Elementary Statistics – Inference for Numerical Data Pt. 2

Dr. Ab Mosca (they/them)

Slides based off slides courtesy of Kaitlyn Cook (https://www.smith.edu/people/kaitlyn-cook)

Plan for Today

- Inference for numerical variables
  - Paired means

Inference for Two Independent Means

Confidence Interval for Difference Between Two Independent Means

For confidence intervals, we use s_1 and s_2 as the best guess of σ_1 and σ_2, so

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

For degrees of freedom, use the smaller of $n_1 - 1, n_2 - 1$

We use SE to compute margin of error for our confidence interval:

$$\left((\bar{x}_1 - \bar{x}_2) - t_{df}^{*} SE, (\bar{x}_1 - \bar{x}_2) + t_{df}^{*} SE\right)$$

Inference for Two Independent Means

Confidence Interval for Difference Between Two Independent Means

For confidence intervals, we use s_1 and s_2 as the best guess of σ_1 and σ_2, so

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

For degrees of freedom, use the smaller of $n_1 - 1, n_2 - 1$

We use SE to compute margin of error for our confidence interval:

$$\left((\bar{x}_1 - \bar{x}_2) - t_{df}^* SE, (\bar{x}_1 - \bar{x}_2) + t_{df}^* SE\right)$$

You suspect WSU and SC students have different numbers of siblings. You perform an experiment to statistically test this suspicion. You sample 100 WSU students and find on average they have 3 siblings. You sample 110 SC students and find on average they have 1 sibling. Calculate a 95% CI for $\mu_{SC} - \mu_{WSU}$ from your $\bar{x}_{SC}$ and $\bar{x}_{WSU}$.

Inference for Two Independent Means

Hypothesis Test for Difference Between Two Independent Means

For a hypothesis test, our null hypothesis will be that there is no difference between means.

We use s_1 and s_2 as the best guess of σ_1 and σ_2, so

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Then, $T = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{SE}$, and df is the smaller of $n_1 - 1, n_2 - 1$

Inference for Two Independent Means

Hypothesis Test for Difference Between Two Independent Means

For a hypothesis test, our null hypothesis will be that there is no difference between means.

We use s_1 and s_2 as the best guess of σ_1 and σ_2, so

$$SE(\bar{x}_1 - \bar{x}_2) = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Then, $T = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{SE}$, and df is the smaller of $n_1 - 1, n_2 - 1$

You suspect WSU and SC students have different numbers of siblings. You perform an experiment to statistically test this suspicion. You sample 100 WSU students and find on average they have 3 siblings. You sample 110 SC students and find on average they have 1 sibling. You perform a hypothesis test with $H_0: \mu_{SC} - \mu_{WSU} = 0, H_A: \mu_{SC} - \mu_{WSU} \neq 0$. Finish the hypothesis test. Calculate T, find the p-value, and compare to an α of 0.05.

Inference for Dependent Means

So far, we have looked at inference for categorical variables, single means, and independent means.

All of the tests we covered had the requirement that observations in the data are independent.

Inference for Dependent Means

So far, we have looked at inference for categorical variables, single means, and independent means.

All of the tests we covered had the requirement that observations in the data are independent.

But what if observations are not independent?

Inference for Dependent Means

So far, we have looked at inference for categorical variables, single means, and independent means.

All of the tests we covered had the requirement that observations in the data are independent.

But what if observations are not independent?

Sometimes, dependency cannot be addresses through a statistical method. However, one specific type of dependency, ***pairing***, can be with tools we already know.

Paired Means

Paired data result from a specific type of experimental structure.

In this type of experiment:

- The observational unit is paired across two levels of the explanatory variable
- For each observational unit, quantitative measures are made on each of the two levels of the explanatory variable. The two measurements are subtracted and only the difference is retained

Paired Means

Paired data result from a specific type of experimental structure.

In this type of experiment:

- The observational unit is paired across two levels of the explanatory variable
- For each observational unit, quantitative measures are made on each of the two levels of the explanatory variable. The two measurements are subtracted and only the difference is retained

Observational Unit	Explanatory Variable	Measurement	Response Variable
Car	Smooth Turn vs. Quick Turn	Amount of tire tread after 1000 miles	Difference in tread

Paired Means

Paired data result from a specific type of experimental structure.

In this type of experiment:

- The observational unit is paired across two levels of the explanatory variable
- For each observational unit, quantitative measures are made on each of the two levels of the explanatory variable. The two measurements are subtracted and only the difference is retained

Observational Unit	Explanatory Variable	Measurement	Response Variable
Car	Smooth Turn vs. Quick Turn	Amount of tire tread after 1000 miles	Difference in tread
Textbook	UCLA vs. Amazon	Price of new text	Difference in price

Are the following data paired? If yes, identify the observational unit, explanatory variable, measurement, and response variable.

- Compare pre- (beginning of semester) and post-test (end of semester) scores of students.
- Assess gender-related salary gap by comparing salaries of randomly sampled men and women.
- Compare artery thicknesses at the beginning of a study and after 2 years of taking Vitamin E for the same group of patients.
- Assess effectiveness of a diet regimen by comparing the before and after weights of subjects.

Observational Unit	Explanatory Variable	Measurement	Response Variable
Car	Smooth Turn vs. Quick Turn	Amount of tire tread after 1000 miles	Difference in tread
Textbook	UCLA vs. Amazon	Price of new text	Difference in price

Inference for Paired Means

Our sample statistic ($\bar{x}_{\text{diff}}$) represents our best guess for the true population parameter, (μ_{diff}). We know this best guess is not perfect; we expect error (variability) due to the sampling process.

Because we can't know the truth directly, we infer the truth via:

1. A confidence interval
2. A hypothesis test

Inference for Paired Means

In either case, we need the sampling distribution for $\bar{x}_{diff}$.

We can approximate it via the central limit theorem as long as:

1. The sample's observations are **independent**
2. The sample size is **large enough**, $n \geq 30$, or clearly normally distributed with no outliers

When these conditions are met, variability of $\bar{x}_{diff}$ is well described by:

$$SE(\bar{x}_{diff}) = \frac{best\ guess\ of\ \sigma_{diff}}{\sqrt{n_{diff}}}$$

Inference for Paired Means

In either case, we need the sampling distribution for $\bar{x}_{diff}$.

We can approximate it via the central limit theorem as long as:

1. The sample's observations are **independent**
2. The sample size is **large enough**, $n \geq 30$, or clearly normally distributed with no outliers

When these conditions are met, variability of $\bar{x}_{diff}$ is well described by:

$$SE(\bar{x}_{diff}) = \frac{best\ guess\ of\ \sigma_{diff}}{\sqrt{n_{diff}}}$$

We typically use s_{diff} (sample variance), as the best guess for σ_{diff} (population variance). However, this is less precise with small samples.
As a solution, we use the t-distribution to model the sampling distribution of $\bar{x}_{diff}$

Confidence Interval for Paired Means

Inference for Paired Means

For confidence intervals, we use $\bar{x}_{diff}$ as the best guess of μ_{diff}, and s_{diff} as the best guess of σ_{diff}, so

$$SE(\bar{x}_{diff}) = \frac{s_{diff}}{\sqrt{n_{diff}}}$$

Confidence Interval for Paired Means

Inference for Paired Means

For confidence intervals, we use $\bar{x}_{diff}$ as the best guess of μ_{diff}, and s_{diff} as the best guess of σ_{diff}, so

$$SE(\bar{x}_{diff}) = \frac{s_{diff}}{\sqrt{n_{diff}}}$$

We use SE to compute margin of error for our confidence interval:

$$\left(\bar{x}_{diff} - t^{*}SE, \bar{x}_{diff} + t^{*}\text{SE}\right)$$

Confidence Interval for Paired Means

Inference for Paired Means

For confidence intervals, we use $\bar{x}_{diff}$ as the best guess of μ_{diff}, and s_{diff} as the best guess of σ_{diff}, so

$$SE(\bar{x}_{diff}) = \frac{s_{diff}}{\sqrt{n_{diff}}}$$

We use SE to compute margin of error for our confidence interval:

$$(\bar{x}_{diff} - t^{*}SE, \bar{x}_{diff} + t^{*}\mathrm{SE})$$

t^{*}_{df} is calculated from a specified percentile on the t-distribution with df.

Ex. 5th percentile of a for a 95% confidence

Confidence Interval for Paired Means

Inference for a Single Mean

For confidence intervals, we use $\bar{x}_{diff}$ as the best guess of μ_{diff}, and s_{diff} as the best guess of σ_{diff}, so

$$SE(\bar{x}_{diff}) = \frac{s_{diff}}{\sqrt{n_{diff}}}$$

We use SE to compute margin of error for our confidence interval:
$(\bar{x}_{diff} - t^{*}SE, \bar{x}_{diff} + t^{*}\mathrm{SE})$

t^{*}_{df} is calculated from a specified percentile on the t-distribution with df.

You suspect that on average WSU students work more during the summer than during the semester. You perform an experiment to statistically test this suspicion. You sample 100 students and calculate the difference in average number of hours worked per week between the summer and semester to be 30. Calculate a 95% CI for μ_{diff} from your $\bar{x}_{diff}$.

Hypothesis Test for Paired Means

Inference for Paired Means

When the conditions are met so that the distribution of $\bar{x}_{\mathrm{diff}}$ can be modeled with a t-distribution, variability of $\bar{x}_{diff}$ is well described by:

$$SE\left(\bar{x}_{diff}\right) = \frac{s_{diff}}{\sqrt{n_{diff}}}$$

Hypothesis Test for Paired Means

Inference for Paired Means

When the conditions are met so that the distribution of $\bar{x}_{\mathrm{diff}}$ can be modeled with a t-distribution, variability of $\bar{x}_{diff}$ is well described by:

$$SE(\bar{x}_{diff}) = \frac{s_{diff}}{\sqrt{n_{diff}}}$$

Because we are using the t-distribution, we will need a T-score to find our p-value:

$$T = \frac{\bar{x}_{diff} - \mu_{0diff}}{SE}$$

Degrees of freedom, $df = n_{diff} - 1$

Inference for Paired Means

Hypothesis Test for Paired Means

When the conditions are met so that the distribution of $\bar{x}_{\text{diff}}$ can be modeled with a t-distribution, variability of $\bar{x}_{diff}$ is well described by:

$$SE(\bar{x}_{diff}) = \frac{s_{diff}}{\sqrt{n_{diff}}}$$

Because we are using the t-distribution, we will need a T-score to find our p-value:

$$T = \frac{\bar{x}_{diff} - \mu_{0diff}}{SE}$$

Degrees of freedom, $df = n_{diff} - 1$

You suspect that on average WSU students work more during the summer than during the semester. You perform an experiment to statistically test this suspicion. You sample 100 students and calculate the difference in average number of hours worked per week between the summer and semester to be 30. Perform a hypothesis test for $H_0: \mu_{diff} = 0, H_A: \mu_{diff} \neq 0$. Use $\alpha = 0.05$.

Review

Work with a small group on the inference-review.pdf problem set under Demos on the course website.