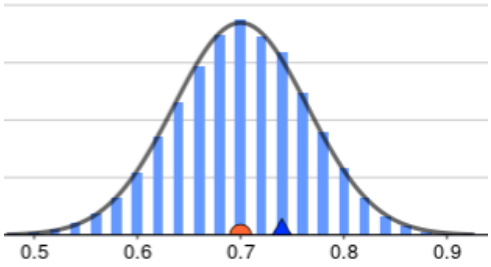
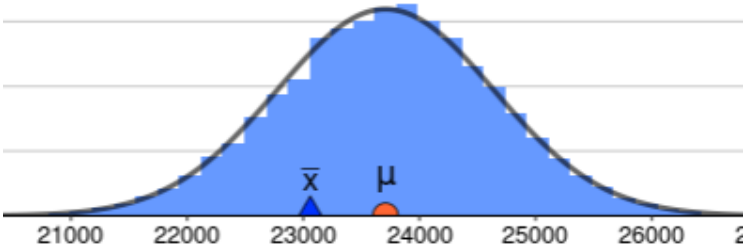
Chapter 4: z-tests and t-tests

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 work on the condition that our Null Model is normally distributed. We then use calculus to find the p-value based on the position of our sample statistic inside the null model.



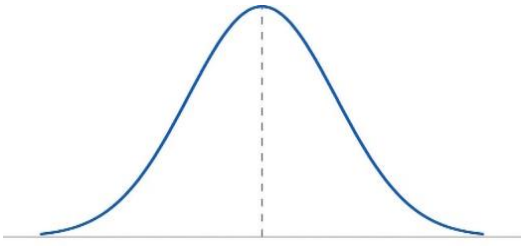


Areas under the Normal Curve

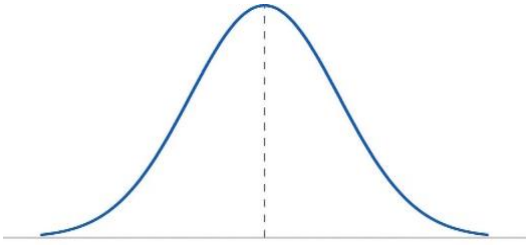
- Every Normal Distributed variable can be defined by two values
  - The _____
  - The _____
- Use the Normal distribution applet from the Art of Stat Web apps page to find the area under the normal distribution for different constraints.

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

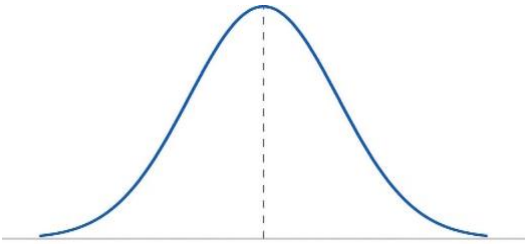
Approximately what percentage of the population is at least 72 inches tall?



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

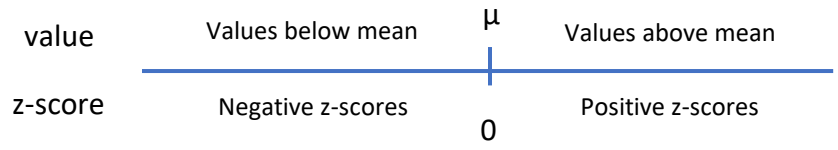


Let's try the second question again, but we'll convert this to a **Standard Normal Distribution** where μ = 0 and σ = 1



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.

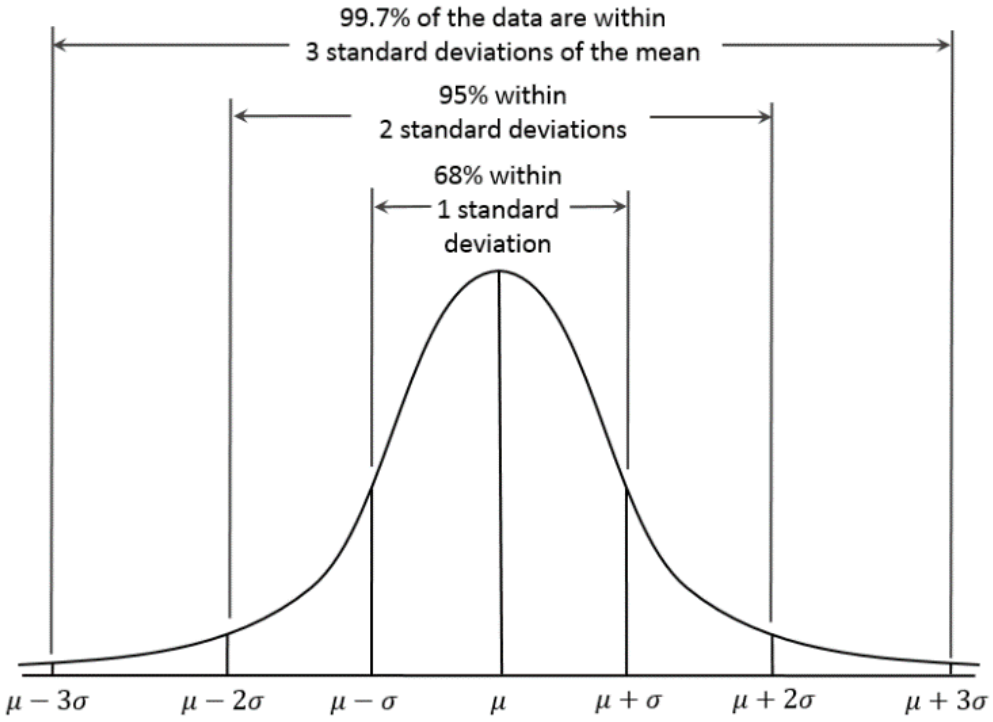


Formula for calculating z-score: $\frac{\text{Observation}-\text{mean}}{\text{standard deviation}} = \frac{x-\mu}{\sigma}$

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?

The Empirical Rule: You don't have to memorize these, but they are here if helpful!



Introducing the z-test

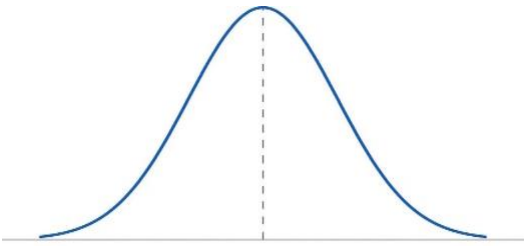
Investigation: A doctor believes that people's systolic blood pressure might increase as the result of taking a new experimental medication. Let's say that this doctor has data from the very large hospital database, showing that the mean systolic blood pressure for the demographic of patients he's targeting is approximately 128.0 with a standard deviation of 10.5. This distribution is approximately normally distributed.

The doctor administers this medication to 25 patients representative of this population demographic. He finds that after 2 weeks on the medication, their average systolic blood pressure is 133.2.

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 $\bar{x}$ to have?

What mean and standard deviation should the distribution of $\bar{x}$ have if assuming the null is true?



Calculate the z-score for our estimate.

Interpreting a z-score as a measure of position for a ***sample mean***

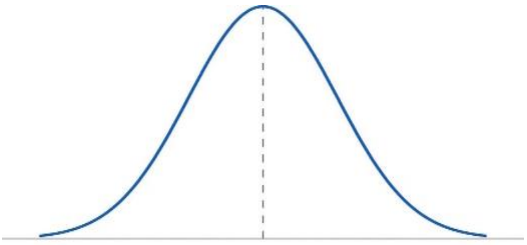
- Our sample mean is _____ *standard errors* below / above the null hypothesized mean.

We should find a p-value of about 0.066 (or 0.66%). Which statement correctly interprets 0.66%?

1. The probability that a **randomly chosen patient** would yield a **single reading** at least as high as 133.2, if the mean systolic reading of all patients who would take this medication were really 128.
2. The probability that a **random sample of 25 patients** would yield a **sample mean** at least as high as 133.2, if the mean systolic reading of all patients who would take this medication were really 128.

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 133.2 or more if assuming the mean systolic reading while taking this medication were still 128.0.



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

Reflection Questions

4.1. When is it appropriate to use a parametric test?

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

4.3. When thinking about a distribution of sample means, what does a sample mean's z-score represent?

4.4. I teach a large class, and I like to estimate what the mean exam 1 score will be early in the exam window (once only about 90 or so students have taken the exam). Distributionally, this exam tends to be somewhat (but not highly) left skewed. My historical average is 84 with a standard deviation of 10. Could I use a z-test to infer whether the mean exam score shows evidence of being different from 84 using my 90 early scores (*if* we assume the early exam takers are a representative sample)? Why or why not?

4.5. Consider my left-skewed exam score distribution. Could I use z-scores and properties of the normal distribution to estimate what proportion of my students will score below an 80? Why or why not?

Reconsider the systolic 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 128 to represent the mean systolic blood pressure for his target population based on conventional wisdom, but we don't have an actual database.

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

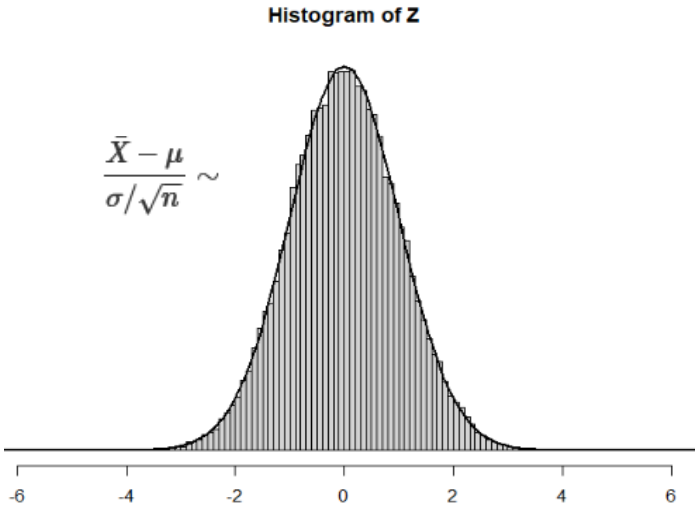
- In many cases, yes!
- As long as the population we're sampling from isn't too asymmetric, and our sample size is sufficiently large, we can depend on the Central Limit Theorem to say that our null model will be _____.

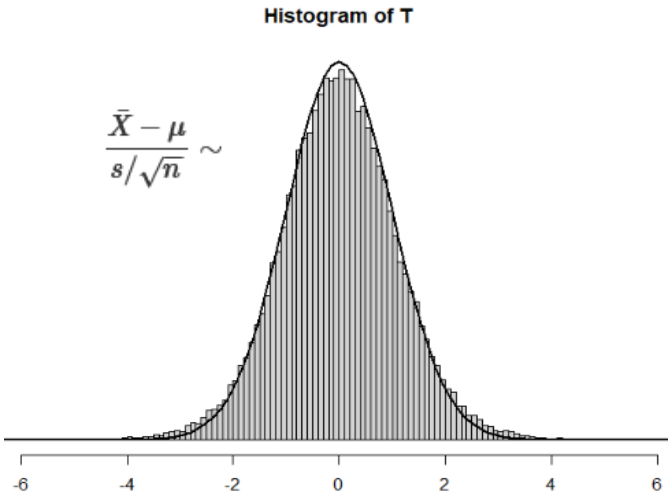
If we don't know the population standard deviation, does that affect our calculation?

- 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:

$$SE_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \approx$$

- 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.



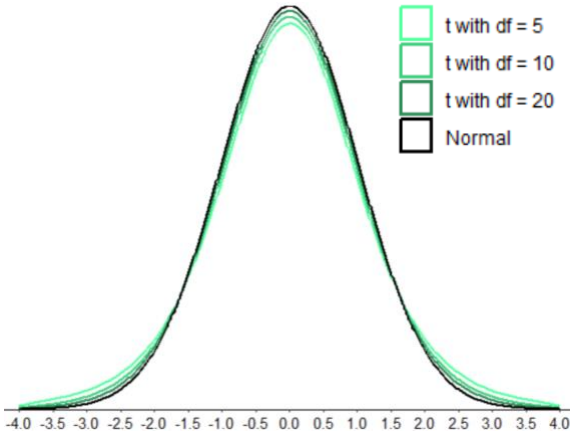


What differences do you notice between these two distributions?

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!
    - For this reason, in practice, a z-test is considered acceptable if using s from a sample size of $n \geq 100$ to estimate σ. Though the t-test is still appropriate and *generally* preferred.

- **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 applet 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 systolic 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 systolic blood pressure from our 25 patients was 10.2. Let's use this to estimate $SE_{\bar{x}}$ and calculate the t-score for our sample mean of 133.2.

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

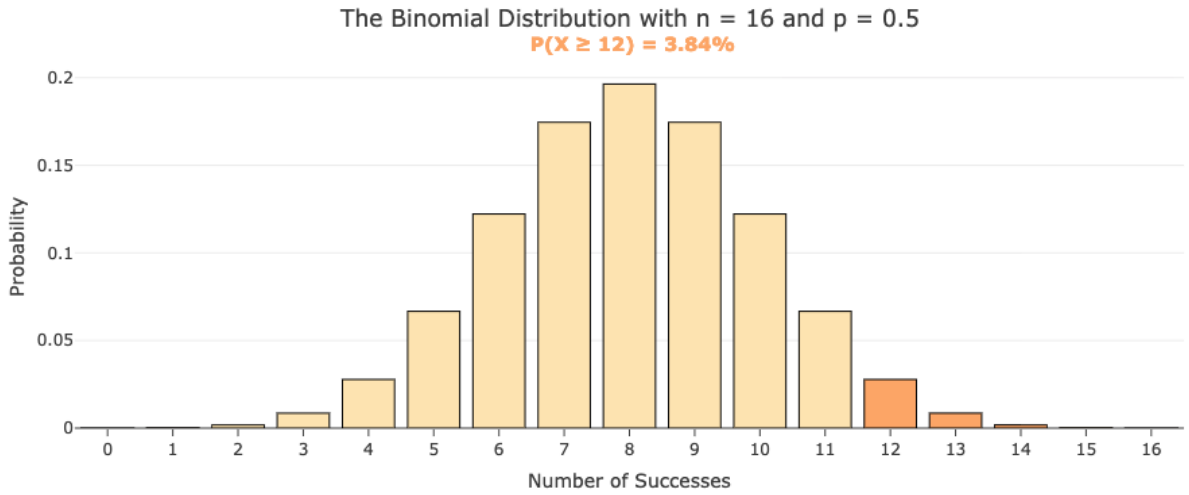
Z-test and T-test for a Mean at a glance

- 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
- Calculate the z-score or t-score for your 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!

Testing proportions using a binomial exact test

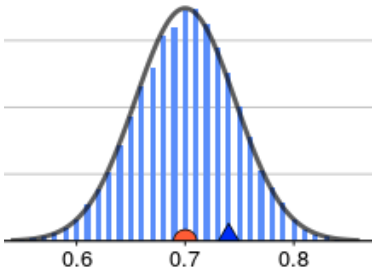
- In Chapter 2, we used a simulation to estimate how often we might see a student guess a certain number of card colors correctly out 16.
- More generally though, we noted that we can use the _____ probability distribution to calculate how often we would see different numbers of "Successes" out of n trials.
  - For example, we might calculate the probability of observing 12 successes out of 16 if each trial has a 50% probability of success. Then identify our p-value as the probability of observing at least 12 successes as our "unusualness" zone.

$$P(X = 12) = \frac{16!}{12!\,(16-12)!} \times 0.5^{12} \times (1-0.5)^{16-12} = 1820 \times 0.5^{12} \times 0.5^{4} = 0.02777$$

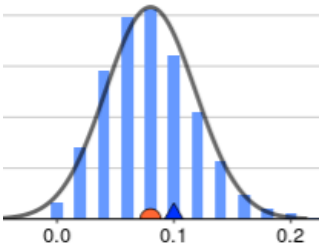


With modern day computational power, an exact binomial test is both quick and precise for testing a proportion! However, in some very large sample size cases, exact binomial tests may be computationally difficult to run. Before advanced computational tools were available, a z-test for proportions was rather common and is still found frequently in scientific literature.

Using the z-test for a proportion



- While sample proportions technically distribute discretely, their distribution may 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 _____.
- For this reason, we might apply the _____ 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.
- …and a larger sample size (we'll say $n \geq 100$).



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

- Proportions are used when we're analyzing binary data.
- 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!

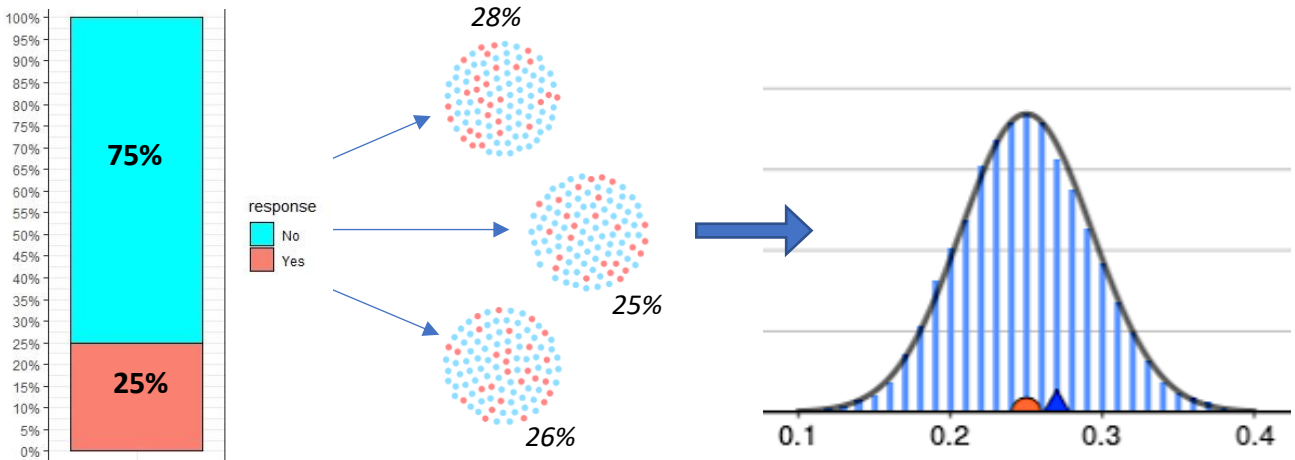
$$\sigma = \sqrt{\pi(1-\pi)}$$

Challenge question: What value for π would maximize σ?

The Standard Error of a sample proportion (when doing a z-test)

- If we treat our data as 0's and 1's, our process for calculating the SE of a sample proportion actually matches that of a sample mean!
- $SE_{\hat{p}} = \frac{\sigma}{\sqrt{n}}$ which for proportions under the null hypothesis is…

Investigation: Researchers are examining whether a new variant of the flu from the Southern Hemisphere has either grown or shrank in commonality as flu season shifts to the Northern Hemisphere. In the Southern Hemisphere last month, 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. Let's ***compare*** the results of a **z-test** to that of a **binomial exact 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?)*

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.

Let's first find the p-value using a z-test using the normal distribution applet:

Then compare it to a p-value if using the exact binomial test:

Which one is more precise?

Using the 4.87% p-value, which of the following might be an appropriate conclusion to make here?

1. If the true proportion is indeed 25%, there is a 4.87% probability of seeing a sample result at least as high as 34%
2. If the true proportion is indeed 25%, there is a 4.87% 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 becomes more evident when testing from a larger sample size.
    - For example, let's say I flipped a dented coin 10,000 times and got 5,176 heads and got a p-value < 0.001. 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 avoid using 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 _____.

Z-test for a Proportion *or* Binomial Exact Test at a glance

- Identify null and alternative hypotheses.
- If doing Binomial Exact Test
  - Plug in number of trials and null hypothesized proportion into applet/software
  - Find p-value: *how often would we observe a result at least as extreme as ours?*
- If doing a z-test
  - Ensure a normal approximation is reasonable: 1) $n \geq 100$ and 2) We expect at least 10 of each response under the null hypothesis: 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!

Summary of when each test is appropriate

Testing a Proportion

- Binomial Exact Tests (always a good choice!)
  - ✓ There are no sample size or distributional shape requirements for this one!
  - ✓ The only time an exact binomial test may not be a good choice is in *very* large sample sizes when the combinatorics calculation is very time-consuming for standard software to handle.
- Z-tests (this is a *reasonable* approach in *large-sample situations,* but has little practical value anymore)
  - ✓ Since sample proportions distribute discretely, then using a continuous normal distribution to approximate the tail areas is only reliable if we have a very dense (large sample) distribution. (10/10 rule and $n \geq 100$).
  - ✓ Many scientists still use z-tests for proportion, but it's no longer of much value with powerful computational tools available. *Just use a binomial exact test!*

Testing a Mean

- T-test (a good choice at any sample size if normality condition met, but *especially* advantageous in *small* samples)
  - ✓ We are using s to approximate σ and accounting for that in the method!
  - ✓ The distribution of $\bar{x}$ is normally distributed (either because the variable we're sampling from is already normally distributed, or because the sample size is large enough for CLT to apply*).
  - ✓ Note that a t-test is not "wrong" to do in large sample situations—rather that its results converge to those of a z-test as sample size increases and as $s \approx \sigma$
- Z-test (generally used in large-sample situations)
  - ✓ σ is known or well approximated (perhaps estimating using s when $n \geq 100$)
  - ✓ The distribution of $\bar{x}$ is normally distributed (either because the variable we're sampling from is already normally distributed, or because the sample size is large enough for CLT to apply*).
- Non-parametric options
  - ✓ If the normality condition isn't met, there are some "Non-parametric test" options, like a *Wilcoxon Signed Rank Test*! Bootstrapping is also quite popular as a simulation-based approach.

**How do we know if n is large enough for CLT to apply?*

- **For this class**, we'll say…
  - Any sample size is fine if the population we're sampling from is normally distributed.
  - $n \geq 10$ is a good benchmark if the population is relatively symmetric, but not normal.
  - $n \geq 30$ is a good benchmark when the population is asymmetric, but only mild skewness
  - We need n very large ($n \geq 100$, sometimes more!) in cases where there is a large skew.
  - *In practice, statisticians might check if the normality condition is met with bootstrapping.*

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.*

Digging Deeper: How do we mathematically derive $\sigma = \sqrt{\pi(1-\pi)}$ for binary data?

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i-\pi)^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2-2x_i\pi+\pi^2}{n}}$$

But recall that all data (x_i) are either 1's or 0's. And since $0^2 = 0$ and $1^2 = 1$, then $x_i^2 = x_i$ for all x_i.

Thus… $\sqrt{\frac{\sum_{i=1}^{n} x_i^2-2x_i\pi+\pi^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} x_i-2x_i\pi+\pi^2}{n}}$

Next, let's break this into 3 terms: $\sqrt{\frac{\sum_{i=1}^{n} x_i-2x_i\pi+\pi^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} x_i}{n} - \frac{2\sum_{i=1}^{n} x_i\pi}{n} + \frac{\sum_{i=1}^{n}\pi^2}{n}}$

Note that $\frac{\sum_{i=1}^{n} x_i}{n} = \pi$. Therefore…$= \sqrt{\pi - 2\pi^2 + \frac{n\pi^2}{n}} = \sqrt{\pi - 2\pi^2 + \pi^2} = \sqrt{\pi - \pi^2} = \sqrt{\pi(1-\pi)}$

Thus, when your data is recorded strictly as 1's and 0's, then π communicates to you not only the center of your distribution, but also how much variation there is in the responses!

Reflection Questions

4.6. 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?

4.7. When working with binary data, which of the two tests that we learned is more precise and generally preferred? In what cases might the other be a reasonable approximation to use?

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

Chapter 4 Additional Practice (Videos available in the Ch 4 module on Canvas!)

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 reasonably 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.

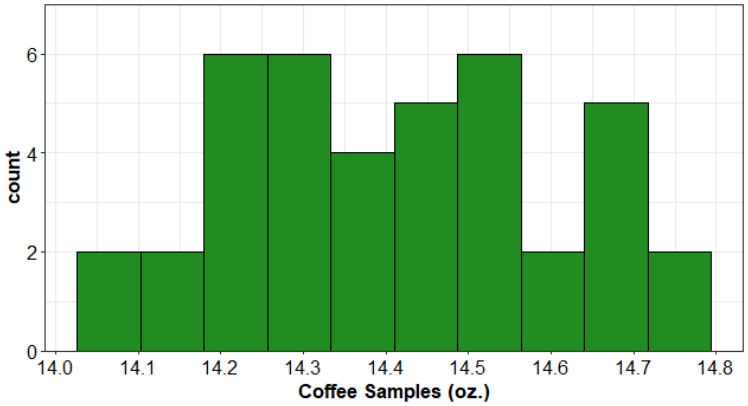
Regardless of whether we use the binomial exact test method or the z-test, we should find a very very low p-value (approaching 0!). Which statement best explains what we found?

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 $\alpha = 0.05$. 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?