Chapter 7: Introduction to Hypothesis Testing

Making an Inference with Hypothesis Testing

- What is an Inference?
  - **Statistical Inference**: a claim about a _____ based only on _____.
  - In many situations, we don't have the whole picture, but we try to piece it back together with the sample of data we do have.
- What is a Hypothesis Test?
  - **Hypothesis Test**: a method of inference that examines whether a _____ _____ using sample information.
  - We typically use a candidate parameter that represents no difference or no change from the status quo, since that is a specific situation we can identify.

Practice: Someone claims to be psychic. We put them to the test by having them guess the **color** of **20** randomly drawn cards in a row (with each card replaced and reshuffled back in the deck). They guess ___ out of **20** correct.

Our Guiding Question: Does this person truly have an advantage, or could this just be some random chance guessing?

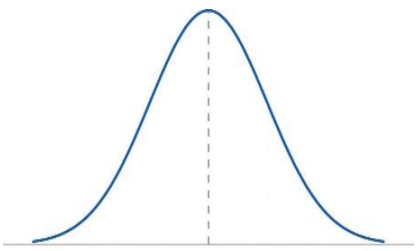
If they had an advantage, it's unclear what their true proportion of guessing was, but if they were guessing with random chance, then there is a clear candidate parameter we could test: p = 0.5

What proportion of sampled cards were guessed correctly? $\hat{p}$ =

For a sample size of 20, what is the standard error for $\hat{p}$ in this situation?

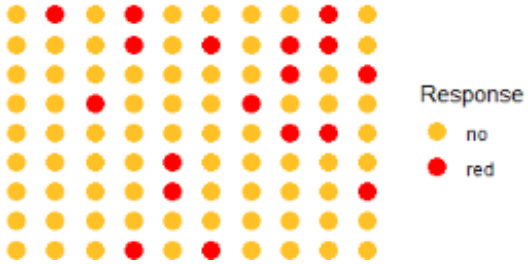
Recall that the distribution of $\hat{p}$ should be normally distributed. What is the z-score for our observed $\hat{p}$ if assuming random chance guessing?

How often would we expect a sample proportion this high or higher if just random chance guessing? (use z-table)

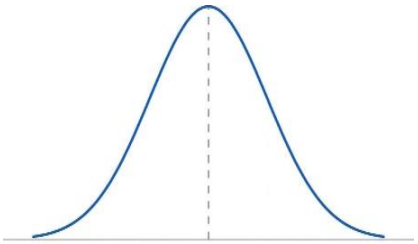


Practice: A particular plant population has a 25% probability for producing red flowers. A biologist is studying whether there is evidence of cross-pollination with another species that produces a *lower* rate of red flowers. She tests this theory by recording the colors of 90 plants in the field in question—finding that 18 of them have red flowers.

Think about it on your own first—does this sample result suggest cross-pollination, or does this seem like a result you would observe if the true proportion were 0.25?



Now let's evaluate this question probabilistically. Choose a candidate parameter to try on, then see how probabilistically unusual it is to see a sample result as far or farther than we did by random chance.



Identify the Hypotheses

- When completing a hypothesis test, there are two scenarios we are weighing:
  - The candidate parameter is true
  - The parameter is some alternative value

- **The Null Hypothesis:** The candidate parameter is true.

  - The term "Null" literally means "Nothing." We use that term for a reason! It is the idea that **nothing** of importance **is happening**. Think of it as the status quo.
  - We abbreviate the null hypothesis with H_0 (pronounced "H not")

- **The Alternative Hypothesis:** The parameter is some alternative value.

  - In many investigations, the alternative hypothesis represents our theory. We are examining a situation because we are curious if there is some departure from the status quo.
  - We abbreviate the alternative hypothesis with H_A

- Tips on identifying the null and alternative hypotheses
  - **Be specific which parameter we are testing.** Are we testing a proportion? A mean? A regression slope?
    - Don't write $H_0 = 0.5$, $H_a > 0.5$
    - Instead write H_0: $p = 0.5$., H_0: $p > 0.5$
  - **Hypotheses are** s**tatements about a parameter.** We are not hypothesizing about the value of a sample statistic (we already know those!).
    - Don't write H_0: $\hat{p} = 0.5$, H_A: $\hat{p} > 0.5$
    - Instead write H_0: $p = 0.5$, H_A: $p > 0.5$
  - **For directional alternatives, it's customary to "mirror" the null.** Even though the null represents the candidate parameter, it completes the range of possibility to also include the other direction not taken by the alternative.
    - H_0: $p \leq 0.5$
    - H_A: $p > 0.5$
  - **Hypotheses can be stated in words or in symbols.**
    - H_0: This person is guessing cards correctly with a probability of 0.5 (or less).
    - H_0: This person is guessing cards correctly with a probability greater than 0.5.

Practice: How would you write the null and alternative hypotheses symbolically in the red flower example?

How would you write them in words?

- **Directional vs. Non-Directional Hypotheses**
  - Both examples we have seen represent _____ hypotheses.
    - In the psychic example, we would only find it interesting/unusual if they were doing better than 0.5 probability of guessing.
    - In the flower example, our theory for cross-pollination only makes sense if the proportion of red is lower than 0.25.
  - In a non-directional situation, we would find a departure in either direction as interesting and noteworthy. Let's look at an example!

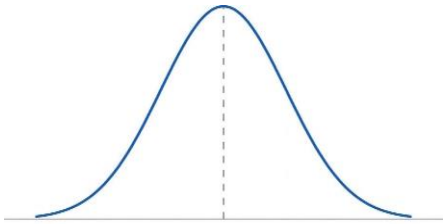
Practice: We hammer a coin into a new shape and then test out how many times it lands heads or tails.

We flip this coin 100 times. We observe 57 heads and 43 tails. Would this be evidence that the coin is now biased? Or is it still plausible the coin might still be fair?

First, write the hypotheses symbolically

Then, find the standard error for $\hat{p}$

Finally, determine how often we would observe a $\hat{p}$ value this far or farther from the null hypothesized parameter if the coin were fair.



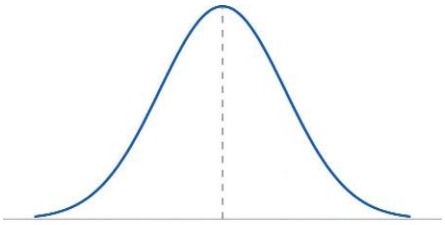
Making a conclusion from a p-value

- **Defining the p-value**
  - **P-value:** The probability of getting a sample result…
  - More informally: a **p-value** measures the **compatibility** between our sample results and the null hypothesis.
  - A p-value of 1 (100%) means we have perfect compatibility between our sample results and the null hypothesis
    - If the psychic guessed half of the cards correctly, that is perfectly compatible with the null that they are guessing randomly.
  - A p-value approaching 0 means that these sample results are almost impossible to observe if the null hypothesis were true
    - If the psychic guessed all 20 cards correctly, that would be an extremely unlikely scenario under the null hypothesis!

- **Making Decisions - Assume the Null unless evidence otherwise!**
  - In some cases, we want to make a simple decision:
    - **Reject the Null:** The Null is unlikely. Our sample results are not compatible with the null hypothesized parameter.
    - **Fail to Reject the Null**: The Null is plausible. Our sample results are compatible enough with the null hypothesized parameter.
  - *Why not "Accept" the Null?*
    - In the biased coin example, let's say we had flipped the coin 20 times and got 10 heads. That is *perfectly* compatible with the null. Does that mean the coin is unbiased?

    - For this reason, we can't make a probabilistic claim about whether the null hypothesis is *exactly* correct. We can only determine if the null is plausible or not.

  - We often set a significance level (represented as α) to represent our cut-off point.
    - A common choice is α = 0.05
    - If the p-value is at or below α…_____
    - If the p-value is above α…_____



What should we decide in the biased coin example if we use α = 0.05 as our benchmark comparison?

- Making Errors
  - **Type I Error:** Incorrectly rejecting the null hypothesis (concluding a difference when there really is none).
  - Our significance level is our preset probability of making a Type I error!
  - **Type II Error:** Incorrectly "failing to reject" the null hypothesis (failing to conclude a difference when there really is a difference).

	Null is really True	*Null is really False*
Fail to Reject Null	Correctly "failing to reject"	
Reject Null		Correctly reject

Practice: An early study looked at the effectiveness of Remdesivir as a drug for treating 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 improvement. What type of error did the small sample study make?

- **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.
    - https://apnews.com/12cf3d07354c47b3b9bb552776071522
    - The table below provides **suggested interpretations** for different p-value ranges.

P-value	**Suggested Interpretation**
P > 10%	Weak or Little evidence (against the null) / (for the alternative)
5-10%	Modest evidence (against the null) / (for the alternative)
1-5%	Strong evidence (against the null) / (for the alternative)
P < 1%	Very strong evidence (against the null) / (for the alternative)

Practice: *According to these suggestions,* How confident are we that the psychic from earlier had at least some advantage in guessing cards?

Hypothesis Testing for a Proportion Summarized

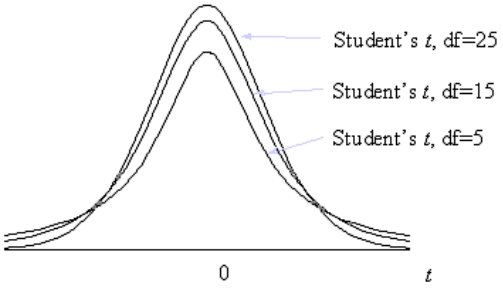
1) Write the null and alternative hypotheses
2) Calculate our sample proportion ($\hat{p}$), our best estimate for the true population proportion (p)
3) Calculate the standard error of $\hat{p}$
4) Calculate our test statistic (z-score) for $\hat{p}$
5) Find the p-value and make a conclusion

Hypothesis testing for means

- Standard Error Estimation
  - When we calculate the standard error for the mean, we will almost always need to calculate it using the standard deviation of our sample (s) since we likely won't know σ.

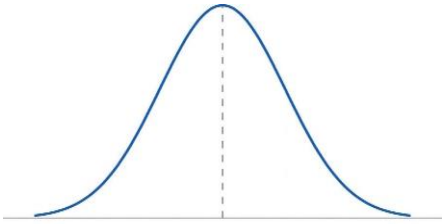
$$SE_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \qquad SE_{\bar{X}} \approx \frac{s}{\sqrt{n}}$$

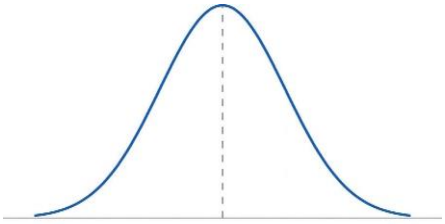
  - This issue doesn't come up with proportions since we can use our null hypothesized proportion value to calculate σ.
  - When hypothesis testing for a mean, we use something called a t-distribution to account for this error in estimation.
- **The t-distribution**
  - Think of the t-distribution as a "corrected" normal distribution. It is slightly wider than a normal distribution to account for the fact that our standard error estimate might be slightly off.
  - For small sample sizes, the correction will be _____, but as n gets larger, our estimate of the standard deviation also gets better.
  - For this reason, there is a different t-distribution for each possible sample size, and the larger the n, the _____ the t-distribution is to the normal distribution.



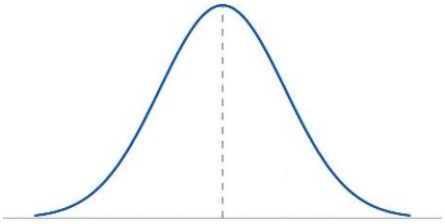
  - **Degrees of Freedom**
    - Instead of identifying t distributions by sample size, we identify them by something called Degrees of Freedom (df).
    - For now, when doing a hypothesis test for a mean, use **df =** _____
  - **Using a t-table**
    - Each row of a t-table represents "critical values" for a t-distribution of a particular df.
      - ❖ Since there is a different t-distribution for each df, there is not as much detailed information as compared to a z-table!
    - What we can do is find the t-score critical value associated with different tail probabilities as a way to "bound" the p-value range.
    - Helpful website for visualizing the t-distribution: https://istats.shinyapps.io/tdist/

If df = 20 (i.e., n = 21), what is the t-score associated with a right tail of 0.05? What is the t-score associated with a *left-tail* of 0.05?



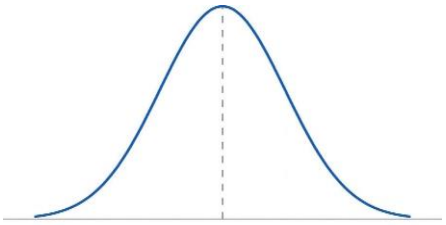


If df = 12 (i.e., n = 13), what is the *left-tail* probability associated with a t-score of -1.2? Report as a bounded range!

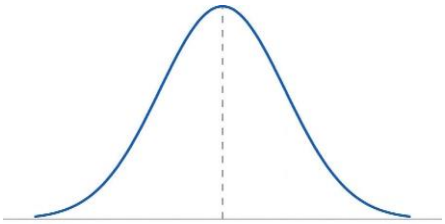


- Notice that as df increases, the t-score associated with each tail probability approaches the matching z-score at that tail probability (z-score listed with the ∞ row).
- Since the t-table doesn't list all values starting at df = 31, it's customary to round df down to the next lowest value on the table.

If df = 45, what is the t-score associated with a right tail of 0.01?



If df = 23, what is the two-tailed probability associated with a t-score of -2, and 2? Report as a bounded range!



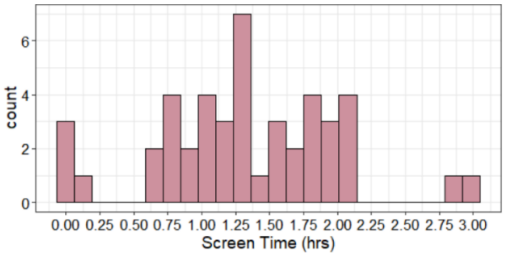
- **Setting up the t-test**
  - A hypothesis test for a mean is conducted in the same way, except that our *test statistic* is now a t-score instead of a z-score.

$$t = \frac{\bar{x} - \mu_0}{\sim SE_{\bar{x}}}$$

  - The only exceptions to this are when
    - We have a known σ to use to find the true SE
    - Our sample size is large enough that the correction is unnecessary (if df > 120)

Practice: Child psychologists recommend that young children should have no more than 1 hour of "screen time" a day. A researcher is studying whether 3-year-olds are getting *too much* screen time on average. She takes a random sample of 45 moms of 3-year-old children and surveys them on their children's screen time. She finds that the average amount of screen time reported was 1.36 hours with a standard deviation of 0.68.

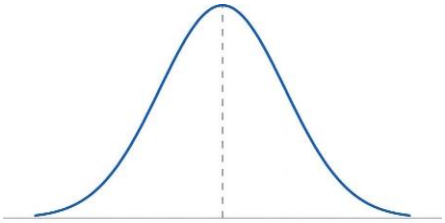
Write the null and alternative hypotheses symbolically *(is this a directional or non-directional question?)*



Identify $\bar{x}$ and calculate the (estimated) standard error in this value as an estimate for μ.

How many standard errors is our sample mean away from the null hypothesized mean*? In other words, what is our test statistic?*

Using the t-table, find a bounded range for the p-value for this hypothesis test.



Hypothesis Testing for a Mean Summarized

1) Write the null and alternative hypotheses
2) Identify our sample mean ($\bar{x}$), our best estimate for the true population mean (μ)
3) Calculate the standard error of $\bar{x}$
4) Calculate our test statistic (t-score or occasionally z-score) for $\bar{x}$
5) Find/bound the p-value and make a conclusion

- **P-value Interpretation pitfalls**
  - **Never "accept" or "conclude" the null hypothesis—even if the p-value is high**
    - Remember that p-values are measuring _____ between our sample results and the null hypothesis. Even a p-value of 100% **doesn't** mean the null is true!
    - Stick to phrases like: The null hypothesis is "plausible" or "reasonable." It could be that there is a small difference, and we just don't have a large enough sample to detect it!

  - **Be cautious with the term "statistically significant"**
    - Describing low p-values as "statistically significant" has led to a lot of confusion in interpreting what they actually tell us.
    - Statistically significant just means that we are confident that the difference we observed is unlikely to occur due to random chance (there is _____).
    - This does **not** tell us if the difference is important or meaningful!

  - **P-values don't tell us the magnitude of a difference**
    - **Example:** A study assessed veterinarians' self-efficacy after completing a professionalism program compared to a control group who did not do this program. The p-value for this difference in mean scores was .0004. Sounds like it made a big difference! Until you realize the mean scores for each group—on a scale of 1 to 5—were 3.74 and 3.83.
    - This study had a sample size of about 600.
    - <u>The larger the sample size, the better you are at detecting differences confidently—including the very small ones (you have more "statistical power")</u>
    - **P-values help us determine if there is *any* departure from the null hypothesis, but p-values alone _____ tell you _____ the departure is.**

Practice: A random sample of 38 cancer survivors were asked if they would consider themselves "physically active." The researchers were interested in whether the proportion who answer "yes" to this question would be different than that of the general population. The researchers completed a z-test and got a p-value of 0.45. They concluded that cancer survivors consider themselves as active as the general population.

Did the researchers phrase their conclusion appropriate, or could this be misleading?

If the study made an error, which type of error could they have made (Type I or Type II)?

When is each test appropriate?

- z-test for a proportion
  - The distribution of $\hat{p}$ is normally distributed
    - We have at least 10 of each response in our sample
- t-test for a mean
  - The distribution of $\bar{x}$ is normally distributed
    - Either the population distribution we are gathering data from is approximately normally distributed
    - …or n > 30, and the population does not have a heavy skew (long tail)
- z-test for a mean
  - σ is known, or approximately known due to large sample size (df > 120).
  - The distribution of $\bar{x}$ is normally distributed
    - *if n > 120, then this is relatively safe to assume!*

Quick summary of Important Terms

Null Hypothesis: A candidate parameter—represents status quo, no effect, no difference

Alternative Hypothesis: An alternative parameter range—represents some difference or effect

Standard Error: The expected error of our sample statistic from the population parameter

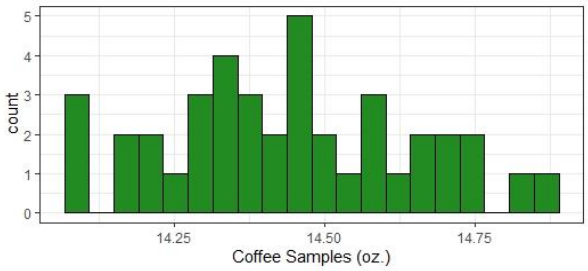
Test Statistic: Our sample statistic converted to a standardized value. In one-sample testing, we may calculate a z-score or t-score as our test statistic.

P-value: The probability of getting a test statistic this or more extreme IF the null hypothesis is true.

Practice: Starbucks "Grande" size coffee has room for 16 ounces of coffee. However, they don't fill the cup completely up to the brim because that would be ridiculous and result in coffee catastrophes! Let's say Starbucks has announced they put in 14.50 ounces of coffee on average. To test this claim, you have randomly selected 40 customers who ordered a grandé coffee from the Starbucks at the bookstore to have their coffee content measured.

We are investigating whether the amount of coffee that Starbucks pours into their Grande drinks is **different from** 14.50 on average.

Write the null and alternative hypotheses in symbols



Of our 40 grandé coffee measurements, the average amount of coffee was 14.43 ounces with a standard deviation of 0.20. Calculate the standard error for our sample mean, and then calculate our test statistic.

Find the appropriate p-value and make a conclusion for our investigation at $\alpha = 0.10$. Do we have evidence to claim the true average pour amount at this Starbucks is different from 14.50oz?

What if we upped our sample size to 100 people instead of 40. If the sample standard deviation and the sample mean remained the same…

Should the Standard Error get bigger or smaller?	Should the test statistic get bigger or smaller?	Should the p-value get bigger or smaller?