Textbook Exercises

3.20 Customized Home Pages A random sample of n = 1675 Internet users in the US in January 2010 found that 469 of them have customized their web browser's home page to include news from sources and on topics that particularly interest them. State the population and parameter of interest. Use the information from the sample to give the best estimate of the population parameter. What would we have to do to calculate the value of the parameter exactly?

Solution

The population is all internet users in the US. The population parameter of interest is p, the proportion of internet users who have customized their home page. For this sample, $\hat{p} = 469/1675 = 0.28$. Unless we have additional information, the best point estimate of the population parameter p is $\hat{p} = 0.28$. To find p exactly, we would have to obtain information about the home page of every internet user in the US, which is unrealistic.

- **3.24 Average Household Size** The latest US Census lists the average household size for all households in the US as 2.61. (A household is all people occupying a housing unit as their primary place of residence.) Figure 3.6 shows possible distributions of means for 1000 samples of household sizes. The scale on the horizontal axis is the same in all four cases.
 - (a) Assume that two of the distributions show results from 1000 random samples, while two others show distributions from a sampling method that is biased. Which two dotplots appear to show samples produced using a biased sampling method? Explain your reasoning. Pick one of the distributions that you listed as biased and describe a sampling method that might produce this bias.
 - (b) For the two distributions that appear to show results from random samples, suppose that one comes from 1000 samples of size n = 100 and one comes from 1000 samples of size n = 500. Which distribution goes with which sample size? Explain.

Solution

- (a) The two distributions centered at the population average are probably unbiased, distributions A and D. The two distributions not centered at the population average ($\mu=2.61$) are biased, dotplots B and C. The sampling for Distribution B gives an average too high, and has large households over-represented. The sampling for Distribution C gives an average too low and may have been done in an area with many people living alone.
- (b) The larger the sample size the lower the variability, so distribution A goes with samples of size 100, and distribution D goes with samples of size 500.
- **3.25 Proportion of US Residents Less than 25 Years Old** The US Census indicates that 35% of US residents are less than 25 years old. Figure 3.7 shows possible sampling distributions for the proportion of a sample less than 25 years olds, for sample of size n = 20, n = 100, and n = 500.
 - (a) Which distribution goes with which sample size?
 - (b) If we use a proportion \hat{p} , based on a sample of size n=20, to estimate the population parameter p=0.35, would it be very surprising to get an estimate that is off by more than 0.10 (that is, the sample proportion is less than 0.25 or greater than 0.45)? How about with a sample of size n=100? How about with a sample of size n=500?
 - (c) Repeat part (b) if we ask about the sample proportion being off by just 0.05 or more.
 - (d) Using parts (b) and (c), comment on the effect that sample size has on the accuracy of an

estimate.

Solution

- (a) As the sample size goes up, the accuracy improves, which means the spread goes down. We see that distribution A goes with sample size n = 20, distribution B goes with n = 100, and distribution C goes with n = 500.
- (b) We see in dotplot A that quite a few of the sample proportions (when n = 20) are less than 0.25 or greater than 0.45, so being off by more than 0.10 would not be too surprising. While it is possible to be that far away in dotplot B (when n = 100), such points are much more rare, so it would be somewhat surprising for a sample of size n = 100 to miss by that much. None of the points in dotplot C are more than 0.10 away from p = 0.35, so it would be extremely unlikely to be that far off when n = 500.
- (c) Many of the points in dotplot A fall outside of the interval from 0.30 to 0.40, so it is not at all surprising for a sample proportion based on n=20 to be more than 0.05 from the population proportion. Even dotplot B has quite a few values below 0.30 or above 0.40, so being off by more than 0.05 when n=100 is not too surprising. Such points are rare, but not impossible in dotplot C, so a sample of size n=500 might possibly give an estimate that is off by more than 0.05, but it would be pretty surprising.
- (d) As the sample size goes up, the accuracy of the estimate tends to increase.
- **3.28 Hollywood Movies** Data 2.7 on page 93 introduces the dataset **HollywoodMovies2011**, which contains information on all the 136 movies that came out of Hollywood in 2011. One of the variables is the budget (in millions of dollars) to make the movie. The figure in the book shows two box plots. One represents the budget data for one random sample of size n = 30. The other represents the values in a sampling distribution of 1000 means of budge data from samples of size 30.
 - (a) Which is which? Explain.
 - (b) From the boxplot showing the data from one random sample, what does one value in the sample represent? How many values are included in the data to make the boxplot? Estimate the minimum and maximum values. Give a rough estimate of the mean of the values and use appropriate notation for your answer.
 - (c) From the box plot showing the data from a sampling distribution, what does one value in the sampling distribution? How many values are included in the data to make the boxplot? Estimate the minimum and maximum values. Give a rough estimate of the value of the population parameter and use the appropriate notation for your answer.

Solution

- (a) We expect means of samples of size 30 to be much less spread out than values of budgets of individual movies. This leads us to conclude that Boxplot A represents the sampling distribution and Boxplot B represents the values in a single sample. We can also consider the shapes. Boxplot A appears to be symmetric and Boxplot B appears to be right skewed. Since we expect a sampling distribution to be symmetric and bell-shaped, Boxplot A is the sampling distribution and the skewed Boxplot B shows values in a single sample.
- (b) Boxplot B shows the data from one sample of size 30. Each data value represents the budget,

in millions of dollars, for one Hollywood movie made in 2011. There are 30 values included in the sample. The budgets range from about 1 million to 145 million for this sample. We see in the boxplot that the median is about 30 million dollars. Since the data are right skewed, we expect the mean to be higher. We estimate the mean to be about 40 million or 45 million. This is the mean of a sample, so we have \bar{x} is approximately 45 million dollars.

- (c) Boxplot A shows the data from a sampling distribution using samples of size 30. Each data value represents the mean of one of these samples. There are 1000 means included in the distribution. They range from about 27 to 79 million dollars. The center of the distribution is a good estimate of the population parameter, and the center appears to be about μ is approximately 53 million dollars, where μ represents the mean budget, in millions of dollars, for all movies coming out of Hollywood in 2011.)
- **3.36 Performers in the Rock and Roll Hall of Fame** From its founding through 2012, the Rock and Roll Hall of Fame has inducted 273 groups or individuals, and 181 of the inductees have been performers while the rest have been related to the world of music in some way other than as a performer. The full dataset is available in **RockandRoll**.
 - (a) What proportion of inductees have been performers? Use the correct notation with your answer.
 - (b) If we took many samples of size 50 from the population of all inductees and recorded the proportion who were performers for each sample, what shape do we expect the distribution of sample proportions to have? Where do we expect it to be centered?

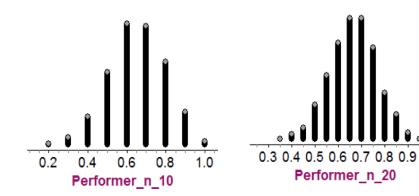
Solution

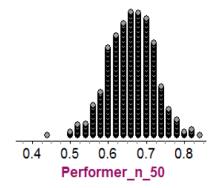
- (a) This is a population proportion so the correct notation is p. We have p = 181/273 = 0.663.
- (b) We expect it to be symmetric and bell-shaped and centered at the population proportion of 0.663.
- **3.38 A Sampling Distribution for Performers in the Rock and Roll Hall of Fame** Exercise 3.36 tells us that 181 of the 273 inductees to the Rock and Roll Hall of Fame have been performers. The data are given in **RockandRoll** Using all inductees as your population:
 - (a) Use StatKey of other technology to take many random samples of size n=10 and compute the sample proportion that are performers. What is the standard error of the sample proportions? What is the value of the sample proportion farthest from the population proportion of p=0.663? How far away is it?
 - (b) Repeat part (a) using samples of size n = 20.
 - (c) Repeat part (a) using samples of size n = 50.
 - (d) Use your answers to parts (a), (b), and (c) to comment on the effect of increasing the sample size on the accuracy of using a sample proportion to estimate the population proportion.

Solution

(a) The standard error is the standard deviation of the sampling distribution (given in the upper right corner of the sampling distribution box in StatKey) and is likely to be about 0.15. Answers will vary, but the sample proportions should go from about 0.2 to about 1.0 (as shown in the dotplot below). In that case, the farthest sample proportion from p = 0.663 is $\hat{p} = 0.2$, and it is 0.663 - 0.2 = 0.463 off from the correct population value.

- (b) The standard error is the standard deviation of the sampling distribution and is likely to be about 0.11. Answers will vary, but the sample proportions should go from about 0.35 to about 0.95 (as shown in the dotplot below). In that case, the farthest sample proportion from p = 0.663 is $\hat{p} = 0.35$, and it is 0.663 0.35 = 0.313 off from the correct population value.
- (c) The standard error is the standard deviation of the sampling distribution and is likely to be about 0.06. Answers will vary, but the sample proportions should go from about 0.44 to about 0.84 (as shown in the dotplot below). In that case, the farthest sample proportion from p = 0.663 is $\hat{p} = 0.44$, and it is 0.663 0.44 = 0.223 off from the correct population value.
- (d) Accuracy improves as the sample size increases. The standard error gets smaller, the range of values gets smaller, and values tend to be closer to the population value of 0.663.





- **3.54 Number of Text Messages a Day** A random sample of n = 755 US cell phone users age 18 and older in May 2011 found that the average number of text messages sent or received per day is 41.5 messages, with standard error about 6.1.
 - (a) State the population and parameter of interest. Use the information from the sample to give the best estimate of the population parameter.
 - (b) Find and interpret a 95% confidence interval for the mean number of text messages.

- (a) The population is all cell phone users age 18 and older in the US. The population parameter of interest is μ , the mean number of text messages sent and received per day. The best point estimate for μ is the sample mean, $\bar{x} = 41.5$.
- (b) The point estimate is x, so a 95% confidence interval is given by:

$$\bar{x}$$
 \pm 2SE
41.5 \pm 2(6.1)
41.5 12.2
29.3 to 53.7

We are 95% confident that the mean number of text messages a day for all cell phone users in the US is between 29.3 and 53.7.

- 3.60 Effect of Overeating for One Month: Average Long-Term Weight Gain Overeating for just four weeks can increase fat mass and weight over two years later, a Swedish study shows. Researchers recruited 18 healthy and normal-weight people with an average age of 26. For a four-week period, participants increased calorie intake by 70% (mostly by eating fast food) and limited daily activity to a maximum of 5000 steps per day (considered sedentary). Not surprisingly, weight and body fat of the participants went up significantly during the study and then decreased after the study ended. Participants are believed to have returned to the diet and lifestyle they had before the experiment. However, two and half years after the experiment, the mean weight gain for participants was 6.8lbs with a standard error of 1.2 lbs. A control group that did not binge had no change in weight.
 - (a) What is the relevant parameter?
 - (b) How could we find the actual exact value of the parameter?
 - (c) Give a 95% confidence interval for the parameter and interpret it.
 - (d) Give the margin of error and interpret it.

- (a) The parameter of interest is μ , the mean effect on weight 2.5 years after a month of overeating and being sedentary.
- (b) The only way to find the exact value would be to have all members of a population overeat and be inactive for a month and then measure the effect 2.5 years later. This is not a good idea!
- (c) The 95% confidence interval using the standard error is $x \pm 2SE = 6.8 \pm 2(1.2) = 6.8 \pm 2.4$. We are 95% sure that the mean weight gain over 2.5 years by people who overeat for a month is between 4.4 and 9.2 pounds.
- (d) The margin of error is ± 2.4 which means we are relatively confident that our estimate of 6.8 pounds is within 2.4 pounds of the true mean weight gain for the population.
- **3.61 Training Fish to Pick a Color** Fish can be trained quite easily. With just seven days of training, golden shiner fish learn to pick a color (yellow or blue) to receive a treat, and the fish will swim to that color immediately. On the first day of training, however, it takes them some time. In the study described under *Fish Democracies* above, the mean time for the fish in the study to reach the yellow mark is $\bar{x} = 51$ seconds with a standard error for this statistic of 2.4 Find and interpret a 95% confidence interval for the mean time it takes a golden shiner fish to reach the yellow mark. Is it plausible that the average time it take fish to find the mark is 60 seconds? Is it plausible that it is 55 seconds?

Solution

Let μ represent the mean time for a golden shiner fish to find the yellow mark. A 95% confidence interval is given by

$$x \pm 2SE$$
 $51 \pm 2(2.4)$
 51 ± 4.8
 $46.2 \text{ to } 55.8$

A 95% confidence interval for the mean time for fish to find the mark is between 46.2 and 55.8

seconds. We are 95% sure that the mean time it would take fish to find the target for all fish of this breed is between 46.2 seconds and 55.8 seconds. In other words, the plausible values for the population mean μ are those values between 46.2 and 55.8. Therefore, 60 is not a plausible value for the mean time for all fish, but 55 is.

3.62 How Often Does the Fish Majority Win? In a school of fish with a minority of stingily opinionated fish wanting to aim for the yello mark and a majority of less passionate fish wanting to aim for the blue mark, as described under *Fish Democracies* above, a 95% confidence interval for the proportion of times the majority wins (they go to the blue mark) is 0.09 to 0.26. Interpret this confidence interval. Is it plausible that fish in this situation are equally likely to go for either of the two options?

Solution

We are 95% confident that schools of fish in this situation will end up going with the majority over the opinionated minority only between 9% and 26% of the time. It is not plausible that the schools of fish in this situation are equally likely to go for either option since that would indicate a proportion of p = 0.5 for each option, and 0.5 is not in the range of plausible values. The highly opinionated fish are definitely having an effect.

- **3.76 Ants on a Sandwich** How many ants will climb on a piece of a peanut butter sandwich left on the ground near an ant hill? To study this, a student in Australia left a piece of a sandwich for several minutes, then covered it with a jar and counted the number of ants. He did this eight times, and the results are shown in the table in the book. (In fact, he also conducted an experiment to see if there is a difference in number of ants based on the sandwich filling. The details of that experiment are given in Chapter 8, and the full dataset is in **SandwichAnts**.)
 - (a) Find the mean and standard deviation of the sample.
 - (b) Describe how we could use eight slips of paper to create one bootstrap statistic. Be specific.
 - (c) What do we expect to be the shape and center of the bootstrap distribution?
 - (d) What is the population parameter of interest? What is the best point estimate for that parameter?
 - (e) A bootstrap distribution of 5000 bootstrap statistics gives a standard error of 4.85. Use the standard error to find and interpret a 95% confidence interval for the parameter defined in part (d).

Solution

- (a) We find for the 8 values in the table that $\bar{x} = 34.0$ and s = 14.63.
- (b) We put the 8 values on the 8 slips of paper and mix them up. Draw one and write down the value and put it back. Mix them up, draw another, and do this 8 times. The resulting 8 numbers form a bootstrap sample, and the mean of those 8 numbers form one bootstrap statistic.
- (c) We expect that the bootstrap distribution will be bell-shaped and centered at approximately 34.
- (d) The population parameter of interest is the mean, μ , number of ants on all possible peanut butter sandwich bits set near this ant hill. There are other possible answers for the population; for example, you might decide to limit it to the time of day at which the student conducted the study. The best point estimate is the sample mean $\bar{x} = 34$.

(e) We have

$$x \pm 2SE$$
 $34.0 \pm 2(4.85)$
 $34.0 9.7$
 $24.3 \text{ to } 43.7$

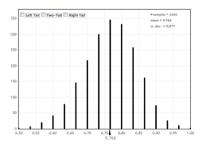
We are 95% confident that the mean number of ants to climb on a bit of peanut butter sandwich left near an ant hill is between 24.3 ants and 43.7 ants.

3.79 Rats with Compassion The phrase "You dirty rat" does rats a disservice. In a recent study, rats showed compassion that surprised scientists. Twenty-three of the 30 rats in the study freed another trapped rat in their cage, even when chocolate served as a distraction and even when the rats would then have to share the chocolate with their freed companion. (Rats, it turns out, love chocolate.) Rats did not open the cage when it was empty or when there was a stuffed animal inside, only when a fellow rat was trapped. We wish to use the sample to estimate the proportion of rats to show empathy in the way. The data are available in the dataset **CompassionateRate**.

- (a) Give the relevant parameter and its point estimate.
- (b) Describe how to use 30 slips of paper to create one bootstrap statistic. Be specific.
- (c) Use *StatKey* or other technology to create a bootstrap distribution. Describe the shape and center of the bootstrap distribution. What is the standard error?
- (d) Use the standard error to find and interpret a 95% confidence interval for the proportion of rats likely to show empathy.

Solution

- (a) The best point estimate for the proportion, p, of rats showing empathy is $\hat{p} = 23/30 = 0.767$.
- (b) On 23 of the slips, we write "yes" (showed empathy) and on the other 7, we write "no". We then mix up the slips of paper, draw one out and record the result, yes or no. Put the slip of paper back and repeat the process 30 times. This set of yes's and no's is our bootstrap sample. The proportion of yes's in the sample is our bootstrap statistic.
- (c) Using technology, we see that the bootstrap distribution is bell-shaped and centered approximately at 0.767. We also see that the standard error is about 0.077.



(d) We have

$$\hat{p}$$
 \pm $2SE$

$$0.767 \pm 2(0.077)$$

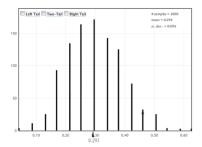
 0.767 ± 0.154
 $0.613 \text{ to } 0.921$

For all laboratory rats, we are 95% confident that the proportion of rats that will show empathy in this manner is between 61.3% and 92.1%.

3.80 Are Female Rats More Compassionate Than Male Rats? Exercise 3.79 describes a study in which rats showed compassion by freeing a trapped rat. In the study, all six of the six female rats showed compassion by freeing the trapped rat while 17 of the 24 male rats did so. Use the results of this study to give a point estimate for the difference in proportion of rate showing compassion, between female rats and male rats. The use StatKey or other technology to estimate the standard error and use it to compute an interval estimate for the difference in proportions. Use the interval to determine whether it is plausible that male and female rats are equally compassionate (i.e., that the difference in proportions is zero). The data are available in the dataset CompassionateRats.

Solution

The sample proportion of females showing compassion is $\hat{p}_F = 6/6 = 1.0$. The sample proportion of males showing compassion is $\hat{p}_M = 17/24 = 0.708$. The best point estimate for the difference in proportions $p_F - p_M$ is $\hat{p}_F - \hat{p}_M = 1.0?0.708 = 0.292$. Using StatKey to create a bootstrap distribution for a difference in proportions using this sample data, we see a standard error of 0.094.



We have

$$(\hat{p}_F - \hat{p}_M)$$
 \pm 2SE
 $(1.0 - 0.708)$ \pm 2(0.094)
 0.292 \pm 0.188
 0.104 to 0.480

Based on this interval the percentage of female rats likely to show compassion is between 10.4% and 48% higher than the percentage of male rats likely to show compassion. Since zero is not in the interval estimate, it is not very plausible that male and female rats are equally compassionate.

Computer Exercises

R problem 1 Load the data set USStates from the textbook into R.

1. Use the ggplot2 library to draw a suitable graph of the Smokers variable, the percentage of residents who smoke. Describe the shape and center of this distribution.

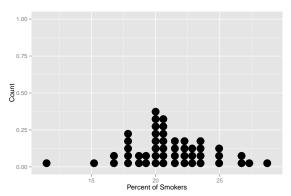
Solution

Below are several graphs which display the distribution of of the Smokers variable. We notice that the distribution is fairly symmetric, and is centered around 20 percent.

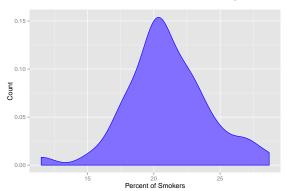
Histogram of Percentage of Residents Who Smoke Dotplot of Percentage of Residents Who Smoke From the USStates Data Set

Percent of Smokers

From the USStates Data Set



Density Plot of Percentage of Residents Who Smoke From the USStates Data Set]



R Code

```
hist(Smokers)
library(ggplot2)
ggplot(USStates, aes(x = Smokers)) + geom_histogram(binwidth=1,color="Navy",fill="Blue")
+ylab("Count")+xlab("Percent of Smokers")
ggplot(USStates,aes(x=Smokers))+geom_density(color="blue",fill="slateblue1")
+ylab("Count")+xlab("Percent of Smokers")
ggplot(USStates,aes(x=Smokers))+geom_dotplot()+ylab("Count")+xlab("Percent of Smokers")
```

2. What is the mean proportion of smokers, averaged evenly across the 50 states?

The mean percent of smokers is $\frac{1050.4}{5} = 21.008$

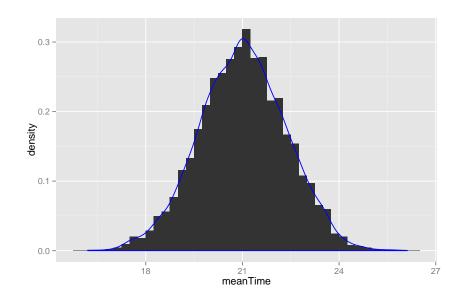
R Code

sum(Smokers)/50

3. Write a function in R that will take 10,000 random samples of 5 states and return the mean proportion of smokers for each sample. This is a simulation of the sampling distribution of the sample mean proportion of smokers. Calculate the mean and standard deviation of this sampling distribution and use ggplot2 to display the distribution with a density plot. Describe the center and shape of the distribution in this plot.

Solution

Using R, I was able to create function which took 10,000 random samples of 5 states and returned the mean proportion of smokers for each sample. The mean of the means was 21.02, and the standard deviation was 1.34. Below is a histogram overlaid with the density curve of the sampling distribution. The plot is symmetric and centered around 21.



R Code

```
sampling.dist <- function(size=10000)
{
    require(Lock5Data)
    data(USStates)
    my.props <- numeric(size)
    for (i in 1:size)
    {
        x <- sample(Smokers,5,replace=FALSE)
        my.props[i] <- sum(x)/5
}</pre>
```

```
mean <- mean(my.props)
    sd <- sd(my.props)
    return(list(mean=mean,sd=sd,my.props=my.props))
}
sampling.dist()$mean
sampling.dist()$sd
means=sampling.dist()$my.props
ggplot(data.frame(mean = sampling.dist()$my.props),aes(x=mean)) +
    geom_histogram(binwidth=0.25,aes(y=..density..)) +
    geom_density(color="blue")</pre>
```

4. What is the standard deviation from the previous part called?

Solution

The standard deviation from the previous part is called the standard error.

5. One sample of five states is Arkansas, Florida, Pennsylvania, California, and Vermont. Find the mean proportion of smokers across these five states and use information from earlier parts to construct a 95% confidence interval for the mean proportion of smokers. Does this confidence interval contain the true mean?

Solution

We find that the mean smoking percentage of the sample consisting of the five states Arkansas, Florida, Pennsylvania, California, and Vermont is 20.66. Previously, we calculated the standard deviation of our sampling distribution to be 1.34. Thus, our 95% confidence interval is

$$\bar{x} \pm 2*SE$$
 $20.66 \pm 2*1.34$
 $(17.99395 , 23.3540)$

We are 95% confident that the true mean of smoking percentages in the 50 US states is between 17.99% and 23.35%. Also, this confidence interval does contain the true mean.

R. Code:

```
Arkansas <- subset(USStates,USStates$State=='Arkansas')
Florida <- subset(USStates,USStates$State=='Florida')
Pennsylvania <- subset(USStates,USStates$State=='Pennsylvania')
California <- subset(USStates,USStates$State=='California')
Vermont <- subset(USStates,USStates$State=='Vermont')
FiveStates <- c(Arkansas$Smokers,Florida$Smokers,Pennsylvania$Smokers,California$Smokers,Vermont$Smokers)
mean(FiveStates)
CI <-c(mean(FiveStates)-2*sampling.dist()$sd,mean(FiveStates)+2*sampling.dist()$sd)</pre>
```

6. What proportion of the 10,000 randomly sampled means are actually within the margin of error of a 95% confidence interval? How close is the actual number to what the expected proportion is?

We find that 9465 of the sample means are within the margin of error. Thus, the proportion of sample means within the margin of error are $\frac{9465}{10000} = 0.9465$. This number is very close to the expected proportion of 0.95.

R Code

 $length(which(17.99395 \le means \& means \le 23.35405))/10000$

R problem 2 Load the data set CommuteAtlanta into R.

1. How many cases are there? What variables are included? Find the mean and standard deviation of each quantitative variable.

Solution

Using R, we are able to determine that there are 500 cases and five variables. The variables are City, Age, Distance, Time, and Sex. The quantitative variables include Age, Distance, and Time. Their means and standard deviations are summarized in the table below.

| Variable | Age | Distance | Time |
|--------------------|----------|----------|----------|
| Mean | 40.242 | 18.156 | 29.11 |
| Standard Deviation | 12.08219 | 13.79828 | 20.71831 |

R Code:

str(CommuteAtlanta)
names(CommuteAtlanta)
mean(Age)
sd(Age)
mean(Distance)
sd(Distance)
mean(Time)
sd(Time)

2. Use R to apply the bootstrap: take 1000 bootstrap samples. Compute the standard deviation of the 1000 sample means distances. Construct a 95% confidence interval for the mean commute distance in Atlanta.

Solution

Using R, I took 1000 samples of size 500 from the variable Distance using replacement. I found the mean of each of these samples. I then computed the mean and standard deviation of the 1000 means. There are as follows.

mean of means
$$= 18.1884$$

sd = 6.101637

Below is a histogram showing my bootstrap distribution.

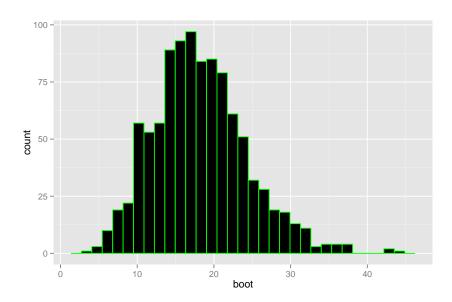


Figure 1: Bootstrap Distribution of Distances from CommuteAtlanta Data

With a standard error of 6.10, we construct a 95% confidence interval as follows

```
\bar{x} \pm 2 * SE
20.66 \pm 2 * 6.10
(5.95 , 30.36)
```

```
R Code
```

```
my.boot <- numeric(1000)
for(i in 1:1000)
{
    x <- sample(Distance,length(CommuteAtlanta),replace=TRUE)
    my.boot[i] <- mean(x)
}
ggplot(data.frame(boot=my.boot),aes(x=boot))+geom_histogram(color="green",fill="black")
hist(my.boot)
mean(my.boot)
sd(my.boot)
CI <- c(mean(Distance)-2*sd(my.boot),mean(Distance)+2*sd(my.boot))</pre>
```

3. Interpret this confidence interval. What important assumptions are you making for this interpretation to be correct?

Solution

We are 95% confident that the true mean distance of commute to work for people in Atlanta is between 5.95 and 30.36.

In order to make this assumption, we are assuming that our original sample was taken randomly and is not biased.