.6	10.2	8.6	7.6	8.2	10.4	11.0	8.5
Extended	3.4	7.8	9.4	4.7	5.4	7.6	5.0	8.0	7.8	9.6

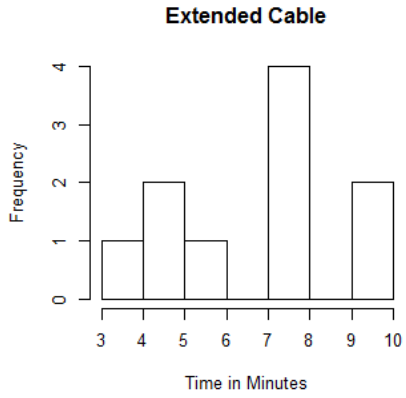
What exploratory analyses would you do?

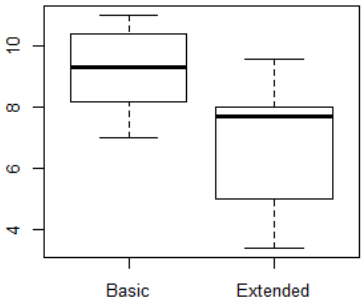
How would you use the bootstrap to compare the time between basic and extended cable?

R code

```
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)

mean(times.Basic); sd(times.Basic); length(times.
Basic)
[1] 9.21
[1] 1.395588
[1] 10

mean(times.Ext); sd(times.Ext); length(times.Ext)
[1] 6.87
[1] 2.102934
[1] 10

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'))
```

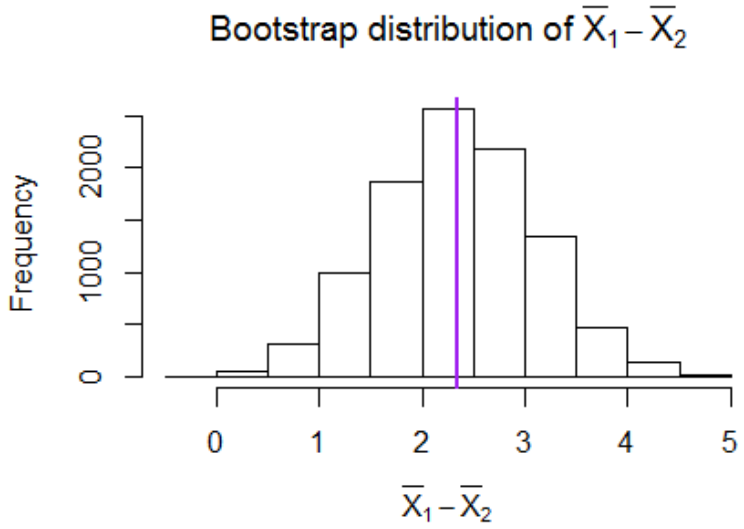


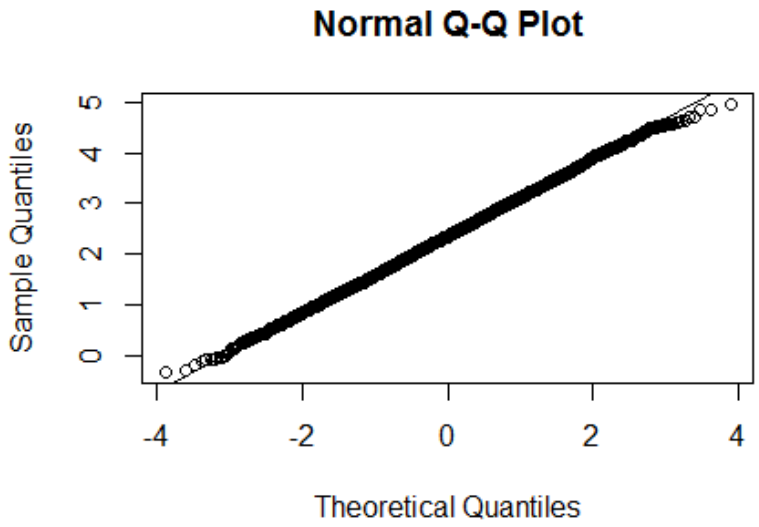




R code cont.

```
# Bootstrap with Independent 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){
   Basic.boot <- sample(times.Basic, n.Basic,
      replace=TRUE)
   Ext.boot <- sample(times.Ext, n.Ext, replace=T)
   times.diff.mean[i] <- mean(Basic.boot) -
      mean(Ext.boot)
}

par(mfrow=c(2,1))

hist(times.diff.mean, main=expression(paste('Boots
trap 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)
```





R code cont.

```
# Compare sample and bootstrap
mean(times.Basic)-mean(times.Ext) #sample difference
[1] 2.34

mean(times.diff.mean) #bootstrap estimated difference
[1] 2.34876

mean(times.diff.mean)-(mean(times.Basic)-mean(times.Ext))  #bias
[1] 0.00876

sd(times.diff.mean) #bootstrap SE
[1] 0.7629355

quantile(times.diff.mean,c(0.025,0.975)) #bootstrap CI
2.5%  97.5%
0.87  3.87
```

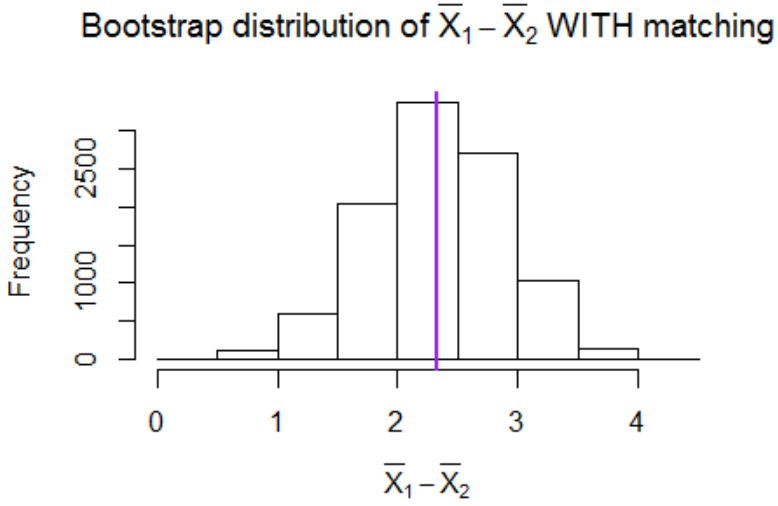
Conclusions?

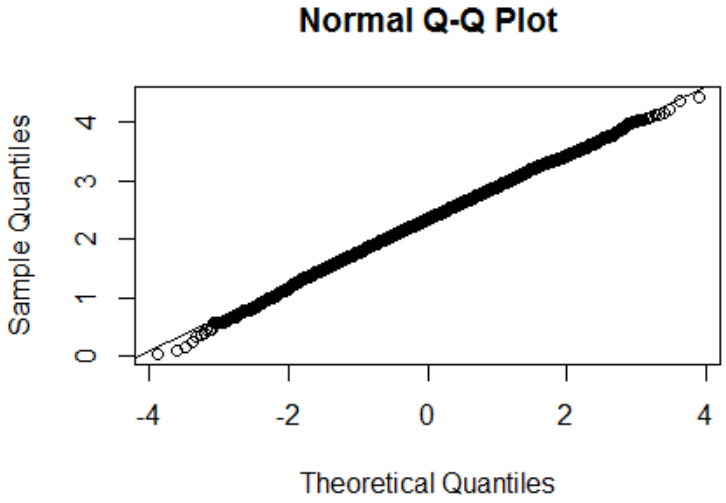
What are alternative approaches to assessing the difference in length of commercial time between basic and extended cable?

How would your analyses change if the commercials were matched in some way, e.g., by hour of day?

R code cont.

```
# Bootstrap with Paired/Matched Samples
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)
}

par(mfrow=c(2,1))

hist(times.diffpair.mean, main=…, xlab=…)

abline(v=mean(times.diffpair.mean), col='purple',
lwd=2)

qqnorm(times.diffpair.mean)
qqline(times.diffpair.mean)

mean(times.diffpair.mean)
[1] 2.333846

quantile(times.diffpair.mean,c(0.025,0.975))
   2.5%   97.5%
1.19975 3.40000
```





Other Statistics and Bias

A major advantage of the bootstrap is the ability to conduct statistical inference when the theoretical distributions are not tractable or easily derived. Once we have drawn a bootstrap sample, we can calculate any statistic for the sample!

Definition 5.2: Bias

The *bias* of an estimator $\hat{\theta}$ is $Bias\left[\hat{\theta}\right] = E\left[\hat{\theta}\right] - \theta$.

The bootstrap estimate of the bias is $\text{Bias}_{\text{boot}}\left[\hat{\theta}^*\right] = E\left[\hat{\theta}^*\right] - \hat{\theta}$.

$E\left[\hat{\theta}^*\right]$ is the mean of the bootstrap definition and $\hat{\theta}$ is the sample estimate.

Source: Chihara and Hesterberg (Pg. 122)

If an estimator, $\hat{\theta}$, tends to over or under estimate the true parameter value, θ, then it is biased. An estimator is unbiased if the bias is zero.

Rule of thumb: If the ratio of bias/SE exceeds ± 0.10, then it could have a substantial effect on the accuracy of confidence intervals.

E. Implementation and Accuracy of the Bootstrap

For one of our first examples, we used all possible bootstrap samples. This was feasible because the sample was small, (1, 3, 4, 6), leading to $4^4 = 256$ possible bootstrap samples.

This quickly becomes infeasible! *Monte Carlo sampling* of the possible bootstrap samples can be used instead. This gives an estimate of the theoretical bootstrap distribution with the larger the number of bootstrap samples drawn resulting in a better estimate.

It's always good to do a sensitivity analysis to see if conclusions change when you repeat your bootstrap analysis.

The permutation and bootstrap frameworks allow us to (1) conduct hypothesis tests and (2) construct confidence intervals, all:

- without assuming a distribution for the sampled random variable
- while working with any statistic or estimator we dream up
- for a wide array of designs

Main differences of permutation vs. bootstrapping:

- sampling without vs. with replacement
- Permutation distribution centered at null hypothesis value; Bootstrap distribution centered at the original mean value (difference, proportion, s.d., etc.) from the data
- Permutation distribution: estimate p-values; Bootstrap distribution: estimate s.e. of estimators, obtain confidence intervals

Advantages 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 (think back to accuracy calculations we did when we discussed the CLT).
- Bootstrap diagnostics are easy to apply to determine amount of skewness in underlying population and its impact on coverage from asymptotic confidence intervals.
- Bootstrap CI and their modifications can provide more accurate coverage (within +/- 10% (relative, not absolute) of the declared $\alpha/2$ x 100% coverage probability (see Hesterberg, 2008, Google Report paper). This can be achieved by generating <u>many</u> bootstrap samples, e.g. up to 10^4 or more of them.

F. How many bootstrap samples?

For good accuracy, 10^4 or more (see above).

How well does bootstrap estimation work with large vs. small sample sizes?

- "In large samples, clearly the bootstrap [is preferred]. In small samples, the classical procedure may be preferred. If the sample size is small, then skewness cannot be estimated accurately from the sample, and it may be better to assume skewness = 0 in spite of the bias, rather than to use an estimate that has high variability."
- Bootstrap CI tend to be too narrow with small sample sizes. One remedy: sample n-1 instead of n from original data

Are there parameters for which bootstrap sampling does not work well?

- The median
- Other quantiles
- Parameters that depend heavily on a small number of observations out of a larger sample

For more information, see the Hesterberg, 2011 review paper in the Paper Repository.