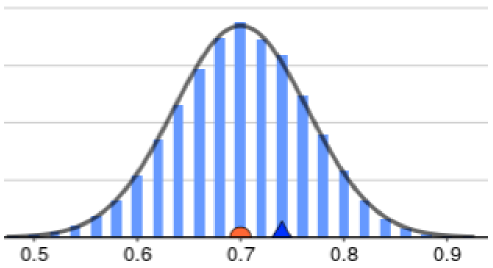
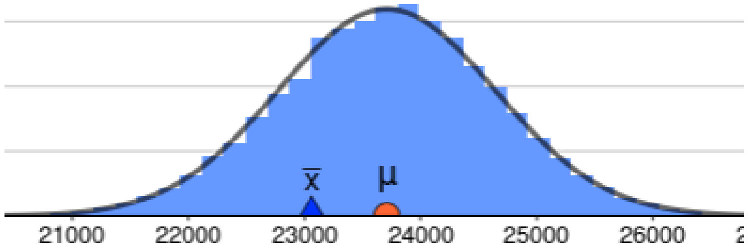
Chapter 4: z-tests and t-tests

Parametric Testing

- When testing a mean or a proportion, our Null Model is very often ________________________.
    - This happens because of the ____________________ Theorem!
- As a result, We can bypass simulation and instead take a shortcut by completing a parametric test.
- _________________ tests assume the Null Model will be normally distributed and use calculus to find the p-value based on the position of our sample statistic inside the null model.



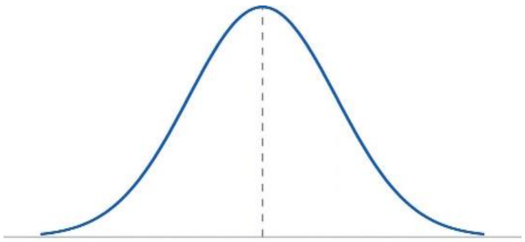


Areas under the Normal Curve

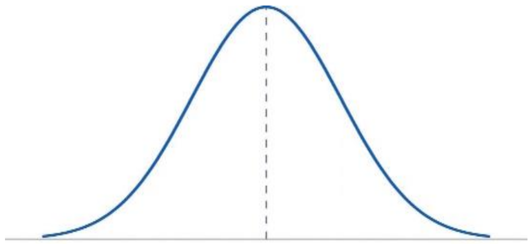
- Every Normal Distributed variable can be defined by two values
    - The __________
    - The ______________________
- Use the Normal distribution simulator from the Art of Stat Web apps page to find the area under the normal distribution for different constraints.

Practice: 18-year old male heights are approximately normally distributed with μ = 69.3 and σ = 2.5.

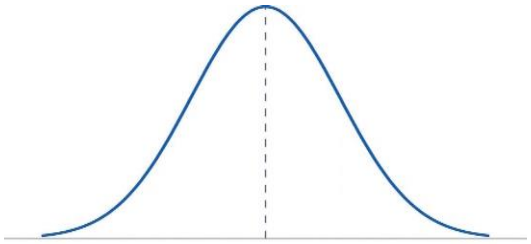
Approximately what percentage of the population is at least 72 inches tall?



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

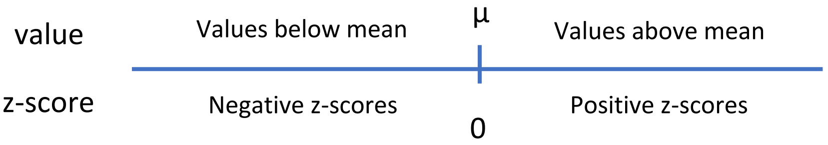


Let's try the second question again, but we'll convert this to a **Standard Normal Distribution** where μ = 0 and σ = 1



Standardization with the z-scale

- **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.
  - The score represents how many standard deviations that value is away from the mean.
  - Negative z-scores mean the data point is __________ the mean, positive z-scores mean it is __________ the mean.



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 z-score at least this far from the mean?

Introducing the z-test

Investigation: A doctor is examining whether people's diastolic blood pressure might increase as the result of taking a new experimental medication.

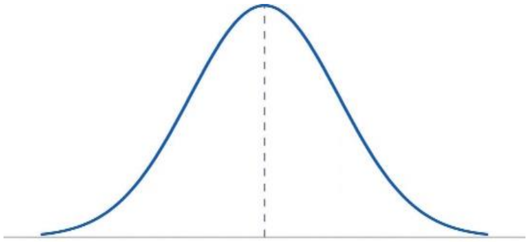
Let's say that this doctor has data from the very large hospital database, showing that the mean diastolic blood pressure for the demographic of patients he's targeting is approximately 80.0 with a standard deviation of 6.17. This distribution is *slightly* right skewed, but reasonably close to symmetric.

The doctor administers this medication to 25 patients representative of this population demographic. He finds that after 2 weeks on the medication, their average diastolic blood pressure is 83.4.

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

What approximate shape should we expect the distribution of $\overline{x}$ to have?

What mean and standard deviation will this distribution have if we use this as our null model?



Calculate the z-score for our estimate.

How do we interpret the z-score for a sample mean in the context of our hypothesis test?

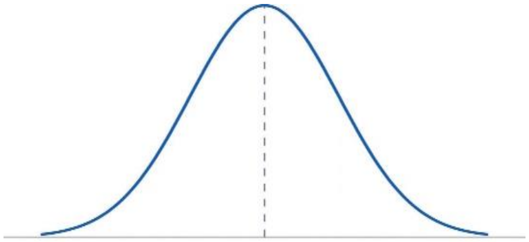
- Our sample mean is __________ *standard errors* below / above the null hypothesized mean.

We should find a p-value of about 0.029 (or 0.29%). Which statement correctly interprets 0.29%?

1. The probability that a **randomly chosen patient** would yield a **single reading** at least as high as 83.4, if the mean diastolic reading of all patients who would take this medication were really 80.0.
2. The probability that a **random sample of 25 patients** would yield a **sample mean** at least as high as 83.4, if the mean diastolic reading of all patients who would take this medication were really 80.0.

What calculation would we make if we were to use the other probability interpretation above?

Let's find the probability that a ____________________________ would yield a __________________ of 83.4 or more if assuming the mean diastolic reading as a result of this medication were still 80.0.



Note that this method only works for individual observations when we assume the population distribution is ***normally distributed!***

Reconsider the diastolic blood pressure investigation from before. However, let’s pretend that this doctor **didn’t** have a reference population to help establish a null model.

How might we still come up with a null hypothesized mean?

- In this case, the null hypothesized mean may be more of an ________________________, or a value that makes sense as a ________________ for comparison.
- Let’s say that our doctor is going to choose 80.0 to represent the mean diastolic blood pressure for his target population based on conventional wisdom and expert consultation.

If we don’t know the exact population shape, can we still infer the shape of our Null Model?

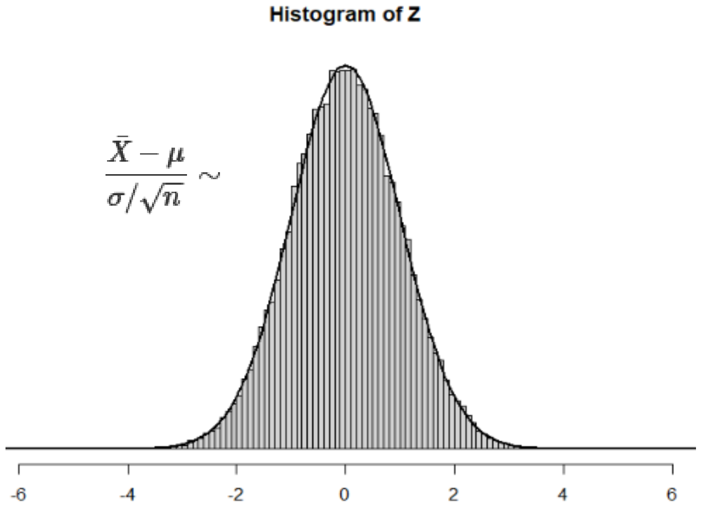
- In many cases, yes!
- As long as the population we’re sampling from isn’t too asymmetric, and our sample size is sufficiently large, we can depend on the Central Limit Theorem to assume our null model will be ___________.

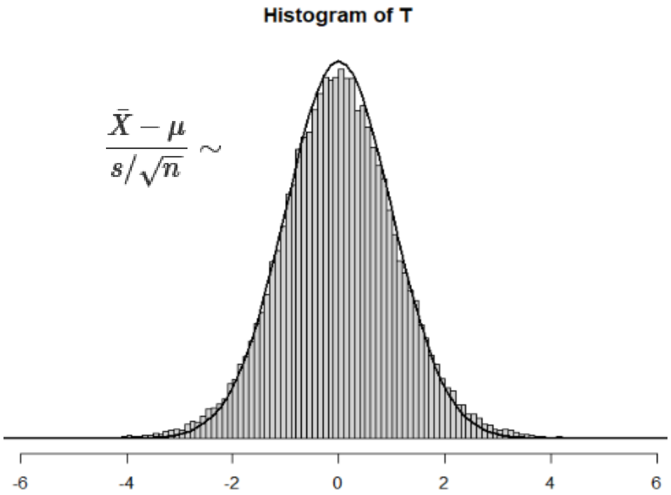
If we don’t know the population standard deviation, does that affect our calculation?

- Our best guess for the population standard deviation will now be the standard deviation of our ________.
- The main change is that to estimate the standard error in our sample mean, we’ll need to use the following approximation:

$$\mathrm{SE}_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx \frac{\square}{\sqrt{n}}$$

- So rather than find the z-score for our sample mean as its position in a standard normal distribution, we’ll calculate something called a t-score that follows a *slightly* different distribution.





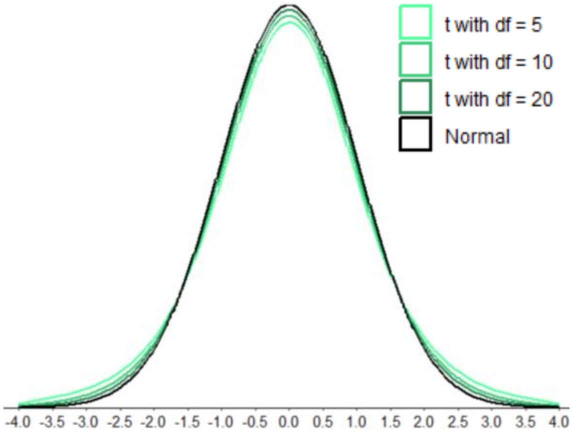
What differences do you notice between these two distributions?

Introducing the t-test

- When estimating the population standard deviation with our sample standard deviation, our method of standardization will follow what we call a t-distribution.
- The t-distribution is much like a standard normal distribution, but with slightly more variability.
  - For small sample sizes, the adjustment will be __________.
  - But as n gets larger, s will become a more consistent estimator for σ, and the t-distribution stabilizes more precisely into a standard normal distribution!

- **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 = n – 1.

- Using the t distribution simulator from the Art of Stat Web Apps page.
  - **Find Probability:** Used to calculate the probability of seeing a t-score this far or farther from 0 under the null hypothesis.
  - **Find Percentile/Quantile**: Used to find the t-score associated with a particular tail probability.



In our diastolic blood pressure question, what would be the degrees of freedom for our t-score calculation?

Back to the Investigation! Let's say that the standard deviation in diastolic blood pressure from our 25 patients was 6.52. Let's use this to estimate $\mathrm{SE}_{\bar{x}}$ and calculate the t-score for our sample mean of 83.4.

Using the t-distribution app, let's find the p-value for our investigation and interpret it in context.

Reflection Questions

4.1. In what contexts is it appropriate to use a parametric test? Can we still use parametric tests when working with non-normally distributed variables? (hint: what did we learn about the Central Limit Theorem?)

4.2. When thinking about a normally distributed variable, what does a data point's z-score represent?

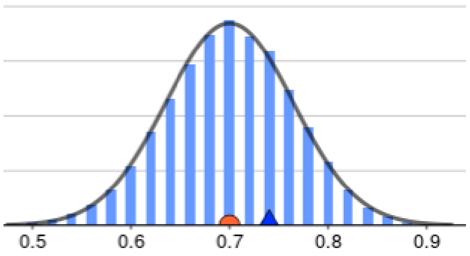
4.3. How might we interpret a z-score for a *sample mean* in the context of a hypothesis test? (hint: what is different about the denominator in that calculation?)

4.4. When completing a *t-test* for a mean (rather than a z-test), what additional uncertainty do we need to factor in? How is this reflected in the shape of a t-distribution as compared to a standard normal distribution?

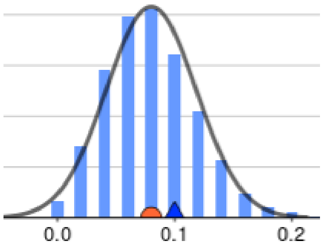
Z-test and T-test for a Mean at a glance

- Identify null and alternative hypotheses.
- Determine if t-test or z-test for mean is appropriate.
    - Z-test: σ is known (or well approximated). T-test: 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 (μ_0) and standard deviation ($SE_{\bar{x}}$) for your null model, where $SE_{\bar{x}} = \sigma/\sqrt{n}$
- Calculate the z-score or t-score for our particular estimate ($\bar{x}$): $z = \frac{\bar{x} - \mu_0}{SE_{\bar{x}}}$ $t = \frac{\bar{x} - \mu_0}{\sim SE_{\bar{x}}}$
- Find p-value: *how often would we observe a z or t-score this or more extreme in the null model?*
- Interpret the p-value and make a conclusion!

Using the z-test for a proportion

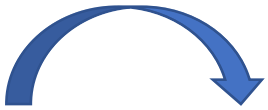


- While sample proportions technically distribute discretely, their distribution can typically be well approximated by the normal distribution with a large enough sample size!
- This wouldn't be a safe assumption in situations where the null hypothesized proportion is very close to 0 or 1 and the sample size is relatively _________.
    - Some apply the 10/10 rule.
        - we need to expect at least 10 "successes" $[n*\pi_0 \geq 1]$ 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!
    - 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.



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

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

- 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?

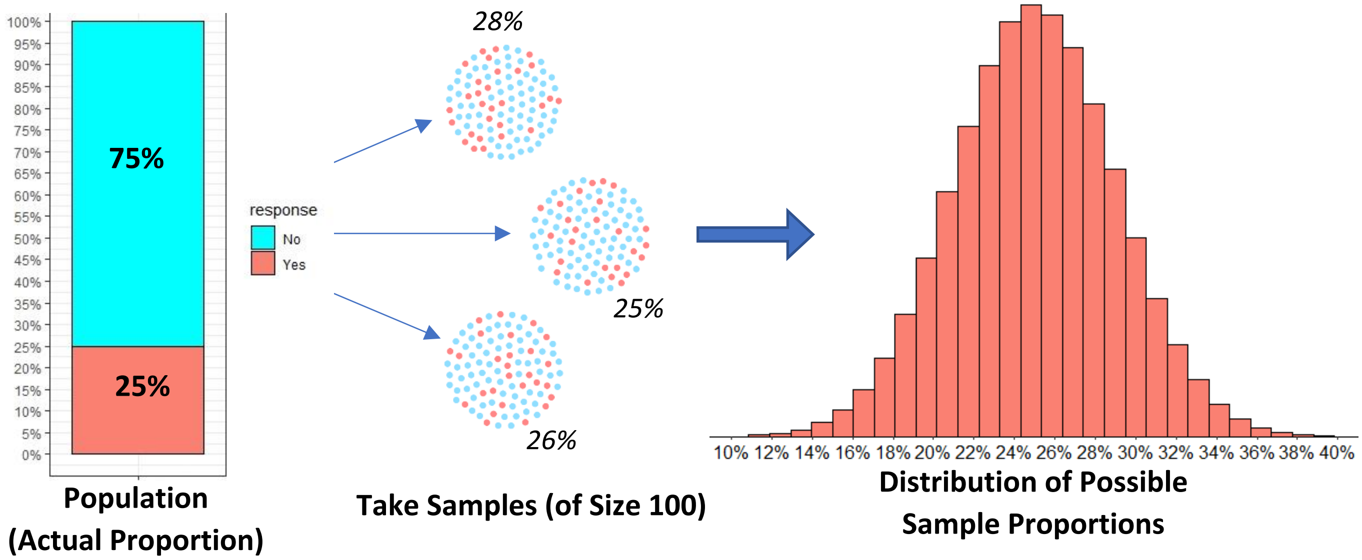
Challenge question: What value for π would maximize σ?

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 $\sigma_{\hat{p}}$
    - $SE_{\hat{p}} = \frac{\sigma}{\sqrt{n}}$ which for proportions under the null hypothesis is…

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 random sample of 100 current 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!



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

Assuming the null is true, what is the standard error for our sample proportion? *(i.e., the standard deviation of our null model)*

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%
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%.

Cautionary Notes about p-values

- **A small p-value doesn't *necessarily* mean a large or meaningful difference.**
    - **Why not?** Because we're simply testing compatibility with the null hypothesis. This can become more evident when testing from a larger sample size.
        - For example, if I flipped a dented coin 10,000 times and got 5,200 heads and got a p-value < 0.0001. What does that tell us? What does that not tell us?
    - The ___________ the sample size, the better you are at detecting differences confidently—including the ______________ differences.
- For this reason**, be careful with the term "Statistically significant."**
    - **What does it mean**? This difference is statistically difficult to explain as random chance.
    - This does **not** tell us if the difference is large or meaningful ("___________ significance")
    - For that reason, many journals and associations have suggested researchers no longer use this term due to its misinterpretation as suggesting more importance than it should.
    - To better estimate *how much* difference there is, we might use a ______________________ or effect size.

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 ($SE_{\hat{p}}$) for your null model, where $SE_{\hat{p}} = \frac{\sqrt{\pi_0(1-\pi_0)}}{\sqrt{n}}$
- Calculate the z-score for our particular estimate ($\hat{p}$) within the null model: $z = \frac{\hat{p}-\pi_0}{SE_{\hat{p}}}$
- 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.

Summary of when each test is appropriate

z-test for a proportion

- ✓ The distribution of $\hat{p}$ is reasonably approximated by a normal distribution.
  - **10/10 rule:** 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 (perhaps from using a very large sample)
- ✓ The distribution of $\bar{x}$ is normally distributed.
  - Population is normally distributed, 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 normally distributed, or sample size large enough for CLT to apply*

**How do we 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 that we need n very large (n ≥ 100) in cases where there is a large skew.

And in general…we assume our sample is representative of the population we are generalizing to! *We'll talk about how we evaluate this argument in more detail in Chapter 11.*

Reflection Questions

4.5. When working with binary data, when would it *not* be appropriate to use a z-test for a proportion?

4.6. When working with numeric data, when would it *not* be appropriate to use a parametric test (z or t-test) for a mean?

4.7. What is technically meant by the term “statistically significant”? How might this term create confusion and misinterpretation for readers?

Chapter 4 Additional Practice (if you need it!)

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 from this healthline article posits that red-headed people are more likely to be left-handed.

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 preferred 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?

Calculate the z-score for our estimate.

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

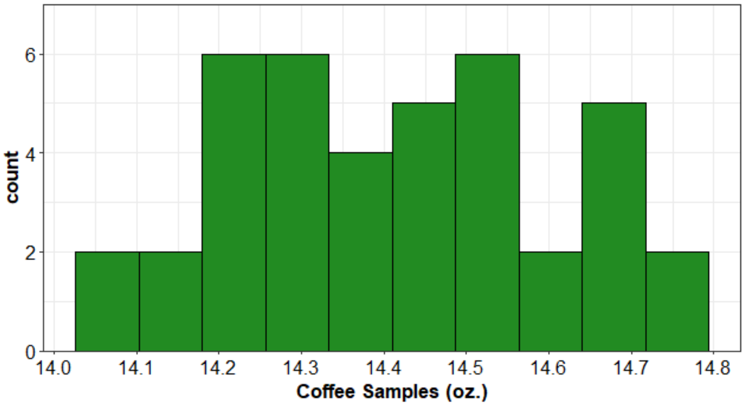
We should find a very very low p-value (approaching 0!). Which statement best explains what we found?

1. The Null Hypothesis is very likely true
2. The Null Hypothesis is very likely false

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?*

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?