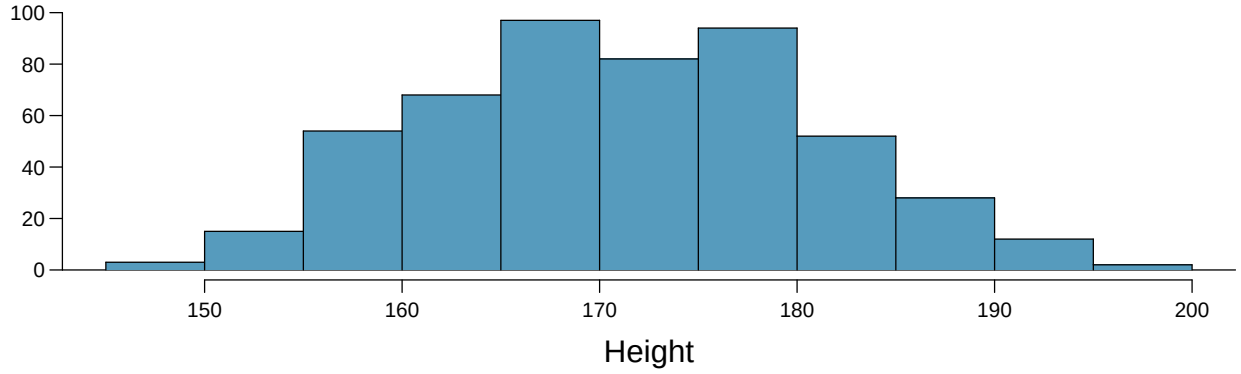
Chapter 5 - Foundations for Inference

Joshua Registe

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?

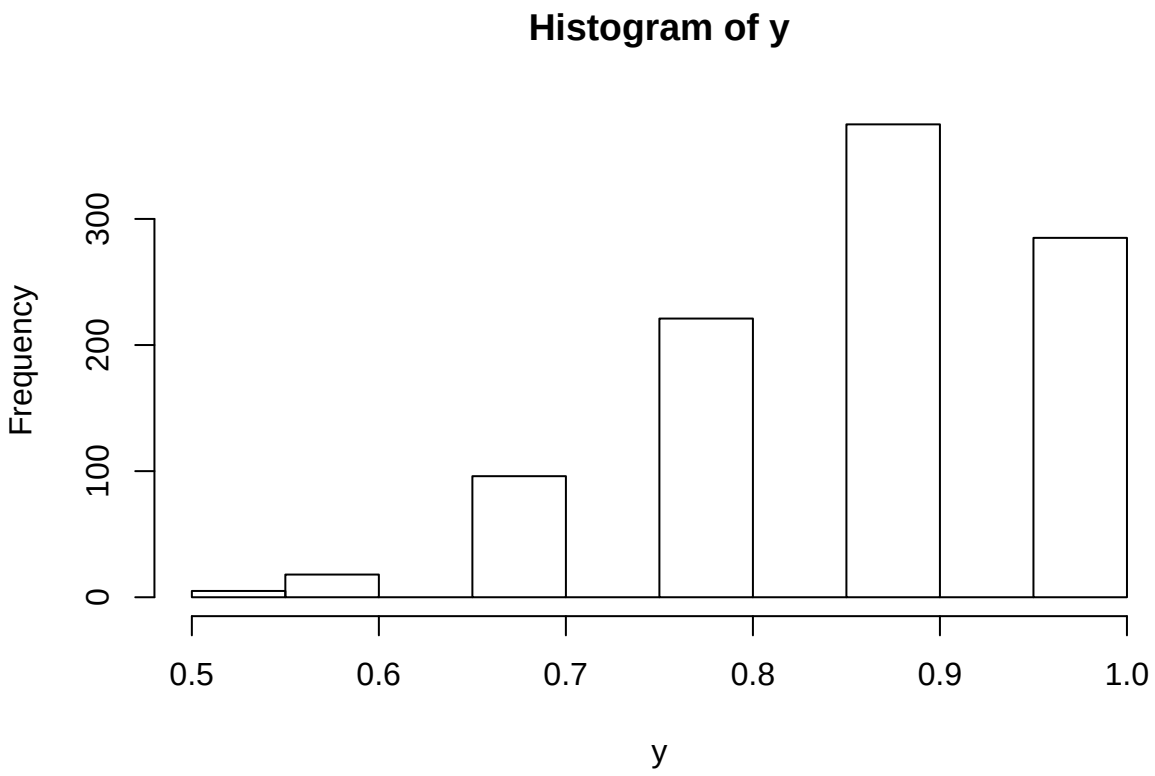
```
paste0("The point estimate for the average height is ",round(mean(bdims$hgt),1))
```

```
## [1] "The point estimate for the average height is 171.1"
```

```
paste0("The point estimate for the median height is ",round(median(bdims$hgt),1))
```

```
## [1] "The point estimate for the median height is 170.3"
```

```
pop = 250000000
all<-c(rep("support",.88*pop),rep("not",.12*pop))
y <- rep(NA,1000)
for (i in 1:1000){

sampled_entries<-sample(all, size= 10)

y[i]<-  sum(sampled_entries=="support")/10



}

hist(y)
```




```
mean(y)
```

```
## [1] 0.8798
```

(b) What is the point estimate for the standard deviation of the heights of active individuals? What about the IQR?

```
paste0("The point estimate for the Interquartile range is",
       round(quantile(bdims$hgt,.75,na.rm = T)-quantile(bdims$hgt,.25,na.rm = T),1))
```

```
## [1] "The point estimate for the Interquartile range is14"
```

(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.

```
?percent
```

```
## starting httpd help server ... done
```

```
paste0("The probability that someone is taller than 180 cm is approximately ",
       percent(round(pnorm(180,mean(bdims$hgt,na.rm = T),sd(bdims$hgt,na.rm = T),lower.tail = F),4),.1))
```

```
## [1] "The probability that someone is taller than 180 cm is approximately 17.3%"
```

```
paste0("The probability that someone is shorter than 155 cm is approximately ",
       percent(round(pnorm(155,mean(bdims$hgt,na.rm = F),sd(bdims$hgt,na.rm = T),lower.tail = T),4),.1))
```

```
## [1] "The probability that someone is shorter than 155 cm is approximately 4.3%"
```

Based on above, I would consider someone shorter than 150 cm unusually short as less than 5% of people from the sample population are this short. However, i would consider taller than 180 cm not to be unusual as this represents greater than 15% of the sample population.

(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.

I would not expect the mean and standard deviation of the new sample to be the same as the ones given above, but I expect the them to be close. If an infinte amount of samples were taken, we would approximate this via the central limit theorem and normal distribution would show that samples drawn will be close to the population mean and sd and it would be less likely that the samples would deviate from this.

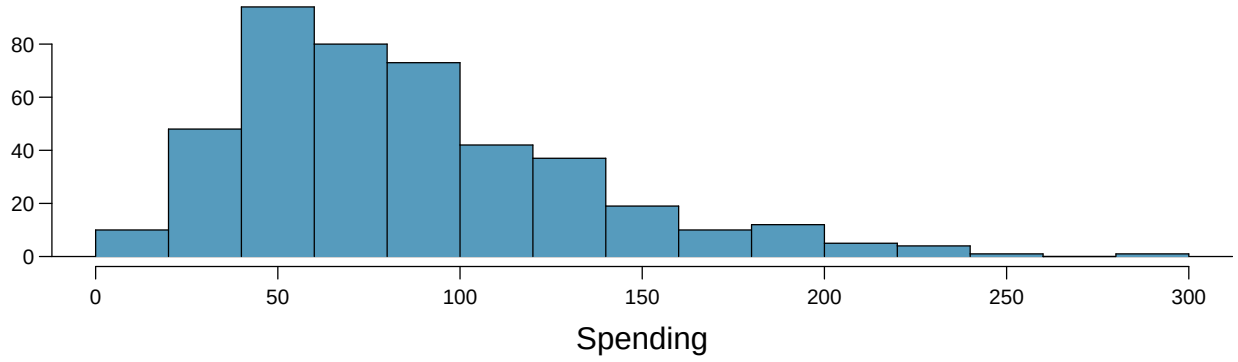
(e) The sample means obtained are point estimates for the mean height of all active individuals, if the sample of individuals is equivalent to a simple random sample. What measure do we use to quantify the variability of such an estimate (Hint: recall that $SD_x = \frac{\sigma}{\sqrt{n}}$)? Compute this quantity using the data from the original sample under the condition that the data are a simple random sample.

We would compute our SE as shown below:

```
(SE = sd(bdims$hgt)/ sqrt(length(bdims$hgt)))
```

```
## [1] 0.4177887
```

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.

```
Average_spending<-mean(tgSpending$spending)

Std_spending<-sd(tgSpending$spending)

Spending_error<- qnorm(.975)*Std_spending/sqrt(length(tgSpending$spending))

(Upper_Est<- Average_spending + Spending_error)
```

```
## [1] 89.11172
```

```
(Lower_Est<- Average_spending - Spending_error)
```

```
## [1] 80.30181
```

True, we are 95% confident that the average spending of these 436 americans are between 80.3 and 89.1

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

False, we can infer that if we took the mean of a bunch of random samples, the distribution would be normal and the confidence interval would be valid for estimating with a level of certainty the population mean.

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

False, the confidence interval is used to estimate the confidence interval of the population mean, not other random sample means.

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

True, we infer from our sample's confidence interval the population mean and that 95% interval lies 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.

```
Average_spending<-mean(tgSpending$spending)

Std_spending<-sd(tgSpending$spending)

Spending_error<- qnorm(.95)*Std_spending/sqrt(length(tgSpending$spending))


(Upper_Est<- Average_spending + Spending_error)
```

```
## [1] 88.40352
```

```
(Lower_Est<- Average_spending - Spending_error)
```

```
## [1] 81.01001
```

True, The lower the confidence interval, the narrower the interval.

(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.

```
Average_spending<-mean(tgSpending$spending)

Std_spending<-sd(tgSpending$spending)

(Spending_error<- qnorm(.975)*Std_spending/sqrt(length(tgSpending$spending)))
```

```
## [1] 4.404957
```

```
(Spending_error<- qnorm(.975)*Std_spending/(sqrt(length(tgSpending$spending)*3^2)))
```

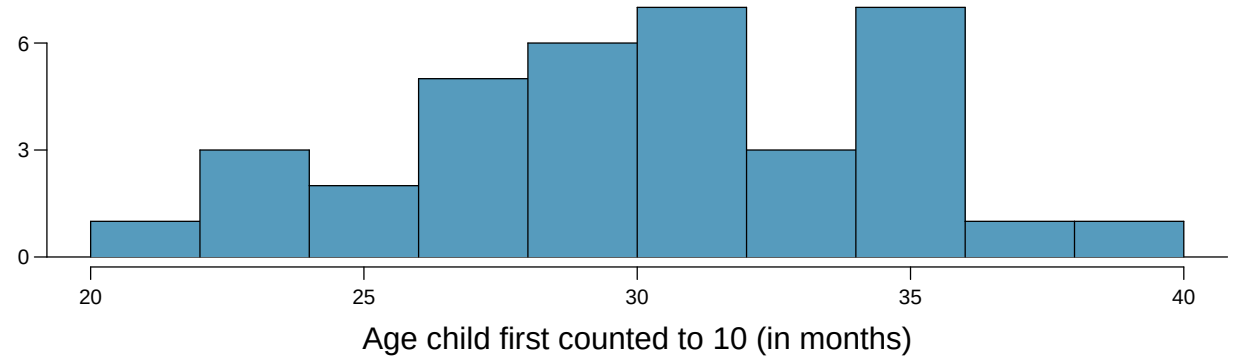
```
## [1] 1.468319
```

False, to reduce the interval to a third of what it is now, we would need to increase the sample size by 3^2 since the sample size is square rooted.

(g) The margin of error is 4.4.

False. This is true only for a 95% confidence interval

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? yes since the this was a random sample, it satisfies the independence test although id does not seem to satisfy the sucess-failure with np>10

(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.

```
population_mean = 32
paste("sample Mean")
```

```
## [1] "sample Mean"
```

```
(sample_mean = mean(gifted$count))
```

```
## [1] 30.69444
```

```
paste("Sample Standard Deviation")
```

```
## [1] "Sample Standard Deviation"
```

```
(sample_sd = sd(gifted$count))
```

```
## [1] 4.314887
```

```
paste("Sample Error")
```

```
## [1] "Sample Error"
```

```r
(sample_error = qnorm(.975)*sample_sd/sqrt(length(gifted$count)))
```

```
## [1] 1.409504
```

```r
paste("The upper limit = ", round(sample_mean+sample_error,1))
```

```
## [1] "The upper limit =  32.1"
```

```r
paste("The lower limit = ", round(sample_mean-sample_error,1))
```

```
## [1] "The lower limit =  29.3"
```

With this, we must reject the null hypothesis that our sample mean is 32 months since our 32 falls within our confidence intervals.

(c) Interpret the p-value in context of the hypothesis test and the data.

```r
z <- (30.69-sample_mean)/sample_sd

pnorm(z)*2
```

```
## [1] 0.9991782
```

because or P value is so high, we do not have sufficient evidence to reject our null hypothesis.

(d) Calculate a 90% confidence interval for the average age at which gifted children first count to 10 successfully.

```r
population_mean = 32
paste("sample Mean")
```

```
## [1] "sample Mean"
```

```r
(sample_mean = mean(gifted$count))
```

```
## [1] 30.69444
```

```r
paste("Sample Standard Deviation")
```

```
## [1] "Sample Standard Deviation"
```

```r
(sample_sd = sd(gifted$count))
```

```
## [1] 4.314887
```

```r
paste("Sample Error")
```

```
## [1] "Sample Error"
```

```r
(sample_error = qnorm(.95)*sample_sd/sqrt(length(gifted$count)))
```

```
## [1] 1.182893
```

```r
paste("The upper limit = ", round(sample_mean+sample_error,1))
```

```
## [1] "The upper limit =  31.9"
```

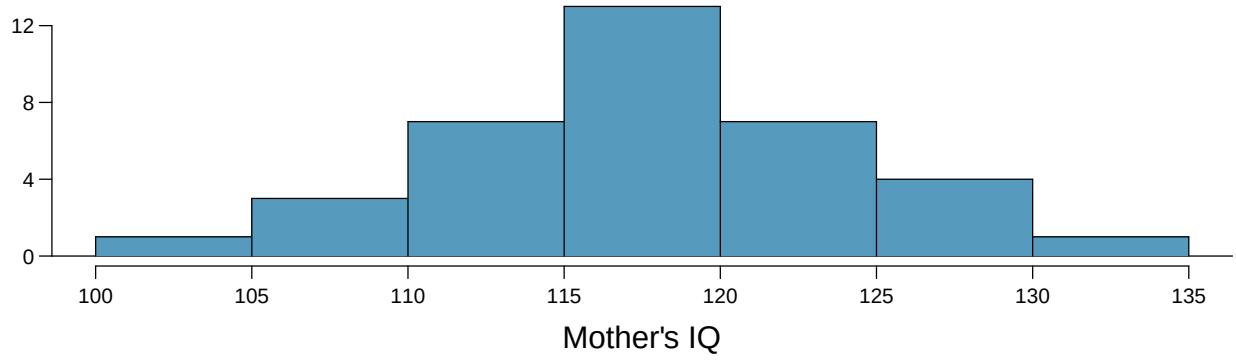
```r
paste("The lower limit = ", round(sample_mean-sample_error,1))
```

```
## [1] "The lower limit =  29.5"
```

(e) Do your results from the hypothesis test and the confidence interval agree? Explain.

yes, the results from the hypothesis test agree, we now see that population mean does not lie within the confidence interval of 10% of the sample mean.

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.

```
population_mean = 100
paste("sample Mean")
```

```
## [1] "sample Mean"
```

```
(sample_mean = mean(gifted$motheriq))
```

```
## [1] 118.1667
```

```
paste("Sample Standard Deviation")
```

```
## [1] "Sample Standard Deviation"
```

```
(sample_sd = sd(gifted$motheriq))
```

```
## [1] 6.504943
```

```
paste("Sample Error")
```

```
## [1] "Sample Error"
```

```r
(sample_error = qnorm(.975)*sample_sd/sqrt(length(gifted$motheriq)))
```

```
## [1] 2.124909
```

```r
paste("The upper limit = ", round(sample_mean+sample_error,1))
```

```
## [1] "The upper limit =  120.3"
```

```r
paste("The lower limit = ", round(sample_mean-sample_error,1))
```

```
## [1] "The lower limit =  116"
```

Based on the upper and lower limits not including the population mean of 100, we fail to reject the null hypothesis that the sample mean is not reflective of the population mean.

```r
cv = qnorm(pnorm(.95))

(pval = pnorm(cv, sample_mean, sample_sd))
```

```
## [1] 6.832766e-73
```

this p value suggest it is very unlikely for us to have a type II error when not rejecting the null hypothesis.

(b) Calculate a 90% confidence interval for the average IQ of mothers of gifted children.

```r
(sample_error = qnorm(.95)*sample_sd/sqrt(length(gifted$motheriq)))
```

```
## [1] 1.78328
```

```r
paste("The upper limit = ", round(sample_mean+sample_error,1))
```

```
## [1] "The upper limit =  119.9"
```

```r
paste("The lower limit = ", round(sample_mean-sample_error,1))
```

```
## [1] "The lower limit =  116.4"
```

(c) Do your results from the hypothesis test and the confidence interval agree? Explain.

yes my results from the hypothesis test and the interval agree. The sample mean is not reflective of the population mean and so we still reject our alternative hypothesis.

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.

sample distribution describes the meanss or summarized when taking many different samples from a population. as the number of samples increase the center of the distribution grows taller relative to the tails and the shape follows a normal distribution. and the spread decreases.

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?

```
1-pnorm(10500,9000,1000)
```

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

the probability is approximately 6.7%

(b) Describe the distribution of the mean lifespan of 15 light bulbs.

```
SE = 1.96*1000/sqrt(15)

9000+SE
```

```
## [1] 9506.07
```

```
9000-SE
```

```
## [1] 8493.93
```

The distribution of the mean lies between 8494 and 9506 lightbulbs @95% confidence interval.

(c) What is the probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10,500 hours?

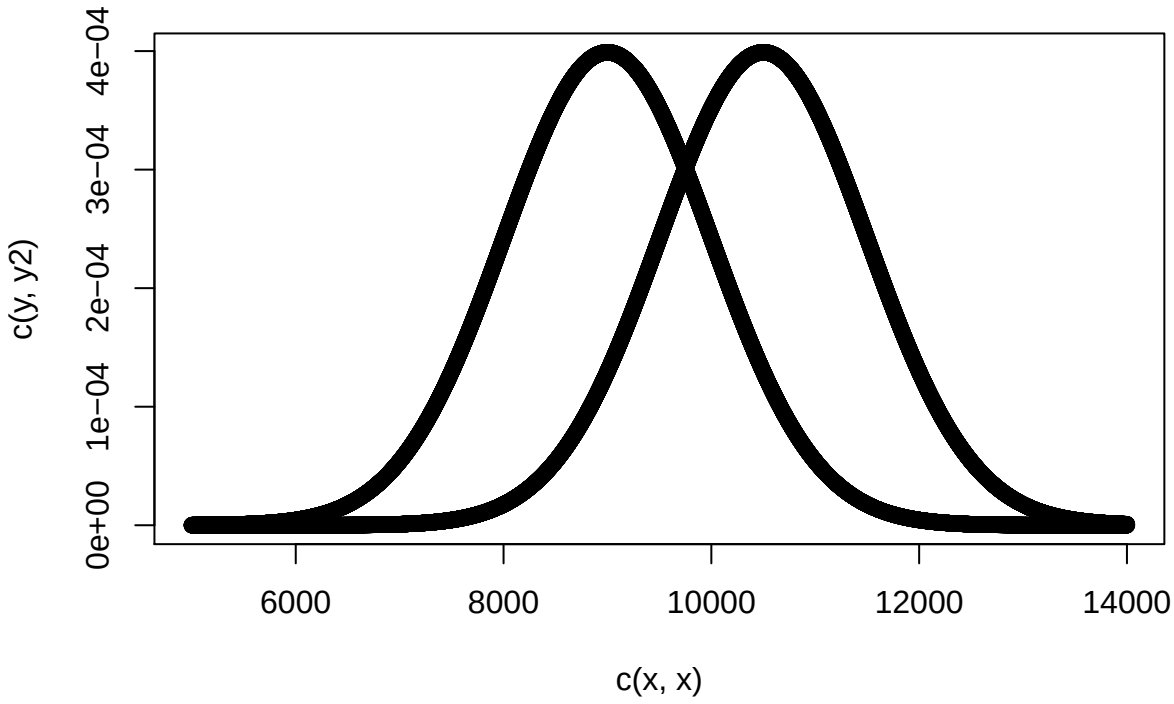
```
pnorm(10500,9000)*2
```

```
## [1] 2
```

the probability of this is approaches 0.

(d) Sketch the two distributions (population and sampling) on the same scale.

```
x <- seq(5000,14000, 1)

y <- dnorm(x, 9000,1000)

y2 <- dnorm(x, 10500,1000)
plot(c(x,x),c(y,y2))
```



(e) Could you estimate the probabilities from parts (a) and (c) if the lifespans of light bulbs had a skewed distribution?

The probability from part a would not be able to be estimated very accurately since we are measuring a points probability within the population. but part C can be estimated because of the central limit theorem when taking many samples from a population this will approximate to a normal distribution so estimating the mean with some confidence level will always be applicable.

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.

Your p value will decrease. Increasing n decreases the standard error. and decreasing the standard error narrows the prediction interval leading to lower significant values