Confidence Intervals

Part 3 - Student's t-distribution

Ryan Miller

Grinnell College
Statistics

Introduction

We've now seen that confidence interval estimates for many different descriptive statistics can be found using the generic formula:

$$\text{point estimate} \pm c * SE$$

- The standard error of our point estimate, SE, can be calculated using information from our sample data and a formula based upon the Central Limit theorem
- We've calibrated the confidence level of the interval by choosing "c" from a standard normal distribution

Central Limit Theorem for Means

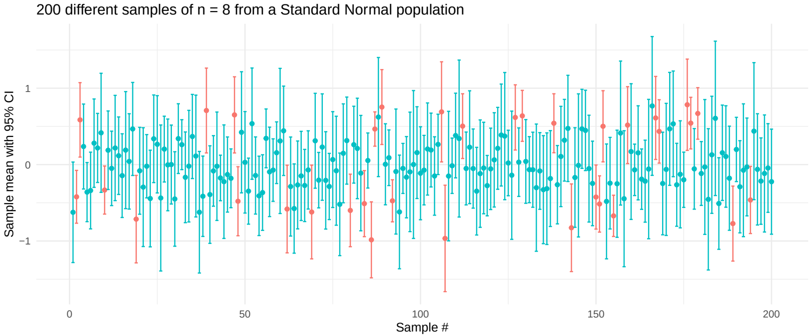
For a *single mean*, CLT suggests:

$$\overline{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- σ is the standard deviation *of the population*

William Gosset and the t-distribution

- Different from our examples involving proportions, the previous CLT result involves a *second unknown parameter*, σ (the population's standard deviation)
  - It seems natural to simply replace this with an estimate from the sample, s, but this is what happens:



Grinnell College Statistics

William Gosset and the t-distribution

- Clearly this 95% CI procedure is *invalid* - too many of these intervals do not contain μ (which is 0)
- William Gosset, a chemist working for Guinness Brewing, became aware of this issue in the 1890s
  - His work evaluating the yields of different barley strains frequently involved small sample sizes

William Gosset and the t-distribution

- Clearly this 95% CI procedure is *invalid* - too many of these intervals do not contain μ (which is 0)
- William Gosset, a chemist working for Guinness Brewing, became aware of this issue in the 1890s
  - His work evaluating the yields of different barley strains frequently involved small sample sizes
- In 1906, Gosset took a leave of absence from Guinness to study under Karl Pearson (developer of the correlation coefficient)
  - Gosset discovered the issue was due to using s interchangeably with σ

William Gosset and the t-distribution

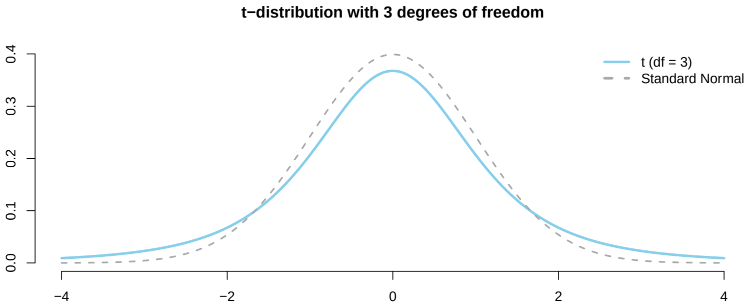
- Treating s as if it were a perfect estimate of σ results in a systematic underestimation of the total amount of variability involved in making the CI
  - To account for the additional variability introduced by estimating σ using s, a modified distribution that's slightly more spread out than the Standard Normal curve must be used

William Gosset and the t-distribution

- Treating s as if it were a perfect estimate of σ results in a systematic underestimation of the total amount of variability involved in making the CI
  - To account for the additional variability introduced by estimating σ using s, a modified distribution that's slightly more spread out than the Standard Normal curve must be used
- Typically the inventor of a new method gets to name it after themselves
  - However, Gosset was forced to publish his new distribution under the pseudonym "student" because Guinness didn't want it's competitors knowing they employed statisticians!
  - Student's t-distribution is now among the most widely used statistical results of all time

The t-distribution

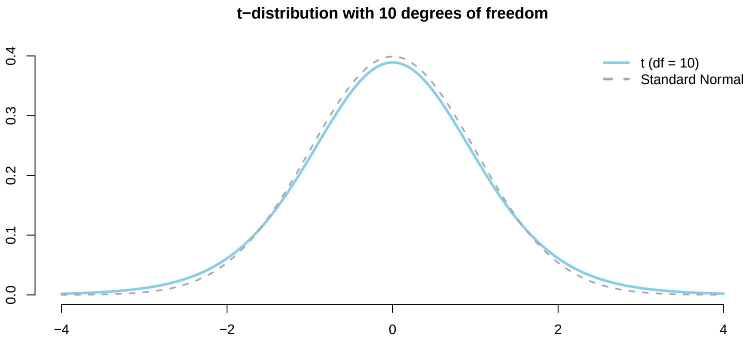
The *t*-distribution accounts the additional uncertainty in small samples using a parameter known as *degrees of freedom*, or *df*:



When estimating a single mean, $df = n - 1$

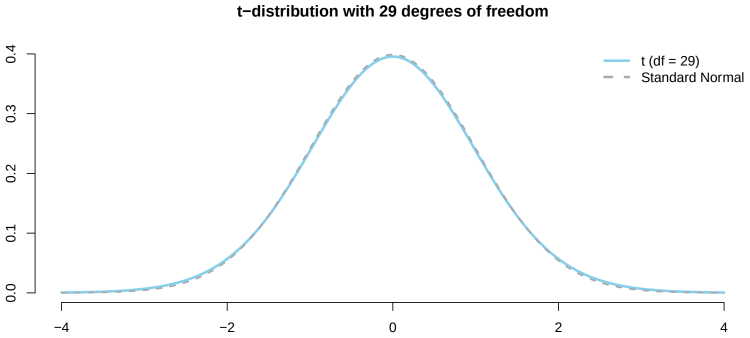
Grinnell College
Statistics

The t-distribution



Grinnell College
Statistics

The t-distribution



Practice

While waiting at an airport, a traveler notices 6 flights to similar a similar part of the country were delayed 6, 10, 13, 23, 45, 55 minutes. The mean delay in this sample was 25.33, with a sample standard deviation of $s = 20.2$. Assuming these data are a representative sample, answer the following:

1) How many degrees of freedom are involved when using the t-distribution to form a CI estimate? What is the value of c that should be used for 95% confidence?
2) What is the 95% CI estimate for the average delay of flights to the part of the country this traveler is heading?

Practice (solution)

1) Because $n = 6$, we'd use $df = n - 1 = 5$. For $df = 5$, $c = 2.571$ defines the middle 95% of the distribution.
2) Point Estimate $\pm$ MOE, Point estimate $= \overline{x} = 25.33$, Margin of error $= c * SE = 2.571 * \frac{20.2}{\sqrt{6}}$
   - 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 (instead of the t-distribution), 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)

When to use the t-distribution

- The t-distribution was designed for small, Normally distributed samples
  - However, it can also be reliably used on large samples, regardless of their shape

	Sample data are approximately Normal	Sample data are non-Normal or skewed
Sample size is large ($n \geq 30$)	Use t-distribution	Use t-distribution
Sample size is small ($n < 30$)	Use t-distribution	*do not* use t-distribution

- Do not fall into the common misconception that the t-distribution requires a certain sample size

Central Limit Theorem for a difference in means

For a *difference of two means*, CLT states:

$$\overline{x}_1 - \overline{x}_2 \sim N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right)$$

- Similar to applications estimating a single mean, the t-distribution should be used when s_1 and s_2 are used as estimates of σ_1 and σ_2
  - Degrees of freedom are complicated for differences in means, so we'll rely on software for these scenarios

Conclusion

- We've now seen how normal approximations and Central Limit theorem allow us to successful construct confidence interval estimates for *one proportion* and a *difference in proportions*
  - With a slight modification, the t-distribution, we can also use CLT for *one mean* and a *difference in means*
  - The t-distribution is *slightly more spread out* than the Normal distribution, which accounts for added the added variability introduced by estimating the population's standard deviation
- Confidence interval estimates can be created for any descriptive statistic, but we'll rely on software like `R` for anything other other than the statistics mentioned above

Grinnell College
Statistics