

## Unit 4: Inference for numerical data

1. Inference using the  $t$ -distribution

Sta 104 - Summer 2018, Term 1

Duke University, Department of Statistical Science

Prof. White

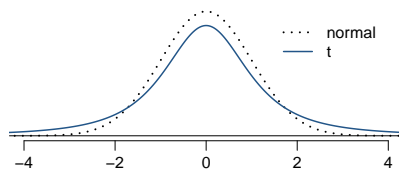
Slides posted at <https://www2.stat.duke.edu/courses/Summer18/sta104.001-1/>

- ▶ Problem Set 3 is due today at 11:55 pm
- ▶ Performance Assessment 3 is due today at 11:55 pm
- ▶ Lab 4 is due tomorrow at 12:45 pm
- ▶ Peer and course evaluation are forthcoming.

1

2.  $T$  corrects for uncertainty introduced by plugging in  $s$  for  $\sigma$ 

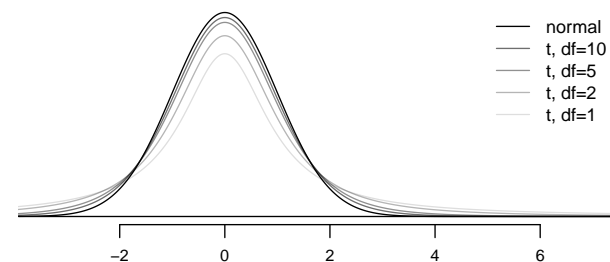
- ▶ CLT says  $\bar{X} \sim N(\text{mean} = \mu, SE = \frac{\sigma}{\sqrt{n}})$ , but, in practice, we use  $s$  instead of  $\sigma$ .
  - Plugging in an estimate introduces additional uncertainty.
  - We make up for this by using a more “conservative” distribution than the normal distribution.
- ▶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



2

 $t$ -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?

Why?

3

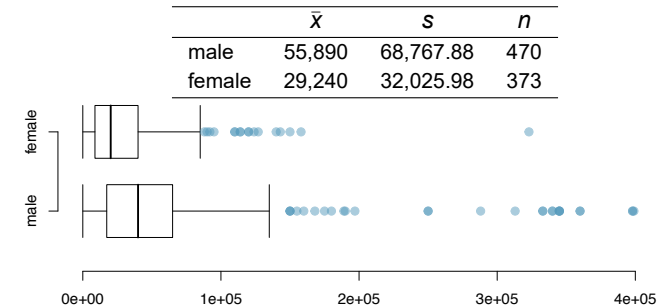
Trace metals in drinking water affect the flavor and an unusually high concentration can pose a health hazard. Ten pairs of data were taken measuring zinc concentration in bottom water and surface water at 10 randomly sampled locations.

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<sup>1</sup> 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:

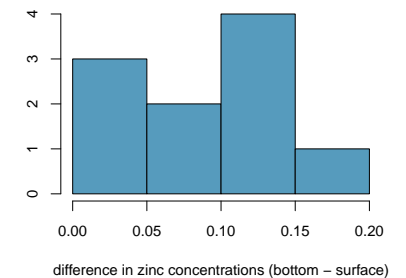


<sup>1</sup>Aside: Surge of media attention in spring 2012 when the House of Representatives voted to eliminate the survey. Daniel Webster, Republican congressman from Florida: “in the end this is not a scientific survey. It’s a random survey.”

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



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

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

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.

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

9

### Standard errors

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

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

- 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

10

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

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

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

**One mean:**

$$df = n - 1$$

**HT:**

$$H_0 : \mu = \mu_0$$

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

**CI:**

$$\bar{x} \pm t_{df}^* \frac{s}{\sqrt{n}}$$

**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:**

$$\bar{x}_{diff} \pm t_{df}^* \frac{s_{diff}}{\sqrt{n_{diff}}}$$

**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:**

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

11