Normal Approximations (part II)

Ryan Miller

The Central Limit Theorem (one mean)

- The Central Limit Theorem (CLT) provides the basis for approximating the sampling distribution of certain statistics using a normal curve
- For a sample mean, $\bar{x}$, it suggests:

$$\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

- Like before, we'll need to replace the population parameters, μ and σ, with suitable estimates from our sample to use this approximation

Confidences Intervals

CLT normal approximation:

$$\bar{x} \sim N(\mu, \tfrac{\sigma}{\sqrt{n}})$$

The approximation above suggests 95% confidence intervals of the form:

$$\bar{x} \pm 2 * \tfrac{s}{\sqrt{n}}$$

Where:

- $\bar{x}$ is the sample mean
- s is the sample standard deviation
- n is the sample size

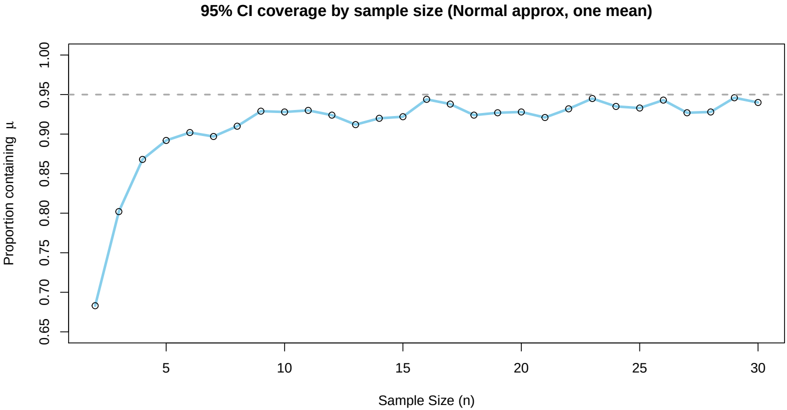
Confidence Interval Coverage

As with other methods, we should study whether these confidence intervals are *valid* (Q: How might we do this?)

1. Draw many random samples from a normally distributed population (which ensures a normal sampling distribution)
2. Construct a 95% confidence interval from each sample using the formula on the previous slide
3. Track the proportion of these intervals containing the actual population mean
4. This time we'll repeat for different sample sizes ranging from $n = 2$ to $n = 30$

Confidence Interval Coverage

Does this normal approximation produce valid 95% confidence intervals?



William Gosset (1876 - 1937)

- Sadly we aren't the first to discover this problem
- William Gosset was an English chemist who worked for Guinness Brewing in the 1890s
  - Gosset's role at Guinness was to statistically evaluate the yield of different varieties of barley
  - These experiments prompted Gosset to question the validity of established statistical procedures under small sample sizes
- In 1906, Gosset took a leave of absence from the brewery to work on the problem with Karl Pearson (inventor of the correlation coefficient)

The t-distribution

- Gosset discovered that plugging in sample standard deviation, s, in place of population standard deviation, σ, produces flawed results when n is small
  - This is because s has its own variability (separate from $\bar{x}$), so treating it like a known entity in the normal approximation leads to intervals that *systematically underrepresent* the amount of uncertainty involved in the confidence interval construction procedure
- Gosset's result, the t-distribution, was published under the name "Student" because Guinness didn't want its competitors knowing that they employed statisticians!
  - The t-distribution has since become of the most widely-used statistical results of all time . . .

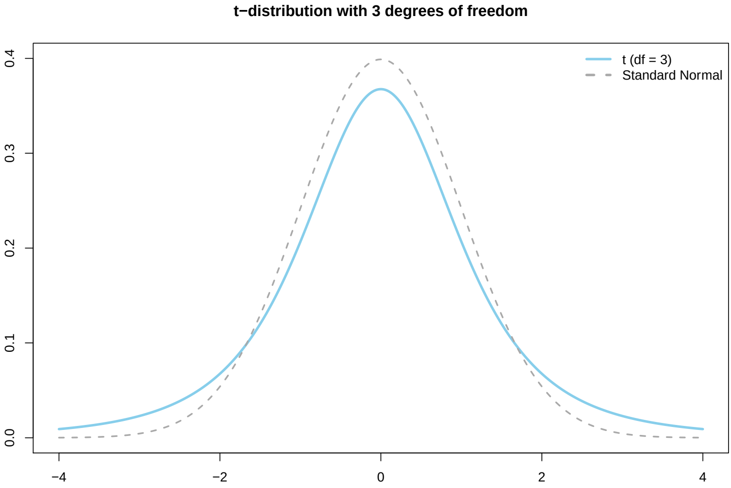
The *t*-distribution

- Unlike the normal distribution, the shape of the t-distribution depends upon the sample size via a parameter named **degrees of freedom** (abbreviated df)
    - In this context, "degrees of freedom" refers to the amount of information available for estimating the standard deviation
    - Once $\bar{x}$ is known, the sum of the deviations, $\sum_{i=1}^{n}(x_i - \bar{x})$, must add up to zero, so not all n elements can vary freely
- Thus, when applying the t-distribution to the mean of a single quantitative variable, $df = n - 1$

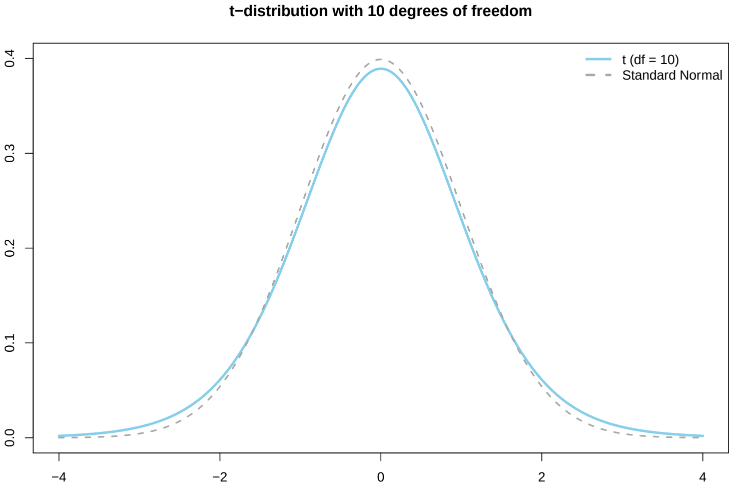
The *t*-distribution

- The *t*-distribution was derived under the assumption of a normally distributed population
  - Generally, population normality can be very difficult to judge from a sample
  - We tend to assume it's a reasonable assumption, unless we can see clear outliers or substantial skew in the sample

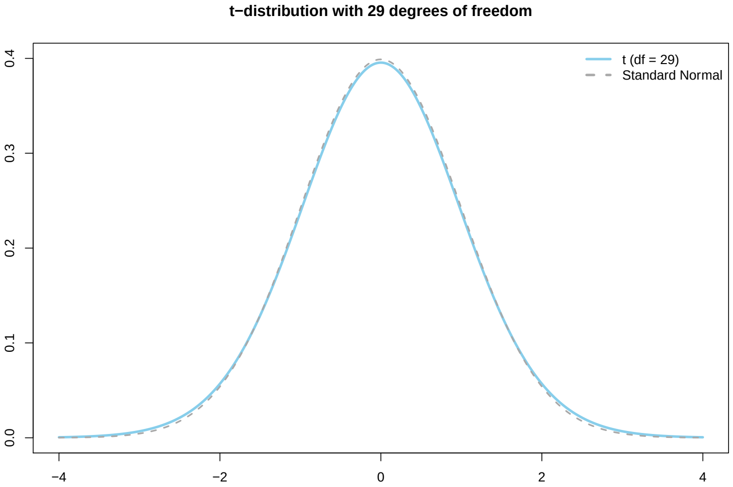
The *t*-distribution



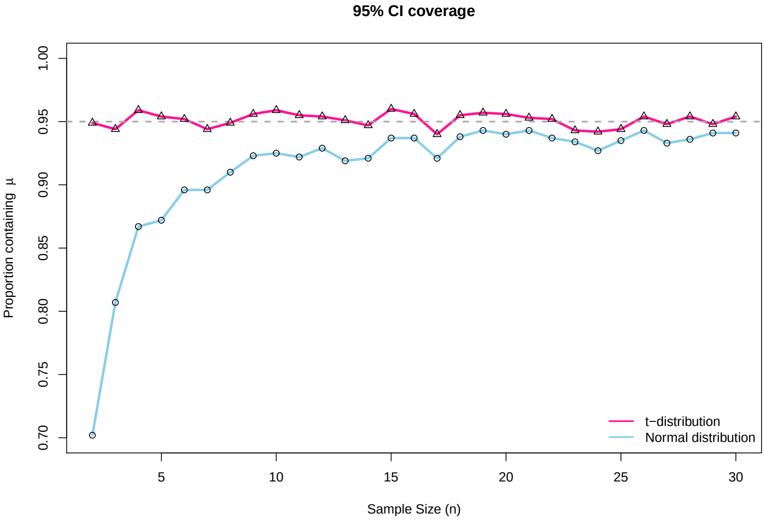
The *t*-distribution



The *t*-distribution



The *t*-distribution



How to use the *t*-distribution

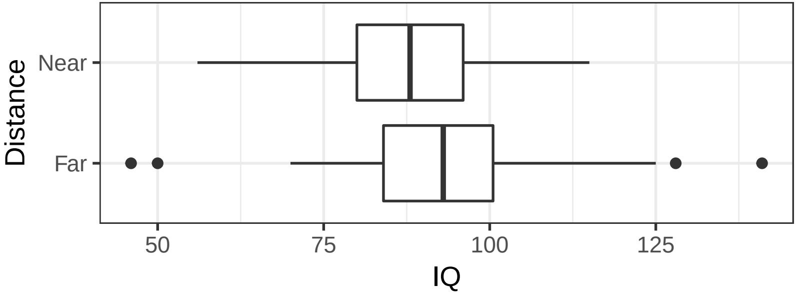
- Gosset's finding requires us to use an extra step when constructing a confidence interval for a mean (or a difference in means)
- For a single mean, we construct a $P\%$ confidence interval via:

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

- Where t^*_{n-1} is a quantile defining the middle $P\%$ of the t-distribution with $n-1$ degrees of freedom

Example - Lead Exposure and IQ

- Researchers in El Paso, TX measured the IQ scores (age-adjusted) of 57 children who lived within 1 mile of a lead smelter and 67 children who lived at least 1 mile away



1. Do these data appear to be normally distributed?
2. Could there be an association between Distance and IQ?

Example - Lead Exposure and IQ

With your group:

1. Download the LeadIQ data (available here on the course website)
2. Construct two separate 95% confidence intervals for each group's mean (use Minitab to tabulate the necessary summary statistics for each group, use StatKey to get t^*_{n-1}, then calculate the intervals by hand)
3. Use these intervals reach a conclusion regarding the impact of distance from the smelter on IQ

Example - Lead Exposure and IQ (solution)

The 95% confidence intervals are shown below:

$$\bar{x}_{\text{near}} \pm t^*_{df=56} * \frac{s_{\text{near}}}{\sqrt{n_{\text{near}}}} = (86.0, 92.4)$$

$$\bar{x}_{\text{far}} \pm t^*_{df=66} * \frac{s_{\text{far}}}{\sqrt{n_{\text{far}}}} = (88.8, 96.6)$$

The large overlap in these intervals suggests we cannot be confident in concluding that there is a population-level difference in IQ

Lead Exposure and IQ - Revisited

The previous approach was sub-optimal, it is better to look at the difference in means (rather than each mean separately). To understand why this is, we'll need a new CLT result:

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

Notice how the standard error of a difference in means is *always less then* sum of the standard errors of each mean separately:

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} < \sqrt{\frac{\sigma_1^2}{n_1}} + \sqrt{\frac{\sigma_2^2}{n_2}}$$

Lead Exposure and IQ - Degrees of Freedom

- This result requires us to use estimates of both σ_1 *and* σ_2, so you might be wondering how to determine the correct degrees of freedom. The answer is quite messy...

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{s_1^2/n_1}{n_1-1} + \frac{s_2^2/n_2}{n_2-1}}$$

- Don't ever calculate this by hand, use software!
- When doing textbook problems, use the smaller of $n_1 - 1$ and $n_2 - 1$
  - This is a *conservative* approach (it underestimates the actual degrees of freedom and leads to wider intervals)

Example - Lead Exposure and IQ

With your group:

1. Construct a 95% confidence interval for the difference in mean IQ (far minus near). Based upon your interval, what do you think of random chance as a possible explanation for the difference in IQ seen in these data?
2. At a lower confidence level, could this study establish that living near a lead smelter *causes* lower IQ?

Example - Lead Exposure and IQ (solution)

1. The 95% CI is found by:

$$\bar{x}_{\text{far}} - \bar{x}_{\text{near}} \pm t^*_{df=56} * \sqrt{\frac{s^2_{\text{far}}}{n_{\text{far}}} + \frac{s^2_{\text{near}}}{n_{\text{near}}}}$$

$$= 92.7 - 89.2 \pm 2.003 * \sqrt{\frac{16.0^2}{66} + \frac{12.2^2}{56}} = (-1.60, 3.88)$$

2. No, this is an observational study, so even if we had been able to rule out random change, confounding variables cannot be ruled out.

Normality Assumptions?

- Recall that the *t*-distribution was derived under the assumption of a normally distributed population
- In the lead-IQ example, it was reasonable to assume the populations that the data came from are normally distribution (why was this reasonable?)
- When the population is not normally distributed, confidence intervals constructed using the *t*-distribution are generally still valid for moderately large samples
  - For a single mean, the rule of thumb is $n \geq 30$
  - For a difference in means, the rule of thumb is $n_1 \geq 30$ *and* $n_2 \geq 30$

Example - Salaries by Sex

- The American Community Survey (ACS) is an ongoing survey conducted by the US Census Bureau
- A sample of these data is available on StatKey under the “Bootstrapping for a Difference in Means” menu as “Employed ACS (Income by Sex)”
- The data are coded such that $1 = $ Male and $0 = $ Female, with salaries reported in thousands of dollars

Example - Salaries by Sex

With your group:

1. Describe the distribution of salary (within each sex)
2. Construct a 95% percentile bootstrap confidence interval for the difference in means
3. *Copy these data into Minitab*, and use the “two-sample t” menu to construct a 95% confidence interval for the difference in means
4. Compare these two intervals, are you surprised?

Example - Salaries by Sex (solution)

1. Salaries within each sex are very right-skewed
2. (-29.2, -9.6)
3. (-28.59, -9.02)
4. These intervals are slightly different (differing by less than 1 unit in either direction), but tell a similar story

The similarity of these confidence intervals despite the high degree of skew is very impressive!

Example - NY vs. NJ Home Prices

The dataset “Home Prices (NY vs NJ)” is available in StatKey and contains home prices (in thousands of dollars) from random samples of 30 homes in New York and 30 homes in New Jersey.

With your group:

1. Describe the distribution of price (within each state)
2. Use bootstrapping to construct a 95% confidence interval for the difference in means
3. *From the summary statistics provided by StatKey*, use the “two-sample t” menu to construct a 95% confidence interval for the difference in means (Hint: use the dropdown menu inside of “two-sample t”)
4. Compare these two intervals, are you surprised?

Example - NY vs. NJ Home Prices (solution)

1. Home prices within each state are very right-skewed
2. (-461.4, 56.3)
3. (-449, 95)
4. These intervals are somewhat different, particularly near the right end-point (which differs by $\sim$40)

The impact of the data's skew on these intervals is much more pronounced (relative to the salaries example), because the sample sizes of $n_1 = 30, n_2 = 30$ are right at the threshold needed for CLT to make up for the lack of normality

The Correlation Coefficient

The last statistic we will discuss CLT results for is the correlation coefficient:

$$r \sim N(\rho, \sqrt{\frac{1-\rho^2}{n-2}})$$

Leading to confidence intervals of the form:

$$r \pm z^* \sqrt{\frac{1-r^2}{n-2}}$$

Will we need to use *t*-distribution? Yes, s_x and s_y are used in place of σ_x and σ_y when calculating r

The Correlation Coefficient

- Because we estimated *two* extra parameters, we'll need to use a t-distribution with $n-2$ degrees of freedom to construct a $P\%$ confidence interval:

$$r \pm t^*_{n-2}\sqrt{\frac{1-r^2}{n-2}}$$

- Again, t^*_{n-2} is the percentile that defines the middle $P\%$ of the t-distribution with $n-2$ degrees of freedom

Example - Mercury vs. pH in Florida Lakes

The dataset "Florida Lakes (Mercury as a function of pH)" is available in StatKey and contains mercury and pH measurements from 53 lakes in Florida.

With your group:

1. Use bootstrapping to construct a 95% confidence interval for the population-level correlation between mercury and pH in these lakes
2. Use the CLT normal approximation from the previous slide to construct a 95% confidence interval (be sure you use the *t*-distribution)
3. Compare these two intervals (noting the shape of the bootstrap distribution)

Example - Mercury vs. pH in Florida Lakes

1. Using bootstrapping, the 95% CI estimate of ρ is (-.72, -.39)
2. Here we should use $t^*_{n-2} = 2.007$, then:
   $-.575 \pm 2.007 * \sqrt{\frac{1-(-.572)^2}{53-2}} = (-.805, -.345)$
3. These intervals are slightly different, likely due to the sampling distribution being slightly skewed

Summary

- Proper use of the Central Limit Theorem for a mean, or a difference in means, or a correlation coefficient, requires the t-distribution
  - This modified distribution is necessary to properly account for the added uncertainty introduced by using the sample standard deviation, s, instead of σ, the population standard deviation
  - This changes the multiplier of SE when determining our interval's margin of error, otherwise the calculation is just like what we've done before (Estimate $\pm c * SE$)

Summary

- The t-distribution depends upon a parameter known as *degrees of freedom*, or *df*
  - For a single mean, we used $n - 1$ degrees of freedom because we needed to estimate 1 extra parameter
  - For a difference in means, this was complicated, but we used the minimum of $n_1 - 1$ and $n_2 - 1$
  - For a correlation coefficient, we used $n - 2$ degrees of freedom because we needed to estimate 2 extra parameters
- Finally, remember that the t-distribution was *derived specifically* for small samples from normally distributed populations
  - You *do not* need a large n to use it (so long as the data appear reasonably normal)

Conclusion

Right now you should...

1. Understand reason why the t-distribution is necessary
2. Know how to construct P% confidence intervals using the t-distribution
3. Know the limitations of these approaches, the assumptions involved, and when to use bootstrapping as an alternative

These notes cover the “CI” parts of Ch 6 from our textbook, I encourage you to read through those subsections of the chapter