Objectives

5.2, 8.1 Inference for a single proportion

- Categorical data from a simple random sample
- Binomial distribution
- Sampling distribution of the sample proportion
- Significance test for a single proportion
- Large-sample confidence interval for $p$
- Choosing a sample size
Categorical Data

So far we have focused on variables whose outcomes are numbers:
- i.e. The height or weight of people/calves
- There are many important situations where the outcome of a variable is categorical (where attaching a number has no real meaning or significance).
- i.e. The car that someone owns, the ethnicity of a person, is a person’s favorite tv channel MTV

We are often interested in the chances associated to each of these outcomes:
- 20% of the people own a Honda, 20% percent own a Ford, 25% own a Toyota and other 35% own something else or no car at all.
- 30% of people classify themselves as ‘white’, 10% of African decent, 20% Hispanic, 5% Asia, 5% mixed and the other 30% classify themselves as other.
- 30% of ‘young’ people between 13-20 said their favorite tv channel was MTV while the other 70% preferred other channels.
Binary Variables

In this section we will focus on categorical variables which have only one of two possible outcomes. Often one outcome is classified as a `success’ or a `failure’ (no meaning should be attached to these names).

- Whether a young person’s favorite channel is MTV (yes or no)
- The birth gender of a statistics student (typically, male or female).
- If a parent is placing pressure on their child (yes or no)

These are examples of **binary** variables.

Typically a survey would be done and several people asked the same question where the response is binary.

- The individual responses in a survey are usually of little interest.
- Normally we want to know the number of people who say `yes’ out of the sample, or equivalently the proportion of the sample who say `yes’.
  - Example: 400 college students were randomly sampled and asked whether their coursework was too heavy. 300 students responded that it was, or, equivalently, 75% of the sample said it was too heavy.

In the first part of the chapter we discuss what sort of **distribution** this data comes from and then how we can use it to do statistical inference.
The binomial distribution explained through genetics

- The loci are locations on a chromosome. For each loci there are two alleles (in layman terms this is a gene). Typically one allele is inherited from each parent. Alleles on different loci can determine different features on an organism.

- Let us consider the following simplified example taken from Drs. Ellison and Reynolds GENE301 notes.
  - Two loci on a chromosome are known to determine the length of a round worm
    - One lab produced a `pure breed’ recessive worm where the relevant alleles aabb and was length 0.8mm (little letters denote recessive).
    - Another lab produced a `pure breed’ dominant worm where the relevant alleles are AABB and was length 1.2mm (capitols denote dominant).
    - These two worms were cross-bred. Since the off-spring gets one allele from each parent (on each loci) it is clear that the off-spring can only have alleles AaBb

<table>
<thead>
<tr>
<th></th>
<th>Dominant/Long Worm</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recessive/short worm</td>
<td>Off spring</td>
</tr>
<tr>
<td>aa</td>
<td>bb</td>
</tr>
</tbody>
</table>
Whole load of combinations.
- It is known that the number of dominant alleles (those with capitol letters) determines the length.
- Length of Worm = 0.8 + 0.1 × Number of dominant alleles.
- The off-springs are cross-bred with each other. In this case there are 16 possible outcomes which lead to different lengths:

<table>
<thead>
<tr>
<th>Worm Length</th>
<th>aabb</th>
<th>aaBb</th>
<th>AabB</th>
<th>aAbB</th>
<th>AAbb</th>
<th>AaBb</th>
<th>aaBB</th>
<th>AAbB</th>
<th>AaBB</th>
<th>AAbb</th>
<th>AABb</th>
<th>AABB</th>
</tr>
</thead>
<tbody>
<tr>
<td>likelihood</td>
<td>0.8</td>
<td>0.9</td>
<td>1.0</td>
<td>1.1</td>
<td>1.2</td>
<td>1.3</td>
<td>1.4</td>
<td>1.5</td>
<td>1.6</td>
<td>1.7</td>
<td>1.8</td>
<td>1.9</td>
</tr>
<tr>
<td>frequency</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
</tr>
</tbody>
</table>

\[
P(X \leq 2) = \frac{16}{16} = 1
\]
Observe each of the possible outcomes are equally likely. Further, the probability of the worm being of a certain length is the same as the probability of that many dominant alleles in the loci.

The distribution, which is the curve given on the previous page is a **Binomial distribution** with \( n=4 \) (maximum number of dominant alleles) and probability \( 1/2 \) – since the probability/frequency of the dominant allele A and B is the same as the probability/frequency of the recessive allele a and b (which is \( 1/2 \)). For short hand we often say the frequency of dominant alleles follows a \( \text{Bin}(4,1/2) \).

So far we have only considered the simple example where we start with breeding AaBb with AaBb – in this case all outcomes are equal.

Let us suppose that the proportion of a dominant allele in the general population is \( p \). eg.

- The proportion of round worms in the population with dominant allele A is \( p=0.8 \).
- Thus the proportion of round worm with recessive allele a has to be 0.2.
- The proportion of round worm in the population with dominant allele B is \( p=0.8 \).
- Thus the proportion of round worm with recessive allele a has to be 0.2.
In this case the same outcomes are possible, just the chance of them occurring is **no longer equal**.

If the chance of a dominant allele is 0.8, then the number of dominant alleles out of 4 follows a binomial distribution with $n=4$ and $p=0.8$, in shorthand we write this as $\text{Bin}(4,0.8)$. The Binomial distribution allows us to determine the proportion of worms with a certain length.

<table>
<thead>
<tr>
<th>worm length likelihood</th>
<th>aabb</th>
<th>AabB</th>
<th>aAbb</th>
<th>AAbb</th>
<th>AABB</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2^4</td>
<td>0.8</td>
<td>0.9</td>
<td>1.0</td>
<td>1.1</td>
<td>1.2</td>
</tr>
<tr>
<td>4 \times 0.2^3 \times 0.8</td>
<td>6 \times 0.2^2 \times 0.8^2</td>
<td>4 \times 0.8^3 \times 0.2</td>
<td>0.8^2</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The above examples illustrate how the binomial distribution is a very useful tool to use in genetics. Before we consider other examples of where the binomial distribution is used, let us briefly discuss the one consequence of the previous example (this will not be examined!).

- It was mentioned that the proportion of an allele in a population is $p$. You may wonder whether in the next generation this proportion stays the same or decreases or increases. In other words will this proportion change over time, will the allele be wiped out?

- Under certain conditions it can be shown that the proportion will be unchanged and will remain $p$ over the generations. This is known as **Hardy-Weinberg equilibrium**.
Topics: The binomial distribution

- Understand the definition of the binomial distribution.
- Interpret the plot of the binomial distribution in terms of chance.
- Be able to construct relevant hypotheses using probabilities.
- Be able to obtain the p-value of a test using the binomial distribution.
- Be able to obtain the p-value of a test using the Statcrunch output.
  - Understand where the standard error for the test comes from.
- Be able to construct a confidence interval using the Statcrunch output.
- Be able to calculate a sample size based on the given margin of error.
Other applications of the Binomial distribution

- **Example 1**: Suppose the proportion of the general public who support gun control is 60%. A random sample of 50 people is taken. We would not expect that exactly 60% of these 50 people to support gun control. This number will vary (it is random). The frequency of these numbers varies according to a Binomial distribution with $n=50$ and $p = 0.6$, i.e. $\text{Bin}(50,0.6)$.

- **Example 2**: Suppose I guessed every question in Midterm 2. There is a 20% chance of my getting a question right. The likelihood of my grade follows Binomial distribution with $n=15$ and $p = 0.2$ i.e. $\text{Bin}(15,0.2)$. 
Applying the binomial to testing hypothesis

- We now show how the Binomial distribution can be used in the context of hypothesis testing.
  - First, we will focus on one sided tests.
    - You will need to know from the context, what the hypothesis of interest is and how to use the correct binomial distribution to do the test.
  - The binomial distribution will give the exact p-values – no normal approximation necessary (compare this with all the sample mean stuff we did in the previous chapters where it was important to determine whether the p-values were reliable).
- However, the normal distribution can also be used to do the test (in this case the p-values are an approximation).
  - This will be one of the few times we actually see how well the normal approximations compare with the true p-values.
An advantage of the normal approximation is that it can be used to construct confidence intervals for the population proportion, which cannot be done using the binomial distribution.

You will observe that this is the one time that the standard errors for the test and the confidence interval will be different. This will be explained later in this chapter.
Example 1: Guessing midterms?

- Suppose a midterm was multiple choice, with five different options for each question and 15 questions.
- Mike scores 5 out of 15, did he know some of the material or was he guessing? We can articulate this as a hypothesis test and use the plot on the previous slide to answer the question.
- Let $p$ denote the probability of getting an answer correct. If $p = 0.2$ then this means the answer is guessed, if $p > 0.2$ then it means some knowledge has been put into answering the question. The test can be written as $H_0: p \leq 0.2$ against $H_A: p > 0.2$.
- Under the null hypothesis he is simply guessing. If this is true, we recall from a previous slide that this means his score follows a Bin(15,0.2).
- To calculate the p-value we look at the Bin(15,0.2) and calculate the chance of scoring 5 or more if this is the correct distribution.
  - Software calculate this for us, just look at the plot of the next slide.
  - **Why 5 or more?:** The alternative hypothesis is pointing to the RIGHT so we calculate the probability to the RIGHT of what is observed (in this case 5). We should include 5 in the calculation.
To calculate the p-value we look at the Bin(15,0.2) and calculate the chance of **scoring 5 or more** if this is the correct distribution.

- Software calculate this for us, just look at the plot of the next slide.
- **Why 5 or more?** The alternative hypothesis is pointing to the RIGHT so we calculate the probability to the RIGHT of what is observed (in this case 5). We should include 5 in the calculation.
We test $H_0: p \leq 0.2$ against $H_A: p > 0.2$.

The software shows that the chance of scoring 5 or more by just guessing the answer is 16.4%. 16.4% is the p-value in this test.

As 16.4% is large (certainly larger than the 5% significance level), we cannot reject the null. There is no evidence that Mike really knew the answer.

In other words, he could have easily have scored 5 out of 15 by simply guessing.

We do not know whether he was guessing or not - may be knew the answer to those 5 questions and not the others. This is why we cannot accept the null!
Example 2: More exams!

- Another multiple choice exam, again with each question having a choice of 5. There are 100 questions in this exam.
- Rick scores 33 out of 100 (in this example, like the last, the person in question gets one third of the questions correct). Is there any evidence to suggest that he knew some of the material?
- Again we are testing $H_0: p \leq 0.2$ against $H_A: p > 0.2$
- The p-value is the probability of scoring 33 or more when he is simply guessing.
We test $H_0: p \leq 0.2$ against $H_A: p > 0.2$ and

We see from the plot that the p-value is the probabilities to the RIGHT of 33 (including 33).
The p-value $= 0.155\%$. This means that on average if 1000 students took this exam and all were guessing about 1.5 of them would score 33 points or more. As this is extremely rare we reject the null and conclude there is evidence that Rick new at least some of the material.

Though we observe that it can happen.
We mentioned previously, we can do the test using the normal distribution. This is done in Statcrunch (Stat -> Proportion Stat -> One sample -> With Summary (put 33 as the number of successes))

Hypothesis test results:
p : proportion of successes for population
H_0 : p = 0.2
H_A : p > 0.2

<table>
<thead>
<tr>
<th>Proportion</th>
<th>Count</th>
<th>Total</th>
<th>Sample Prop.</th>
<th>Std. Err.</th>
<th>Z-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>33</td>
<td>100</td>
<td>0.33</td>
<td>0.04</td>
<td>3.25</td>
<td>0.0006</td>
</tr>
</tbody>
</table>

\[ \text{sample prop} = \hat{p} = 0.33 \]

The p-value given in the output is 0.06%. The p-value using the Binomial distribution is 0.13%. The discrepancy between p-values is due to the normal approximation of the binomial distribution.
There is a difference between the exact p-value of 0.15% and the approximate p-value 0.06%.
However, both are very small and we would reject the null, and determine Rick had some knowledge regardless of the method used.
The binomial morphs into a normal for large \( n \)

We now look at the distribution of grades when there is 50% chance of getting a question correct by guessing and the number of questions in the paper is 100. The plot does do quite normal.

The second plot gives the distribution of grades when there is a 20% chance of getting a question correct by guessing and there are 100 questions in the paper. Despite the slight right skew the plot does look quite normal.

How close the binomial is to the normal depends on:

1. The size of the sample (in this case 100)
2. How close \( p \) is to 0.5. The closer to 0.5 the less skewed and the more normal.
The distribution of the sample proportion

- Here we give the distribution of the sample proportion when:
  - $n = 10$ and $p = 0.5$
  - We see that it is not close to normal.

- Here we give the distribution of the sample proportion when:
  - $n = 1000$ and $p = 0.5$
  - We see that it is very close to normal.
Normal approximation

- We estimate the sample proportion with

\[
\hat{p} \text{ (proportion estimate)} = \frac{\text{number of success}}{\text{sample size}}
\]

- The variability of the estimate is the standard error:

\[
\text{standard error} = \sqrt{\frac{p(1 - p)}{n}}
\]

- This closely resembles the standard error for the sample mean!

- And of course, for large sample sizes the distribution closely resembles a normal distribution (look at the previous slides). Approximately

\[
\hat{p} \sim N(p, \sqrt{\frac{p(1 - p)}{n}})
\]

- We will show how close the probabilities are using the binomial distribution and the normal in the next slide.
Example 3: Even more exams

- Suppose a student scores 63% in a true or false exam with 100 questions. Let $p$ denote the probability of getting a question correct.

- We want to test $H_0: p \leq 0.5$ against $H_A: p > 0.5$.

  - 63% is the same as scoring 63 out of 100 in the exam. To test the above hypothesis we see how likely it is to score 63 or MORE (remember the alternative is pointing RIGHT) by simply guessing.

  - We see that the chance of scoring 63 or more is 0.6%.
  
  This means if 1000 students took the exam and just guessed, about 6 of them would score 63 or over.

  As this is tiny and less than the 5% level, there is evidence that the student was doing more than just random guessing.
We now analyse the score Statcrunch. To summarize: Statistics -> Proportion -> One Sample -> Summary.

Observe that the p-value using the normal approximation is 0.47% which is close to the true p-value of 0.6%. They are very close because:

1. The sample size $n = 100$ is quite large.

2. $p=0.5$ which means the distribution is not all skewed (the CLT kicks in quite fast)
- We do the calculations under the null (that the student was randomly guessing).

- The formula for the standard error is

\[
s.e. = \sqrt{\frac{p(1 - p)}{n}}
\]

- \( n = 100 \) and the under the null \( p = 0.5 \). Therefore the standard error is

\[
s.e. = \sqrt{\frac{0.5 \times 0.5}{100}} = 0.05
\]

- Observe this matches the standard error in the Statcrunch output.
The z-transform is

\[ z = \frac{\hat{p} - p}{s.e} = \frac{0.63(\text{Sample prop}) - 0.5(\text{null})}{0.05(\text{std error})} = 2.6 \]

Using the normal distribution, the area to the right of 2.6 is 0.47%

- The true probability 0.6% and its approximation 0.47% are not the same, but they are very close.

- One confusing aspect is that we do not use the t-distribution. I will try to remind you of this, I don’t want you to be confused by such trivialities.
Example 4: Polls on `gay marriage’

- A recent Gallup poll found that 64.2% (90 people) of sample of 140 individuals were `pro-gay marriage’.
- These are some news headlines. Are the following headlines accurate?
  - The New York Times: Survey suggests that the majority of the general public now support Gay marriage.
  - The Wall Street Journal: Poll suggest that over 60% of the population support Gay marriage.
- Are these reports accurate?
- To answer this question we write them as a hypothesis tests and do the test.
  - We will focus on using the Statcrunch output (using the normal distribution) and later compare the approximate p-values to the true ones.
Example 4a: The New York times

Majority means a proportion is over 50%. This means we are testing $H_0 : p \leq 0.5$ against $H_A : p > 0.5$.

The normal calculation

The standard error $= \sqrt{(0.5 \times 0.5/140)} = 0.04$.

The z-transform $= (0.64-0.5)/0.4 = 3.38$

As this is a one-sided test pointing to the RIGHT we calculate the area to the right of 3.38. Looking up z-tables this gives 0.036%.

Therefore we can reject the 5% level, the survey suggest that the majority of the public do support Gay marriage.
The exact p-value is:

![Binomial Calculator](image)

\[ P(X \geq 90) = 0.00045656 \]

This is the exact p-value

![Normal Calculator](image)

\[ P(X \geq 0.64225714) = 0.00038069 \]

This is the normal approximation p-value.

Even using the exact p-value there is strong evidence to support the alternative that the majority support gay marriage.

Remember the p-value gives us information about the plausibility of the null given the data. The very small p-value tell us that is very difficult to get 90 people out of 140 saying they support gay marriage when the general public is divided 50:50.
Example 4b: The Wall Street Journal reporting

Over 60% support gay marriage. This means we are testing $H_0 : p \leq 0.6$ against $H_A : p > 0.6$.

The calculation

The standard error = $\sqrt{(0.6 \times 0.4/140)} = 0.041$

The z-transform = $(0.64 - 0.6)/0.041 = 1.03$.

Again looking up the area to the RIGHT of 1.03 in the z-tables gives the p-value = 15.15%. As this is greater than 5%, even though the proportion of the sample is greater than 60%, the evidence is not enough to suggest that the proportion of the public that support gay marriage is over 60% (the data is consistent with the null being true).

The Wall Street Journal had incorrectly assumed their sample was the entire population.
The exact p-value is about 17.15%, which again is consistent with the null. There isn’t any evidence to support the alternative that over 60% support Gay marriage.

In this case the p-value is telling us that the data is consistent with the null being true (it does not tell us the null is true). In other words, it is easily possible to obtain a sample where 90 out of 140 support Gay marriage when public opinion is divided 60:40 in favor of Gay marriage (this is the null).
Example 5: Is letter writing a lost art?

- NPR wanted to know whether less than 20% of the public have written a letter in the past 18 months. They randomly sampled 200 people and found that 25 of them had written a letter. Is there evidence to support their claim?

- The hypothesis of interest is $H_0 : p \geq 0.2$ against $H_A : p < 0.2$.

We see that both the true p-value (0.36%) and the approximate p-value (0.4%) back the alternative hypothesis; i.e. There is evidence that less than 20% of the population writes letters these days (sadly!).
Example 6: Changes in viewing habits

- Last year 60% of the general public watched TV for at least one hour a day. Have viewing habits changed this year? To see if there is any evidence of this 1000 people were sampled and 560 claimed watch TV for at least one a day. Does the data support the hypothesis?

- The hypothesis of interest is $H_0 : p = 0.6$ against $H_A : p \neq 0.6$.

<table>
<thead>
<tr>
<th>Proportion</th>
<th>Count</th>
<th>Total</th>
<th>Sample Prop.</th>
<th>Std. Err.</th>
<th>Z-Stat</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>560</td>
<td>1000</td>
<td>0.56</td>
<td>0.015491933</td>
<td>-2.5819889</td>
<td>0.0098</td>
</tr>
</tbody>
</table>

The p-value for the two sided test – which is the area to the LEFT of $-2.58$ times two is 0.98%.

This is pretty small and we reject the null at the 5% level and determine there has been a change in viewing habits.
Example 1: Confidence intervals and television habits

- Previously we focused on testing. However, we are often interested in understanding where the true proportion lies.
  - For example, given that out of 1000 people randomly surveyed 560 said they watched TV for more than an hour a day. Where does the true proportion lie?
  - Again we answer this question through Statcrunch.

<table>
<thead>
<tr>
<th>Proportion</th>
<th>Count</th>
<th>Total</th>
<th>Sample Prop.</th>
<th>Std. Err.</th>
<th>L. Limit</th>
<th>U. Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>p</td>
<td>560</td>
<td>1000</td>
<td>0.56</td>
<td>0.015697133</td>
<td>0.52923418</td>
<td>0.59076582</td>
</tr>
</tbody>
</table>

The proportion of people who watch more than an hour of TV is somewhere between $[0.529,0.59] = [52.9,59]%$ with 95% confidence.
Example 1: The calculation

95\% confidence interval results:

\( p \) : Proportion of successes
Method: Standard-Wald

<table>
<thead>
<tr>
<th>Proportion</th>
<th>Count</th>
<th>Total</th>
<th>Sample Prop.</th>
<th>Std. Err.</th>
<th>L. Limit</th>
<th>U. Limit</th>
</tr>
</thead>
<tbody>
<tr>
<td>( p )</td>
<td>560</td>
<td>1000</td>
<td>0.56</td>
<td>0.0156</td>
<td>0.529</td>
<td>0.590</td>
</tr>
</tbody>
</table>

Where does this calculation come from?

- The standard error is
  \[ s.e. = \sqrt{\frac{p(1 - p)}{n}} \]

- Since \( p \) is unknown we use the best guess
  \[ \hat{p} = \frac{560}{1000} = 0.56 \]

- This gives the standard error
  \[ \sqrt{\frac{0.56(1 - 0.56)}{1000}} = 0.0157 \]

- Using the normal distribution the 95\% confidence interval for the proportion is
  \[ [0.56 \pm 1.96 \times 0.0157] = [0.529, 0.590] \]
Example 2: Confidence intervals for Gay marriage

- Recall that the purpose of a confidence interval is to locate the true proportion. We now try to find where the opinion of the public on Gay Marriage lies.

<table>
<thead>
<tr>
<th>95% confidence interval results:</th>
</tr>
</thead>
<tbody>
<tr>
<td>p : Proportion of successes</td>
</tr>
<tr>
<td>Method: Standard-Wald</td>
</tr>
<tr>
<td>Proportion</td>
</tr>
<tr>
<td>p</td>
</tr>
</tbody>
</table>

The Calculation

The standard error (as are not testing) = \( \sqrt{0.642 \times 0.368/140} \) = 0.0405

We use the normal approximation, to give the 95% confidence interval

\[ 0.642 \pm 1.96 \times 0.0405 \] = [0.56, 0.72].

Thus based on the data, the proportion of the public who support gay marriage is between 56% and 72% (with 95% confidence). This interval will get narrower as we increase the sample size.
Standard errors for CI’s and p-values of proportions

- Again we note that unlike the case of sample means we cannot directly deduce p-values from confidence intervals. This is because the standard errors are different in both cases. For example in the Gay marriage example, where 90 out of a random sample of 140 supported Gay marriage.

- If we are testing \( H_0: p \leq 0.5 \) against \( H_A: p > 0.5 \) then the standard error is

  \[
  0.042 = \sqrt{\frac{0.5 \times 0.5}{140}}
  \]

  Since the p-value is done under the null being true, we need to calculate the standard error under the null being true too.

- On the other hand, if we wanted to construct a confidence interval for the proportion of the population who supported Gay marriage, we don’t have a clue what \( p \) could be and instead we use the estimate 0.642, and the standard error is

  \[
  0.04 = \sqrt{\frac{0.642 \times (1 - 0.642)}{140}}
  \]

These standard errors are different. Because one is constructed when we have given value of \( p \) and the other when we estimate it from data.
The standard error for proportions

- As in everything we have done and will do in the rest of this course, the reliability of an estimator is determined by its **standard error**.
- The standard error for sample proportions resembles (in fact can be considered the same) as the standard error for the sample mean:

  \[ s.e. = \sqrt{\frac{p(1-p)}{n}} \]

- Just like the sample mean, you have no control over the numerator \( p(1-p) \), this is analogous to \( \sigma \) in the standard error of the sample mean.
  - We show on the next slide that \( p(1-p) \) will be largest when \( p = 0.5 \) and smallest when \( p \) is close to zero or one.
- But you can control for the sample size. The larger the sample size, the smaller the standard error.
- The above observations allow us to choose the sample size to ensure the margin of error has a certain length (see the next few slide).
Newspaper reporting

- A newspaper would report the results of the “gay marriage” survey as follows.

- A recent survey suggests that between [56.2,72.2]% of the general public now support gay marriage (based on a 95% confidence interval).

- From this interval we can immediately deduce:
  - The proportion who supported gay marriage in the sample was 64.2% (since it is half of the interval).
  - The Margin of Error is (64.2-56.2)% = 8%
  - The standard error is 8/1.96 = 4.08%
  - The sample size is the solution of
    \[
    0.0408 = \sqrt{\frac{0.642 \times (1 - 0.642)}{n}}
    \]
    Which is \( n = 140 \).

- Thus the confidence interval gives us all information about the sample collected.

- Observe that a MoE of 8% is very large and not that informative.
Here we show that if the sample size is fixed (just set $n = 1$, for simplicity) the standard error will be largest when $p = 0.5$. 

![Plot of $\sqrt{\frac{p(1 - p)}{500}}$]
Plot shows that the estimated proportion is more variable if the true proportion is close to 0.5 but the variability reduces as the proportion decreases.

**Explanation why** If you always obtain close to a 100% in a exam, $p = 1$ and there isn’t any variability from exam to exam.

If you are an “average” student getting on average about 70%, your exam result from exam to exam will be more variable.

This observation will be useful when selecting a sample size according to a pre-specified margin of error.
Sample size for a desired margin of error

The general formula for the 95% confidence interval for the population proportion is

\[
\hat{p} \pm 1.96 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}, \quad \hat{p} + 1.96 \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}
\]

Thus the margin of error for the 95% confidence interval for where the true proportion will lie is half of the length of the above interval which is

\[
MoE = 1.96 \times \sqrt{\frac{p(1 - p)}{n}}
\]

We use the this formula when designing an experiment. In particular, to choose a sample size which will achieve a pre-specified maximum margin of error.
Suppose a polling company is estimating the proportion of people who will vote for a candidate. They are interested in constructing a confidence interval for the proportion of the population who would vote for a certain candidate. They choose the sample size such that it has a desired margin of error (it cannot be larger than the desired amount).

Solving the above equation in terms of \( n \):

\[
M_oE = 1.96 \times \sqrt{\frac{p(1-p)}{n}}
\]

\[
n = \left( \frac{1.96}{M_oE} \right)^2 \times p(1-p)
\]

However, we encounter a problem. Before we conduct the experiment we do not know the proportion \( p \), making it impossible to solve the above equation.
We must choose the sample size which ensures that the **margin of error is at most a certain value** (and no more).

From the above plot $p = 0.5$ gives the largest standard error.

- Use $p=0.5$ in the margin of error calculation. This is the most conservative choice and once the data has been collected the margin of error may be smaller.

- If prior information on the probability is available use that
  - If we know $p$ will be at **most** 0.3. Then use $p=0.3$ in the margin of error calculation (since the MoE is less for $p < 0.3$).

- If we know that $p$ will be at **least** 0.7 then use $p = 0.7$ in the margin of error calculation (since the MoE is more for $p > 0.7$).
Example 1 (Margin of Error calculation)

- Suppose we want to construct a confidence interval for the proportion of the population who are pro-gay marriage. We want this confidence interval to have an margin of error no larger than 2% (same as 0.02).
- What sample size should we use?
- We have no prior information on the proportions. But we know that for a 95% CI the largest maximum of error is (using \( p =0.5 \))

\[
\text{MoE} = 1.96 \times \sqrt{\frac{0.5 \times 0.5}{n}} = 1.96 \sqrt{\frac{0.25}{n}}
\]

- Therefore solving the above with \( \text{MoE} = 0.02 \) gives

\[
n = 0.25 \times \left( \frac{1.96}{0.02} \right)^2 = 2401
\]

- Observe that a very large sample size is required to obtain a small margin of error.
Example 2 (Margin of Error Calculation)

- We want to construct a confidence interval for the proportion of the population who support vaccinations. It is known this proportion is greater than 0.7.
- How large a sample size should we choose to ensure that the margin of error of a 99% CI is at most 0.01 (=1%, note this has nothing to do with p-values)?
- Since the proportion is known to be somewhere between [0.7,1]. We use 0.7 in the margin of error calculation as this leads to largest margin error over the range \( p = [0.7,1] \).

\[
n = \left( \frac{2.57}{MoE} \right)^2 p(1-p) = \left( \frac{2.57}{0.01} \right)^2 0.7 \times 0.3 = 13870
\]

- We see that we need a sample size of at least 13870 to obtain a margin of error that is at most 0.01 (1%).
- If the true \( p = 0.9 \), then when with a sample size \( n = 13870 \) the margin of error will be smaller:

\[
MoE = 2.57 \times \sqrt{\frac{0.1 \times 0.9}{13870}} = 0.0065(= 0.65\%)
\]
Examples 3 (margin of error calculation)

What sample size would we need in order to achieve a margin of error no more than 0.02 (2%) for a 90% confidence interval for the population proportion of arthritis patients taking ibuprofen who suffer some “adverse side effects”?

We could use 0.50 for our guessed $p^*$. However, since the drug has been approved for sale over the counter, we can safely assume that no more than 15% of patients should suffer “adverse symptoms” (a better guess than 50%).

For a 90% confidence level, $z^* = 1.645$. Computing the required sample size $n$:

$$n = \left( \frac{z^*}{m} \right)^2 \times p^* (1 - p^*) = \left( \frac{1.645}{0.02} \right)^2 \times (0.15)(0.85) = 863.$$

To obtain a margin of error no more than 2%, we would need a clinical study with a sample size $n$ of at least 863 arthritis patients.
Accompanying problems associated with this Chapter

- Quiz 15
- Homework 7 (Questions 4 and 8)
- Homework 8 (Questions 3 and 4)