Unit 4: Inference for numerical data

1. Inference using the t-distribution

GOVT 3990 - Spring 2018

Cornell University

1. Main ideas

    1. T corrects for uncertainty introduced by plugging in s for σ
    2. When comparing means of two groups, details depend on paired or independent
    3. All other details of the inferential framework is the same…

Outline

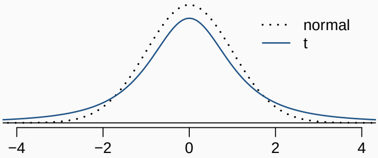
1. Main ideas
    1. T corrects for uncertainty introduced by plugging in s for σ
    2. When comparing means of two groups, details depend on paired or independent
    3. All other details of the inferential framework is the same...

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

- CLT says $\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$, but, in practice, we use s instead of σ.
  - Plugging in an estimate introduces additional uncertainty.
  - We make up for this by using a more “conservative” distribution than the normal distribution.

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

- CLT says $\bar{x} \sim N\left(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}\right)$, but, in practice, we use s instead of σ.
  - Plugging in an estimate introduces additional uncertainty.
  - We make up for this by using a more “conservative” distribution than the normal distribution.
- t-distribution also has a bell shape, but its 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.
  - Extra thick tails help mitigate the effect of a less reliable estimate for the standard error of the sampling distribution.



t-distribution

- Always centered at zero, like the standard normal (z) distribution

t-distribution

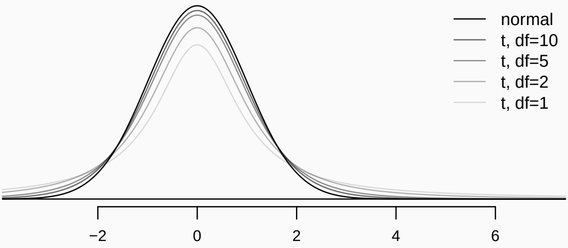
- Always centered at zero, like the standard normal (z) distribution
- Has a single parameter, *degrees of freedom* (df), that is tied to sample size.
  - one sample: $df = n - 1$
  - two (independent) samples: $df = min(n_1 - 1, n_2 - 1)$

- Always centered at zero, like the standard normal (z) distribution
- Has a single parameter, *degrees of freedom* (df), that is tied to sample size.
  - one sample: $df = n - 1$
  - two (independent) samples: $df = min(n_1 - 1, n_2 - 1)$

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

- Always centered at zero, like the standard normal (z) distribution
- Has a single parameter, *degrees of freedom* (df), that is tied to sample size.
  - one sample: $df = n - 1$
  - two (independent) samples: $df = min(n_1 - 1, n_2 - 1)$

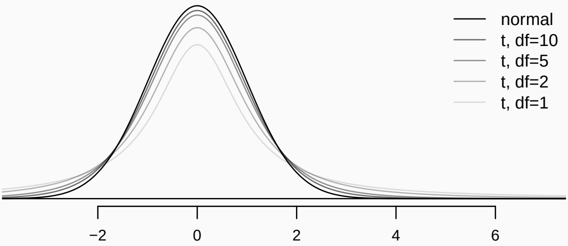
What happens to shape of the t-distribution as df increases?



t-distribution

- Always centered at zero, like the standard normal (z) distribution
- Has a single parameter, *degrees of freedom* (df), that is tied to sample size.
  - one sample: $df = n - 1$
  - two (independent) samples: $df = min(n_1 - 1, n_2 - 1)$

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

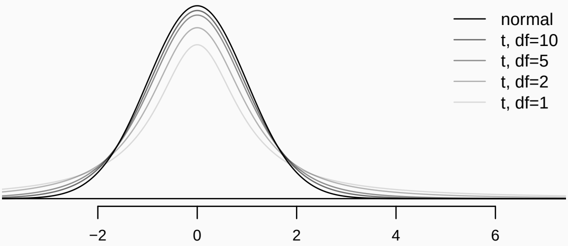


Approaches normal.

t-distribution

- Always centered at zero, like the standard normal (z) distribution
- Has a single parameter, *degrees of freedom* (df), that is tied to sample size.
  - one sample: $df = n - 1$
  - two (independent) samples: $df = min(n_1 - 1, n_2 - 1)$

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



Approaches normal.

Why?

Outline

1. Main ideas
    1. T corrects for uncertainty introduced by plugging in s for σ
    2. When comparing means of two groups, details depend on paired or independent
    3. All other details of the inferential framework is the same...

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

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*

Since 2005, the American Community Survey polls $\sim$3.5 million households yearly...

Example 2: Gender gap in salaries

Since 2005, the American Community Survey polls $\sim$3.5 million households yearly...

Aside: Surge of media attention in spring 2012 when the House of Representatives voted to eliminate the survey. Daniel Webster, Republican congressman from Florida had an interesting opinion:

Example 2: Gender gap in salaries

Since 2005, the American Community Survey polls ∼3.5 million households yearly...

Aside: Surge of media attention in spring 2012 when the House of Representatives voted to eliminate the survey. Daniel Webster, Republican congressman from Florida had an interesting opinion:

The New York Times

The Beginning of the End of the Census?

“This is a program that intrudes on people’s lives, just like the Environmental Protection Agency or the bank regulators,” said Daniel Webster, a first-term Republican congressman from Florida who sponsored the relevant legislation.

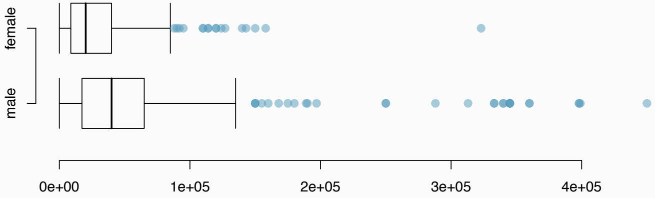
“We’re spending $70 per person to fill this out. That’s just not cost effective,” he continued, “especially since in the end this is not a scientific survey. It’s a random survey.”

In fact, the randomness of the survey is precisely what makes the survey scientific, statistical experts say.

Example 2: Gender gap in salaries

Since 2005, the American Community Survey 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



How are the two examples different from each other? How are they similar to each other?

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

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*

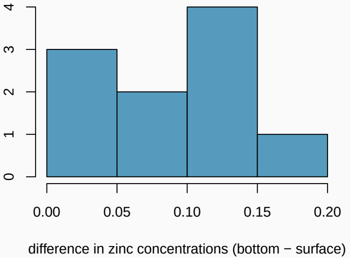
Suppose we want to compare the average zinc concentration levels in the bottom and surface:

- 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

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



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

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.

$$\mu_{diff}$$

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.

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

For comparing average salaries in two independent groups

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.

$$\bar{x}_m - \bar{x}_f$$

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

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

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

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

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

Outline

1. Main ideas

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

    2. When comparing means of two groups, details depend on paired or independent

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

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

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

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

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

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

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

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

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

One mean:

$df = n - 1$

HT:

$H_0 : \mu = \mu_0$
$T_{df} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

CI:

$\bar{x} \pm t^\star_{df} \frac{s}{\sqrt{n}}$

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

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

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

One mean:

$df = n - 1$

HT:

$H_0 : \mu = \mu_0$

$T_{df} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

CI:

$\bar{x} \pm t^{\star}_{df} \frac{s}{\sqrt{n}}$

Paired means:

$df = n_{diff} - 1$

HT:

$H_0 : \mu_{diff} = 0$

$T_{df} = \frac{\bar{x}_{diff} - 0}{\frac{s_{diff}}{\sqrt{n_{diff}}}}$

CI:

$\bar{x}_{diff} \pm t^{\star}_{df} \frac{s_{diff}}{\sqrt{n_{diff}}}$

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

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

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

One mean:

$df = n - 1$

HT:

$H_0 : \mu = \mu_0$

$T_{df} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$

CI:

$\bar{x} \pm t^{\star}_{df} \frac{s}{\sqrt{n}}$

Paired means:

$df = n_{diff} - 1$

HT:

$H_0 : \mu_{diff} = 0$

$T_{df} = \frac{\bar{x}_{diff} - 0}{\frac{s_{diff}}{\sqrt{n_{diff}}}}$

CI:

$\bar{x}_{diff} \pm t^{\star}_{df} \frac{s_{diff}}{\sqrt{n_{diff}}}$

Independent means:

$df = min(n_1 - 1, n_2 - 1)$

HT:

$H_0 : \mu_1 - \mu_2 = 0$

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

CI:

$\bar{x}_1 - \bar{x}_2 \pm t^{\star}_{df} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$