Confidence Intervals for Means

Ryan Miller

Introduction

- The *fundamental goal* of statisticians is to use information from sample data to make *reliable* statements about a population
  - This idea is called **statistical inference**

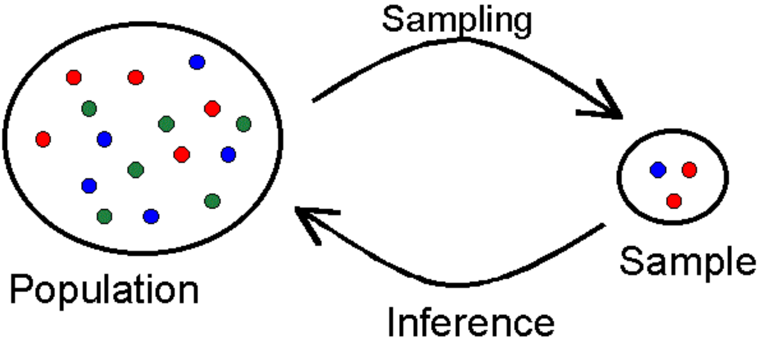


Image credit: http://testofhypothesis.blogspot.com/2014/09/the-sample.html

Interval Estimation

- Confidence intervals are an important part of statistical inference, as they allow us to account for uncertainty in our estimates in a meaningful way
  - So far, we've used Normal models as the basis for 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:

$$\hat{p} \pm z^* SE$$

Interval Estimation for Quantitative Data

- The Normal model/formula we've used for proportions is based upon the Central Limit theorem, a result describing the distribution of sample averages
  - We should expect a similar formula to apply when estimating a population mean, μ:
$$\bar{x} \pm z^* SE$$

Interval Estimation for Quantitative Data

- The Normal model/formula we've used for proportions is based upon the Central Limit theorem, a result describing the distribution of sample averages
  - We should expect a similar formula to apply when estimating a population mean, μ:
$$\bar{x} \pm z^* SE$$
- 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 $\hat{p}$

Interval Estimation for Quantitative Data

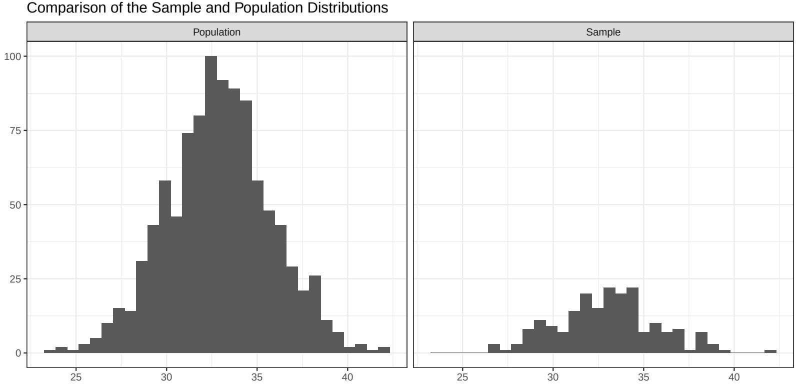
- The Normal model/formula we've used for proportions is based upon the Central Limit theorem, a result describing the distribution of sample averages
  - We should expect a similar formula to apply when estimating a population mean, μ:

$$\bar{x} \pm z^* SE$$

- 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

- One 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

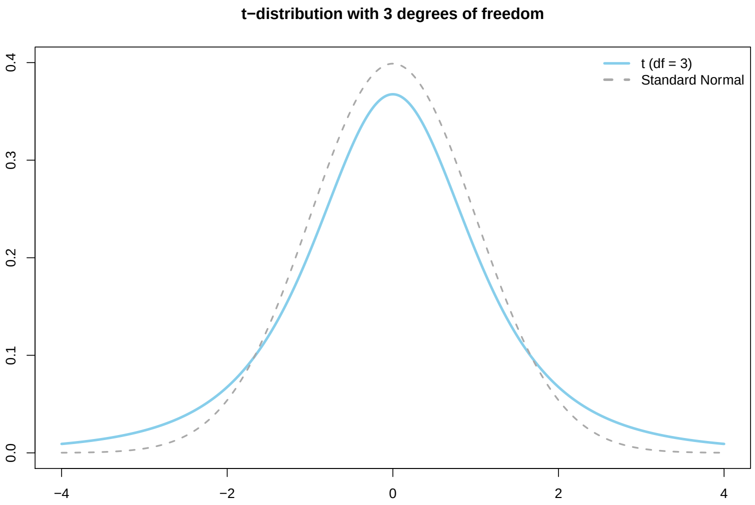
Student's t-distribution

- 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)

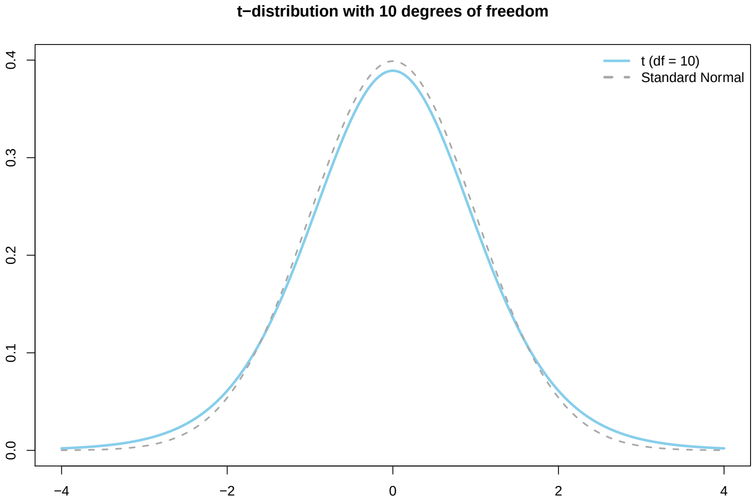
Student's t-distribution

- 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

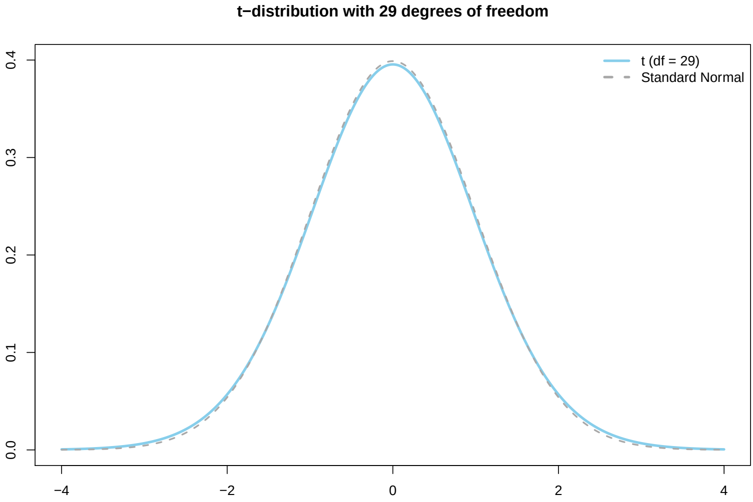
The t-distribution



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 see 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 at this airport to the part of the country that you are traveling to

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)

Conclusion

- This lecture introduced the t-distribution, a necessary modification to the Normal model in scenarios involving a means
  - The standard error in these situations 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 (where the assumptions of these models are violated) in the last video of this week