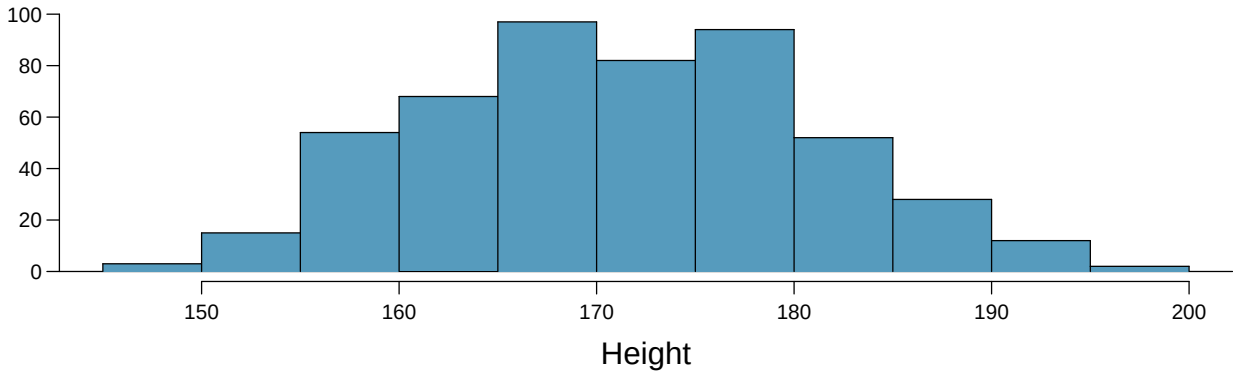
Chapter 5 - Foundations for Inference

Heights of adults. (7.7, p. 260) Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender, for 507 physically active individuals. The histogram below shows the sample distribution of heights in centimeters.



(a) What is the point estimate for the average height of active individuals? What about the median?
(b) What is the point estimate for the standard deviation of the heights of active individuals? What about the IQR?
(c) Is a person who is 1m 80cm (180 cm) tall considered unusually tall? And is a person who is 1m 55cm (155cm) considered unusually short? Explain your reasoning.
(d) The researchers take another random sample of physically active individuals. Would you expect the mean and the standard deviation of this new sample to be the ones given above? Explain your reasoning.
(e) The sample means obtained are point estimates for the mean height of all active individuals, if the sample of individuals is equivalent to a simple random sample. What measure do we use to quantify the variability of such an estimate (Hint: recall that $SD_x = \frac{\sigma}{\sqrt{n}}$)? Compute this quantity using the data from the original sample under the condition that the data are a simple random sample.

```
# Answer a - The mean is 171.1 and the median is 170.3
summary(bdims$hgt)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   147.2   163.8   170.3   171.1   177.8   198.1
```

```
# Answer b - 9.407205 and IQR is 14

sd(bdims$hgt)
```

```
## [1] 9.407205
```

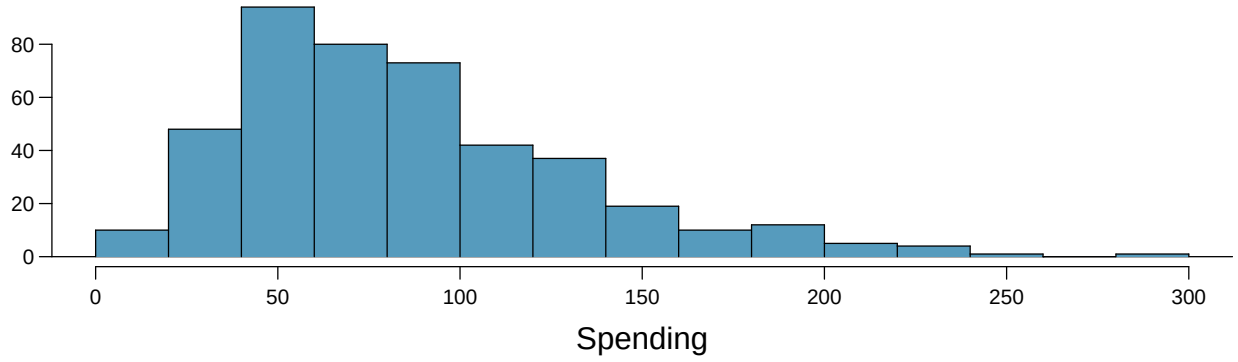
```
IQR(bdims$hgt)
```

```
## [1] 14
```

```
# Answer c - not unsually tall since he is within 1 SD but short is because it is
# more than 1 SD from the mean.


# Answer d - For another random sample the point estimates could be similar,
#but not same

# Answer e- We will Compute Standard Error for sample mean using
# $SD_x = \frac{\sigma}{\sqrt{n}}$)
9.4/sqrt(507)
```

```
## [1] 0.4174687
```

Thanksgiving spending, Part I. The 2009 holiday retail season, which kicked off on November 27, 2009 (the day after Thanksgiving), had been marked by somewhat lower self-reported consumer spending than was seen during the comparable period in 2008. To get an estimate of consumer spending, 436 randomly sampled American adults were surveyed. Daily consumer spending for the six-day period after Thanksgiving, spanning the Black Friday weekend and Cyber Monday, averaged $84.71. A 95% confidence interval based on this sample is ($80.31, $89.11). Determine whether the following statements are true or false, and explain your reasoning.



(a) We are 95% confident that the average spending of these 436 American adults is between $80.31 and $89.11.

(b) This confidence interval is not valid since the distribution of spending in the sample is right skewed.

(c) 95% of random samples have a sample mean between $80.31 and $89.11.

(d) We are 95% confident that the average spending of all American adults is between $80.31 and $89.11.

(e) A 90% confidence interval would be narrower than the 95% confidence interval since we don't need to be as sure about our estimate.

(f) In order to decrease the margin of error of a 95% confidence interval to a third of what it is now, we would need to use a sample 3 times larger.

(g) The margin of error is 4.4.

```
# Answer a- False.Confidence interval is used for population and not for sample.

# Answer b -FALSE, the sample is enough and the confidence interval is valid when
# the sample distribution is skewed.


# Answer C- False.Confidence interval is used for population and not for sample.

# Answer d- calculations shows that number is pretty close as per below
#calculation.


# input sample size, sample mean, and sample standard deviation

n <- 436
xbar<-mean(thanksgiving_spend$spending)
s<-sd(thanksgiving_spend$spending)
```

```r
#The following code shows how to calculate a 95% confidence interval for the true population mean
#calculate margin of error
margin <- qt(0.975,df=n-1)*s/sqrt(n)


#calculate lower and upper bounds of confidence interval
low <- (xbar - margin)
low
```

```
## [1] 80.28952
```

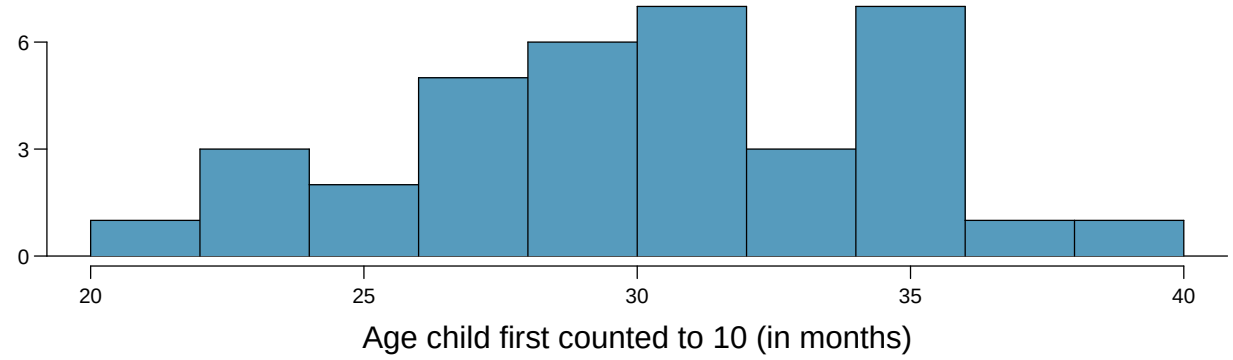
```r
high <- (xbar + margin)
high
```

```
## [1] 89.12401
```

```r
# Answer e - As a general rule of thumb, a small confidence interval is better.
# The confidence interval will narrow as your sample size increases,
# which is why a larger sample is always preferred. This can be proved from above
# my replacing .95 with .90 and seeing the new results.

# Answer F-FALSE.The margin of error for sample estimates will shrink with the square
# root of the sample size.So it needs to be 9 times larger not 3.

# Answer g : TRUE. MARGIN OF ERROR -
(89.11- 80.31)/2
```

```
## [1] 4.4
```

Gifted children, Part I. Researchers investigating characteristics of gifted children collected data from schools in a large city on a random sample of thirty-six children who were identified as gifted children soon after they reached the age of four. The following histogram shows the dis- tribution of the ages (in months) at which these children first counted to 10 successfully. Also provided are some sample statistics.



n	36
min	21
mean	30.69
sd	4.31
max	39

(a) Are conditions for inference satisfied?
(b) Suppose you read online that children first count to 10 successfully when they are 32 months old, on average. Perform a hypothesis test to evaluate if these data provide convincing evidence that the average age at which gifted children fist count to 10 successfully is less than the general average of 32 months. Use a significance level of 0.10.
(c) Interpret the p-value in context of the hypothesis test and the data.
(d) Calculate a 90% confidence interval for the average age at which gifted children first count to 10 successfully.
(e) Do your results from the hypothesis test and the confidence interval agree? Explain.

```
# Answer a
# Yes,  we have good sample size and sample was randomly selected

# Answer b
x <- 32
n <- 36
min <- 21
mean <- 30.69
sd <- 4.31
max <- 39

#Null hypothesis (H0): The average development for a child is mean=32 months.

#Alternate hypothesis (HA): The average development for a child is mean not equal to 32 months.


SE <- sd/sqrt(n)
Z <- (mean - x)/(SE)
p <- pnorm(Z)
p
```

```
## [1] 0.0341013
```

```
# since the the p-value = 0.0342026 is less than  significance level(of .1)
# Hence we reject the Null hypothesis (H0).



# Answer c:
# A p-value less than 0.05  is
# statistically significant. It indicates strong evidence against the null
# hypothesis,

z_score = (mean - x) / SE
P_value=pnorm(z_score)
P_value
```

```
## [1] 0.0341013
```

```
# Answer d:
# 90% CI.
# This means alpha = .10 We can get z(alpha/2) = z(0.05) from R:

print(z90<-qnorm(.95))
```

```
## [1] 1.644854
```

```
print(lower <- mean-z90 *SE)
```

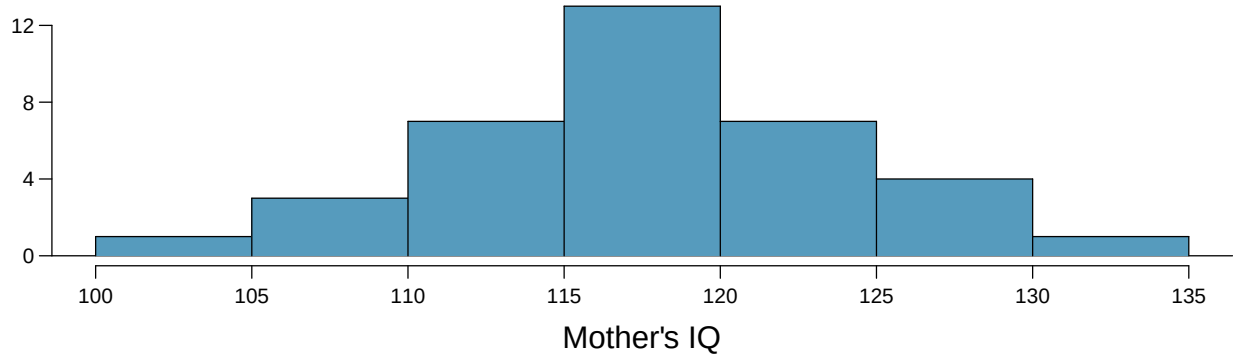
```
## [1] 29.50845
```

```
print(upper<-  mean+z90 *SE)
```

```
## [1] 31.87155
```

```
#Answer e:

#The results from the hypothesis test and the confidence interval
#agree because the confidence interval says that 90% chance the true mean
#for gifted children
#the 90% Confidence interval is from 29.51 to 31.87.months.
```

Gifted children, Part II. Exercise above describes a study on gifted children. In this study, along with variables on the children, the researchers also collected data on the mother's and father's IQ of the 36 randomly sampled gifted children. The histogram below shows the distribution of mother's IQ. Also provided are some sample statistics.



n	36
min	101
mean	118.2
sd	6.5
max	131

(a) Perform a hypothesis test to evaluate if these data provide convincing evidence that the average IQ of mothers of gifted children is different than the average IQ for the population at large, which is 100. Use a significance level of 0.10.
(b) Calculate a 90% confidence interval for the average IQ of mothers of gifted children.
(c) Do your results from the hypothesis test and the confidence interval agree? Explain.

```
#Answer a
x <- 100
mean = 118.2
n=36
sd=6.5
se=sd/sqrt(n)
SignificanceLevel=0.10

#Null hypothesis (H0): The average of mother's IQ of gifted children = population's IQ average.
#Alternate hypothesis (HA): The average of mother's IQ of gifted children not equal to population's IQ

Z <- (mean - x)/(SE)
p <- 1 - pnorm(Z)
p
```

```
## [1] 0
```

```
#since the the p-value = 0 is less than  significance level(of .1)
#Hence we reject the Null hypothesis (H0).

#Answer b
#This means alpha = .10 We can get z(alpha/2) = z(0.05) from R:

print(z90<-qnorm(.95))
```

```
## [1] 1.644854
```

```
print(lower <- mean-z90 *se)
```

```
## [1] 116.4181
```

```
print(upper<-  mean+z90 *se)
```

```
## [1] 119.9819
```

```
#Answer c
#Yes, these results agree since we have already rejected the Null hypothesis; that is that our p-value
```

CLT. Define the term "sampling distribution" of the mean, and describe how the shape, center, and spread of the sampling distribution of the mean change as sample size increases.

Answer: #The sampling distribution of the mean is the mean of the population from which #the scores were sampled So, if a population has a mean mu, then the mean of #the sampling distribution of the mean is also mu.

CFLBs. A manufacturer of compact fluorescent light bulbs advertises that the distribution of the lifespans of these light bulbs is nearly normal with a mean of 9,000 hours and a standard deviation of 1,000 hours.

(a) What is the probability that a randomly chosen light bulb lasts more than 10,500 hours?
(b) Describe the distribution of the mean lifespan of 15 light bulbs.
(c) What is the probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hours?
(d) Sketch the two distributions (population and sampling) on the same scale.
(e) Could you estimate the probabilities from parts (a) and (c) if the lifespans of light bulbs had a skewed distribution?

```
#Answer a-The probability that a randomly chosen light bulb lasts more than
#10,500 hours is 6.68%.

mu <- 9000
sd <- 1000
x <- 10500

zscore<- (x - mu)/sd

z = (10500 - 9000)/1000

p<- 1 - pnorm(z)
p
```

```
## [1] 0.0668072
```

```
#Answer b: Because the distribution of lifespans, X, are normal, then the sampling distribution
#of the mean lifespan of n = 15 lightbulbs is also normal.

SE = sd/sqrt(15)
SE
```

```
## [1] 258.1989
```

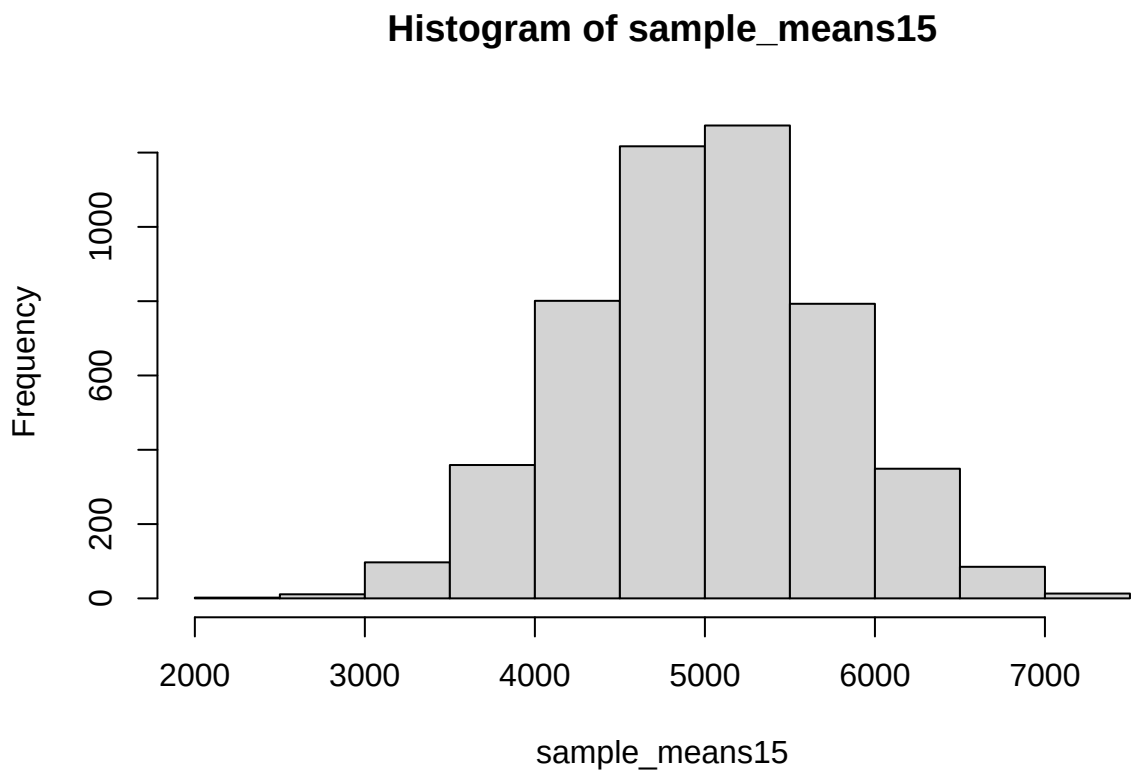
```
#Answer c:258.1989
sd<-1000
n<-15

se=sd/sqrt(n)
se
```

```
## [1] 258.1989
```

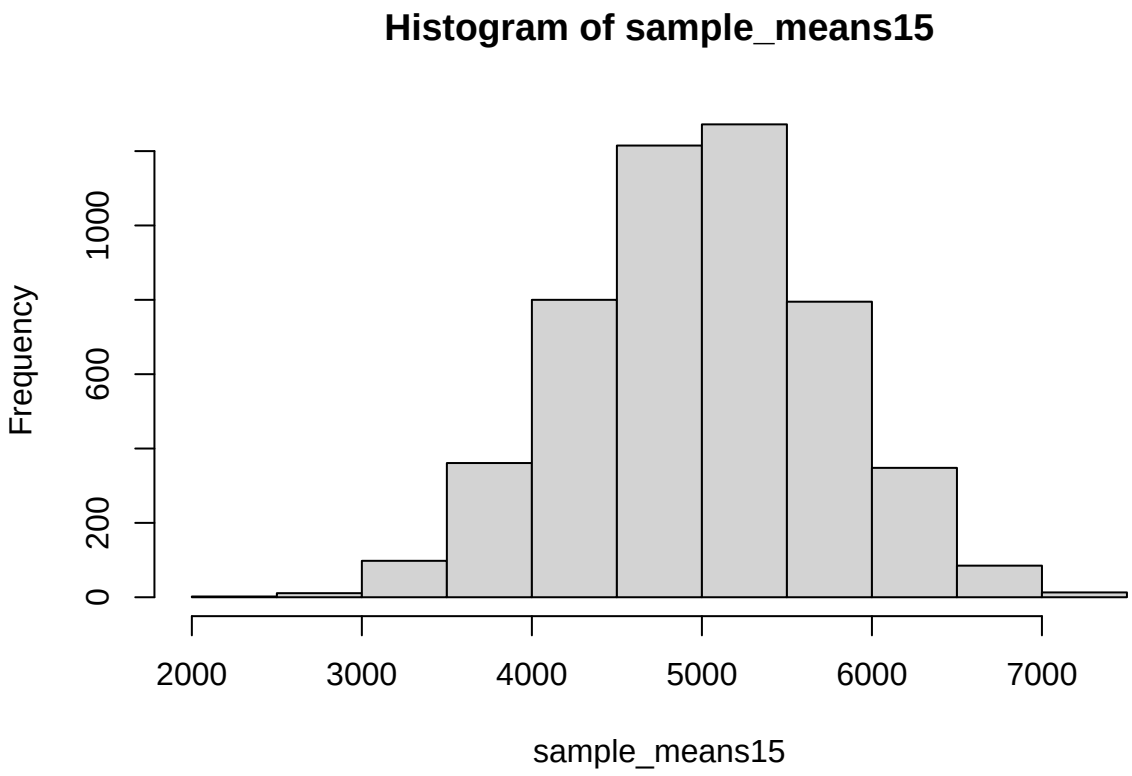
```
#Answer c: it is 0%

n <- 15
x <- 10500
mu <- 9000
sd <- 1000

z = (x - mu)/(sd/sqrt(n))
p<- 1 - pnorm(z)
p
```

```
## [1] 3.133452e-09
```

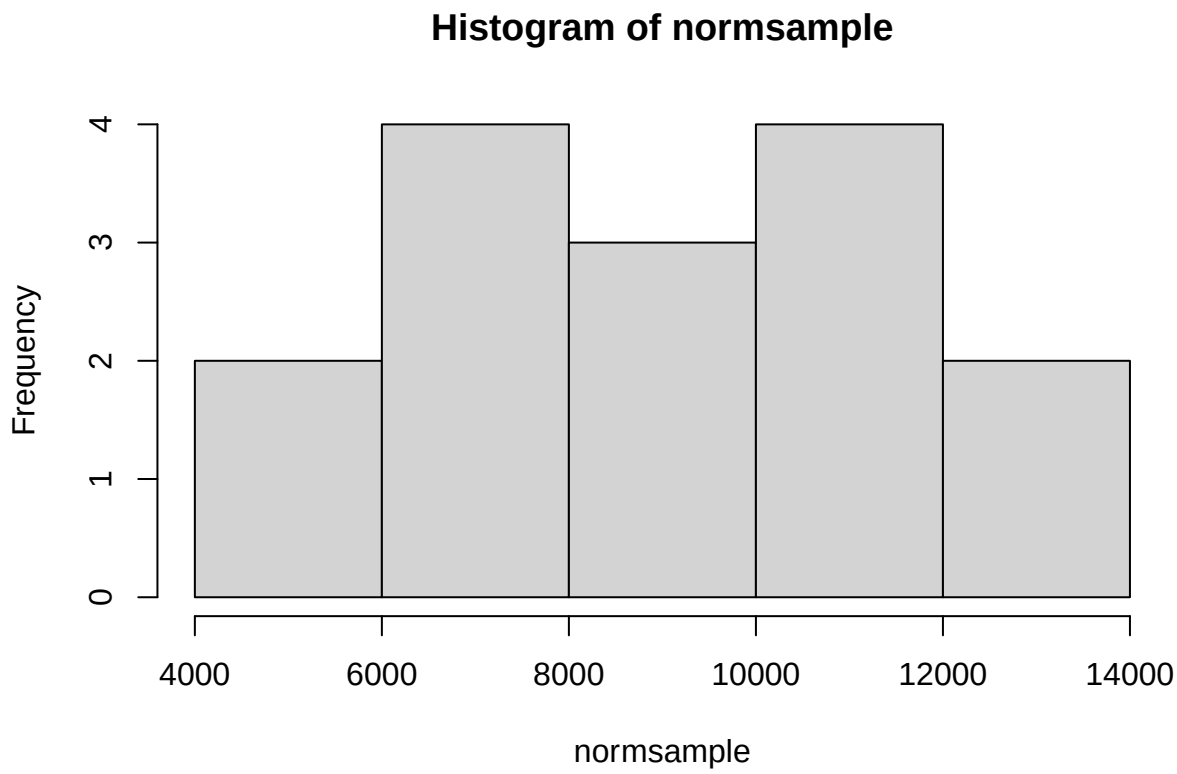
```
#Answer d:

#Population-N(9000,1000^2),Sampling-(9000,1000^2/15)

p <- 10000

sample_means15 <- rep(NA, 5000)

for(i in 1:5000){
   samp <- sample(p, 15)
   sample_means15[i] <- mean(samp)
   }

hist(sample_means15)
```



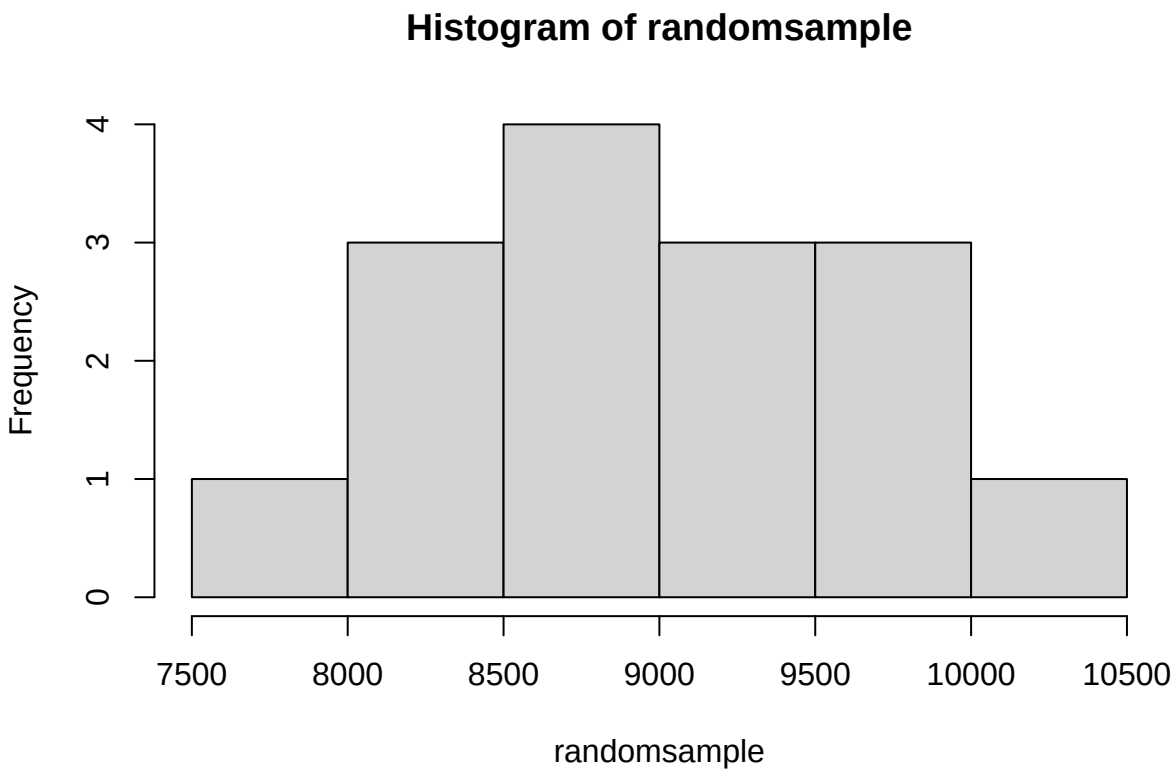

```
for(i in 1:15){
   samp <- sample(p, 15)
   sample_means15[i] <- mean(samp)
   }

hist(sample_means15)
```



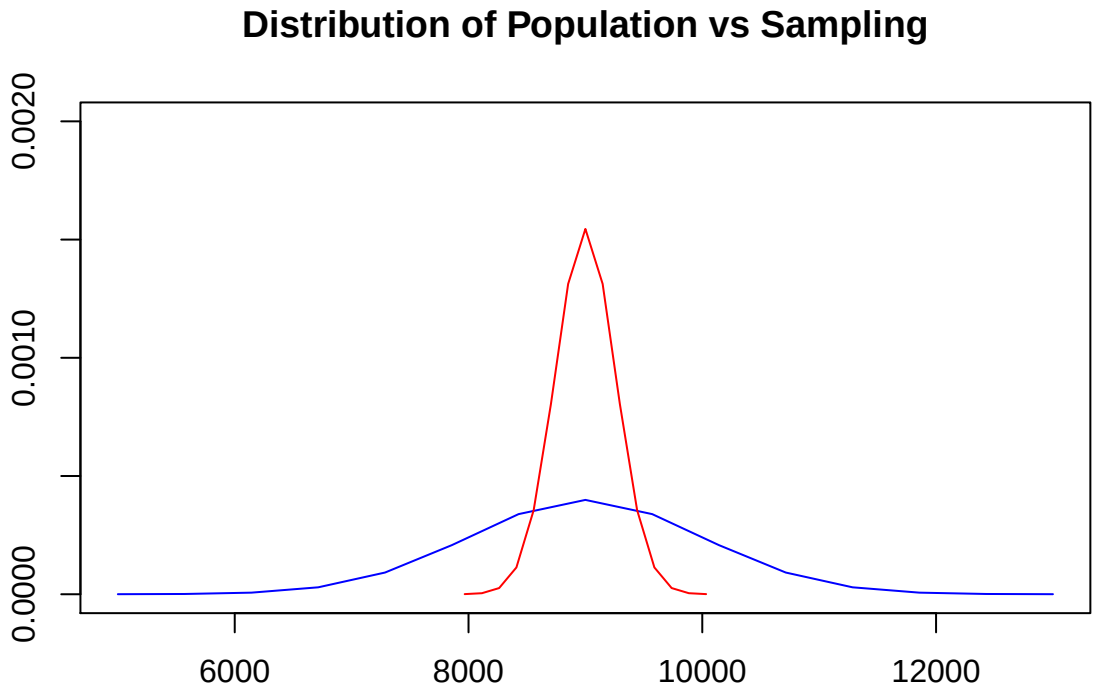

```r
#ANOTHER WAY

sd <- 1000
mean <- 9000
se <- 1000/sqrt(15)

normsample <- seq(mean - (4 * sd), mean + (4 * sd), length=15)
randomsample<- seq(mean - (4 * se), mean + (4 * se), length=15)
hnorm <- dnorm(normsample,mean,sd)
hrandom<- dnorm(randomsample,mean,se)

hist(normsample)
```




```
hist(randomsample)
```




```
plot(normsample, hnorm, type="l",col="blue",
xlab="",ylab="",main="Distribution of Population vs Sampling", ylim=c(0,0.002))
lines(randomsample, hrandom, col="red")
```




```
#Answer e:
#If the lifespans of light bulbs had a skewed distribution, we cannot estimate
#the probabilities because one of the assumptions in order to perform parts
#(a) and (c) calculations is that it must be a normal distribution
```

Same observation, different sample size. Suppose you conduct a hypothesis test based on a sample where the sample size is n = 50, and arrive at a p-value of 0.08. You then refer back to your notes and discover that you made a careless mistake, the sample size should have been n = 500. Will your p-value increase, decrease, or stay the same? Explain.

```
x <- 100
mean = 118.2
sd=6.5


n1 <- 50

se1<-sd/sqrt(n1)
se1
```

```
## [1] 0.9192388
```

```
Z1 <- (mean - x)/(se1)
Z1
```

```
## [1] 19.79899
```

```
n2 <- 500

se2<-sd/sqrt(n2)
se2
```

```
## [1] 0.2906888
```

```
Z2 <- (mean - x)/(se2)
Z2
```

```
## [1] 62.6099
```

```
#And by calculating the respective probabilities, we find that
#the p value will always change if the sample size changes. in the above
#case the p-value will decrease
```