Unit 4: Inference for numerical data

1. Inference using the t-distribution

Sta 101 - Fall 2015

Duke University, Department of Statistical Science

Dr. Çetinkaya-Rundel

Slides posted at *http://bit.ly/sta101_f15*

1. Housekeeping

2. Main ideas
    1. T corrects for uncertainty introduced by plugging in s for σ
    2. When comparing means of two groups, ask if paired or independent

-

1. Housekeeping

2. Main ideas
    1. T corrects for uncertainty introduced by plugging in s for σ
    2. When comparing means of two groups, ask if paired or independent

1. Housekeeping

2. Main ideas
    1. T corrects for uncertainty introduced by plugging in s for σ
    2. When comparing means of two groups, ask if paired or independent

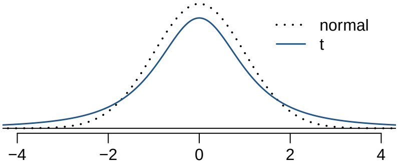
2. T corrects for uncertainty introduced by plugging in s for σ

- CLT says $\bar{x} \sim N(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}$, 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

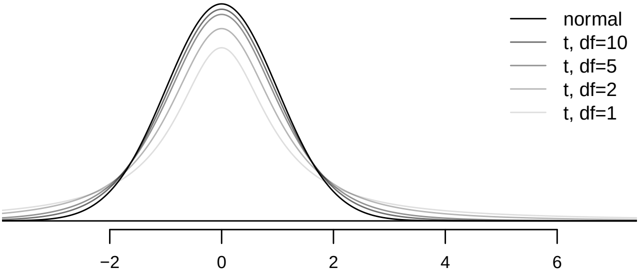
- CLT says $\bar{x} \sim N(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}$, but, in practice, we use s instead of σ.
  - Plugging in an estimate introduces additional uncertainty.
  - We make up for this by using a more "conservative" distribution than the normal distribution.
- Also has a bell shape, but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

- CLT says $\bar{x} \sim N(mean = \mu, SE = \frac{\sigma}{\sqrt{n}}$, 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.
- 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 are helpful for mitigating the effect of a less reliable estimate for the standard error of the sampling distribution.

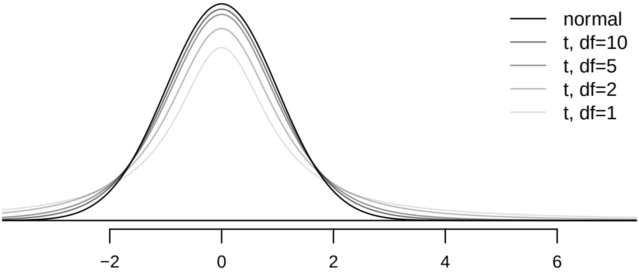


- Always centered at zero, like the standard normal (z) distribution

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

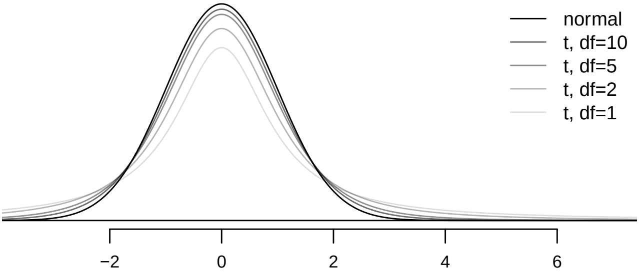


- Always centered at zero, like the standard normal (z) distribution
- Has a single parameter: *degrees of freedom* (df)
  - one sample: $df = n - 1$
  - two (independent) samples: $df = min(n_1 - 1, n_2 - 1)$



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)
  - 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.

1. Housekeeping

2. Main ideas
    1. T corrects for uncertainty introduced by plugging in s for σ
    2. When comparing means of two groups, ask if paired or independent

2. When comparing means of two groups, ask if paired or independent

- 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)

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

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

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*

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:

- When two sets of observations have this special correspondence (not independent), they are said to be *paired*.

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

- When two sets of observations have this special correspondence (not independent), they are said to be *paired*.
- To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations.

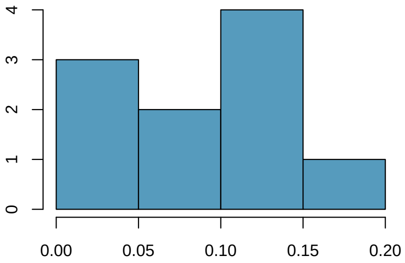
Suppose we want to compare the average zinc concentration levels in the bottom and surface:

- When two sets of observations have this special correspondence (not independent), they are said to be *paired*.
- To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations.
- It is important that we always subtract using a consistent order.

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

- When two sets of observations have this special correspondence (not independent), they are said to be *paired*.
- To analyze paired data, it is often useful to look at the difference in outcomes of each pair of observations.
- It is important that we always 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



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

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.

$$\mu_{diff}$$

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.

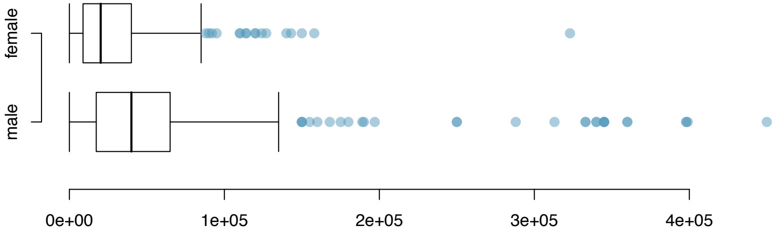
$$\mu_{diff}$$

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

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

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



ACS: 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.”

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