Confidence Intervals for Numerical Data

Ryan Miller

Confidence Intervals

- Previously, we covered a couple different methods for constructing confidence interval estimates for categorical data (proportions)
- These approaches relied upon finding a reasonable *probability model* for the **sampling distribution** (of the sample proportion)
  - Binomial distribution (Clopper-Pearson intervals)
  - Normal distribution (CLT intervals)

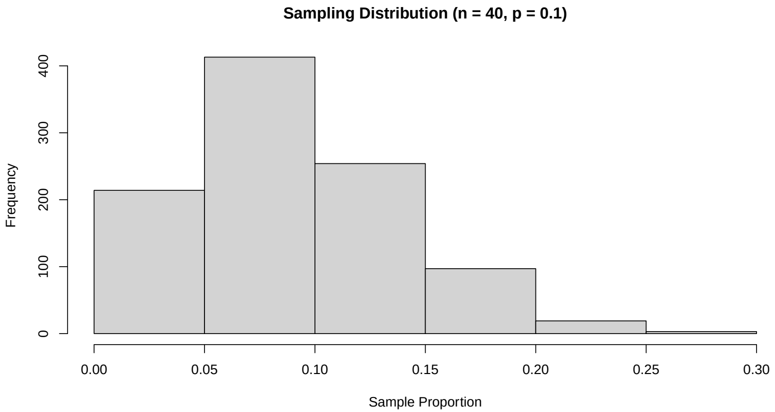
CLT Confidence Intervals for a Proportion

- Recall that the CLT approach was only valid when the conditions $n\hat{p} \geq 10$ and $n(1 - \hat{p}) \geq 10$ were both met
  - Thus, there are two scenarios where the CLT approach fails

1) The sample proportion is too close a boundary of either 0 and 1 (relative to the sample size)
2) The sample size is too small (and the proportion is away from the boundaries)

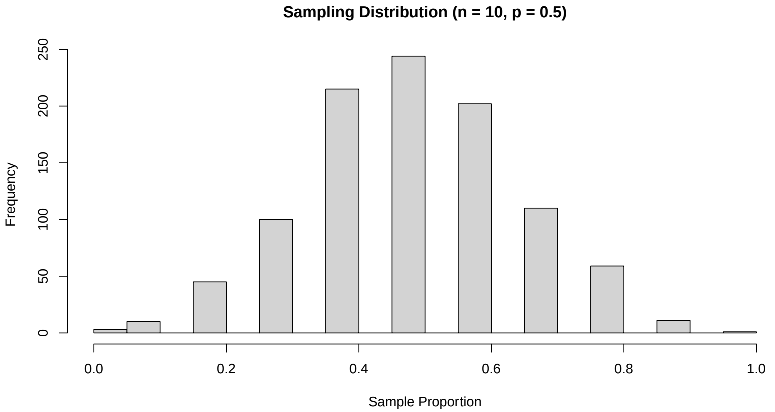
CLT Confidence Intervals for a Proportion

- In this scenario ($n = 40$ and $p = 0.1$), notice $n\hat{p} \approx 4$ and the sampling distribution is *skewed right*



CLT Confidence Intervals for a Proportion

- In this scenario ($n = 10$ and $p = 0.5$), notice $n * \hat{p} \approx 5$ and the sampling distribution is symmetric, but it is too discrete to be properly represented by a Normal curve



Validity

- Suppose the 1,000 different samples shown in the previous two histograms were each used to construct a 95% confidence interval estimate of p, how many would you expect to contain p?

Validity

- Suppose the 1,000 different samples shown in the previous two histograms were each used to construct a 95% confidence interval estimate of p, how many would you expect to contain p?
  - In the first scenario ($n = 40$ and $p = 0.1$), 918 of 1000 these "95% confidence" intervals contain p
  - In the second scenario ($n = 10$ and $p = 0.5$), 871 of 1000 "95% confidence" intervals contain p

Validity

- Suppose the 1,000 different samples shown in the previous two histograms were each used to construct a 95% confidence interval estimate of p, how many would you expect to contain p?
  - In the first scenario ($n = 40$ and $p = 0.1$), 918 of 1000 these "95% confidence" intervals contain p
  - In the second scenario ($n = 10$ and $p = 0.5$), 871 of 1000 "95% confidence" intervals contain p
- Fortunately, we could use the binomial distribution to calculate Clopper-Pearson (Exact Binomial) intervals in situations like these

How About Numerical Data?

- Now let's consider a numeric random variable, X (ie: an individual's height, income, cholesterol, etc.)
  - Central Limit Theorem suggests the following sampling distribution for $\bar{x}$, the sample average of X

$$\bar{x} \sim N(E(X) = \mu, \sqrt{Var(X)/n} = \tfrac{\sigma}{\sqrt{n}})$$

- This result suggests confidence intervals of the form:

$$\bar{x} \pm z^* \tfrac{\sigma}{\sqrt{n}}$$

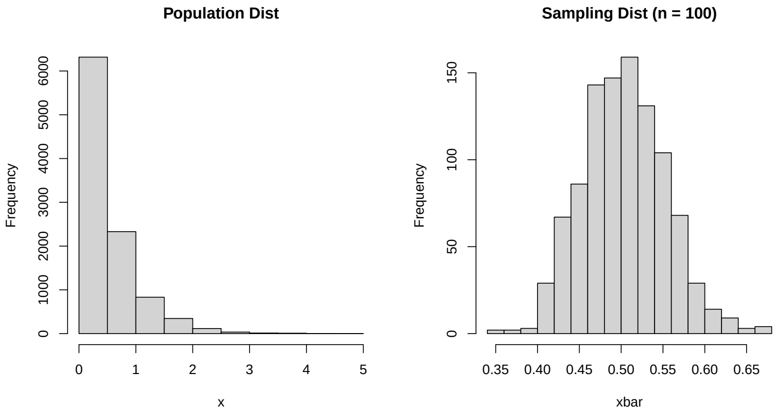
CLT Confidence Intervals for a Mean

- As we saw with categorical data, these CLT-based intervals will only be valid when the *sampling distribution* is approximately Normal
  - For numerical data, there are two ways this could occur:

1) The sample size is relatively large ($n \geq 30$), regardless of how the population is distributed
2) The sample size is small, but the *population distribution* is Normal

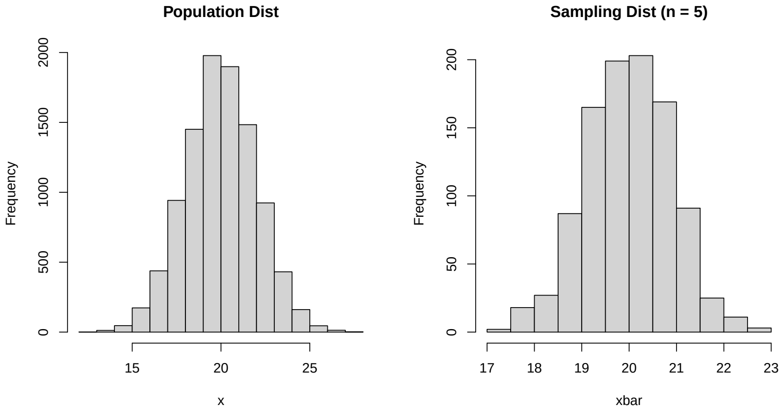
CLT Confidence Intervals for a Mean

- The sampling distribution of $\bar{x}$ is approximately Normal for $n = 100$, even when the population is heavily right-skewed



CLT Confidence Intervals for a Mean

- The sampling distribution of $\bar{x}$ is approximately Normal for very small samples ($n = 5$) if the population is Normally distributed



Validity

- Now consider using the 1,000 different samples shown in the previous histograms to construct a 95% confidence interval estimate of μ (the population mean)
  - This would look like $\bar{x} \pm 1.96\frac{s}{\sqrt{n}}$ for each sample (recognize that we're using the *sample standard deviation*, s, as an estimate of the population standard deviation, σ)
- How many of the 1,000 intervals would you expect to contain μ?

Validity

- Now consider using the 1,000 different samples shown in the previous histograms to construct a 95% confidence interval estimate of μ (the population mean)
  - This would look like $\bar{x} \pm 1.96\frac{s}{\sqrt{n}}$ for each sample (recognize that we're using the *sample standard deviation*, s, as an estimate of the population standard deviation, σ)
- How many of the 1,000 intervals would you expect to contain μ?
  - In the first scenario ($n = 100$ and right-skewed population), 953 of 1000 these "95% confidence" intervals contain μ
  - In the second scenario ($n = 5$ and Normal population), 924 of 1000 "95% confidence" intervals contain μ

Validity

- Something is clearly going wrong in the second scenario, and we aren't the only ones to have observed this problem

Validity

- Something is clearly going wrong in the second scenario, and we aren't the only ones to have observed this problem
- 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

Validity

- Something is clearly going wrong in the second scenario, and we aren't the only ones to have observed this problem
- 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)
  - Not accounting for this additional variability is the flaw in the previously constructed confidence intervals ($n = 5$, Normal population scenario)

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)
  - Not accounting for this additional variability is the flaw in the previously constructed confidence intervals ($n = 5$, Normal population scenario)
- 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

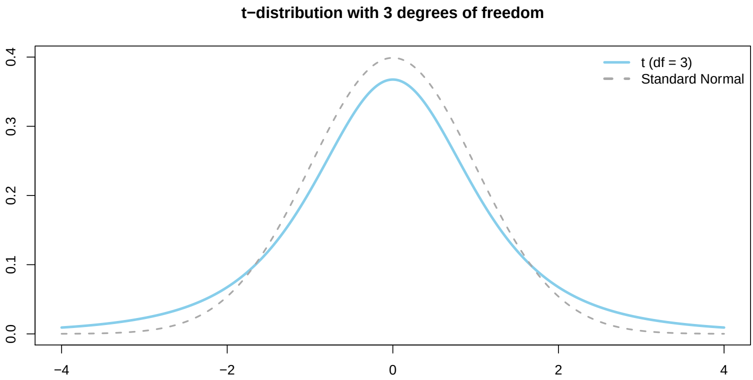
The t-distribution

- The t-curve looks much like a standard Normal curve, but it has *thicker tails* to properly account for additional variability that results from estimating σ using s
  - The amount of additional variability is linked to the sample size through a parameter known as *degrees of freedom* (df)

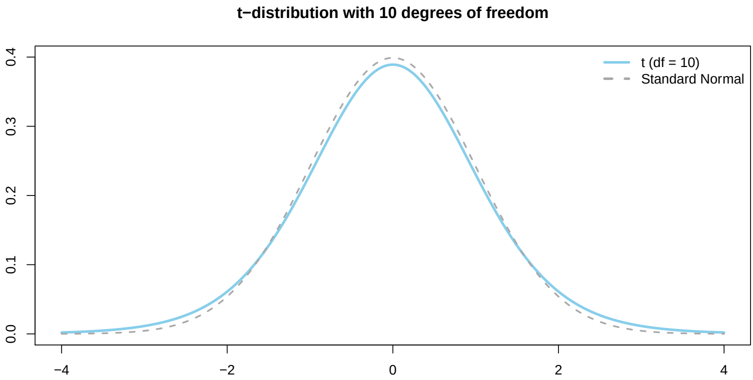
The t-distribution

- The t-curve looks much like a standard Normal curve, but it has *thicker tails* to properly account for additional variability that results from estimating σ using s
  - The amount of additional variability is linked to the sample size through a parameter known as *degrees of freedom* (df)
- Similar to Chi-square testing, this is a reference to the number of unique pieces of information available to estimate σ
  - If $\bar{x}$ is assumed known, the sum of deviations $\sum_{i=1}^{n}(x_i - \bar{x})$ must add up to zero, thus only $n-1$ deviations are necessary to know everything
- So, when applying the t-distribution to the mean of a single numeric variable, $df = n - 1$

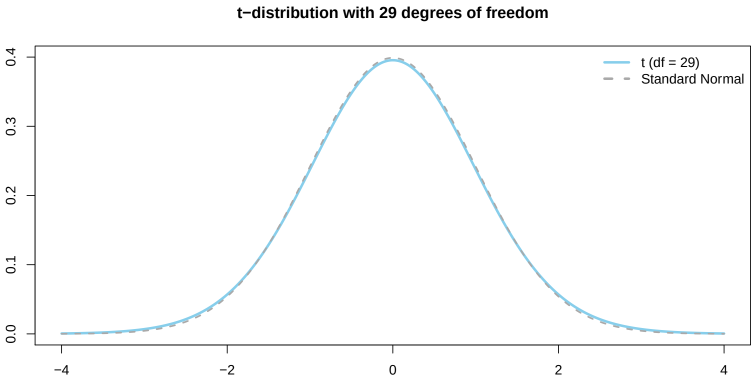
The t-distribution



The t-distribution



The t-distribution



How to use the t-distribution

- For a single mean, we construct a $P\%$ confidence interval via:

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

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

```r
qt(.975, df = 10)
```

```
## [1] 2.228139
```

```r
qt(.975, df = 100)
```

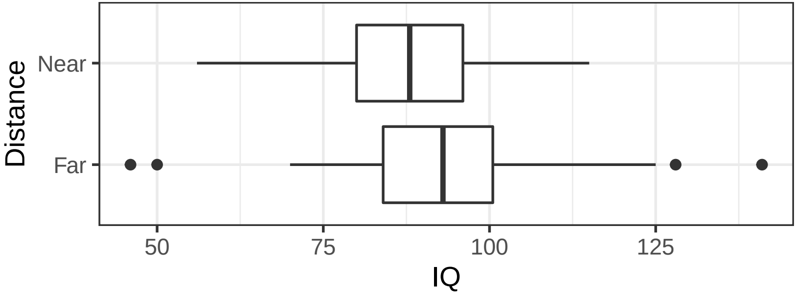
```
## [1] 1.983972
```

```r
qt(.975, df = 1000)
```

```
## [1] 1.962339
```

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

- Load these data into R using the following code
- Then use the t-distribution to come up with separate 90% confidence intervals for the population IQ in each group (Near and Far)

```r
data <- read.csv("https://remiller1450.github.io/data/LeadIQ.csv")
near <- data$IQ[data$Distance == "Near"]
far <- data$IQ[data$Distance == "Far"]
```

Example (solution, “Near” group)

```
xbar_near = mean(near)
sd_near = sd(near)
n_near = length(near)
ts_near = qt(.95, df = n_near - 1)

lower_end = xbar_near - ts_near*sd_near/sqrt(n_near)
upper_end = xbar_near + ts_near*sd_near/sqrt(n_near)
c(lower_end, upper_end)
```

```
## [1] 86.49585 91.89012
```

Example (solution - “Far” group)

```
xbar_far = mean(far)
sd_far = sd(far)
n_far = length(far)
ts_far = qt(.95, df = n_far - 1)

lower_end = xbar_far - ts_far*sd_far/sqrt(n_far)
upper_end = xbar_far + ts_far*sd_far/sqrt(n_far)
c(lower_end, upper_end)
```

```
## [1] 89.43076 95.94238
```

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 (difference in means)

- This result requires 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!

Lead Exposure and IQ - Degrees of Freedom

The `t.test` function provides a proper 90% CI estimate for the difference in means (far - near)

```
t.test(x = far, y = near, conf.level = .90)$conf.int
```

```
## [1] -0.702856  7.690025
## attr(,"conf.level")
## [1] 0.9
```

Comments on the t-distribution

- The t-distribution should *always* be used when using a sample of numerical data to estimate a population mean (or difference in means)
  - That said, when $n \geq 30$ there isn't much of a difference relative to using the Normal distribution

Comments on the *t*-distribution

- The *t*-distribution should *always* be used when using a sample of numerical data to estimate a population mean (or difference in means)
  - That said, when $n \geq 30$ there isn't much of a difference relative to using the Normal distribution
- When the sample is small *and* the population is Normally distributed, the *t*-distribution is necessary for valid statistical inference
  - When the sample is small *and* the population is skewed, we're out of luck (at least for now)

Next Steps

- Next we'll use the t-distribution as the basis for hypothesis testing of population means (this is called the t-test)
    - After that, we'll revisit the situation where the sample is small and the population is skewed and explore a creative approach to achieving valid statistical inference

Next Steps

- Next we'll use the t-distribution as the basis for hypothesis testing of population means (this is called the t-test)
  - After that, we'll revisit the situation where the sample is small and the population is skewed and explore a creative approach to achieving valid statistical inference
- For now, the main takeaway from this lecture is the t-distribution
  - You should know when to use it (numerical data)
  - You should know why it's important (extra uncertainty using s as an estimate of σ)
  - You should know how to use it (the `qt` function, degrees of freedom, etc.)