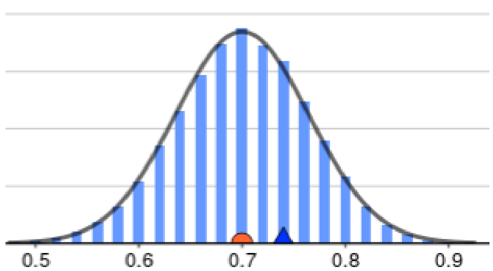
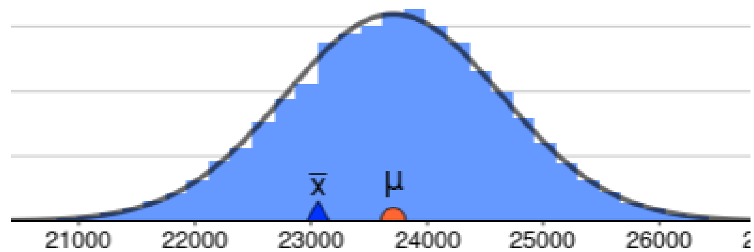
Chapter 4: z-tests and t-tests

Parametric Testing

- When testing a mean or a proportion, our Null Model is very often Normally Distributed.
    - This happens because of the Central Limit Theorem!
- As a result, We can bypass simulation and instead take a shortcut by completing a parametric test.
- Parametric tests assume the Null Model will be normally distributed and use calculus to find the p-value precisely based on the position of our sample statistic relative to the distribution.



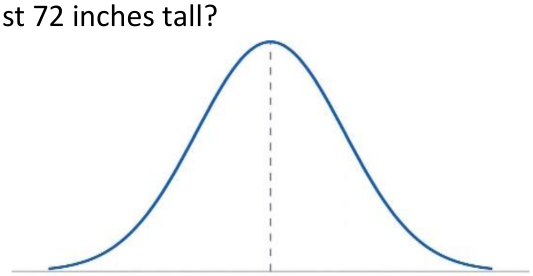


Areas under the Normal Curve

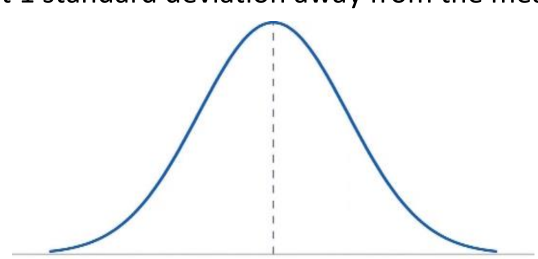
- Every Normal Distributed variable can be defined by two values
    - The mean (μ)
    - The standard deviation (σ)
- Use the Art of Stat simulation linked here to find the area under the normal distribution for different constraints - https://istats.shinyapps.io/NormalDist/

Practice: 18-year old male heights are approximately normally distributed with $\mu = 69.3$ and $\sigma = 2.5$.

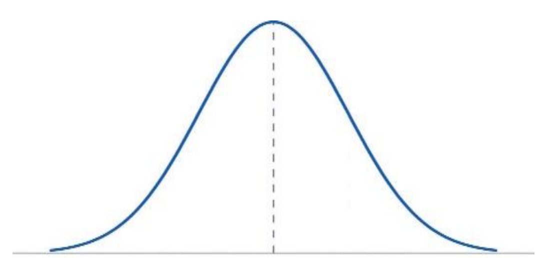
Approximately what percentage of the population is at least 72 inches tall?



Approximately what percentage of the population is at least 1 standard deviation away from the mean?



Let's try the second question again, but we'll convert this to a **Standard Normal Distribution** where $\mu = 0$ and $\sigma = 1$



Standardization with the z-scale

- Since every variable has its own mean and standard deviation, it's common to first convert your value of interest to a "standardized value" and find its position on a standard normal curve.
  - **Standardize:** To relate measurements and values to a scale that can be referenced across contexts with different units.
- **Z-scores** represent the standardized position of values in a normal distribution
  - Negative z-scores mean the data point is <u>below</u> the mean, positive z-scores mean it is <u>above</u> the mean.
  - The z-score value represents how many standard deviations that point is away from the mean

value	Values below mean	μ	Values above mean
z-score	Negative z-scores	0	Positive z-scores

Formula for calculating z-score: $\frac{\text{Observation} - \text{mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma}$

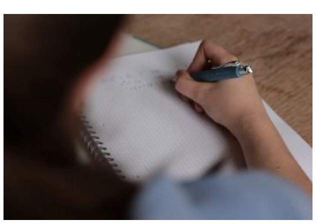
Practice: Consider an 18-year old male with a height of 65.8 inches.

What is the z-score for this height?

How often would we observe a value at least this far from the mean?

Introducing the z-test

Investigation: An instructor is examining whether the mean exam score might change when allowing students to bring a cheatsheet for the exam. He thinks it's plausible it could go up or down. Let's say that this instructor has a large collection of data showing a mean exam score of 87.2 with a standard deviation of 5.17 when no cheatsheet was allowed. The distribution of exam scores is somewhat left skewed, but not highly skewed.



The instructor administers the exam again with 48 students, allowing them to bring a cheatsheet, and finds the mean exam score for this group to be 85.9.

What is the null hypothesized parameter? What is our estimate? How much absolute error is there between them?

What is the Null and Alternative hypothesis in this investigation? *Is this directional or non directional?*

What distribution shape will the Null Model likely take and Why?

Normal, because of CLT. N is 48 and no large skew

What mean and standard deviation would it have?

$\sigma_{\bar{x}} = SE_{\bar{x}} = 5.17/\text{sqrt}(48)$

Draw Normal Curve and label

$\mu_0 = 87.2$

Calculate the z-score for our estimate.

Our sample mean is ________ standard errors below / above the null hypothesized mean

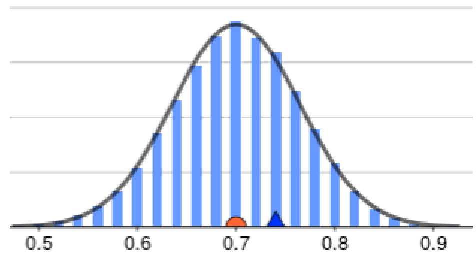
We should find a p-value of about 0.0815 *(let's verify that with the normal distribution web app*!). Which statement correctly interprets our p-value?

1. The probability that a **randomly chosen test taker** would score an 85.9 or lower, if the true mean score is 87.2, is 8.15%
2. The probability that a **randomly chosen test taker** would score at least this far from the 87.2 if 87.2, if the true mean score, is 87.2 is 8.15%
3. The probability that the **class average** would be 85.9 or lower, if the true mean score is 87.2, is 8.15%
4. The probability that the **class average** would be at least this far from the 87.2, if 87.2 if the true mean score is 87.2, is 8.15%

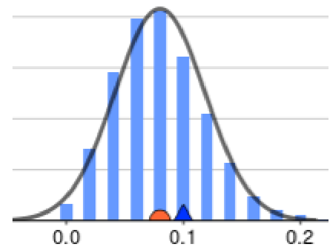
Z-test for a Mean at a glance

- Identify null and alternative hypotheses
- Determine if z-test for mean is appropriate
  - σ is known or well approximated by s (e.g., n is very large, like > 100)
  - The distribution of $\bar{x}$ should be normally distributed *(see box at the end of the chapter!)*
- Identify mean and $\sigma_{\bar{x}}$ for your null model. $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- Calculate the z-score for our particular estimate ($\bar{x}$) within the null model: $z = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}}$
- Find p-value: *how often would we observe a z-score this or more extreme in the null model?*
- Interpret the p-value and make a conclusion!

Using the z-test for a proportion



- While sample proportions technically distribute discretely, their distribution can typically be well approximated by the normal distribution with a large enough sample size!
- This wouldn't be a safe assumption in situations where the null hypothesized proportion is very close to 0 or 1 and the sample size is relatively small.
  - Some apply the 10/10 rule—we need to expect at least 10 "successes"…$n*\pi_0 \geq 10$…and 10 "failures"…$n*(1-\pi_0) \geq 10$…under the null hypothesis.
  - Otherwise, we run into a floor or ceiling problem with the distribution of $\hat{p}$ not quite symmetrically distributing around π!



Converting successes and failures to 0's and 1's

- Proportions are used when we're analyzing binary data
  - The category names might be anything, but we generically refer to them as "successes" and "failures."
- Analytically though, we treat these as 0's and 1's to statistically summarize our results!

No	0
Yes	1
Yes	1
Yes	1
Yes	1
Yes	1
No	0
Yes	1

  - The proportion of successes would then be equivalent to the mean if we assume our data is 0's and 1's.
    - $\mu = \frac{0+1+1+1+1+1+0+1}{8} = \frac{6}{8} = \pi$
  - Likewise, we can also calculate the standard deviation of binary data by taking the standard deviation of 0's and 1's.

$$\sigma = \sqrt{\frac{\sum(x_i - \pi)^2}{n}}$$

  - But…since proportions are both a measure of center and a measure of variability, the standard deviation can be rewritten solely as a function of π. Which is very convenient!

$$\sigma = \sqrt{\pi(1-\pi)}$$

Challenge question: If the true proportion of successes was either 0 or 1, what would σ come out to be?

0. Since all responses are the same, there is no variability!

Challenge question: What value for π would maximize σ?

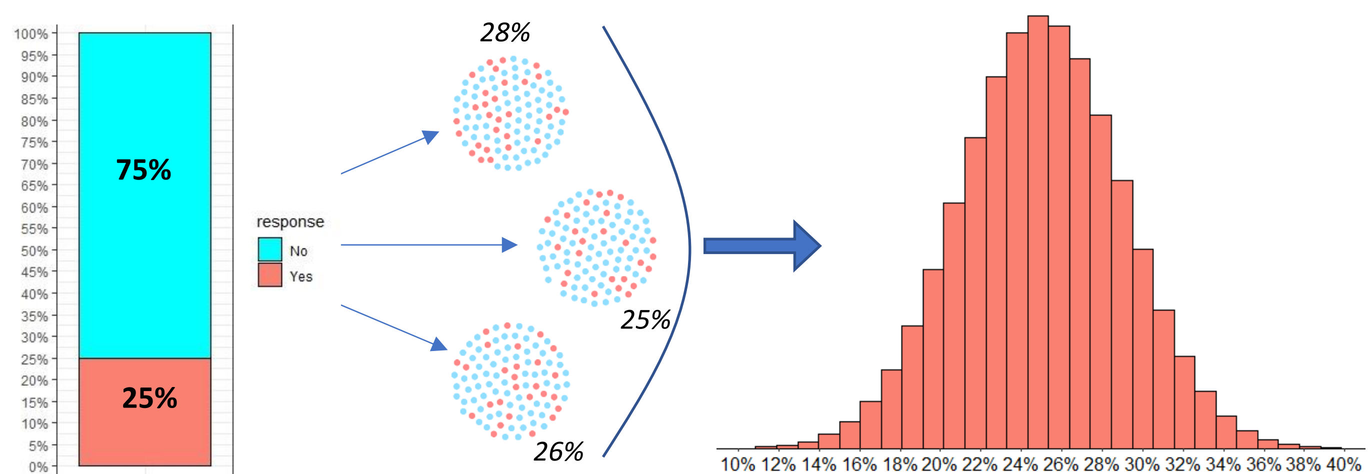
0.5. This would happen if half were 0's and half were 1's. Calculate using shortcut formula.

Setting up the z-test for a proportion

- Just like with means, a z-test is for a proportion is assuming the distribution of possible $\hat{p}$ values is normally distributed.
  - Our Null Model is a normal distribution centered at π_0 with a standard deviation of $\sigma_{\hat{p}}$
  - $\sigma_{\hat{p}} = \frac{\sigma}{\sqrt{n}}$ which for proportions under the null hypothesis, is = $\frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{n}}$

Investigation: Researchers are examining whether a new variant of the flu might be emerging as **the new dominant strain** in the Northern Hemisphere. In the Southern Hemisphere, it was responsible for 25% of all flu cases. Researchers take a sample of size 100 flu cases and find that 34 of them are infected with the new variant.

The distribution of $\hat{p}$ will be *approximately* normally distributed, so let's proceed with a z-test!



Population (Actual Proportion) | **Take Samples (of Size 100)** | **Distribution of Possible Sample Proportions**

Identify the Null and Alternative Hypotheses for this investigation *(is this a directional investigation?)*

<u>(directional, since we'd like to know if it is becoming more common)</u>

Assuming the null is true, what is the standard error for our sample proportion?

Calculate the z-score for our estimate. Our sample proportion is ________ standard errors below / above the null hypothesized proportion

The p-value for our investigation comes out to be…

Let's use the normal distribution app to find the p-value for this investigation. Which of the following might be an appropriate conclusion to make here?

1. Approximately 1.89% of all flu cases are infected with the new variant
2. The probability that exactly 25% of flu cases are of the new variant is 1.89%.
3. If the true proportion is indeed 25%, there is a 1.89% probability of seeing a sample result at least as high as 34% by random chance
4. If the true proportion is indeed 25%, there is a 1.89% probability of seeing a sample result at least this far from 25% by random chance.

Z-test for a Proportion at a glance

- Identify null and alternative hypotheses
- Determine if Null Model is well approximated as normally distributed
    - We expect at least 10 of each response under the null hypothesis (10/10 rule). So $n*\pi_0 \geq 10$ and $n*(1-\pi_0) \geq 10$
- Identify "mean" (π_0) and standard deviation ($\sigma_{\hat{p}}$) for your null model. $\sigma_{\hat{p}} = \frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{n}}$
- Calculate the z-score for our particular estimate ($\hat{p}$) within the null model: $z = \frac{\hat{p}-\pi_0}{\sigma_{\hat{p}}}$
- Find p-value: *how often would we observe a z-score this or more extreme in the null model?*
- Interpret the p-value and make a conclusion!

Read on your own

Simulation vs. Parametric Testing Approach

- With a **simulation-based** approach to testing, we simulate taking different possible samples, recording the relevant sample statistic, and then calculating what proportion are as far or farther from the center than our observed sample statistic.
- With a **parametric testing** approach, we make an assumption that the Null Model will be normally distributed and skip the simulation process. We just find the area of the normal curve that's more extreme than our sample statistic.
- Comparing the two
    - Parametric testing is more precise (if the assumption holds!)
        - Simulation-based testing provides approximate p-values, while parametric testing skips ahead to identify what your simulated results are converging to.
    - Simulation-based testing is more flexible (we don't need the normality assumption!)
        - Numeric distributions with heavy skewness may not distribute normally
        - Proportions distribute discretely, so small samples may be a "bumpy" approximation of the normal distribution. It will also have some skew if too close to 0 or 1.

Investigation: An American football should be inflated to a minimum pressure of 12.5 psi. There is, however, variability in individual footballs due to environmental conditions. This became a hot topic after the 2015 AFC Championship game between the New England Patriots and the Indianapolis Colts, when it was discovered that the footballs provided by the New England Patriots were of consistently lower pressure values.

The 11 balls were tested at halftime, and the measurements for each of those balls is recorded here. *There were actually two independent sets of recordings with two pumps. This represents the higher set of measurements. This same pump also found each of the Colts' footballs to be at or above 12.50.*

#1	11.80
#2	11.20
#3	11.50
#4	11.00
#5	11.45
#6	11.95
#7	12.30
#8	11.55
#9	11.35
#10	10.90
#11	11.35

Is there evidence that the New England Patriots were inflating their footballs below the 12.5 psi regulation on average, or could this discrepancy be explained as random chance?

Can we simulate possible sample statistics?

- This isn't like dice where the "fair die" distribution is theoretically known, or the male height situation, where we have historical data to know the parameters of the population.
- As a result, we couldn't simulate from a null model because we don't quite know what it looks like!

Can we use a z-test?

- Even if we don't know the shape of the population, the CLT says that sample means tend to distribution normally with a standard deviation of $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$
- Only problem—we don't know σ here either!

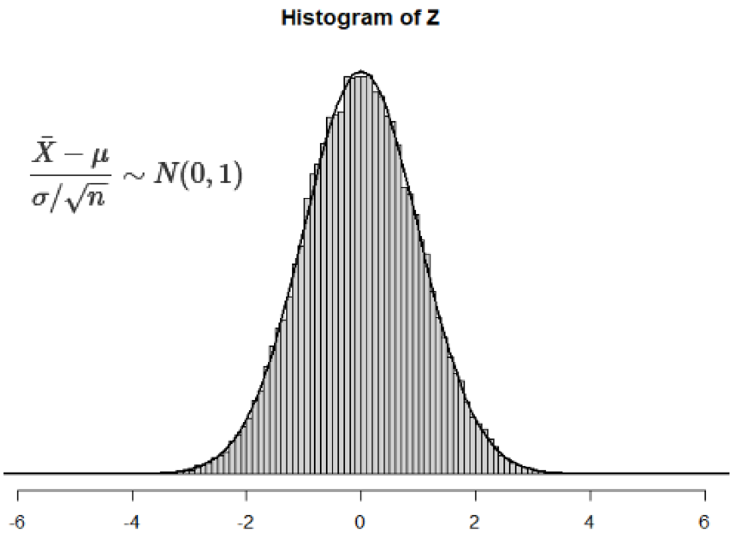
What *can* we do?

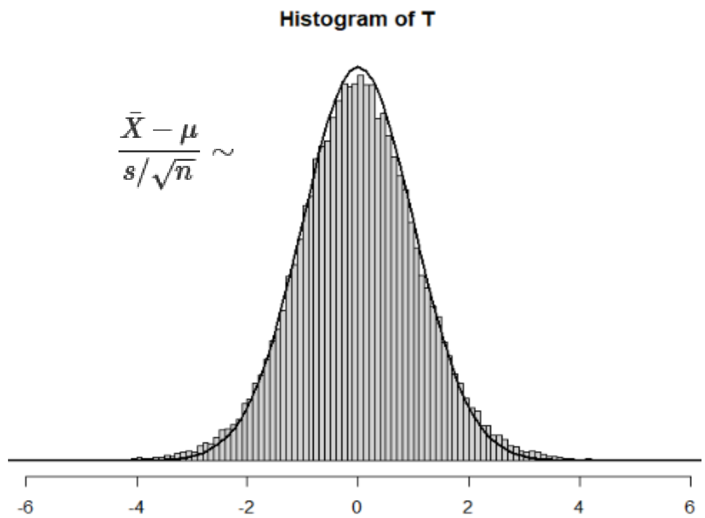
- We do have our <u>sample standard deviation (s)</u>, so we could make an approximate calculation with $\underline{s_{\bar{x}} = \frac{s}{\sqrt{n}}}$
- However, this creates some uncertainty, especially when our sample size is small.

The t-distribution

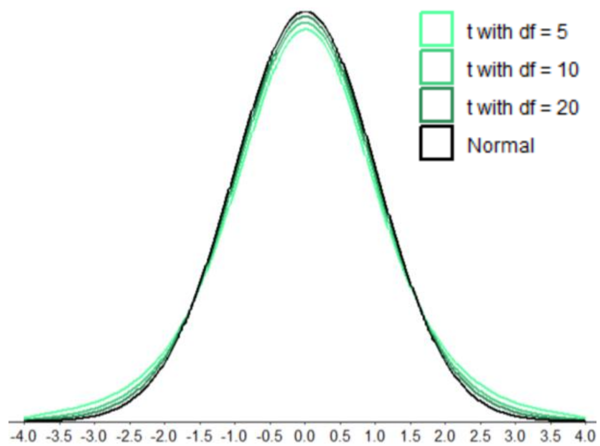
- The t-distribution is much like a standard normal distribution, but with slightly more variability.
  - When σ is known, we can convert any sample mean to a z-score
  - When σ is estimated by s, we introduce uncertainty, and this calculation will be very close to following a standard normal, but with slightly more variability.

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1) \qquad \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim \underline{T(df = n-1)}$$





For small sample sizes, the adjustment will be <u>wider</u>, but as n gets larger, the t-distribution slowly converges to a normal distribution in shape!



- **Degrees of Freedom**
  - Instead of identifying t distributions by sample size, we identify them by something called Degrees of Freedom (df).
  - When doing a hypothesis test for a mean, use df = <u>n – 1</u>

- **Using the t-distribution simulation** - https://istats.shinyapps.io/tdist/
  - **Find Probability:** Calculate the probability of seeing a t-score this far or farther from 0 under the null hypothesis.
  - **Find Percentile/Quantile**: Find the t-score associated with a particular cumulative or tail probability.

In our football investigation, what would be the degrees of freedom in our calculation of $s_{\bar{x}}$?

Back to the Investigation! In our Football example, the mean psi among the 11 Patriots footballs was $\bar{x}$ = 11.486 with s = 0.410.

What is the null and alternative hypotheses in this investigation? *Do you think this should be a directional or non-directional investigation?*

While the sample size is small, intuition says this variable should be somewhat normally distributed, and we'd thus expect $\bar{x}$ to distribute normally.

What would the mean and *approximate* standard deviation of our null model be?

Approximately normal: mean of 12.5 and SE of 0.410/sqrt(11)

Calculate the t-score for our estimate.

Using the t-distribution app, find the p-value for our investigation and make a conclusion at $\alpha = 0.01$.

T-test for a Mean at a glance

- Identify null and alternative hypotheses
- Determine if t-test for mean is appropriate
  - σ is unknown and estimated by s
  - The distribution of $\bar{x}$ should be normally distributed *(see box at the end of the chapter!)*
- Identify mean and approximate $\sigma_{\bar{x}}$ for your null model with $s_{\bar{x}} = \frac{s}{\sqrt{n}}$
- Calculate the t-score for our particular estimate ($\bar{x}$) within the null model: $t = \frac{\bar{x} - \mu_0}{\sim\sigma_{\bar{x}}}$
- Find p-value: *how often would we observe a t-score this or more extreme in the null model?*
- Interpret the p-value and make a conclusion!

Common Hypothesis Testing Mistakes

- **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 at least some difference.
  - This does **not** tell us if the difference is practically significant.

- **P-values don't tell us the magnitude of a difference**
  - **P-values** help us determine if there is ***any*** **departure** from the **null hypothesis,** but p-values alone cannot tell you how large the departure is.
  - The larger the sample size, the better you are at detecting differences confidently—including the very small differences.
  - For this reason, never mistake a very small p-value with a very large departure from the Null Hypothesis. *This is where confidence intervals or effect sizes can help!*

When is each test appropriate?

z-test for a proportion

- The distribution of $\hat{p}$ is reasonably approximated by a normal distribution
  - We expect at least 10 of each response under the null hypothesis (10/10 rule). So $n*\pi_0 > 10$ and $n*(1-\pi_0) > 10$

z-test for a mean

- σ is known or well approximated with our sample (for example, when n > 100)
- The distribution of $\bar{x}$ is normally distributed
  - Population is normal, or sample size large enough for CLT to apply

t-test for a mean

- σ is unknown and estimated by s
- The distribution of $\bar{x}$ is normally distributed
  - Population is normal, or sample size large enough for CLT to apply

In general…how do I know if n is large enough for CLT to apply?

- At a more advanced level, this can be checked with "bootstrapping" for any case.
- **For this class**, we'll say n > 30 is a good benchmark when we have no reason to think there is a large skew, and n > 100 is a reasonable benchmark if there is a large skew.

If parametric assumptions aren't met…

- For proportions, consider a simulation-based test or use binomial exact test.
- For means there are several options: https://www.statisticshowto.com/probability-and-statistics/statistics-definitions/parametric-and-non-parametric-data/

Chapter 4 Additional Practice

Practice: Many people express a hand preference (e.g., being "right-handed" or "left-handed"), but what determines that? While part of hand preference may be environmentally conditioned, there is also theory that genetics play a role. One theory posits that red-headed people are more likely to be left-handed! https://www.healthline.com/health/red-hair-blue-eyes#other-insights

In Western Countries, only about 12% of the population identifies as "left-handed." Based on genetic theory, we'd like to see if red-headed people might be more likely to be left-handed than the general population. Let's say that we take a random sample of 125 red-headed people. We found that 40 of them had a preference for their left-hand.

What is the null hypothesized parameter? What is our estimate?

Identify the Null and Alternative Hypotheses for this investigation

Can we approximate the Null Model with a normal distribution? If so, what mean and standard deviation would that distribution have?

<u>Approximately normal: 0.12*125, 0.88*125 (10/10) rule. Also look at the sim?)</u>

<u>$\pi_0 = 0.12$ $\sigma_{\hat{p}} = 0.029$</u> <u>Draw Normal Curve and label</u>

Calculate the z-score for our estimate.

Our sample proportion is ________ standard errors below / above the null hypothesized proportion

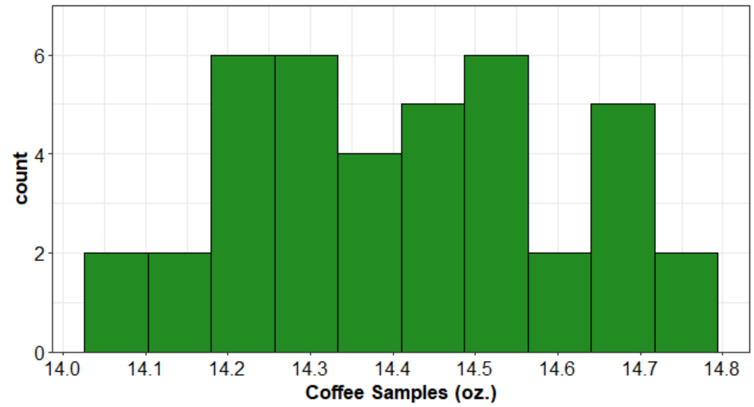
We should find a very very low p-value (approaching 0!). Which statement best explains what we found?

1. The Null Hypothesis is very likely true
2. <u>The Null Hypothesis is very likely false</u>

Practice: Starbucks "Grandé" 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 claims 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 there is evidence that Starbucks pours **less than** 14.50 oz on average into their grandé coffee drinks.

Write the null and alternative hypotheses



Of our 40 grandé coffee measurements, the average amount of coffee was 14.42 ounces with a standard deviation of 0.19. Calculate the standard error for our sample mean, and then calculate our test statistic. *Should we do a z-test or t-test?*

<u>Note we should do t-test since we need to estimate σ with s when testing a mean, and n is not particularly large (not greater than 120)</u>

Find the appropriate p-value and make a conclusion for our investigation at α = 0.05. Do we have evidence to claim the true average pour amount at this Starbucks is less than 14.50oz?

Let's say 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 closer to 0 or farther from 0?	Should the p-value get bigger or smaller?
<u>Smaller! Since we expect our sample mean to be more precise</u>	<u>Farther! Since SE is getting smaller, then a discrepancy of 0.08 will be more SE's away now.</u>	<u>Smaller! As our test statistic shows, we are more SE's away from the null, suggesting less likely to be random chance</u>

Chapter 4 Learning Goals

After this chapter, you should be able to…

- Understand "parametric tests" like z-tests and t-tests as precise methods for finding p-values when we can safely assume our Null Hypothesized Model is normally distributed
- Recognize the standard normal distribution as a normal distribution that has been transformed to set the mean to 0 and the standard deviation to 1.
- Recognize situations when parametric tests may **not** be appropriate
    - For proportions: when the true proportion is very close to 0 or 1 and our sample size is relatively small. *See box at the end of the chapter.*
    - For means: when our population distribution is skewed, and our sample size is too small to assume the Central Limit Theorem would apply. *See box at the end of the chapter.*
- Complete a z-test for a mean
    - Identify the null hypothesized mean (μ_0), our sample estimate ($\bar{x}$), and the absolute error between them.
    - Identify the null and alternative hypotheses for our investigation
    - Calculate the standard error of the sample mean: $\sigma_{\bar{x}}$
    - Calculate the z-score for our sample mean *(how many standard errors away from the null)*
    - Interpret the p-value provided and make a conclusion
- Complete a z-test for a proportion
    - Identify the null hypothesized proportion (π_0), our sample estimate ($\hat{p}$), and the absolute error between them.
    - Identify the null and alternative hypotheses for our investigation
    - Calculate the standard error of the sample proportion: $\sigma_{\hat{p}}$
    - Calculate the z-score for our sample proportion *(how many standard errors away from the null)*
    - Interpret the p-value provided and make a conclusion
- Recognize situations where we would use a t-test rather than a z-test-- testing a mean in which σ is approximated by s.
- Understand the t-distribution as an "adjusted" standard normal distribution (slightly wider).
- Complete a t-test for a mean
    - All steps involving a z-test for a mean, but with an approximation for $\sigma_{\bar{x}}$