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 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:

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

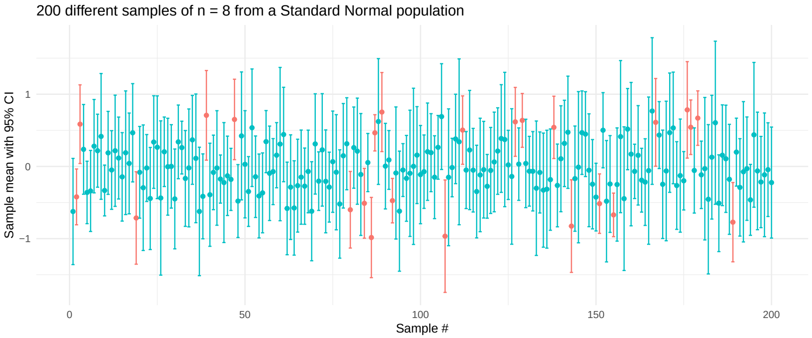
Because σ is the standard deviation *of the population*, we must estimate it using the sample standard deviation in the confidence interval formula:

$$\overline{x} \pm c * \frac{s}{\sqrt{n}}$$

Grinnell College
Statistics

William Gosset and the *t*-distribution

Different from our last lab involving a single proportion, this formula involves a *second unknown parameter*, σ. This is what happens when we estimate σ via s rather than using its true value to calculate the confidence interval's margin of error:



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 statistical chemist working for Guinness Brewing, became aware of this issue in the late 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 statistical chemist working for Guinness Brewing, became aware of this issue in the late 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

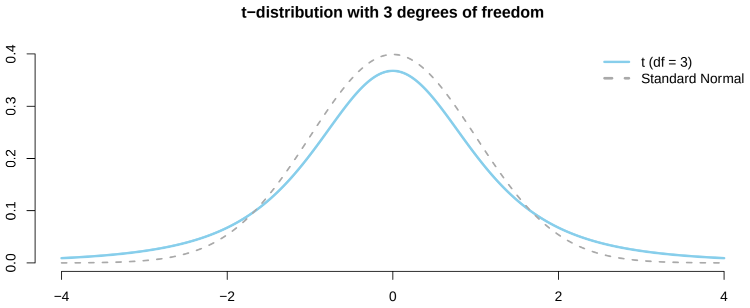
- Treating s as if it were a perfect estimate of σ results in a systematic underestimation of the total amount of variability involved in the estimation procedure
  - To account for the additional variability introduced by estimating σ using s, a modified distribution that's slightly more spread out than the Normal distribution 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 the estimation procedure
  - To account for the additional variability introduced by estimating σ using s, a modified distribution that's slightly more spread out than the Normal distribution 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 its competitors knowing they were using statistical analyses!
  - Student's t-distribution is now among the most widely used statistical results of all time

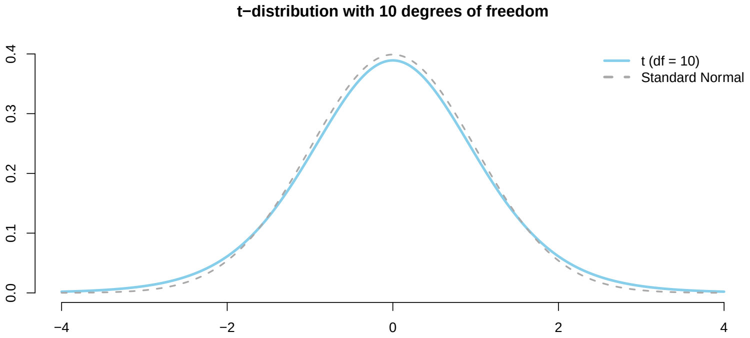
The t-distribution

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



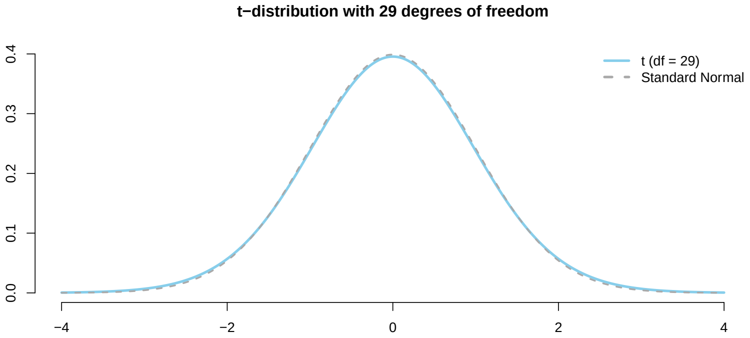
When estimating a single mean, $df = n - 1$

The t-distribution



Grinnell College
Statistics

The *t*-distribution



Grinnell College
Statistics

When to use the t-distribution

- The t-distribution was designed for small samples of quantitative data drawn from a Normally distributed population
  - 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

Conclusion

We've now encountered a few different probability models used in calculating confidence intervals:

- For a single proportion, we can use a Normal approximation if the sample size is large, otherwise we should use the exact binomial distribution.
- For a single mean or a difference in means, we should use the t-distribution for small samples from a Normal distributed population, or for large samples from any population.

Today's lab will summarize the formulas and `R` functions that should be used for each of the descriptive statistics we've covered so far this semester.