Statistical Inference for One-sample quantitative data

Ryan Miller

Outline

- Video #1
  - Confidence intervals and the t-distribution
- Video #2
  - The one-sample t-test
- Video #3
  - Model assumptions for using the t-distributions
- Video #4
  - Simulation-based alternatives to statistical inference on quantitative data

Introduction

So far, we've covered two major types of *statistical inference*:

1) **Confidence interval estimation** - using sample data to derive a range of plausible estimates for some characteristic of a population
2) **Hypothesis testing** - using sample data to evaluate a conjecture about a population

Introduction

So far, we've covered two major types of *statistical inference*:

1) **Confidence interval estimation** - using sample data to derive a range of plausible estimates for some characteristic of a population
2) **Hypothesis testing** - using sample data to evaluate a conjecture about a population

- We've now seen numerous examples of these approaches for *one-sample categorical data*, or scenarios that can be summarized by a *single proportion*
- This week, we'll see how the approach differ for *one-sample quantitative data*, or scenarios that can be summarized by a *single mean*

Interval Estimation

- Recall that we've used Normal probability models as the basis for constructing $P\%$ confidence intervals:

$$\text{Point Estimate} \pm z^* SE$$

- Until now, the only population characteristic we've considered estimating is the *population proportion*, or p:

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p} * (1 - \hat{p})}{n}}$$

- Recall that the SE formula was based upon the results of the *Central Limit theorem*

CLT for Means

- As a refresher, CLT states that the distribution of the sample average will converge to a Normal distribution with a mean equal to the expected value of an individual data-point and a standard error equal to the standard deviation of an individual data-point divided by $\sqrt{n}$

CLT for Means

- As a refresher, CLT states that the distribution of the sample average will converge to a Normal distribution with a mean equal to the expected value of an individual data-point and a standard error equal to the standard deviation of an individual data-point divided by $\sqrt{n}$
- For proportions, the *standard error*, SE, was the square root of the variance of a single data-point divided n
    - $SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$
    - This formula is easy to use since n is known and p is estimated by $\hat{p}$

CLT for Means

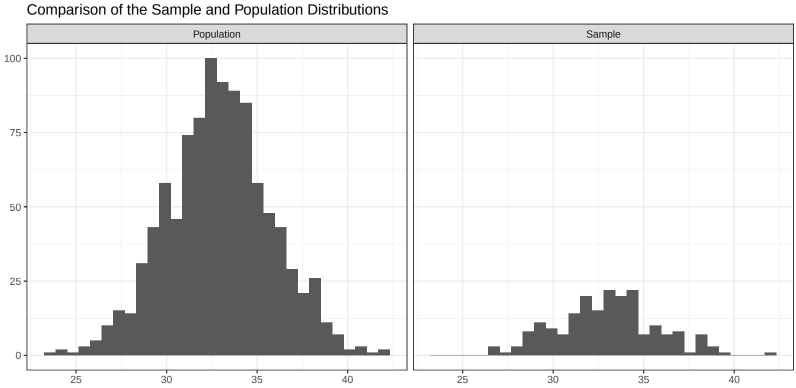
- As a refresher, CLT states that the distribution of the sample average will converge to a Normal distribution with a mean equal to the expected value of an individual data-point and a standard error equal to the standard deviation of an individual data-point divided by $\sqrt{n}$
- For proportions, the *standard error*, SE, was the square root of the variance of a single data-point divided n
  - $SE(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$
  - This formula is easy to use since n is known and p is estimated by $\hat{p}$
- For means, $SE(\bar{x}) = \sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$
  - This formula is more challenging because we don't know σ (the standard deviation of cases in the population)

Interval Estimation for Quantitative Data

- A simple solution is to estimate σ (the standard deviation of cases in the population) using the sample data
  - The standard deviation of the cases in the sample is denoted by s, but is it really valid to use s in place of σ when estimating the population mean?

Interval Estimation for Quantitative Data

- A simple solution is to estimate σ (the standard deviation of cases in the population) using the sample data
  - The standard deviation of the cases in the sample is denoted by s, but is it really valid to use s in place of σ when estimating the population mean?



William Gosset

- William Gosset was an English chemist working for Guinness Brewing in the 1890s
  - At Guinness, Gosset's role was to statistically evaluate the yields of different varieties of barley
  - Through his work, Gosset began to question the validity of the Central Limit Theorem's results for small samples

William Gosset

- William Gosset was an English chemist working for Guinness Brewing in the 1890s
  - At Guinness, Gosset's role was to statistically evaluate the yields of different varieties of barley
  - Through his work, Gosset began to question the validity of the Central Limit Theorem's results for small samples
- In 1906, Gosset took a leave of absence to go work with Karl Pearson (creator of the correlation coefficient) on the problem

Student's t-distribution

- Gosset discovered the flaw was due to using the sample standard deviation, s, in place of the population standard deviation, σ
  - As you'd expect, s is not a perfect estimate of σ, especially when the sample size is small

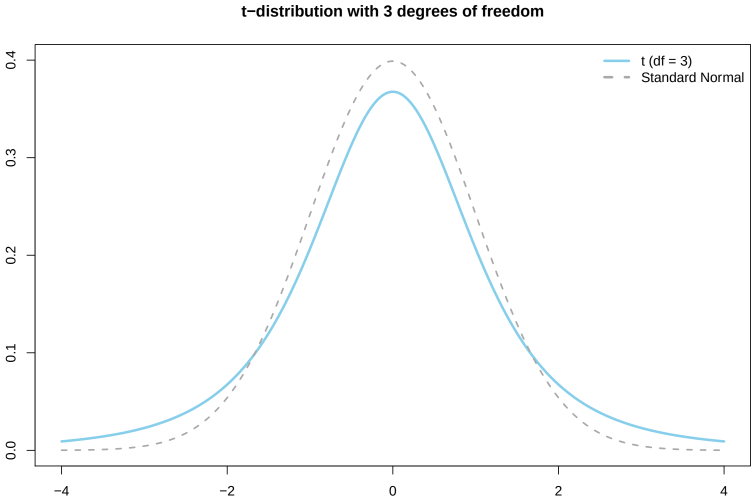
Student's t-distribution

- Gosset discovered the flaw was due to using the sample standard deviation, s, in place of the population standard deviation, σ
  - As you'd expect, s is not a perfect estimate of σ, especially when the sample size is small
- Simply "plugging in" s into the CLT result introduces a new source of variability (due to the imperfect estimation of σ)

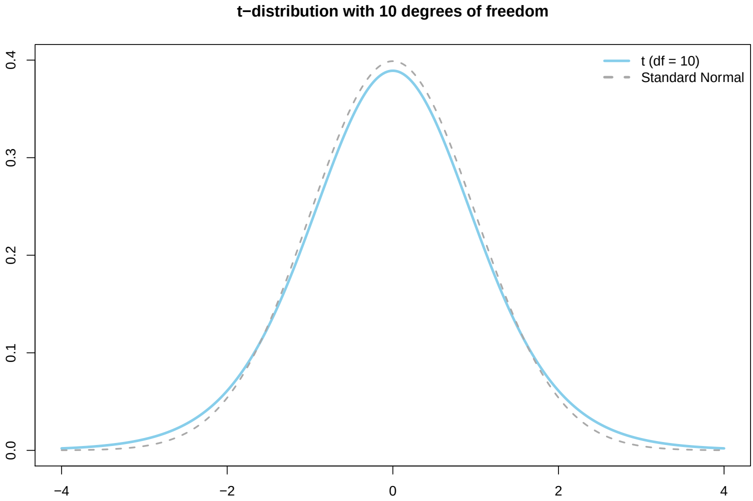
Student's t-distribution

- Gosset discovered the flaw was due to using the sample standard deviation, s, in place of the population standard deviation, σ
  - As you'd expect, s is not a perfect estimate of σ, especially when the sample size is small
- Simply "plugging in" s into the CLT result introduces a new source of variability (due to the imperfect estimation of σ)
- Usually the person who discovers an important results gets to name it
  - However, Gosset had to publish his work under the name "Student" because Guinness didn't want competitors knowing it employed statisticians!
  - Gosset's result, called Student's t-distribution, is among the most widely-used statistical results of all time

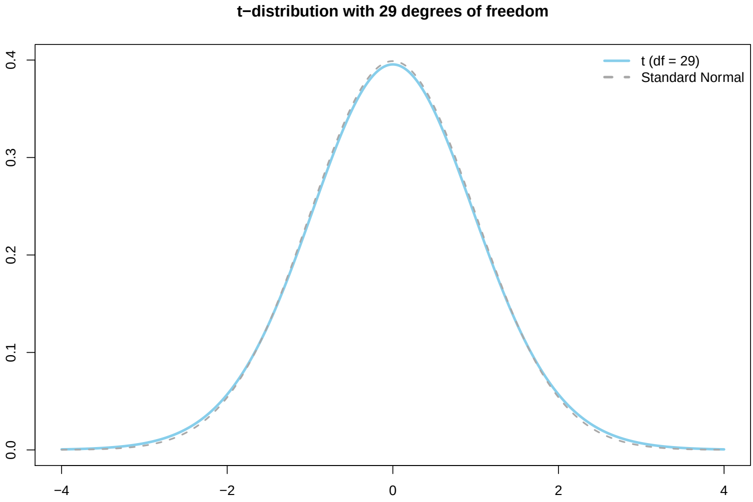
The t-distribution



The t-distribution



The t-distribution



How to use the t-distribution

When estimating a single mean, we use the t-distribution to construct a $P\%$ confidence interval via:

$$\bar{x} \pm t^*_{n-1} \frac{s}{\sqrt{n}}$$

- t^*_{n-1} is a percentile from the t-distribution with $n-1$ degrees of freedom defining the middle $P\%$ of the distribution
- $\frac{s}{\sqrt{n}}$ is the *standard error* (SE) of the sample mean, $\bar{x}$

Example

- While waiting at an airport, a passenger notices 6 flights to similar a similar part of the country were delayed 6, 10, 13, 23, 45, 55 minutes
  - The mean delay of this sample was 25.33
  - The standard deviation of delays in the sample was 20.2
- Assuming these data are representative, use them to come up with a 95% confidence interval estimate for the average flight delay from this airport to your destination

Example (solution)

- 95% CI for a population mean: Point Estimate $\pm$ MOE
  - Point estimate $= \bar{x} = 25.33$
  - Margin of error $= t^*_{df=5} * SE = 2.571 * \frac{20.2}{\sqrt{6}}$

Example (solution)

- 95% CI for a population mean: Point Estimate $\pm$ MOE
  - Point estimate $= \bar{x} = 25.33$
  - Margin of error $= t^*_{df=5} * SE = 2.571 * \frac{20.2}{\sqrt{6}}$
- All together, 95% CI: $25.33 \pm 2.571 * \frac{20.2}{\sqrt{6}} = (4.1, 46.5)$
  - We are 95% confident the average delay is somewhere between 4.1 minutes and 46.5 minutes

Example (solution)

- 95% CI for a population mean: Point Estimate $\pm$ MOE
  - Point estimate $= \bar{x} = 25.33$
  - Margin of error $= t^*_{df=5} * SE = 2.571 * \frac{20.2}{\sqrt{6}}$
- All together, 95% CI: $25.33 \pm 2.571 * \frac{20.2}{\sqrt{6}} = (4.1, 46.5)$
  - We are 95% confident the average delay is somewhere between 4.1 minutes and 46.5 minutes
- Note: if we'd erroneously used a Normal model, we'd get an interval that is much narrower (9.2, 41.5), but this interval wouldn't have the confidence level we are advertising (ie: it wouldn't really be a 95% CI because it would miss too often)

Closing Remarks

- We've now introduced the t-distribution, a necessary modification to the Normal model in scenarios involving a quantitative data
  - In these situations, the standard error required estimating an extra parameter, thus the t-distribution modifies the Normal model to account for this added uncertainty
- In this class, you should generally expect to use the t-distribution for means and the Normal distribution for proportions (aside a few exceptions where assumptions aren't met)
  - We will cover exceptions in a later video

Introduction

Previously, we introduced the Z-test:

1) State the null hypothesis (a conjecture about the population that would be useful to disprove)
2) Use the null hypothesis (and corresponding null model) to find a Z-value describing the *sample estimate*
3) Locate the Z-value on the Standard Normal curve to find the p-value
4) Use the p-value to make a decision regarding the null hypothesis

Introduction

Previously, we introduced the Z-test:

1) State the null hypothesis (a conjecture about the population that would be useful to disprove)
2) Use the null hypothesis (and corresponding null model) to find a Z-value describing the *sample estimate*
3) Locate the Z-value on the Standard Normal curve to find the p-value
4) Use the p-value to make a decision regarding the null hypothesis

This approach needs to be adjusted for scenarios involving means (since there is added uncertainty induced by estimating an extra parameter), the resulting procedure is known as the T-test

The T-test

- Procedurally, only difference between the T-test and Z-test is the probability distribution used to calculate the p-value
  - When analyzing one-sample *categorical* data, the Z-test compares $z = \frac{\hat{p}-p}{SE}$ to the Standard Normal distribution
  - When analyzing one-sample *quantitative* data, the T-test compares $t = \frac{\bar{x}-\mu}{SE}$ to a t-distribution with $df = n - 1$

Example

- According to thr Australian government, the mean birthweight of all babies born in Australia is 7.86 lbs
- A hospital in Missouri reports the average birthweight of 112 born there last year was 7.68, with a sample standard deviation of 1.31
- Assuming this Missouri hospital is a representative sample of all babies born in the US, do these data support the hypothesis that birthweight of US babies is different from that of Australian babies?

Example (solution)

1) $H_0 : \mu = 7.86$ vs $H_A : \mu \neq 7.86$

Example (solution)

1) $H_0 : \mu = 7.86$ vs $H_A : \mu \neq 7.86$
2) Noting that we observed $\bar{x} = 7.68$, and $SE = \frac{s}{\sqrt{n}}$, we calculate $t = \frac{7.68 - 7.86}{1.31/\sqrt{112}} = -1.45$

Example (solution)

1) $H_0 : \mu = 7.86$ vs $H_A : \mu \neq 7.86$
2) Noting that we observed $\bar{x} = 7.68$, and $SE = \frac{s}{\sqrt{n}}$, we calculate $t = \frac{7.68-7.86}{1.31/\sqrt{112}} = -1.45$
3) We next must locate $t = -1.45$ on a t-distribution with $df = n - 1 = 111$ using StatKey

Example (solution)

1) $H_0 : \mu = 7.86$ vs $H_A : \mu \neq 7.86$
2) Noting that we observed $\bar{x} = 7.68$, and $SE = \frac{s}{\sqrt{n}}$, we calculate $t = \frac{7.68-7.86}{1.31/\sqrt{112}} = -1.45$
3) We next must locate $t = -1.45$ on a t-distribution with $df = n - 1 = 111$ using StatKey
4) The two-sided p-value is 0.15, so we conclude insufficient evidence to believe the mean birthweight of babies in the US differs from that of Australia

Comparison vs. The Z-test

- As previously mentioned, the only difference between the T-test and Z-test is null distribution
  - We use the T-test to account for the small amount of additional uncertainty introduced when using the sample standard deviation, s, to estimate σ, the standard deviation of the population

Comparison vs. The Z-test

- As previously mentioned, the only difference between the T-test and Z-test is null distribution
  - We use the T-test to account for the small amount of additional uncertainty introduced when using the sample standard deviation, s, to estimate σ, the standard deviation of the population
- Thus, we know the p-value of a T-test will always be higher than the corresponding Z-test
  - Here, the p-value is 0.15 comparing our T-value of -1.45 to the $t_{df=111}$ distribution
  - If we compared -1.45 to the Standard Normal curve, we'd get a p-value of 0.148

Closing Remarks

- The T-test is a modified version of the Z-test that makes the Normal results of CLT suitable for statistical inference on quantitative data
  - You should expect to use the Z-test for hypothesis testing on proportions (categorical data)
  - You should expect to use the T-test for hypothesis testing on means (quantitative data)

Conditions for the Normal model (one proportion)

When performing statistical inference on a *proportion*, we've used the following Normal model:

$$\hat{p} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$$

- This model works well when $n * p \geq 10$ and $n * (1 - p) \geq 10$

Conditions for the Normal model (one proportion)

When performing statistical inference on a *proportion*, we've used the following Normal model:

$$\hat{p} \sim N(p, \sqrt{\tfrac{p(1-p)}{n}})$$

- This model works well when $n * p \geq 10$ and $n * (1 - p) \geq 10$
- In hypothesis testing, we use this model to determine what might have been observed in our sample if H_0 were true
  - For this reason, we use value specified in H_0 in place of the unknown population parameter, p

Conditions for the Normal model (one proportion)

When performing statistical inference on a *proportion*, we've used the following Normal model:

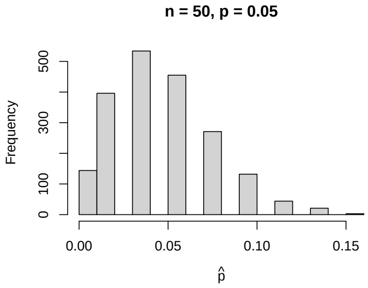
$$\hat{p} \sim N(p, \sqrt{\tfrac{p(1-p)}{n}})$$

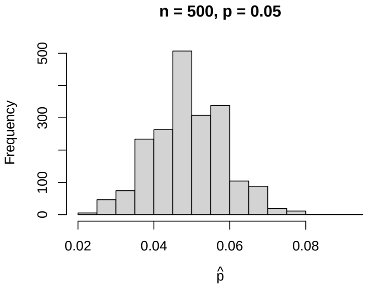
- This model works well when $n * p \geq 10$ and $n * (1-p) \geq 10$
- In hypothesis testing, we use this model to determine what might have been observed in our sample if H_0 were true
  - For this reason, we use value specified in H_0 in place of the unknown population parameter, p
- In confidence interval estimation, we use this model to determine the variability of possible sample proportions
  - For this reason, we used our best estimate of p, which is the sample proportion $\hat{p}$

Examples of Violations (proportion)

The conditions $n * p \geq 10$ and $n * (1 - p) \geq 10$ can be violated in two ways:

1) p is too close to a boundary value (a proportion of 0 or 1) relative to the sample size

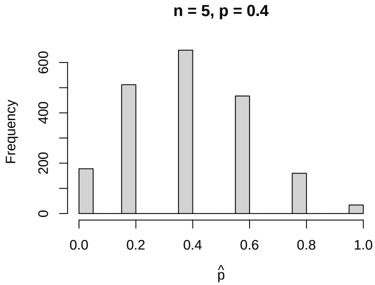


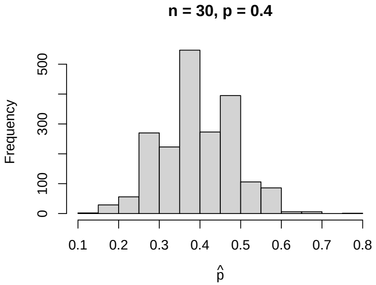


Examples of Violations (proportion, part 2)

The conditions $n * p \geq 10$ and $n * (1 - p) \geq 10$ can be violated in two ways:

2) p isn't near a boundary, but n is too small





Conditions for the t-distribution (one mean)

When performing statistical inference on a *mean*, we've used the t-distribution:

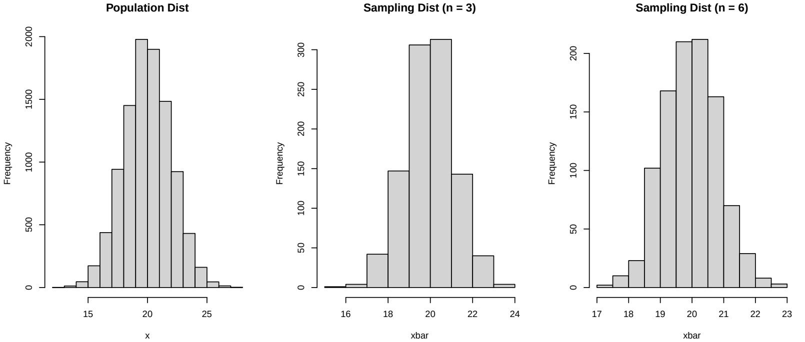
$$\frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

This model works well in two situations:

1) the population we sampled from is Normally distributed (regardless of sample size)
2) the sample size is large ($n \geq 30$)

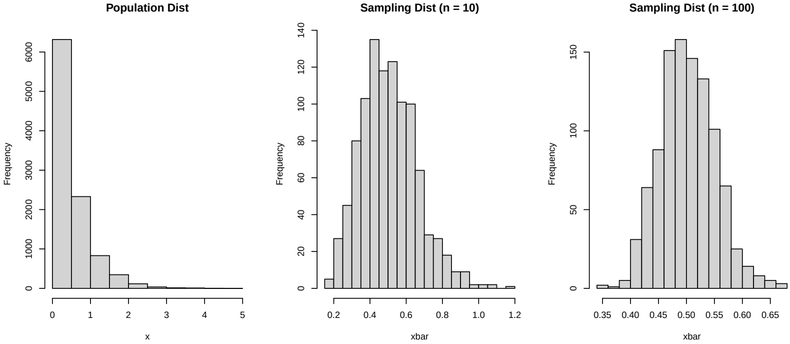
Examples of Violations (mean)

An illustration of the first situation (Normal population, any sample size):



Examples of Violations (mean, part 2)

An illustration of the second situation (Skewed population, large samples):



Comments

- Each of the examples used in the lecture are *hypothetical* in the sense that we'd never be able to see thousands of replications of any real-world study
  - That said, they illustrate the importance of checking the conditions that are recommended for the models we've using

Comments

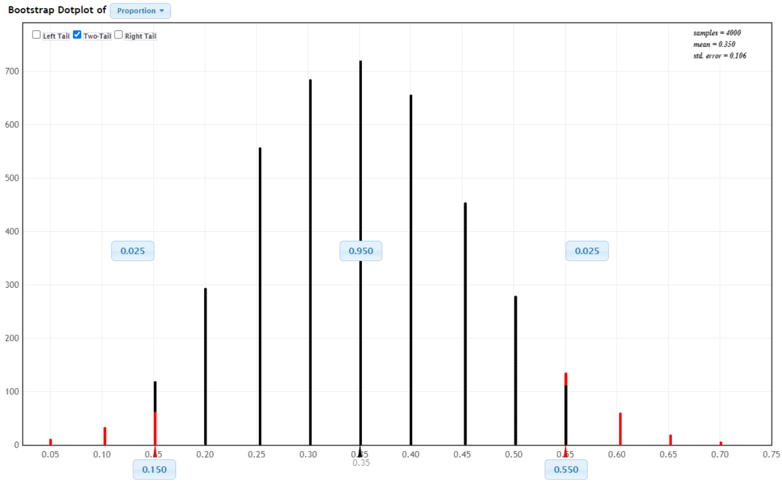
- Each of the examples used in the lecture are *hypothetical* in the sense that we'd never be able to see thousands of replications of any real-world study
  - That said, they illustrate the importance of checking the conditions that are recommended for the models we've using
- But can we still do inference when these conditions aren't met?
  - The answer is yes, but we'll need to estimate the sampling/null distribution in another way (simulation)

Simulation for One Proportion (CI)

- Consider a large calculus class at a University
  - In a survey of 20 students from this class, only 7 report getting an A or B on a midterm exam
- Can these data be used to estimate the proportion of the *entire* class who received an A or B?
  - Notice $n * \hat{p} = 20 * \frac{7}{20} = 7$, which does not meet the conditions for using a Normal model

Simulation for One Proportion (CI) - solution

Using simulation via StatKey, we estimate with 95% confidence that between 15% and 55% of the class got an A or B

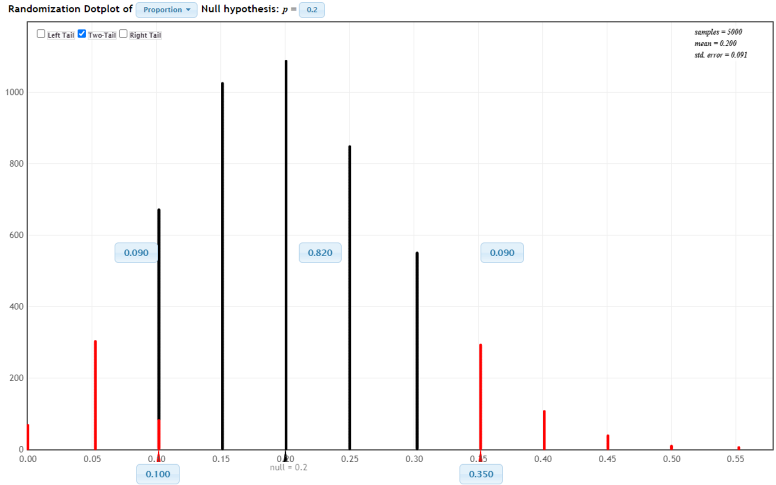


Simulation for One Proportion (Testing)

- Does this sample (where only 7 of 20 reported getting an A or B) provide convincing evidence that more than 20% of the class got an A or B?
  - $H_0 : p = 0.2$
  - Notice $n * p = 20 * 0.2 = 4$, which does not meet the conditions for using a Normal model

Simulation for One Proportion (Testing) - solution

Using simulation via StatKey, the two-sided p-value of this test is approximately 0.18

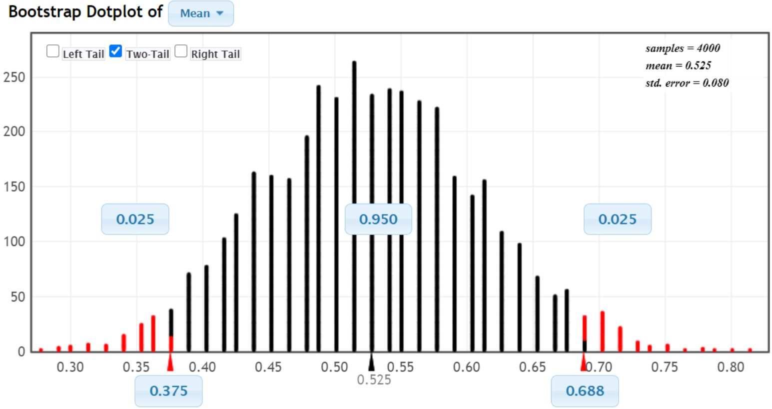


Simulation for One Mean (CI)

- The EPA recommends homeowners take action when radon levels above 0.4 pCi/L are consistently present
  - Suppose the basement of a home is tested on 8 randomly selected dates, and resulting in the following measurements $\{2, .7, .3, .9, .5, .3, .7, .6\}$
- Can these data be used to estimate the true radon levels of this home?
  - Notice the sample size is small and we aren't sure if the population being sampled is Normally distributed

Simulation for One Mean (CI)

Using simulation via StatKey, the 95% *bootstrapped* confidence interval is (0.375, 0.688)

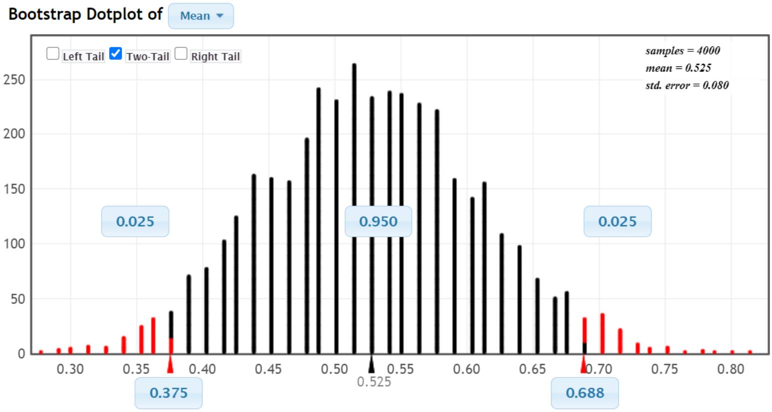


Simulation for One Mean (Testing)

- The EPA recommends homeowners *requires* action if random levels are above 4 pCi/L
  - Suppose the basement of a home is tested on 8 randomly selected dates, and resulting in the following measurements $\{2, .7, .3, .9, .5, .3, .7, .6\}$
  - Do these 8 measurements provide sufficient evidence that the EPA *does not* need to intervene? (ie: evidence that $\mu < 4$)

Simulation for One Mean (Testing)

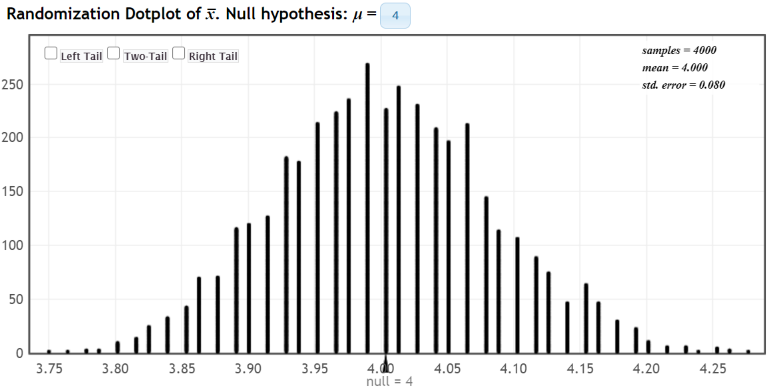
Using simulation via StatKey, the 95% *bootstrapped* confidence interval is (0.375, 0.688)



- Are these samples *convincing evidence* that the basement's radon levels are below 4 pCi/L?
  - Again, recognize the sample size is small and that we aren't

Simulation for One Mean (Testing)

Using simulation via StatKey, a *randomization test* provides a p-value of essentially zero (recall $\bar{x} = 0.525$)



Conclusion

- This lecture reviewed the conditions necessary for responsibly using probability models inspired by the Central Limit theorem for statistical inference

Conclusion

- This lecture reviewed the conditions necessary for responsibly using probability models inspired by the Central Limit theorem for statistical inference
- It also introduced simulation-based alternatives that can be used when these conditions are not met
  - In this class, I am less concerned with you being able to execute these simulation-based approaches, and more concerned with your ability to identify situations when they are warranted (ie: violated conditions)

Conclusion

- This lecture reviewed the conditions necessary for responsibly using probability models inspired by the Central Limit theorem for statistical inference
- It also introduced simulation-based alternatives that can be used when these conditions are not met
  - In this class, I am less concerned with you being able to execute these simulation-based approaches, and more concerned with your ability to identify situations when they are warranted (ie: violated conditions)
- Recognize that *p*-values and confidence intervals obtained via simulation are interpreted identically to those obtained using more traditional methods
  - That is, a confidence interval always describes a range of plausible values for a population parameter
  - A *p*-value always measures how compatible the sample data are with a null hypothesis