Unit 4: Inference for numerical data
1. Inference using the *t*-distribution

Sta 101 - Spring 2019

Duke University, Department of Statistical Science

Dr. Ellison

Slides posted at
https://www2.stat.duke.edu/courses/Spring19/sta101.001/

Outline

1. Housekeeping
2. Main ideas
    1. Problem: Extra uncertainty is introduced into CLT hypothesis testing and confidence intervals when we plug in s for σ.
        Old Solution (from Unit 3): When we don't know σ, only proceed with CLT hypothesis testing (or confidence interval) if n>30. But we still have some unaccounted for uncertainty of approximating σ with s.
        Better Solution (Use from Now on):
        - Use T-distribution instead of Z-distribution when you plug in s for σ
    2. Other Hypothesis Tests and Confidence Intervals you Can Make:
        - When comparing means of two groups, details depend on paired or independent
        - All other details of the inferential framework is the same...

Announcements

Coming up…

- Lab Assignment 6 is due **Thursday just before your lab section time.**
- Peer Evaluations is due **Thursday 2/28 11:55pm** *(part of your participation grade)*
- Read over project statement before **Thursday 2/28**
- Data Exploration Project is due **Thursday 3/7**

Outline

1. Housekeeping
2. Main ideas
    1. Problem: Extra uncertainty is introduced into CLT hypothesis testing and confidence intervals when we plug in s for σ.
        Old Solution (from Unit 3): When we don't know σ, only proceed with CLT hypothesis testing (or confidence interval) if n>30. But we still have some unaccounted for uncertainty of approximating σ with s.
        Better Solution (Use from Now on):
        - Use T-distribution instead of Z-distribution when you plug in s for σ
    2. Other Hypothesis Tests and Confidence Intervals you Can Make:
        - When comparing means of two groups, details depend on paired or independent
        - All other details of the inferential framework is the same...

Outline

Central Limit Theorem

Under the right conditions, the distribution of the sample means (sampling dist) is well approximated by a normal distribution:

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

Outline

Central Limit Theorem

Under the right conditions, the distribution of the sample means (sampling dist) is well approximated by a normal distribution:

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

Outline

Central Limit Theorem

Under the right conditions, the distribution of the sample means (sampling dist) is well approximated by a normal distribution:

$$\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$$

Problem: Extra uncertainty is introduced into CLT hypothesis testing and confidence intervals when we plug in *s* for *σ*.

Ex: $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$

.

Outline

Problem: Extra uncertainty is introduced into CLT hypothesis testing and confidence intervals when we plug in *s* for *σ*.

Ex: $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$

Old Solution (from Unit 3):

When we don't know σ, only proceed with CLT hypothesis testing (or confidence interval) if n>30 *(even if the population was normal)*.

"Old Rules" from Unit 3…

Outline

When can we make a CLT confidence interval or hypothesis test?

What we know from Unit 3

Outline

Making a Confidence Interval for μ with CLT

Independence

1. ✓Random sampling/assignment is used.
2. ✓ Sample size n < 10% of population

✓ One of the available **Sample Size/Skewness "Scenarios"** is met

SCENARIOS	σ is known	σ is not known (have s)
Scenarios: n >30	$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm z^* \frac{s}{\sqrt{n}}$
Scenarios: • n ≤ 30 AND • population distribution IS approximately normal.	$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$	✗
Scenarios: • n ≤ 30 AND • population distribution IS NOT approximately normal.	✗	✗

2

What we know from Unit 3

Outline

Hypothesis Testing for μ with CLT

Independence

1. ✓Random sampling/assignment is used.
2. ✓ Sample size n < 10% of population

✓ One of the available **Sample Size/Skewness "Scenarios"** is met

SCENARIOS	σ is known	σ is not known (have s)
Scenarios: n >30	Test Stat $Z = \frac{\bar{x} - (null\ value)}{\sigma/\sqrt{n}}$	Test Stat $Z = \frac{\bar{x} - (null\ value)}{s/\sqrt{n}}$
Scenarios: • n ≤ 30 AND • population distribution IS approximately normal.	Test Stat $Z = \frac{\bar{x} - (null\ value)}{\sigma/\sqrt{n}}$	✗
Scenarios: • n ≤ 30 AND • population distribution IS NOT approximately normal.	✗	✗

2

Outline

1. Housekeeping
2. Main ideas
   1. Problem: Extra uncertainty is introduced into CLT hypothesis testing and confidence intervals when we plug in s for σ.

      Old Solution (from Unit 3): When we don't know σ, only proceed with CLT hypothesis testing (or confidence interval) if n>30. But we still have some unaccounted for uncertainty of approximating σ with s.

      Better Solution (Use from Now on):
      - 🔍👫⚙ Use T-distribution instead of Z-distribution when you plug in s for σ

   2. Other Hypothesis Tests and Confidence Intervals you Can Make:
      - 🔍 When comparing means of two groups, details depend on paired or independent
      - 🔍 All other details of the inferential framework is the same…

Outline

Problem: Extra uncertainty is introduced into CLT hypothesis testing and confidence intervals when we plug in s for σ.

Ex: $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$ → $\bar{x} \pm z^* \frac{s}{\sqrt{n}}$

Old Solution (from Unit 3):

When we don't know σ, only proceed with CLT hypothesis testing (or confidence interval) if n>30.

Issues:

- We want to make confidence intervals and hypothesis tests for n ≤ 30.
- We still have some unaccounted for uncertainty of approximating σ with s.

Outline

Unit 4 onward…

Using the T-distribution can gives us more flexibility.

Unit 4 Outline

Making a Confidence Interval for μ with CLT

Independence

1. ✓Random sampling/assignment is used.
2. ✓ Sample size n < 10% of population

✓ One of the available **Sample Size/Skewness "Scenarios"** is met

SCENARIOS	σ is known	σ is not known (have s)
Scenarios: n >30	$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm t^*_{n-1} \frac{s}{\sqrt{n}}$
Scenarios: • n ≤ 30 AND • population distribution IS approximately normal.	$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm t^*_{n-1} \frac{s}{\sqrt{n}}$ *or not extremely skewed
Scenarios: • n ≤ 30 AND • population distribution IS NOT approximately normal.	✗	✗

2

Unit 4 Outline

Hypothesis Testing for μ with CLT

Independence

1. ✓Random sampling/assignment is used.
2. ✓ Sample size n < 10% of population

✓ One of the available **Sample Size/Skewness "Scenarios"** is met

SCENARIOS	σ is known	σ is not known (have s)
Scenarios: n >30	Test Stat $Z = \frac{\bar{x} - (null\ value)}{\sigma/\sqrt{n}}$	Test Stat $T_{n-1} = \frac{\bar{x} - (null\ value)}{s/\sqrt{n}}$
Scenarios: • n ≤ 30 AND • population distribution IS approximately normal.	Test Stat $Z = \frac{\bar{x} - (null\ value)}{\sigma/\sqrt{n}}$	Test Stat $T_{n-1} = \frac{\bar{x} - (null\ value)}{s/\sqrt{n}}$ *or not extremely skewed
Scenarios: • n ≤ 30 AND • population distribution IS NOT approximately normal.	✗	✗

2

Outline

Problem: Extra uncertainty is introduced into CLT hypothesis testing and confidence intervals when we plug in s for σ.

Ex: $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$ → $\bar{x} \pm z^* \frac{s}{\sqrt{n}}$

Old Solution (from Unit 3):

When we don't know σ, only proceed with CLT hypothesis testing (or confidence interval) if n>30.

Issues:

- We want to make confidence intervals and hypothesis tests for n ≤ 30.
- We still have some unaccounted for uncertainty of approximating σ with s.

Outline

Unit 4 onward…

Using the T-distribution can incorporate the uncertainty of using s when we don't know σ.

Unit 4

Outline

Making a Confidence Interval for μ with CLT

Independence

1. ✓Random sampling/assignment is used.
2. ✓ Sample size n < 10% of population

✓ One of the available **Sample Size/Skewness "Scenarios"** is met

SCENARIOS	σ is known	σ is not known (have s)
Scenarios: n >30	$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm t^*_{n-1} \frac{s}{\sqrt{n}}$
Scenarios: • n ≤ 30 AND • population distribution IS approximately normal.	$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$	$\bar{x} \pm t^*_{n-1} \frac{s}{\sqrt{n}}$ *or not extremely skewed
Scenarios: • n ≤ 30 AND • population distribution IS NOT approximately normal.	✗	✗

2

Unit 4

Outline

Hypothesis Testing for μ with CLT

Independence

1. ✓Random sampling/assignment is used.
2. ✓ Sample size n < 10% of population

✓ One of the available **Sample Size/Skewness "Scenarios"** is met

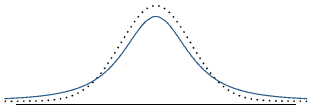
SCENARIOS	σ is known	σ is not known (have s)
Scenarios: n >30	Test Stat $Z = \frac{\bar{x} - (null\ value)}{\sigma/\sqrt{n}}$	Test Stat $T_{n-1} = \frac{\bar{x} - (null\ value)}{s/\sqrt{n}}$
Scenarios: • n ≤ 30 AND • population distribution IS approximately normal.	Test Stat $Z = \frac{\bar{x} - (null\ value)}{\sigma/\sqrt{n}}$	Test Stat $T_{n-1} = \frac{\bar{x} - (null\ value)}{s/\sqrt{n}}$ *or not extremely skewed
Scenarios: • n ≤ 30 AND • population distribution IS NOT approximately normal.	✗	✗

2

Outline

Properties of the T-distribution:

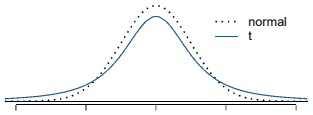
How is it similar/different to the normal distribution?



2. T corrects for uncertainty introduced by plugging in s for σ

T-distribution is more "conservative" distribution than the normal distribution.

- *t*-distribution also has a bell shape, but:
  - Peak is *lower* than the normal model's
  - Tails are *thicker* than the normal model's
    - Observations are more likely to fall beyond two SDs from the mean than under the normal distribution.

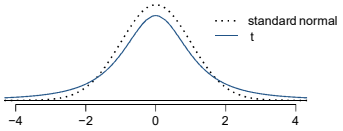


2

2. T corrects for uncertainty introduced by plugging in s for σ

T-distribution is more "conservative" distribution than the normal distribution.

- *t*-distribution also has a bell shape, but:
  - Peak is *lower* than the normal model's
  - Tails are *thicker* than the normal model's
    - Observations are more likely to fall beyond two SDs from the mean than under the normal distribution.
  - Always centered at zero, like the standard normal (*z*) distribution

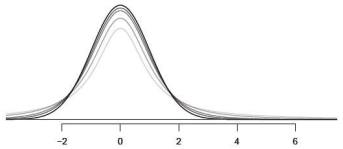


2

Outline

Properties of the T-distribution:

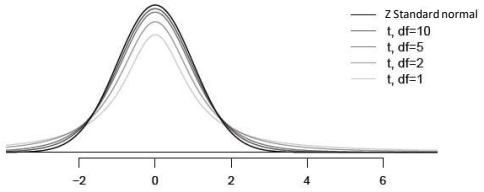
What is the "parameter" that determine the tail thickness/peak height of t-distribution?



t-distribution

- Has a single parameter, *degrees of freedom* (df), that is tied to sample size. Determines tail thickness, peak height.

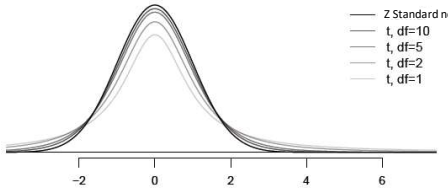
What happens to shape of the t-distribution as df increases?



t-distribution

- Has a single parameter, *degrees of freedom* (df), that is tied to sample size. Determines tail thickness, peak height.

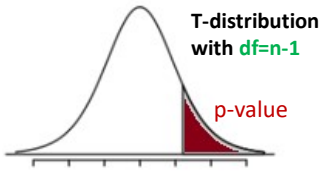
What happens to shape of the t-distribution as df increases?



df↑
- thickness of tails↓
- peak ↑
- approaches standard normal dist

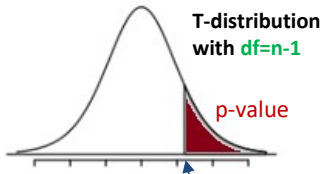
Outline

How do we use the t-distribution for hypothesis testing for one population mean?



Outline

How do we use the t-distribution for hypothesis testing for one population mean?

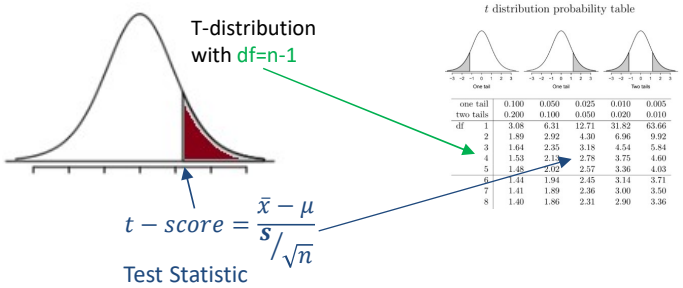


$$t - score = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

Test Statistic

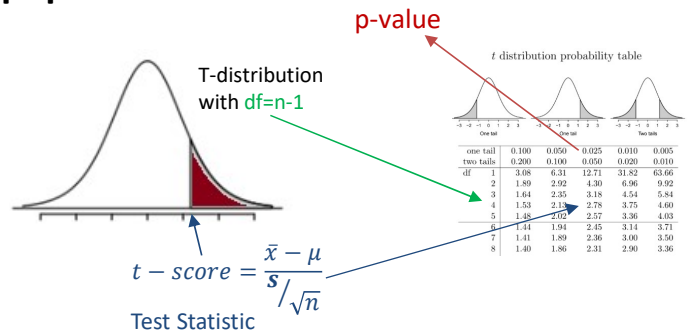
Outline

How do we use the t-distribution for hypothesis testing for one population mean?



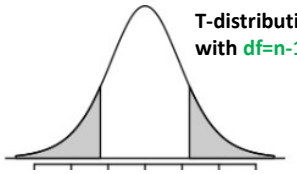
Outline

How do we use the t-distribution for hypothesis testing for one population mean?



Outline

How do we use the t-distribution for confidence intervals for one population mean?

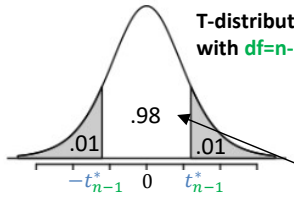


98% Confidence Interval

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

Outline

How do we use the t-distribution for confidence intervals for one population mean?

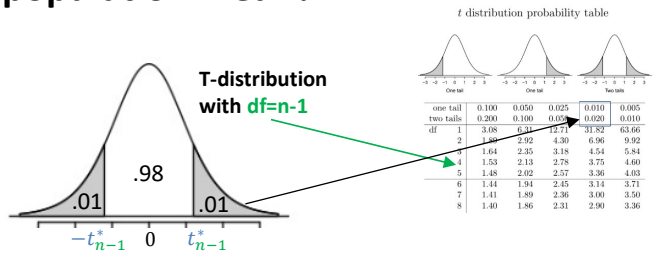


98% Confidence Interval

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

Outline

How do we use the t-distribution for confidence intervals for one population mean?

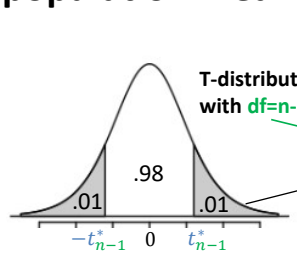


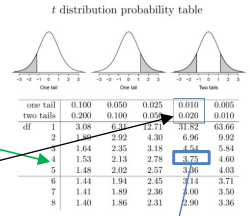
98% Confidence Interval

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

Outline

How do we use the t-distribution for confidence intervals for one population mean?





98% Confidence Interval

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

Unit 4 onward…

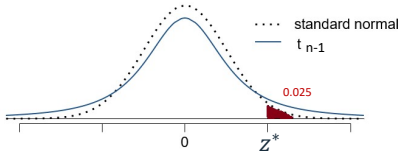
Outline

How/why does the T-distribution incorporate the uncertainty of using s when we don't know σ?

Clicker question

The critical value z^* for a 95% confidence interval constructed using $\bar{x} \pm z^*\frac{\sigma}{\sqrt{n}}$ is shown below. Will a 95% confidence interval constructed using $\bar{x} \pm t^*_{n-1}\frac{\sigma}{\sqrt{n}}$ be wider or narrower?

a.) wider
b.) narrower

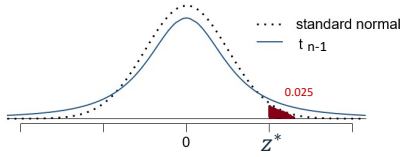


Clicker question

The critical value z^* for a 95% confidence interval constructed using $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$ is shown below. Will a 95% confidence interval constructed using $\bar{x} \pm t^*_{n-1} \frac{\sigma}{\sqrt{n}}$ be wider or narrower?

a.) wider

b.) narrower

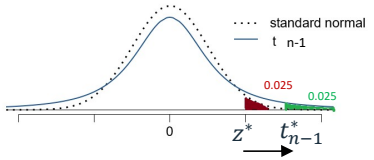


Clicker question

The critical value z^* for a 95% confidence interval constructed using $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$ is shown below. Will a 95% confidence interval constructed using $\bar{x} \pm t^*_{n-1} \frac{\sigma}{\sqrt{n}}$ be wider or narrower?

a.) wider

b.) narrower



Outline

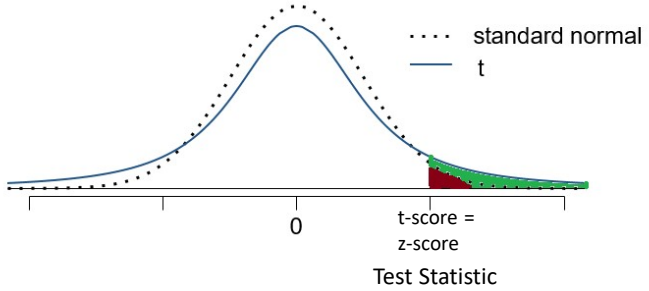
For large confidence levels, the T-distribution's thicker tails lead to…

→ wider confidence intervals

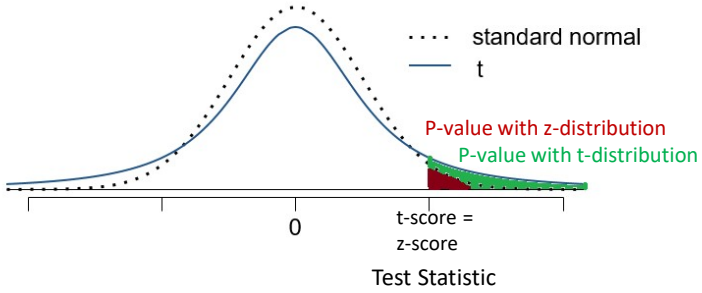
→ more uncertainty about pop. param.

Outline

Hypothesis Testing



Hypothesis Testing



For large z-scores/t-scores, the T-distribution's thicker tails lead to…

→ higher p-values

→ harder to reject the null hypothesis.

1. Housekeeping
2. Main ideas
    1. Problem: Extra uncertainty is introduced into CLT hypothesis testing and confidence intervals when we plug in s for σ.
        Old Solution (from Unit 3): When we don't know σ, only proceed with CLT hypothesis testing (or confidence interval) if n>30. But we still have some unaccounted for uncertainty of approximating σ with s.
        Better Solution (Use from Now on):
        - 🔍 👫 ☼ Use T-distribution instead of Z-distribution when you plug in s for σ

    2. Other Hypothesis Tests and Confidence Intervals you Can Make:
        - 🔍 When comparing means of two groups, details depend on paired or independent
        - 🔍 All other details of the inferential framework is the same...

Confidence Intervals and Hypothesis Testing for Other Population Parameters:

μ_{diff}

Population "mean difference" of paired observations

Example 1: Zinc in water

Trace metals in drinking water affect the flavor and an unusually high concentration can pose a health hazard. Ten pairs of data were taken measuring zinc concentration in bottom water and surface water at 10 randomly sampled locations.

Location	bottom	surface
1	0.43	0.415
2	0.266	0.238
3	0.567	0.39
4	0.531	0.41
5	0.707	0.605
6	0.716	0.609
7	0.651	0.632
8	0.589	0.523
9	0.469	0.411
10	0.723	0.612

4

Example 1: Zinc in water

Trace metals in drinking water affect the flavor and an unusually high concentration can pose a health hazard. Ten pairs of data were taken measuring zinc concentration in bottom water and surface water at 10 randomly sampled locations.

Location	bottom	surface
1	0.43	0.415
2	0.266	0.238
3	0.567	0.39
4	0.531	0.41
5	0.707	0.605
6	0.716	0.609
7	0.651	0.632
8	0.589	0.523
9	0.469	0.411
10	0.723	0.612

Water samples collected at the same location, on the surface and in the bottom, cannot be assumed to be independent of each other, hence we need to use a *paired* analysis.

Source: https://onlinecourses.science.psu.edu/stat500/node/51

4

Example 1: Zinc in water

Trace metals in drinking water affect the flavor and an unusually high concentration can pose a health hazard. Ten pairs of data were taken measuring zinc concentration in bottom water and surface water at 10 randomly sampled locations.

Pairing = Location	bottom	surface
1	0.43	0.415
2	0.266	0.238
3	0.567	0.39
4	0.531	0.41
5	0.707	0.605
6	0.716	0.609
7	0.651	0.632
8	0.589	0.523
9	0.469	0.411
10	0.723	0.612

Identifying a Paired Means Test

- Each observation in one population has a corresponding observation in the other population. The problem usually talks about this correspondence/pairing.

Source: https://onlinecourses.science.psu.edu/stat500/node/51

4

Example 1: Zinc in water

Trace metals in drinking water affect the flavor and an unusually high concentration can pose a health hazard. Ten pairs of data were taken measuring zinc concentration in bottom water and surface water at 10 randomly sampled locations.

Location	bottom	surface
1	0.43	0.415
2	0.266	0.238
3	0.567	0.39
4	0.531	0.41
5	0.707	0.605
6	0.716	0.609
7	0.651	0.632
8	0.589	0.523
9	0.469	0.411
10	0.723	0.612

Identifying a Paired Means Test

- Each observation in one population has a corresponding observation in the other population. The problem will talk about this correspondence/pairing.
- The sample sizes of two groups HAVE to be the same.

Source: https://onlinecourses.science.psu.edu/stat500/node/51

4

Example 1: Zinc in water

Trace metals in drinking water affect the flavor and an unusually high concentration can pose a health hazard. Ten pairs of data were taken measuring zinc concentration in bottom water and surface water at 10 randomly sampled locations.

Location	bottom	surface
1	0.43	0.415
2	0.266	0.238
3	0.567	0.39
4	0.531	0.41
5	0.707	0.605
6	0.716	0.609
7	0.651	0.632
8	0.589	0.523
9	0.469	0.411
10	0.723	0.612

Source: https://onlinecourses.science.psu.edu/stat500/node/51

Identifying a Paired Means Test

- Each observation in one population has a corresponding observation in the other population. The problem will talk about this correspondence/pairing.
- The sample sizes of two groups HAVE to be the same.
- Common paired-means test examples:
  - Before/after data
  - Couples/twins

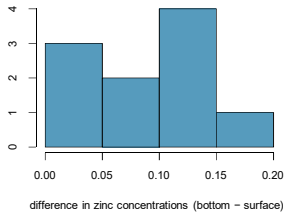
4

Analyzing paired data

Suppose we want to compare the average zinc concentration levels in the bottom and surface:

- Two sets of observations with a special correspondence (not independent): *paired*
- Synthesize down to differences in outcomes of each pair of observations, subtract using a consistent order

Location	bottom	surface	difference
1	0.43	0.415	0.015
2	0.266	0.238	0.028
3	0.567	0.39	0.177
4	0.531	0.41	0.121
5	0.707	0.605	0.102
6	0.716	0.609	0.107
7	0.651	0.632	0.019
8	0.589	0.523	0.066
9	0.469	0.411	0.058
10	0.723	0.612	0.111



7

Parameter and point estimate for paired data

For comparing average zinc concentration levels in the bottom and surface when the data are paired:

8

Parameter and point estimate for paired data

For comparing average zinc concentration levels in the bottom and surface when the data are paired:

- *Parameter of interest:* Average difference between the bottom and surface zinc measurements of *all* drinking water.

$$\mu_{diff}$$

8

Parameter and point estimate for paired data

For comparing average zinc concentration levels in the bottom and surface when the data are paired:

- *Parameter of interest:* Average difference between the bottom and surface zinc measurements of *all* drinking water.

$$\mu_{diff}$$

- *Point estimate:* Average difference between the bottom and surface zinc measurements of drinking water from the *sampled* locations.

$$\bar{x}_{diff}$$

8

Parameter and point estimate for paired data

For comparing average zinc concentration levels in the bottom and surface when the data are paired:

- *Parameter of interest:* Average difference between the bottom and surface zinc measurements of *all* drinking water.

$$\mu_{diff}$$

- *Point estimate:* Average difference between the bottom and surface zinc measurements of drinking water from the *sampled* locations.

$$\bar{x}_{diff}$$

- *Standard deviation:* Standard deviation of the differences between the bottom and surface zinc measurements of drinking water from the *sampled* locations.

$$s_{diff}$$

8

Standard errors

- Dependent (paired) groups (e.g. pre/post weights of subjects in a weight loss study, twin studies, etc.)

$$SE_{\bar{x}_{diff}} = \frac{s_{diff}}{\sqrt{n_{diff}}}$$

3. All other details of the inferential framework is the same...

$$HT: test\ statistic = \frac{point\ estimate - null}{SE}$$

$$CI: point\ estimate \pm critical\ value \times SE$$

Making a Confidence Interval for μ_{diff} with CLT

Independence

1. ✓Random sampling/assignment is used to collect the pairs of data
2. ✓ Sample size n < 10% of population of pairs.

✓ One of the available **Sample Size/Skewness "Scenarios"** is met

SCENARIOS	σ_{diff} is known	σ_{diff} is not known (have s_{diff})
Scenarios: n >30 n=# of pairs	$\bar{x}_{diff} \pm z^* \frac{\sigma_{diff}}{\sqrt{n}}$	$\bar{x}_{diff} \pm t^*_{n-1} \frac{s_{diff}}{\sqrt{n}}$
Scenarios: • n ≤ 30 AND • population distribution of differences IS approximately normal. n=# of pairs	$\bar{x}_{diff} \pm z^* \frac{\sigma_{diff}}{\sqrt{n}}$	$\bar{x}_{diff} \pm t^*_{n-1} \frac{s_{diff}}{\sqrt{n}}$ *or not extremely skewed
Scenarios: • n ≤ 30 AND • population distribution of differences IS NOT approximately normal. n=# of pairs	✗	✗

Hypothesis Testing for μ_{diff} with CLT

Independence

1. ✓Random sampling/assignment is used to collect the pairs of data
2. ✓ Sample size n < 10% of population of pairs.

✓ One of the available **Sample Size/Skewness "Scenarios"** is met

SCENARIOS	σ_{diff} is known	σ_{diff} is not known (have s_{diff})
Scenarios: n >30 n=# of pairs	Test Stat $Z = \frac{\bar{x}_{diff} - (null\ value)}{\sigma_{diff}/\sqrt{n}}$	Test Stat $T_{n-1} = \frac{\bar{x}_{diff} - (null\ value)}{s_{diff}/\sqrt{n}}$
Scenarios: • n ≤ 30 AND • population distribution of differences IS approximately normal. n=# of pairs	Test Stat $Z = \frac{\bar{x}_{diff} - (null\ value)}{\sigma_{diff}/\sqrt{n}}$	Test Stat $T_{n-1} = \frac{\bar{x}_{diff} - (null\ value)}{s_{diff}/\sqrt{n}}$ *or not extremely skewed
Scenarios: • n ≤ 30 AND • population distribution of differences IS NOT approximately normal. n=# of pairs	✗	✗

Confidence Intervals and Hypothesis Testing for Other Population Parameters:

$$\mu_{group1} - \mu_{group2}$$

Difference of Population Means from Independent Populations

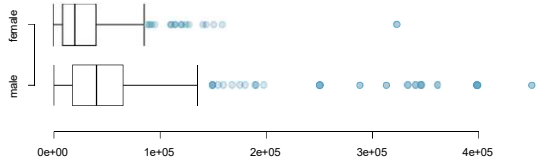
For comparing average salaries in two independent groups

https://www.wsj.com/articles/parsing-the-gender-pay-gap-1542917969

Example 2: Gender gap in salaries

Since 2005, the American Community Survey[1] polls ~3.5 million households yearly. The following summarizes distribution of salaries of males and females from a random sample of individuals who responded to the 2012 ACS:

	$\bar{x}$	s	n
male	55,890	68,767.88	470
female	29,240	32,025.98	373

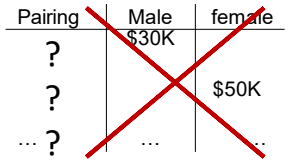


[1]Aside: Surge of media attention in spring 2012 when the House of Representatives voted to eliminate the survey. Daniel Webster, Republican congressman from Florida: "in the end this is not a scientific survey. It's a random survey."

Example 2: Gender gap in salaries

Since 2005, the American Community Survey[1] polls ~3.5 million households yearly. The following summarizes distribution of salaries of males and females from a random sample of individuals who responded to the 2012 ACS:

	$\bar{x}$	s	n
male	55,890	68,767.88	470
female	29,240	32,025.98	373



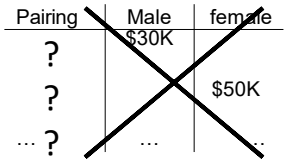
Identifying an Independent Means Test

- Observations in one population have no explicit or obvious pairing with another observation in the other population.

Example 2: Gender gap in salaries

Since 2005, the American Community Survey[1] polls ~3.5 million households yearly. The following summarizes distribution of salaries of males and females from a random sample of individuals who responded to the 2012 ACS:

	$\bar{x}$	s	n
male	55,890	68,767.88	470
female	29,240	32,025.98	373



Identifying an Independent Means Test

- Observations in one population have no explicit or obvious pairing with another observation in the other population.
- If the sample sizes are different → NOT a paired means test!

Parameter and point estimate for independent data

For comparing average salaries in two independent groups

- *Parameter of interest:* Average difference between the average salaries of *all* males and females in the US.

$$\mu_m - \mu_f$$

Parameter and point estimate for independent data

For comparing average salaries in two independent groups

- *Parameter of interest:* Average difference between the average salaries of *all* males and females in the US.

$$\mu_m - \mu_f$$

- *Point estimate:* Average difference between the average salaries of *sampled* males and females in the US.

$$(\bar{x}_1 - \bar{x}_2)$$

9

Standard errors

- Independent groups (e.g. grades of students across two sections)

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

An extension of the CLT says:
Under the right conditions,

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

- For the same data, $SE_{paired} < SE_{independent}$, so be careful about calling data paired

10

3. All other details of the inferential framework is the same...

$$HT: test\ statistic = \frac{point\ estimate - null}{SE}$$

$$CI: point\ estimate \pm critical\ value \times SE$$

11

2. T corrects for uncertainty introduced by plugging in s for σ

Making a Confidence Interval for $\mu_1 - \mu_2$ with CLT

Independence between Groups
✓both populations are independent.

Independence within Groups
1. ✓Random sampling/assignment is used for both samples
2. ✓ Sample size n1 < 10% of population 1 and sample size n2 < 10% of population 2.

✓ One of the available **Sample Size/Skewness "Scenarios"** is met

SCENARIOS	σ1 and σ2 known	σ1 and σ2 not known
Condition: n1 >30 AND n2 >30	$(\bar{x}_1 - \bar{x}_2) \pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$(\bar{x}_1 - \bar{x}_2) \pm t^*_{\min(n1-1, n2-1)} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
Condition: n1 ≤ 30 OR n2 ≤ 30 and both population distributions are approximately normal.	$(\bar{x}_1 - \bar{x}_2) \pm z^* \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$	$(\bar{x}_1 - \bar{x}_2) \pm t^*_{\min(n1-1, n2-1)} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ *Or not extremely skewed (if using t-distribution)
Condition: n1 ≤ 30 OR n2 ≤ 30 and both population distributions are NOT approximately normal.	✗	✗

2. T corrects for uncertainty introduced by plugging in s for σ

Making a Hypothesis Test for $\mu_1 - \mu_2$ with CLT

Independence between Groups

✓both populations are independent.

Independence within Groups

1. ✓Random sampling/assignment is used for both samples
2. ✓ Sample size n1 < 10% of population 1 and sample size n2 < 10% of population 2.

✓ One of the available **Sample Size/Skewness "Scenarios"** is met

SCENARIOS	σ1 and σ2 known	σ1 and σ2 not known
Condition: n1 >30 AND n2 >30	Test Stat $z = \frac{(\bar{x}_1 - \bar{x}_2) - (\text{null value})}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	Test Stat $t_{\min(n1-1,n2-1)} = \frac{(\bar{x}_1 - \bar{x}_2) - (\text{null value})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
Condition: n1 ≤ 30 OR n2 ≤ 30 and both population distributions are approximately normal.	Test Stat $z = \frac{(\bar{x}_1 - \bar{x}_2) - (\text{null value})}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$	Test Stat $t_{\min(n1-1,n2-1)} = \frac{(\bar{x}_1 - \bar{x}_2) - (\text{null value})}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$ *Or not extremely skewed
Condition: n1 ≤ 30 OR n2 ≤ 30 and both population distributions are NOT approximately normal.	✗	✗

Clicker question

Suppose an independent means test has a sample size of n_1 in the first group and n_2 in the second group. What degrees of freedom would you use in the t-distribution if you were constructing a confidence interval or calculating a p-value, using a t-distribution?

a) $\min(n_1 - 1, n_2 - 1)$
b) $(n_1 - 1) + (n_2 - 1)$
c) $n_1 + n_2 - 1$
d) $n-1$

Clicker question

Suppose an independent means test has a sample size of n_1 in the first group and n_2 in the second group. What degrees of freedom would you use in the t-distribution if you were constructing a confidence interval or calculating a p-value, using a t-distribution?

a) $\boldsymbol{\min(n_1 - 1, n_2 - 1)}$
b) $(n_1 - 1) + (n_2 - 1)$
c) $n_1 + n_2 - 1$
d) $n-1$

Outline

1. Housekeeping
2. Main ideas
   1. Problem: Extra uncertainty is introduced into CLT hypothesis testing and confidence intervals when we plug in s for σ.
      Old Solution (from Unit 3): When we don't know σ, only proceed with CLT hypothesis testing (or confidence interval) if n>30. But we still have some unaccounted for uncertainty of approximating σ with s.
      Better Solution (Use from Now on):
      - Use T-distribution instead of Z-distribution when you plug in s for σ

   2. Other Hypothesis Tests and Confidence Intervals you Can Make:
      - When comparing means of two groups, details depend on paired or independent
      - All other details of the inferential framework is the same...

3. All other details of the inferential framework is the same...

$$HT: test\ statistic = \frac{point\ estimate - null}{SE}$$

11

3. All other details of the inferential framework is the same...

$$HT: test\ statistic = \frac{point\ estimate - null}{SE}$$

$$CI: point\ estimate \pm critical\ value \times SE$$

11