Model Assumptions and Alternative Apporaches to Inference

Ryan Miller

Introduction

- Lately we've been relying upon probability models as the basis for statistical inference
  - We've used the Normal distribution as a model for the distribution of a *sample proportion*
  - We've used the T-distribution as a model for the distribution of a *sample mean*

Introduction

- Lately we've been relying upon probability models as the basis for statistical inference
  - We've used the Normal distribution as a model for the distribution of a *sample proportion*
  - We've used the T-distribution as a model for the distribution of a *sample mean*
- This lecture will recap the underlying conditions necessary for those models to be a reasonable approximation of reality
  - We will also introduce *simulation* as an alternative approach to inference when these conditions are not satisfied

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:

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

- 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:

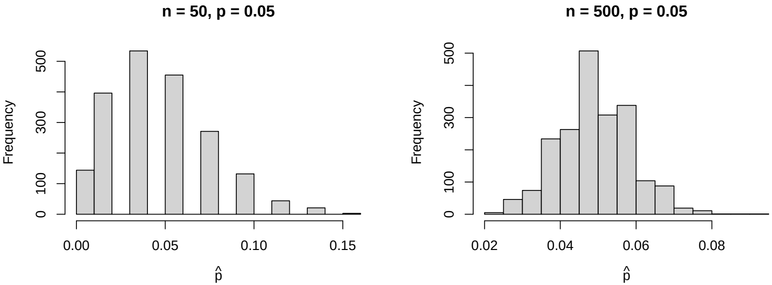
$$\hat{p} \sim N(p, \sqrt{\frac{p(1-p)}{n}})$$

- 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:

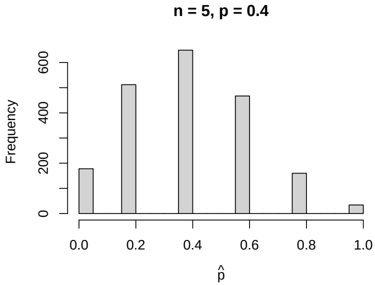
1) p is too close to a boundary value (a proportion of 0 or 1) relative to the sample size

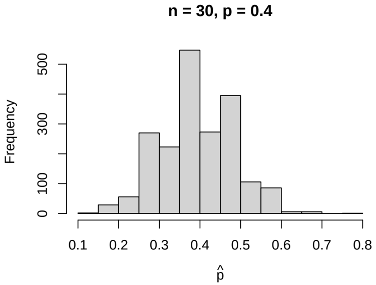


Examples of Violations (proportion, part 2)

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:

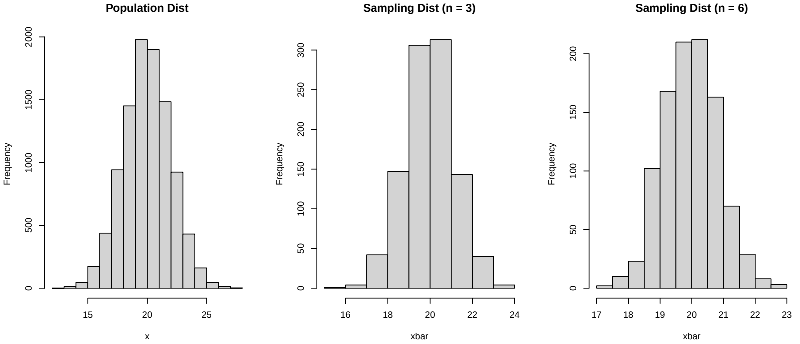
$$\frac{\bar{x} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

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 ($n \geq 30$)

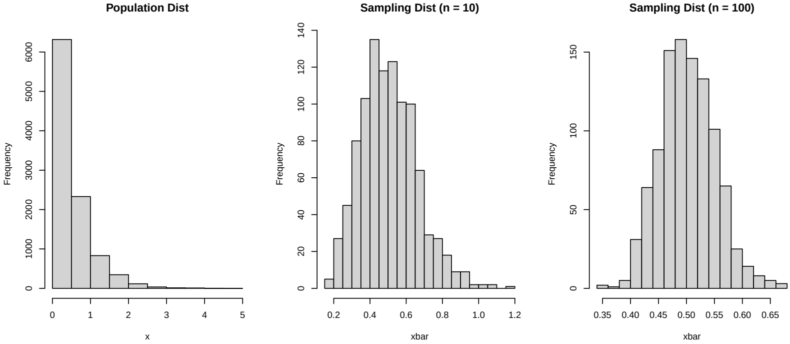
Examples of Violations (mean)

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

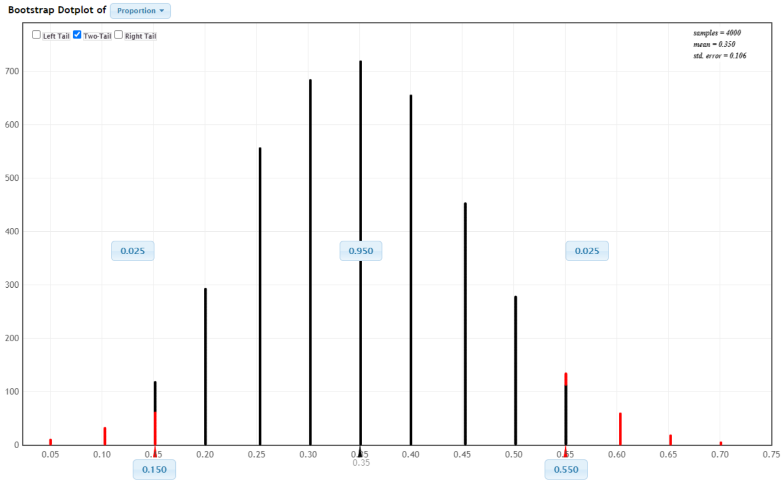
- Each of the examples used in the lecture are *hypothetical* in the sense that we'd never be able to see thousands of replications of any real-world study
  - That said, they illustrate the importance of checking the conditions that are recommended for the models we've using
- But can we still do inference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

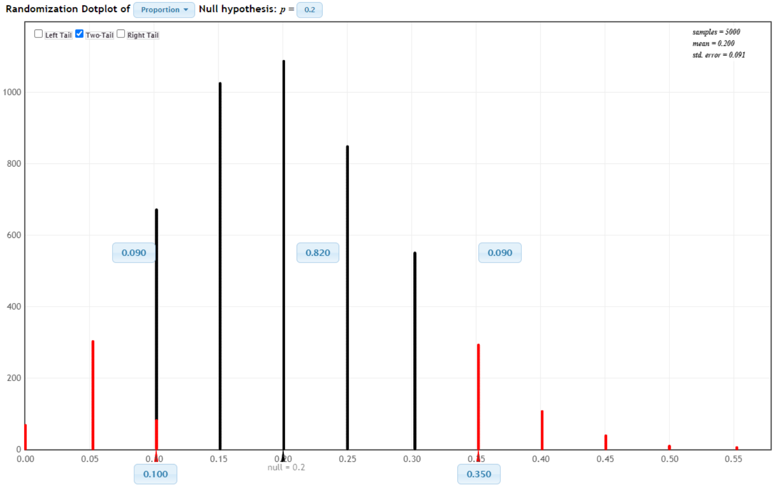


Simulation for One Proportion (Testing)

- 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

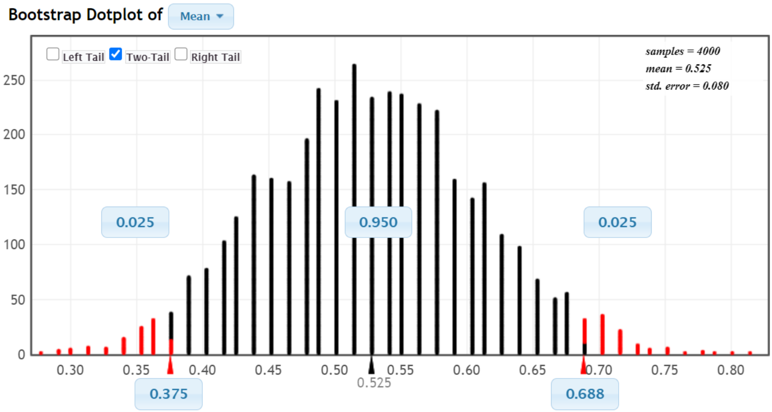


Simulation for One Mean (CI)

- 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)

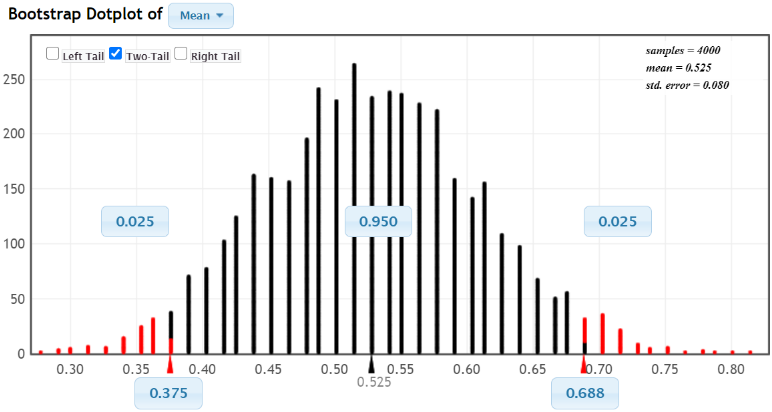


Simulation for One Mean (Testing)

- 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)

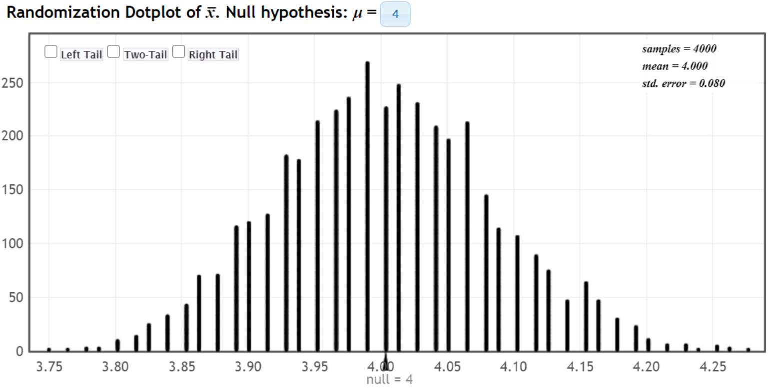
Using simulation via StatKey, the 95% *bootstrapped* confidence interval is (0.375, 0.688)



- 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