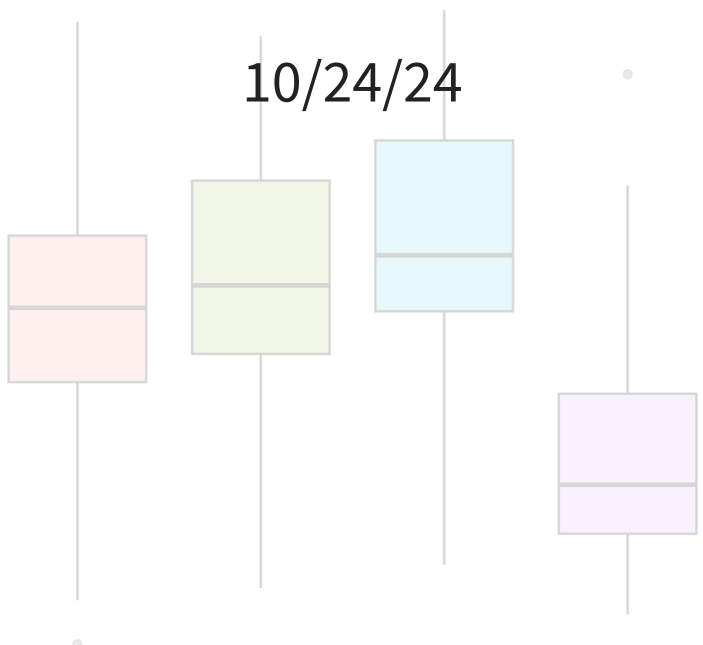
Confidence Intervals for Means

10/24/24



Recap

- Central Limit Theorem: if we have a sufficiently large sample of n independent observations from a population with mean μ and standard deviation σ, then $\bar{X} \dot{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$
- When considering the special case of sample proportions, if success-failure condition is met, we have $\hat{p} \dot{\sim} N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$
- To obtain a $\gamma \times 100\%$ CI for a mean, we use

$$\text{point estimate} \pm \text{critical value} \times \text{SE}$$

  - We needed to replace the standard error with an estimate

Checking normality

- Remember, CLT requires a sufficiently large sample size n or assumption of Normality of the underlying data.
- No perfect way to check Normality, but rule of thumb:
  - If $n < 30$ small: check that there are no clear outliers
  - If $n \geq 30$ large: check that there are no particularly extreme outliers

CI for a single mean

CI for a single mean (known variance)

Suppose we want a $\gamma \times 100\%$ CI for population mean μ.

What would your "best guess" point estimate for μ be?

- If CLT holds, then we know

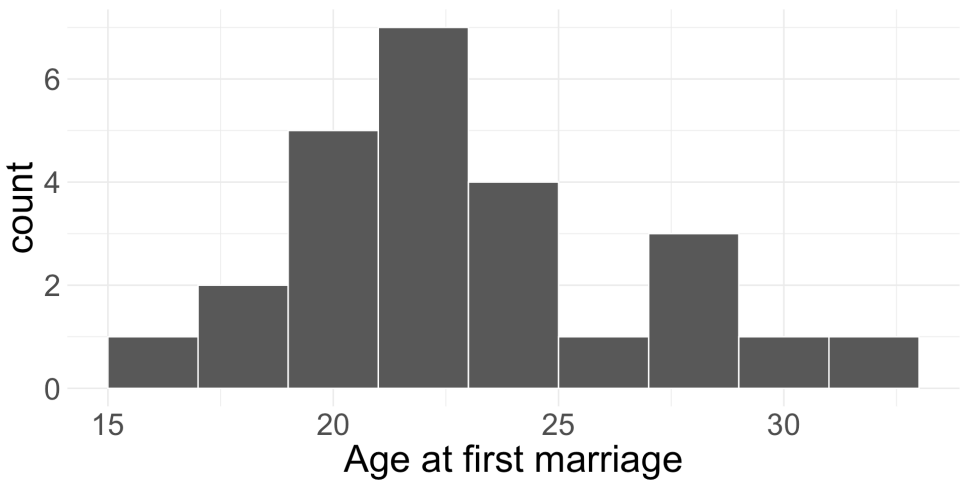
$$\bar{X} \dot{\sim} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- So our $\gamma \times 100\%$ CI for μ is:

$$\text{point estimate} \pm \underbrace{\text{critical value} \times \text{SE}}_{\text{Margin of Error}} = \bar{x} \pm z^*_{(1+\gamma)/2} \times \frac{\sigma}{\sqrt{n}}$$

Example: age at marriage

In 2006-2010, the CDC conducted a thorough survey asking US women their age at first marriage. The standard deviation of the responses is 4.72 years. Suppose we randomly sample 25 US women and ask them their age at first marriage (plotted below). Their average age at marriage was 23.32.



We will obtain an 80% confidence interval for the mean age of US women at first marriage.

- Are conditions of CLT met?
- If so, what does CLT tell us?

What is/are the population parameter(s)? What is the statistic?

Example: age at marriage (cont.)

Obtain an 80% confidence interval for the mean age of US women at first marriage.

By CLT:

$$\bar{X} \dot{\sim} N\left(\mu, \frac{4.72}{\sqrt{25}}\right) = N(\mu, 0.944)$$

Collect necessary components:

1. Point estimate: $\bar{x} = 23.32$
2. Standard error: 0.944
3. Critical value: $z^*_{0.9} =$ `qnorm(0.9, 0, 1)` $= 1.28$

So our 80% confidence interval is $23.32 \pm 1.28 \times 0.944 = (22.11, 24.53)$

Interpret this interval!

Utility of this model

- The previous formula for the confidence interval for μ relies on knowing σ
- But wait…
  - Want to construct a CI for μ because we don't know its value
  - If we don't know μ, it seems highly unlikely that we would know σ!
- So in practice, we will have to estimate standard error for $\bar{X}$:

$$\text{SE} \approx \frac{s}{\sqrt{n}}$$

where s is the observed sample standard deviation

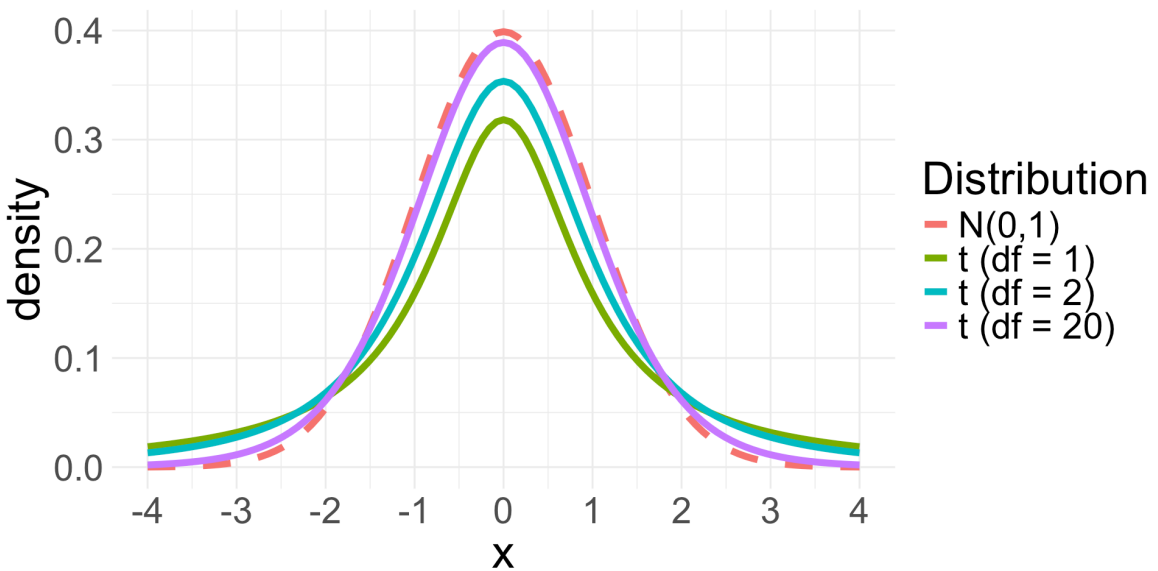
- Recall we did something similar for CI for p

Variance issue

- Replacing s for σ works well enough when n is extremely large so we can estimate σ accurately
- However, estimating variance is extremely difficult when n is small, and still not great for large n
- So if σ is unknown, we ***cannot*** use the Normal approximation to model $\bar{X}$ for inferential tasks
- Instead, we will use a new distribution for inference calculations, called the t-distribution

t-distribution

- The **t-distribution** is symmetric and bell-curved (like the Normal distribution)
- Has "thicker tails" than the Normal distribution (the tails decay more slowly)



- t-distribution is always centered at 0
- One parameter: **degrees of freedom (df)** defines exact shape of the t
  - Denoted t_{df} (e.g. t_1 or t_{20})

- As df increase, t resembles the $N(0, 1)$. When $df \geq 30$, the t_{df} is nearly identical to $N(0, 1)$

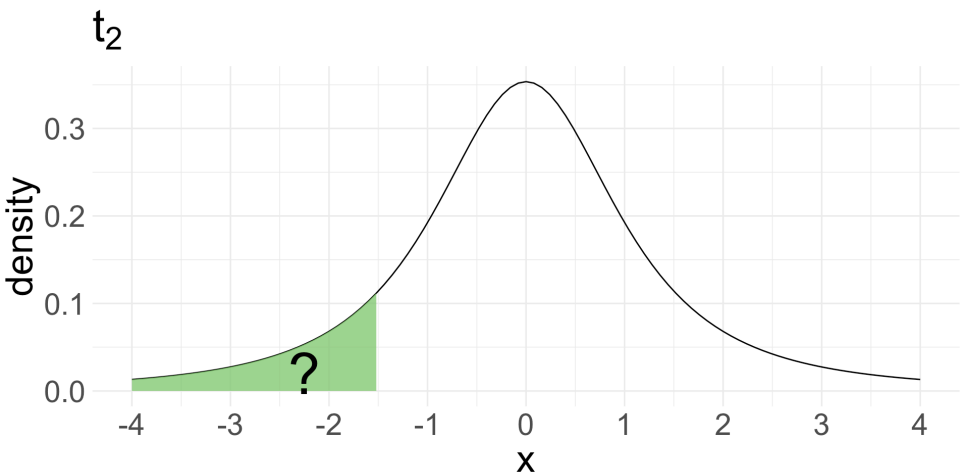
Working with t distribution

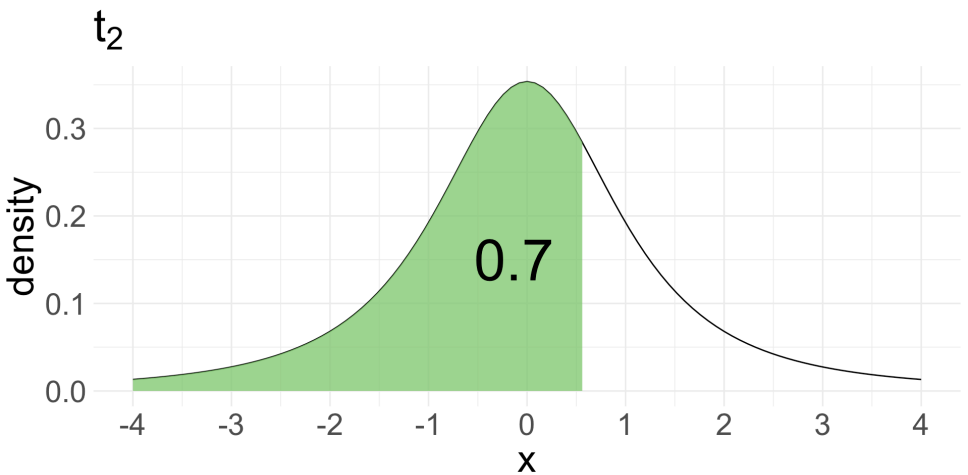
Let's draw pictures for the following:

- What proportion of the t_2-distribution falls below -1.5?
- What value of the t_2-distribution has 70% area lying below it?

t distribution in R

- `pnorm(x, mean, sd)` and `qnorm(%, mean, sd)` used to find probabilities and percentiles for the Normal distribution
- Analogous functions for t-distribution: `pt(x, df)` and `qt(%, df)`





`pt(-1.5,df =2)` = 0.1361966

`qt(0.7, df =2)` = 0.6172134

CI for a single mean (unknown variance)

- Still require independent observations and the Normality condition for CLT
- General formula for $\gamma \times 100\%$ CI is the same, but we simply change what goes into the margin of error.

$$\text{point estimate} \pm t^*_{df,(1+\gamma)/2} \times \widehat{\text{SE}}$$

$$\bar{x} \pm t^*_{df,(1+\gamma)/2} \times \frac{s}{\sqrt{n}}$$

- $df = n - 1$
- critical value $t^*_{df,(1+\gamma)/2} = (1 + \gamma)/2$ percentile of the t_{df} distribution

Example: age at marriage (cont.)

Let's return to the age at marriage example. Once again let's obtain an 80% confidence interval for the average age of first marriage for US women, but now suppose we **don't know** σ.

In our sample of $n = 25$ women, we observed a sample mean of 23.32 years and a sample standard deviation of $s = 4.03$ years.

1. Point estimate: $\bar{x} = 23.32$
2. Standard error: $\widehat{\text{SE}} = \frac{s}{\sqrt{n}} = \frac{4.03}{\sqrt{25}} = 0.806$
3. Critical value:
   - $df = n - 1 = 24$
   - t^*_{24} = `qt(0.9, df =24)` = 1.32

So our 80% confidence interval for μ is:

$$23.32 \pm 1.32 \times 0.806 = (22.26, 24.38)$$

Comparing CIs

Known variance:

80% CI: (22.11, 24.53)

Unknown variance:

80% CI: (22.26, 24.38)

- How do the two intervals compare?
- Interpretation of CI does not change even if we use a different model!

Examples

Assume that all conditions necessary for inference are satisfied.

`qnorm(0.90)` = 1.28 `qt(0.90, df = 35)` = 1.31 `qt(0.90, df = 36)` = 1.31

`qnorm(0.95)` = 1.64 `qt(0.95, df = 35)` = 1.69 `qt(0.95, df = 36)` = 1.69

`qnorm(0.975)` = 1.96 `qt(0.975, df = 35)` = 2.03 `qt(0.975, df = 36)` = 2.03

1. A 90% confidence interval for a population mean μ is given as $(18.985, 21.015)$. The interval was obtained based on a SRS for 36 observations. Calculate the sample mean and sample standard deviation.
2. The standard deviation for students at particular Ivy League college is 250 points. Two students, Raina and Luke, want to estimate the average SAT score of students at this college. They want their margin of error to be no more than 25 points.
    a. Raina wants to use a 90% confidence level. How large a sample does Raina need to collect?
    b. Luke wants to use a 95% confidence level. Without calculations, determine whether Luke's sample should be larger or smaller than Raina's. Explain your reasoning.
    c. Calculate the minimum sample size for Luke.

CI for paired mean

Paired data

Suppose we have two sets of observations/data $\boldsymbol{x} = (x_1, x_2, \ldots x_n)$ and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$

- The data are considered **paired data** if each x_i corresponds to exactly one y_i
- Example: your score on the midterm and your score on the final
- When analyzing paired data, we are typically interested in the difference in outcomes of each pair of observations

Paired differences

- Let $d_i = y_i - x_i$ for each $i = 1, \ldots, n$ be the observed differences
- The d_i come from larger population with true mean difference μ_d and standard deviation of differences σ_d
- The sample mean difference and sample standard deviation of the differences are

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \qquad\qquad s_d = \frac{1}{n-1}\sum_{i=1}^{n}(d_i - \bar{d})^2$$

CLT for mean difference in pairs

- Suppose the n observational units are independent and the distribution of the differences is approximately normal. Then CLT says:

$$\bar{d} \dot{\sim} N\left(\mu_d, \frac{\sigma_d}{\sqrt{n}}\right)$$

- We are usually interested in performing inference for μ_d when both μ_d and σ_d unknown
- Our formula for $\gamma \times 100\%$ CI for μ_d is analogous to the formula for one mean when σ unknown:

$$\text{point estimate} \pm t^*_{df,(1+\gamma)/2} \times \widehat{\text{SE}}$$

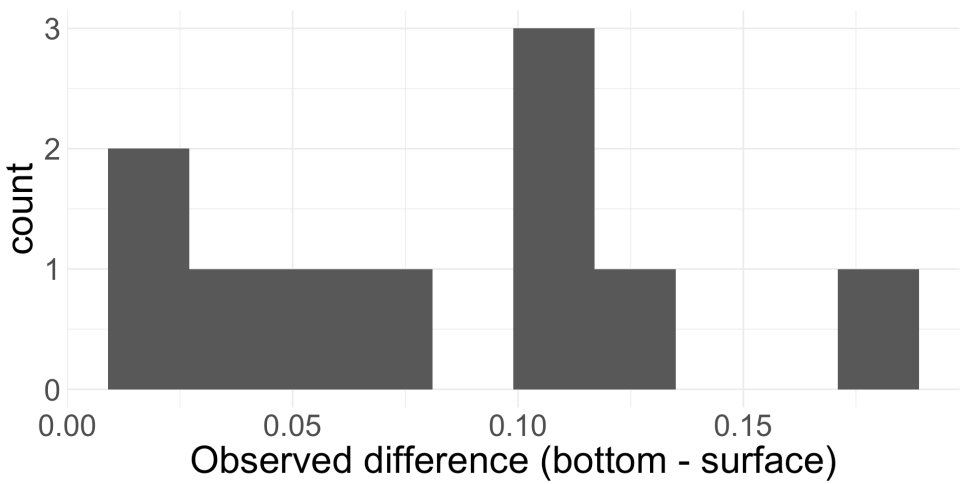
$$\bar{d} \pm t^*_{df,(1+\gamma)/2} \times \frac{s_d}{\sqrt{n}}$$

where $df = n - 1$

Example: zinc

Data consist of measured zinc concentrations in bottom water and surface water at 10 randomly sampled wells:

Do the data suggest that the true average concentration in the bottom water is different than that of surface water? Let's answer this using a 95% confidence interval.



	bottom	surface
1	0.430	0.415
2	0.266	0.238
3	0.567	0.390
4	0.531	0.410
5	0.707	0.605
6	0.716	0.609

Are the data paired? Does CLT apply?

Example: zinc (cont.)

```
zinc <- zinc |>
  mutate(d = bottom - surface)
d_bar <- mean(zinc$d)
d_bar
```

```
[1] 0.0804
```

```
s_d <- sd(zinc$d)
s_d
```

```
[1] 0.05227321
```

1. point estimate: $\bar{d} = 0.0804$
2. SE $\approx \frac{s_d}{\sqrt{n}} = \frac{0.052}{\sqrt{10}} = 0.016$
3. critical value: what code would you write?
   - $df = n - 1 = 9$
   - $t^*_{9,0.975} =$ `qt(0.975,9)` $= 2.26$

So our 95% confidence interval is:

$$0.0804 \pm 2.26(0.016) = (0.044, 0.117)$$

Do the data suggest that the true average concentration in the bottom water is different than that of surface water? Explain.

CIs for difference in two means

Difference of two means

Now consider two populations under the condition that the data/populations are not paired.

We might be interested in learning about whether or not the means of each population are equal (think about the voice jitter homework problem)!

- Let μ_1 and μ_2 represent the population means for the two populations 1 and 2
- Samples of size n_1 and n_2 from each population, respectively
- We might think it reasonable to use $\bar{x}_1 - \bar{x}_2$ as a point estimate for $\mu_1 - \mu_2$

Conditions for inference

Now that we have two populations, conditions for CLT and use of the t-distribution for inference will look slightly different:

1. **Independence** (extended): need data within *and* between the two groups
    - e.g.the two data sets come from independent random samples or from a randomized experiment
2. **Normality**: we need to check for approximate normality for *both* groups separately

CI for difference in two means

If the conditions hold, then our usual formula for $\gamma \times 100\%$ CI still holds:

$$\text{point estimate} \pm \text{critical value} \times \text{SE}$$

1. Point estimate

If σ_1 and σ_2 known:

2. $\text{SE} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

3. critical value: $z^*_{(1+\gamma)/2}$

- $(1+\gamma)/2$ percentile of $N(0,1)$

If σ_1 and σ_2 unknown:

2. $\text{SE} \approx \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$

3. critical value: $t^*_{df,(1+\gamma)/2}$

- $(1+\gamma)/2$ percentile of t_{df}
- $df = \min\{n_1 - 1, n_2 - 1\}$

Example: voice shimmer

Let's consider the voice shimmer of PD vs non-PD patients from last week's homework.

Convince yourself that this data isn't paired!

- Population 1: people with Parkinson's Disease
- Population 2: people without Parkinson's Disease

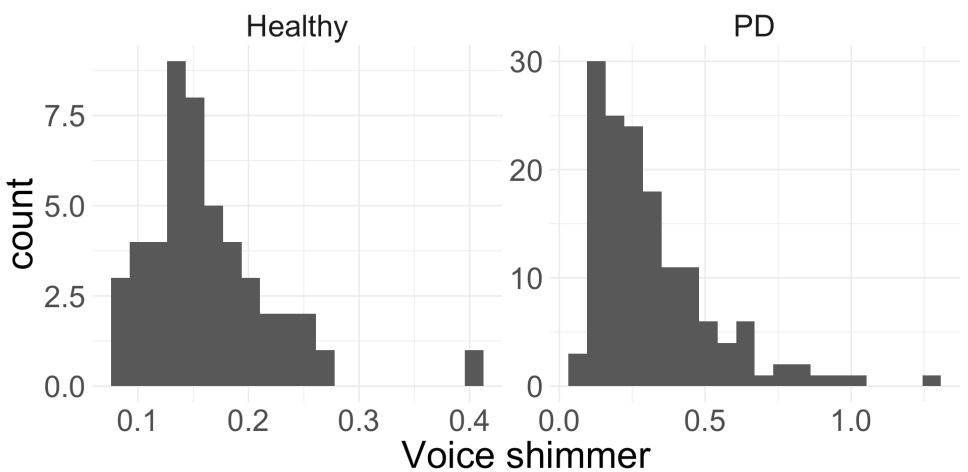
Research question: are average voice shimmers different between people with and without Parkinson's? Create a 95% confidence interval to answer this question.

- We care about the difference in means $\mu_{\text{PD}} - \mu_{\text{H}}$

Example: voice shimmer (cont.)

Are average voice shimmers different between people with and without Parkinson's? Create a 95% confidence interval to answer this question.

status	n	xbar	s
Healthy	48	0.163	0.058
PD	147	0.321	0.208



Do assumptions for CLT hold?

- Independence: random sample!
- Normality condition: $n \geq 30$ in both groups with no particularly extreme outliers

Set-up/find the following:

1. Point estimate
2. Standard error
3. Code for critical value

Example: voice shimmer (cont.)

1. Point estimate: $\bar{x}_{\text{PD}} - \bar{x}_{\text{H}} = 0.32 - 0.16 = 0.158$
2. SE $\approx \sqrt{\frac{s^2_{\text{PD}}}{n_{\text{PD}}} + \frac{s^2_{\text{H}}}{n_{\text{H}}}} = \sqrt{\frac{0.21^2}{147} + \frac{0.06^2}{48}} = 0.019$
3. Critical value:
   - $df = \min\{n_{\text{PD}} - 1, n_{\text{H}} - 1\} = \min\{147 - 1, 48 - 2\} = 47$
   - Want 0.975-th percentile of t_{47} distribution: `qt(0.975, df =47)` = 2.01

Putting everything together, our 95% CI for $\mu_{\text{PD}} - \mu_{\text{H}}$ is:

$$0.158 \pm 2.01 \times 0.019 = (0.12, 0.196)$$

- Interpret this CI in context. Note: direction of difference matters!
- Are average voice shimmers different between people with and without Parkinson's? Briefly explain why or why not.