# charleyferrari_week4hw

**4.4, Heights of Adults**

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 of active individuals? What about the median?

The point estimate of the average height would be the mean of the sample: 171.1. The median estimate would be 170.3.

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

The point estimate for the standard deviation is the sample standard deviation: SD = 9.4. The IQR would be the range between the 1st Quartile and the 3rd quartile, so the range between 163.8 and 177.8.

c. Is a person who is 180cm tall considered unusually tall? And is a person who is 155cm unusually short? Explain your reasoning.

We're given a sample IQR of 163.8 and 177.8. Based on these numbers alone, I'd say that 180cm and 155cm are unusual because they are outside of this range. Confidence intervals could perhaps be calculated for these IQRs (as they can for any point estimate, as shown in section 4.4) that might further clarify how unusual these measurements are.

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 wouldn't expect the mean and standard deviation to be exactly the same, because the sample will include different individuals. Samples can only provide an estimation of the true underlying dynamics of the population. We can talk about their distribution, but we cannot predict the exact values of a sample mean or standard deviation.

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 random sample. What measure do we use to quantify the variability of such an estimate (hint: recall that $SD_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$)

The variability of this estimate is the standard error, which is defined above as the population standard deviation over radical n. If the sample size is large enough, the standard deviation of the sample can be assumed to be the same as the population. For smaller sample sizes, this cannot be calculated.

**4.14: Thanksgiving Spending, Part I.**

The 2009 Holiday retail season, which kicked off on November 27, 2009, had been marked by somewhat lower self-reported consumer spending then was previously 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.

False, we are 100% confident of the average spending of these 436 Americans. The confidence interval corresponds to our guess of the underlying population based on this sample.

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

Unclear. We'd have to do more sophisticated testing to see if this data is skewed enough to produce unreliable results according to the CLT.

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

False. This confidence interval corresponds to a guess as to where the population mean lies. We can't speak about how far the particular sample we collected is away from the population mean, which is what this sort of question would depend on.

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

True, this is the correct way to use a confidence interval.

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.

True, as $\alpha$ goes up, the interval length narrows, because we can be less sure that the mean is in the confidence 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.

The margin of error MOE is defined as:

$$MOE = Z \times SE$$

And the standard error is defined as:

$$\frac{\sigma_{sample}}{\sqrt{n}}$$

So, if n increases by 3 times, the standard error, and thus the MOE, would only decrease by a factor of $\sqrt{3}$

g. The margin of error is 4.4

Since the margin of error is defined as:

$$MOE = z \times SE$$

And the confidence interval is:

$$\bar{x} \pm z \times SE$$

We can use this information to solve for $z \times SE$

$$\bar{x} - z \times SE = CI_{lower}$$

$$84.71 - 1.96 \times SE = 80.31$$

This does indeed solve for $z \times SE = 4.4$. This statement is true.

### 4.24: 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 as soon as they rerached the age of four. The following histogram shows the distribution of the ages (in months) at which these children first counted to 10 successfully. Also provided are some sample statistics.

a. Are conditions for inference satisfied?

It would appear so based on the sample size. We would have to assume that the data isn't too skewed (it doesn't look like it from the histogram) and that the children were chosen randomly.

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 first count to 10 successfully is less than the general average of 32 months. Use a significance level of 0.10.

First lets design our experiment:

$$H_o : \mu = 32$$
$$H_a : \mu \neq 32$$

Lets calculate the p-value of reading at 32 months. We have a sample mean of 30.69, and we're looking for what the z-score of 32 would be.

```
xbar <- 30.69

sd <- 4.31

n <- 36

se <- sd/sqrt(n)

z <- (32 - xbar)/se

z
```

```
## [1] 1.823666
```

This is a two tailed test, so to get our p-value, we're going to calculate the area to the right of 1.82 and to the left of -1.82. I'll use the pnorm function to calculate this.

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

pvalue
```

```
## [1] 0.0682026
```

This is below our $\alpha$ of 0.1, so we're rejecting the null hypothesis.

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

This p-value represents the area under the tails of a normal distribution with the mean and standard error calculated for the sample. Given a hypothesis, in this case, the article that states that children start reading at 32 months, we can use this p-value to test out whether or not to reject it. In this case, the area under the tails was 0.07. 32 is not close enough to our sample mean to justify the null hypothesis.

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

```
interval <- c(xbar - z*se, xbar + z*se)

interval
```

```
## [1] 29.38 32.00
```

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

### 4.26: Gifted children, Part II

Exercise 4.24 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.

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 significance level 0.1.

$$H_o : \mu = 100$$
$$H_a : \mu \neq 100$$

Lets calculate our confidence interval. First, the z-score. This is a two-tailed test, so we want 10% under the tails, and 5% under each tail.

```
z <- qnorm(0.95)
xbar <- 118.2
sd <- 6.5
n <- 36
se <- sd/sqrt(n)
interval <- c(xbar-z*se, xbar+z*se)
interval
```

```
## [1] 116.4181 119.9819
```

We can say with 90% confidence that the average IQ of gifted mothers is higher than the average IQ of the general population. More specifically, we're saying that with 90% confidence we can say the average IQ of gifted mothers is not 100.

We can also solve this using p-values. Lets calculate a new z-score, which represents the z-score of the null hypothesis, z-score of 100:

```
z <- (100-xbar)/se
```

and then we can calculate our p-value:

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

pvalue
```

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

The p-value is very small, which matches our conclusion before, of the confidence interval being pretty far away from the mean of the total population.

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

This was done above!

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

They do agree, they are two ways of arriving at the same answer. You can tell this because the p-value is very small, and the confidence interval around the population mean is far from the population mean given.

### 4.34: 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.

Sampling distribution is the distribution of several $\bar{x}$'s with different sample sizes (n). For small sample sizes, the distribution of repeated $\bar{x}$'s are less normal, and may vary based on the population distribution. When the sample size is increased, the distribution of the sample means become more and more normal, no matter what the underlying distribution is.

### 4.40: CFLBs

A manufacturer of compace fluorescent light bulbss advertises that the distribution of the lifespans of these light bulbs is nearly normal with a mean of 9000 hours and a standard deviation of 1000 hours.

a. What is the probability that a randomly chosen light bulb lasts more than 10500 hours?

Here, we need to find the area of the normal curve centered around 9000 with an SD of 1000, to the right of 10500. This is 1 - pnorm(10500)

```
1 - pnorm(10500,mean=9000,sd=1000)
```

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

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

The mean of the distribution of mean lifespans will be the population mean of 9000. Calculating the standard error as $\frac{\sigma}{\sqrt{n}}$ usually works with the sample standard deviation if the sample size is above 30. In this case, we have the reported population standard deviation, so we'll use that. The set of means will be distributed normally with mean 9000 and sd as the SE:

```
se <- 1000/sqrt(15)
se
```

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

c. What is the probability that the mean lifespan of 15 randomly chosen light bulbs is more than 10500 hours?

We essentially have a population of normally distributed data, with a mean of 9000 and a standard deviation of 258. So, we can just find the area under the curve to the right of 10500 hours.

```
1-pnorm(10500,mean=9000,sd=258)
```

```
## [1] 3.050719e-09
```

In this case, the chance of getting a mean of 10500 is very small.

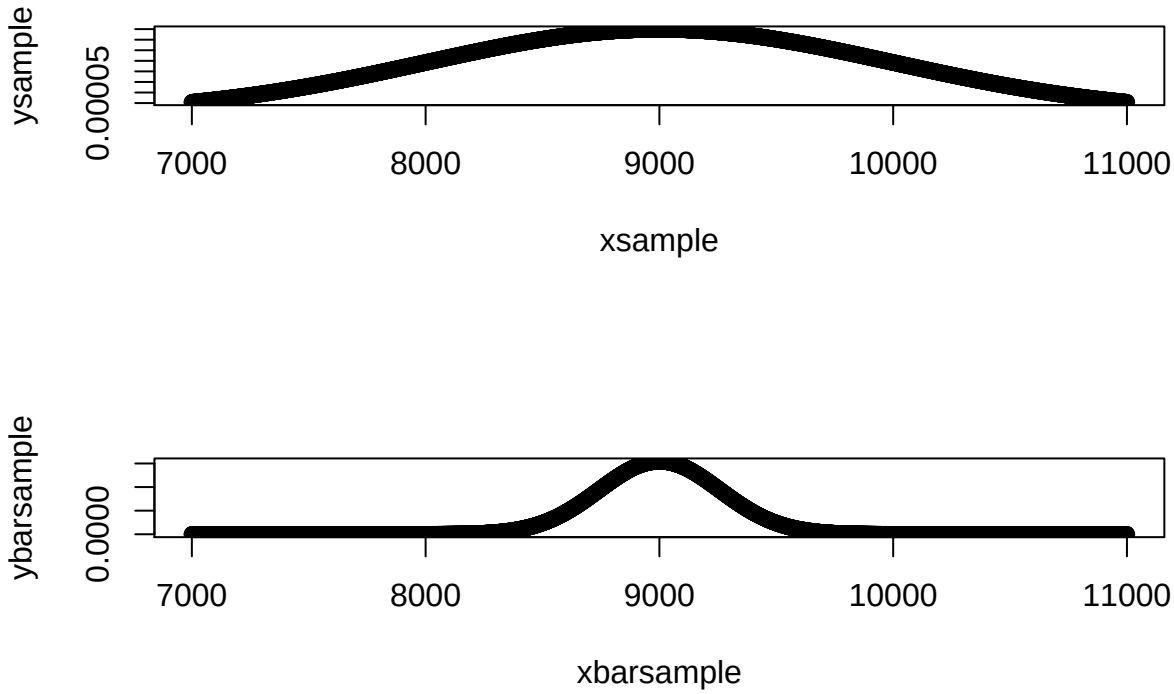
d. Sketch the two distributions (population and sampling) on the same scale.

```
par(mfrow = c(2, 1))

xsample <- 7000:11000
ysample <- dnorm(xsample,mean=9000,sd=1000)

xbarsample <- 7000:11000
ybarsample <- dnorm(xbarsample,mean=9000,sd=se)

plot(xsample,ysample)
plot(xbarsample, ybarsample)
```




The top graph is the sample distribution, while the bottom graph is the distribution of $\bar{x}$ for sample sizes of 15.

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

If the underlying distribution of the population was known, and still skewed, a would be solveable by calculating the probabilities of the given distribution. This would not be possible for part c, since it depends on the central limit theorem, which depends on the

**4.48: 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.

Calculating the relevant z-score of the p-value of a variable x can be done below:

$$z = \frac{relevantmeasure}{SE}$$

With SE defined as:

$$SE = \frac{\sigma}{\sqrt{n}}$$

so, as n increases, SE decreases, and the z-score increases. If the z-score (or absolute value of the z-score) increases, the p-value decreases. With a higher z-score, the p-value will represent less of the tail area.

**4.13: Waiting at an ER, Part I.**

A hospital administrator hoping to improve wait times decides to estimate the average emergency room waiting time at her hospital. She collects a random sample of 64 patients and determines the time (in minutes) between when they checked in to the ER until they were first seen by a doctor. A 95% confidence interval based on this sample is (128 minutes, 147 minutes), which is based on the normal model for the mean. Determine whether the following statements are true or false, and explain your reasoning.

a. This confidence interval is not valid since we do not know if the population distribution of tthe ER wait times is nearly normal.

False. As long as the data isn't too heavily skewed, the point estimates of any distribution will be normally distributed according to the central limit theorem.

b. We are 95% confident that the average waiting time of these 64 emergency room patients is between 128 and 147 minutes?

False. We're 100% certain of the mean since we calculated it for this sample of 64 emergency room patients.

c. We are 95% confident that the average waiting time of all patients at this hospital's ER is between 128 and 147 minutes.

True. This is what the confidence interval is used for. Using our sample, we're estimating the underlying statistics of the total population.

d. 95% of random samples havea sample mean between 128 and 147 minutes.

False. This is a range around a sample mean, and the confidence interval is a measure of whether or not the population mean is within this range. Because this method doesn't show us exactly how far away this is from the population mean, we cannot know how different sample means relate to each other.

e. A 99% confidence interval would be narrower than the 95% confidence interval since we need to be more sure of our estimate.

False. THe 99% confidence interval is actually wider. Because we need to be more sure of our answer, we make our interval wider to allow more variance.

f. The margin of error is 9.5 and the sample mean is 137.5.

The center of our interval is the sample mean, so lets calculate that first:

```
samplemean <- mean(c(128,147))

samplemean
```

```
## [1] 137.5
```

So,our sample mean is indeed 137.5.

The margin of error MOE is defined as:

$$MOE = Z \times SE$$

We can solve for this equation using the CI interval already calculated

$$CI_{lower} = \bar{x} - 1.96 \times SE$$
$$128 = 137.5 - 1.96 \times SE$$

Solving for MOE, we have $1.96 \times SE = 9.5$ Both of these statements are correct.

g. In order to decrease the margin of error of a 95% confidence interval to half of what it is now, we would need to double the sample size

False.

$$MOE = Z \times SE$$

This can b e restated as:

$$MOE = Z \times \frac{\sigma_{population}}{\sqrt{n}}$$

If we wanted halve the MOE, doubling the sample size would only decrease it by a factor of $\sqrt{n}$.

### 4.23: Nutrition Labels

The nutrition label on a bag of potato chips says that a one ounce serving of potato chips has 130 calories and contains ten grams of fat, with three grams of saturated fat. A random sample ofo 35 bags yielded a sample mean of 134 calories with a standard deviation of 17 calories. Is there evidence that the nutrition label does n ot provide an accurate measure of calories in the bags of potato chips? We have verified the independence, sample size, and skew conditions are satisfied.

$H_o$: $\mu = 130$ $H_a$: $\mu \neq 130$

Lets find out a 95% confidence interval for our sample mean.

$$CI_{95} = \bar{x} \pm Z \times SE$$

```
se <- 17/sqrt(35)
interval <- c(134-1.96*se, 134+1.96*se)

interval
```

```
## [1] 128.3679 139.6321
```

Our 95% confidence interval includes 130, so we don't have enough evidence to reject the null hypothesis.

### 4.25: Waiting at an ER, Part III

The hospital administrator mentioned in 4.13 randomly selected 64 patients and measured the time (in minutes) between when they checked into the ER and the time they were first seen by a doctor. The average time is 137.5 minutes and the standard deviation is 39 minutes. She is getting grief from her supervisor on the basis that the wait times in the ER has increased greatly from last year's average of 127 minutes. However, she claims that the increase is probably just due to chance.

a. Are conditions for infrerence met? Note any assumptions you must make to proceed.

The only condition we know is met is the sample size. It is indeed greater than 30. We would also have to assume that the sample is randomly chosen and that the underlying distribution of wait times is not too skewed.

b. Using a significance level of $\alpha = 0.05$, is the change in wait times statistically significant? Use a two-sided test since it seems the supervisor had to inspect the data before she suggested an increase occurred.

We'll use z = 1.96 for this once again.

$H_o$: $\mu = 127$ $H_a$: $\mu \neq 127$

```
z <- 1.96

se <- 39/sqrt(64)

interval <- c(137.5-z*se, 137.5+z*se)

interval
```

```
## [1] 127.945 147.055
```

Looks like we have just enough evidence to reject the null hypothesis. The mean is between 127.95 and 147.055 with 95% certainty. Assuming the last mean was closer to 127.0, evidence would suggest mean wait times have increased.

c. Would the conclusion of the hypothesis test change if the significance level was changed to $\alpha = 0.01$?

Probably, but lets check. First, lets calculate our z-score.

This is a two-tailed test, so we want 1% underneath both tails, and .05% underneath either tail. Lets use this information to calculate our z-score, then our intervals using the same method above.

```
z <- qnorm(1-0.005)

se <- 39/sqrt(64)

interval <- c(137.5-z*se, 137.5+z*se)

interval
```

```
## [1] 124.9428 150.0572
```

We are 99% confident that the mean falls between 124.94 and 150.06. With this chosen confidence interval, we can now say that it's possible that the difference was due to chance.