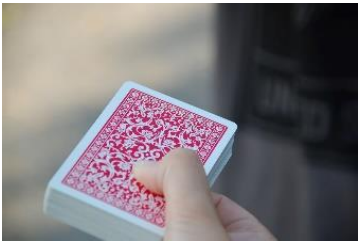


Chapter 2: Testing for a Proportion



Investigation: Someone claims to have Extra-sensory perception (ESP). They claim they can perceive things before they happen. We’d like to put their ESP to the test by having them guess the **color** of **16** randomly drawn cards in a row.

In a regular playing deck, half of the cards are red and half of the cards are black. Let’s collect some data to see how many cards this individual can guess (perceive?) correctly before we flip them!

Table 1. Trial Results

Trial	Correct
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	
15	
16	

They guess ____ out of **16** correct.

What percent of the time do ***you think*** someone would guess ***at least*** that many cards correctly by chance if they were guessing randomly for each card?

Based on the data, is there evidence to support this person’s claim that they have at least some advantage in guessing the card?

What is the **minimum** number of cards out of 16 that someone would have to guess correctly in order for **you** to believe they really had an advantage in guessing cards today? 11? 12? 13? 14? 15? All 16?

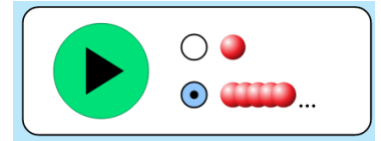
Chapter 2: Testing for a Proportion

Let's use "[Plinko Probability](#)" to help us navigate this question.

- Use this link, or web search "Phet, plinko probability."
Choose the "Lab screen" and play around with it for a minute.

Mini investigation: Set Rows to 2 (creating bins labeled 0, 1, and 2) and keep binary probability at 0.50. *Note that you can choose to run balls one at a time, or let them run continuously.*

- **What percentage of plinko balls land in the 0 bin? The 1 bin? The 2 bin?**
Why do you think they land where they do as often as they do?

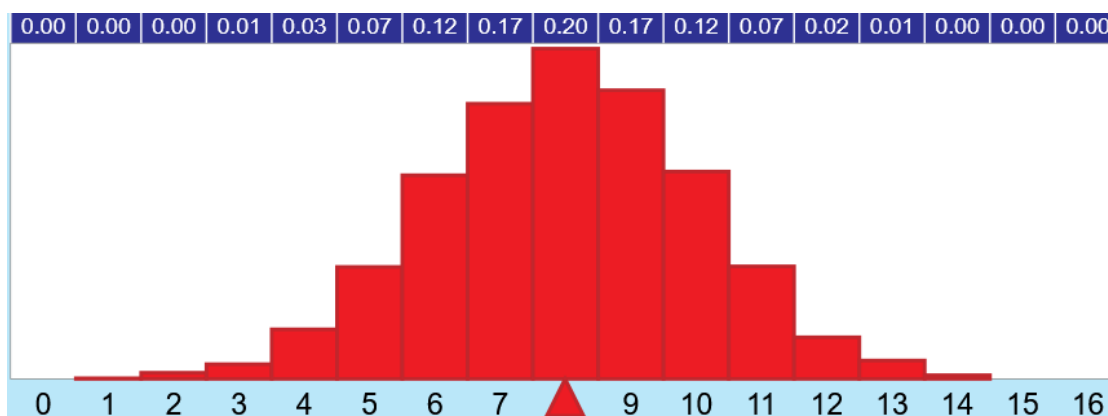


Mini investigation: Now adjust to 4 rows. What do you notice about where the plinko balls land? **Are they equally likely to land in any bin, or are some more likely than others?**

How might we use this simulation to help us estimate **what percentage of the time** a student would perform **at least as well** as what we observed in our classroom demonstration? Try it out!

The Language of Hypothesis Testing

- **Null Hypothesis:** Null means _____. A Null Hypothesis assumes no change, no difference, or no association in the situation we're studying.
 - **In our investigation,** the "Null Hypothesis" would be that...
- **Null Model:** A set of possible results that could happen under a Null Hypothesis.
 - **In our investigation,** our histogram was a null model, representing how often someone would guess various numbers of cards correctly if each card is a _____ guess.



- The **P-value** is a *probability* representing how compatible our sample result is with the null hypothesis.
 - In our class, we'll typically try to incorporate these pieces.

Generically

The probability of getting a **sample** result at least this far from expectation

if we assume the Null Hypothesis is true

is X%

In our investigation...

The probability that someone would guess at least _____ cards correctly out of 16

if we assume they had a _____ accuracy rate

is _____

We may sometimes shorten our interpretation by substituting the "if the null hypothesis is true" piece with "by random chance."

Digging Deeper: What do we mean by "model"?

What examples come to mind when you think of a "model"?

At its essence, a model is meant to be a _____ **of a possible reality!** When we do statistics, we'll use the term "model" in a few different contexts. In hypothesis testing, we use a "null model" to represent possible outcomes when assuming the null hypothesis. Later in the course, we'll talk about a model as a proposed relationship between variables that we will attempt to represent with an equation or graph!

Digging Deeper: Where do we get these probabilities?

We generated these probabilities from simulation, but we can also model this question with a function that accounts for this particular probabilistic situation.

The **binomial distribution** can be used to determine the likelihood of observing x “successes” out of n trials when there is a p probability of success on each trial. *This would be the case if we assume each trial is independent and the probability of success remains constant on each trial.*

$$p_X(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}, x = 0, 1, \dots, n$$

For example, if we wanted to know how often someone would guess *exactly* 12 out of 16 cards correctly by random chance, we would calculate:

$$p_X(12) = \frac{16!}{12!(16-12)!} 0.5^{12} (0.5)^4 = 0.0278$$

Using the [binomial distribution applet](#) and looking at the “formulas and properties” tab, we can calculate the p-value more precisely by adding up the probability of $x = 12, 13, 14, 15$, and 16 .

Reflection Questions

2.1: Consider if you tossed a fair coin 20 times. Are you just as likely to get 10 tails out of 20 as you are to get 5 tails out of 20? Why or why not? **Hint:** *What did we learn while using the plinko simulation?*

2.2: When testing a research question with a hypothesis test, what general answer does a null hypothesis represent? **Hint:** *what does the word “null” mean?*

2.3: One way that you might have used the plinko simulation to address our ESP investigation was to set rows to 16, run plinko balls down the board, and observe what bin they landed in. How did that help us assess whether the student’s performance could reasonably have occurred through random chance guessing?

2.4: If our classroom card-guesser had guessed *fewer* cards correctly, would the p-value have gone up or gone down? In general, what does a higher or lower p-value tell us?

Investigation: We hammer a quarter into a new shape, creating a visible dent. We'd like to test whether this coin now has a bias and favors one side more than the other. We flip this coin 100 times. We observe 58 heads and 42 tails. Would this be evidence of a bias, or might this be a reasonable result to observe if the coin were still fair?

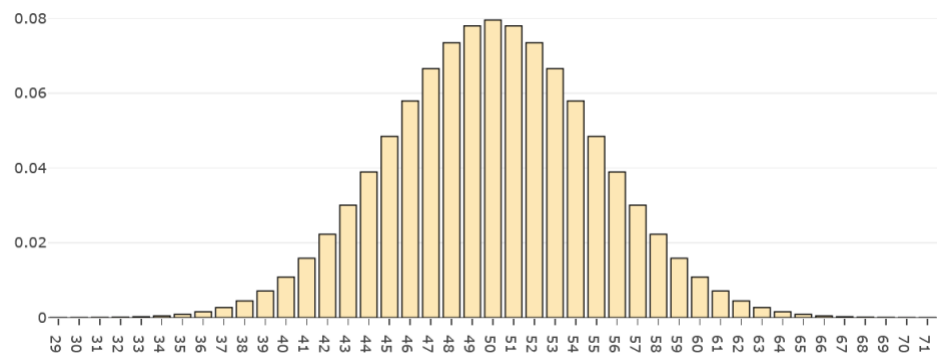


Let's start by describing what the Null Hypothesis would be for this investigation!

Now that we have a sample size of 100, we need to use a larger simulation. Let's use the [Binomial Distribution Simulation](#) from "Art of Stat web apps" page.

- Go to the "Find Probabilities" Tab up top.
- Adjust "Bernoulli trials" to our sample size.
- Adjust "Probability of success" to match our Null Hypothesis of 0.5.
- Set "Type of Probability" to reflect an upper tail for now.
- Set "Value of x" to our particular sample result.

According to the Null Model, how often would we see *at least* 58 heads out of 100 by random chance?



Investigation Reconceived: Truthfully, we started this investigation without knowing whether the dent would favor heads or tails. In other words, a deviation from fairness in either direction would support claim of bias. How is that different from the card-guessing question earlier?

Should this change how we evaluate the unusualness of our sample result? In which situations would I find a result *at least* as supportive of a bias as this 58 heads situation?

Making sense of Directional and Non-Directional Investigations

- **Directional Investigations:** Our results would only matter (or support our theory) if they are specifically lower or specifically higher than Expectation.



- **ESP Example:** If someone claims to have ESP, then we'd only find the sample results compelling if it was _____ than expectation.

- **Non-Directional Investigations:** Our results could matter (or support our theory) if they demonstrate a departure in either direction from Expectation.



- **Dented Coin Example:** We would have found a result of 42 heads and 58 heads _____ compelling in suggesting a possible bias.

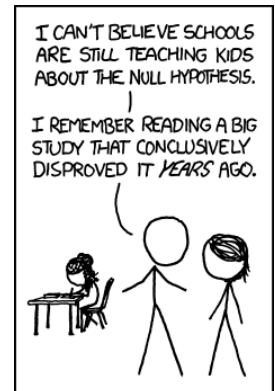
Writing Null and Alternative Hypotheses

- Think about posing a research question ("Does this person have an advantage guessing cards? Does this coin have a bias?). There are two possible answers to each question.

- The Null Hypothesis is True.
 - The Null Hypothesis is False and some Alternative is True.

- **The Null Hypothesis:** The parameter we are testing follows the status quo.

- There is no change, no difference, no association. **Nothing** of importance is happening.
 - We abbreviate the null hypothesis with H_0 (pronounced "H not").



- **The Alternative Hypothesis:** The parameter is some alternative value. There is at least _____.

- In many investigations, the **alternative** hypothesis represents **our theory**.
 - If we pose our research question as "Has there been some change? Or is there some departure from the status quo?" The alternative says _____ while the null says _____.
 - We abbreviate the alternative hypothesis with H_A .

- Remember that **Hypotheses** are **statements** about a **parameter**.

- Did this person guess more than half of these 16 cards correctly? (right / wrong)
 - Does this person seem to be guessing cards above a 50% success rate? (right / wrong)
 - We are not hypothesizing about the value of our sample proportion, _____ (we know that!). Instead, we are asking about _____.

Non-Directional Case: Dented Coin**Directional Case: ESP Test**

- **P-values for Binary Decisions – Assume the Null unless evidence otherwise!**

- In some cases, we want to make a simple decision:
 - **Reject the Null:** Our p-value is _____. The Null is not compatible with our sample results.
 - **Fail to Reject the Null:** Our p-value is _____. Our sample results could reasonably take place under the Null Hypothesis.
- We often set a significance level (represented as α) to determine our decision. Think of α as an unusualness threshold we have to pass in order to decide the null is likely wrong.
 - A common choice is $\alpha = 0.05$
 - If the p-value is at or below α ...
 - If the p-value is above α ...

Practice: What should we decide in the biased coin example if we use $\alpha = 0.05$ as our benchmark comparison?

- **P-values as insights**

- There are times to make binary decisions from hypothesis tests, but we can still regard lower p-values as stronger evidence and higher p-values as weaker evidence.
 - [The American Statistical Association's statement about p-values](#) highlights this issue.
 - The table below provides **suggested interpretations** for different p-value ranges.

Table 2. P-value Interpretations

P-value	Suggested Interpretation
$P > 10\%$	Weak or Little evidence against the null / for the alternative
5-10%	Mild evidence against the null / for the alternative
1-5%	Moderate evidence against the null / for the alternative
0.1-1%	Strong evidence against the null / for the alternative
$< 0.1\%$	Very strong evidence against the null / for the alternative

Chapter 2: Testing for a Proportion

- Making Errors
 - **Type I Error:** Incorrectly rejecting the null hypothesis (concluding a difference when there really is none).
 - When the Null hypothesis is true, the probability of making a Type I error will be α .
 - *That's because we will incorrectly reject the null whenever we happen to have a p-value below α by random chance!*
 - **Type II Error:** Incorrectly “failing to reject” the null hypothesis (failing to conclude a difference when there really is a difference).
 - Several things can make the probability of a Type II error more likely.
 - Our sample size is _____.
 - The true departure from the Null is _____.
 - We set a significance level very low.

Table 3. Type I and Type II Error

	<i>Null is really True</i>	<i>Null is really False</i>
<i>Fail to Reject Null</i>	Correctly “failing to reject”	
<i>Reject Null</i>		Correctly reject

Practice: An early study looked at the effectiveness of Remdesivir in lowering the mortality rate among those hospitalized with COVID-19. The small sample study did not have a low enough p-value to conclude it was more effective than standard treatment, but a larger study conclusively found Remdesivir truly lowered the mortality rate.

The null hypothesis for the small sample study was...

Did the small sample study make an error?

Many researchers have studied the use of Chicken Noodle Soup as a potential cure for the common cold. One researcher found that those eating chicken noodle soup had faster recovery times than the general population of cold sufferers, finding a p-value below $\alpha = 0.05$. But in reality, chicken noodle soup offers no benefit.

The null hypothesis for the small sample study was...

Did they make an error?

In the dented coin example, let's say the dent has not actually changed the probability of getting heads or tails. In our investigation, we failed to reject the null hypothesis.

Did we make an error?

Read on your own



We typically say we “fail to reject” the null hypothesis when our p-value is above α . **Why not “conclude” or “accept” the null hypothesis?**

- The null represents a very specific value or situation, and we would really need data on the entire population to confirm if it’s technically correct!
- For example, our 58 heads out of 100 tosses simply suggested that this *could* still be a fair coin, but the coin could still have a rather small bias, like a 52/48 heads to tails split.
- Rejecting the null and concluding the alternative is still possible though. The null represents a specific value, and the alternative is a range of possibilities!

When getting a low p-value from a **non-directional test**, can we still make a **directional claim**?

- **YES!** Non-directional questions simply affect how we calculate our p-value. However, we **can** make a **directional claim when we’re done** by noticing which way the difference falls.
- Non-directional tests simply raise the bar for concluding the alternative. Your p-value will be twice as high as if you did a directional test.
- In the dented coin example, I may not know ahead of time which way the coin is biased. If I got 80 heads out of 100, my p-value is the probability of seeing 80 or more + 20 or fewer heads out of 100.
- But now that I’ve shown strong evidence that this is unlikely explained as random chance, I can look at my data to see the direction of the bias. There is no mystery at that point!

Reflection Questions

2.5: When doing a hypothesis test, do we write our hypotheses with statistic symbols or parameter symbols? Why does that difference matter?

2.6: Why in some cases should we do a non-directional test rather than a directional test?

2.7: What is the name of the p-value threshold we use to determine if we have evidence to reject the null? What does it mean if our p-value is above that threshold?

Chapter 2 Additional Practice (Videos available for these in Canvas Ch 2 module)

Bottlenose Dolphins are considered among the most intelligent animals. A group of researchers would like to test their color recognition and memory.

19 dolphins were shown a panel of four buttons of different colors. One of the four would light up, and then the dolphin would be guided to swim to the other side of the pool to an identical panel of four colored buttons. The dolphin was given a treat if it pressed the same colored button from the other side. This was repeated several times to help the dolphin make the association.



The next day, they repeated this activity with the 19 dolphins—for each dolphin, one of the four colors lit up, and then the dolphin was directed to the other identical panel to see if the first button pressed was of the same color. The biologist would like to know if the dolphins are performing *better* than what we would expect by random guessing.

The sample result showed that 12 out of 19 dolphins guessed on the first time correctly.

Write the null and alternative hypotheses for this investigation? *Is it directional or non-directional?*

Let's again use the [Binomial Distribution Simulation](#) to create the null model:

- Go to the "Find Probabilities" Tab up top.
- Adjust "Bernoulli trials" to our sample size. Our sample size is...
- Adjust "Probability of success" to match our Null Hypothesis. The null hypothesized proportion is...
- Set "Type of Probability" to reflect the direction of our investigation. Type of probability should be...
- Set "Value of x" to our particular sample result. The sample number correct is...

What is the p-value for our investigation? What would be an appropriate conclusion to make?

Now, imagine that we made a mistake, and it was actually 13 dolphins, not 12, who guessed correctly. Would this change our p-value? (*i.e., is the null hypothesis more plausible or less plausible?*)

