STAT 22000 Lecture Slides
Inference on Population Means Using t-Distributions

Yibi Huang
Department of Statistics
University of Chicago

Outline

This set of slides covers Section 5.1 in the text.

- t-distributions
- t-tests
- t-confidence intervals

What if σ is Unknown?

Recall that if $X_1, X_2, \ldots, X_n$ are i.i.d. (or a SRS) from a population with **unknown mean** μ and standard deviation σ, then when n is large

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \quad \text{is nearly } N(0,1)$$

We use this fact to construct confidence intervals and do hypothesis test about μ in Section 4.2 and 4.3.

As the population SD σ is UNKNOWN, we replace it with the sample SD s. with the restrictions that

- the sample size n need to be big enough, and
- the population distribution cannot be too skewed (no outlier).

This lecture we are going to introduce one way remove the restriction on sample size.

Student's t-Distributions

If $X_1, X_2, \ldots, X_n$ are i.i.d. from $N(\mu, \sigma)$, the *t-Statistic* defined as

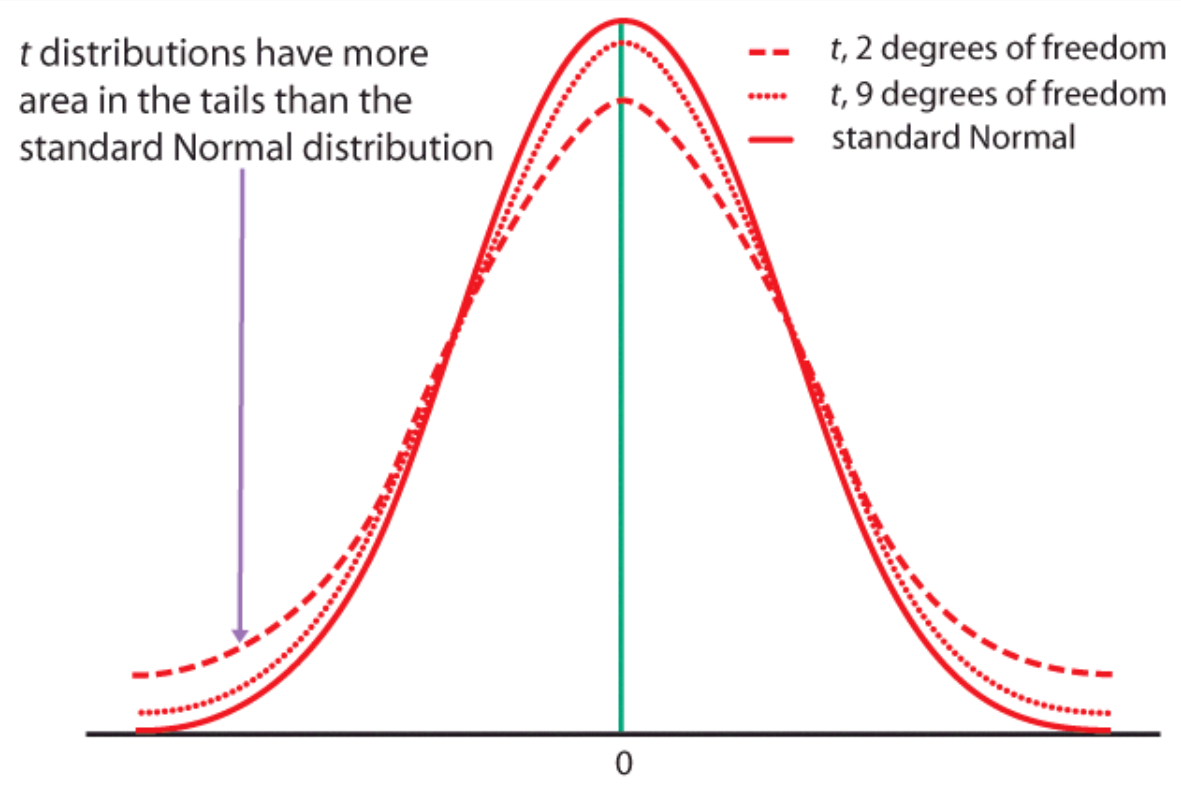
$$t = \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

can be shown to have a *t distribution* with *$n - 1$ degrees of freedom.*

What's is a t-distribution?

Density Curves of t-Distributions

- Bell-shaped, symmetric about 0
- more spread out than normal — *heavier tails*
- Shape of the curves determined by the *degrees of freedom (df).* The larger the df, the lighter the tails, the closer the t-curve to the $N(0, 1)$ curve
- As df $= \infty$, t-curve = standard normal curve



The Extra Variability of t-Distribution Makes Sense

$t = \dfrac{\overline{X} - \mu}{s/\sqrt{n}}$ has greater variability than $z = \dfrac{\overline{X} - \mu}{\sigma/\sqrt{n}}$ because

- knowing the population SD σ means we have more info about the population, and hence we can make a more accurate inference about the population mean.
- not knowing the population SD σ means we are less certain about the distribution we are sampling from, so that extra uncertainty needs to be accounted for, and our inference will be less accurate
- As we have more data, we have more information about the distribution we are sampling from, so our inferences should act more like the case when we know σ.
  So as df increase, t-curve approaches $N(0, 1)$ curve

t Probability Table (p.430-431 in Text)

one tail	0.1	0.05	0.025	0.01	0.005
two tails	0.2	0.10	0.050	0.02	0.010
df 1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05
13	1.35	1.77	2.16	2.65	3.01
14	1.35	1.76	2.14	2.62	2.98
15	1.34	1.75	2.13	2.60	2.95
16	1.34	1.75	2.12	2.58	2.92
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
21	1.32	1.72	2.08	2.52	2.83
22	1.32	1.72	2.07	2.51	2.82
23	1.32	1.71	2.07	2.50	2.81
24	1.32	1.71	2.06	2.49	2.80
25	1.32	1.71	2.06	2.49	2.79
26	1.31	1.71	2.06	2.48	2.78
27	1.31	1.70	2.05	2.47	2.77
28	1.31	1.70	2.05	2.47	2.76
29	1.31	1.70	2.05	2.46	2.76
30	1.31	1.70	2.04	2.46	2.75

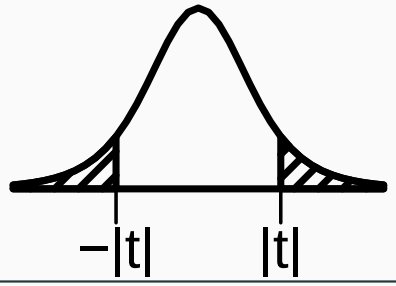
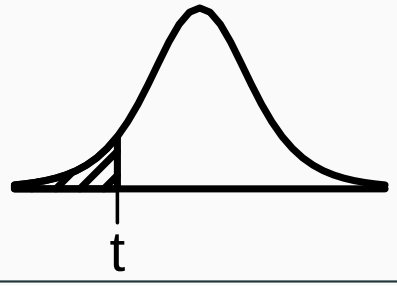
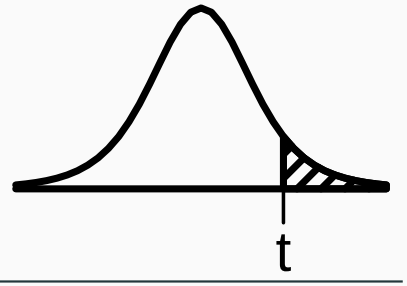
one tail	0.100	0.050	0.025	0.010	0.005
two tails	0.200	0.100	0.050	0.020	0.010
df 31	1.31	1.70	2.04	2.45	2.74
32	1.31	1.69	2.04	2.45	2.74
33	1.31	1.69	2.03	2.44	2.73
34	1.31	1.69	2.03	2.44	2.73
35	1.31	1.69	2.03	2.44	2.72
36	1.31	1.69	2.03	2.43	2.72
37	1.30	1.69	2.03	2.43	2.72
38	1.30	1.69	2.02	2.43	2.71
39	1.30	1.68	2.02	2.43	2.71
40	1.30	1.68	2.02	2.42	2.70
41	1.30	1.68	2.02	2.42	2.70
42	1.30	1.68	2.02	2.42	2.70
43	1.30	1.68	2.02	2.42	2.70
44	1.30	1.68	2.02	2.41	2.69
45	1.30	1.68	2.01	2.41	2.69
46	1.30	1.68	2.01	2.41	2.69
47	1.30	1.68	2.01	2.41	2.68
48	1.30	1.68	2.01	2.41	2.68
49	1.30	1.68	2.01	2.40	2.68
50	1.30	1.68	2.01	2.40	2.68
60	1.30	1.67	2.00	2.39	2.66
70	1.29	1.67	1.99	2.38	2.65
80	1.29	1.66	1.99	2.37	2.64
90	1.29	1.66	1.99	2.37	2.63
100	1.29	1.66	1.98	2.36	2.63
150	1.29	1.66	1.98	2.35	2.61
200	1.29	1.65	1.97	2.35	2.60
300	1.28	1.65	1.97	2.34	2.59
400	1.28	1.65	1.97	2.34	2.59
500	1.28	1.65	1.96	2.33	2.59
∞	1.28	1.65	1.96	2.33	2.58

One-sample t Test of a Population Mean

Similar to the normal z test, the t-statistic for examining H_0 : $\mu = \mu_0$ is

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \sim t_{n-1}$$

The p-value depends on H_a

H_a	$\mu \neq \mu_0$	$\mu < \mu_0$	$\mu > \mu_0$
p-value			

Then we reject H_0 when P-value $< \alpha$.

The bell curve above is the t-curve with df $= n - 1$, not the normal curve.

How to Use the t-Table to Find p-Values?

one tail	0.1	0.05	0.025	0.01	0.005
two tails	0.2	0.10	0.050	0.02	0.010
df 19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
21	1.32	1.72	2.08	2.52	2.83
22	1.32	1.72	2.07	2.51	2.82
23	1.32	1.71	2.07	2.50	2.81
24	1.32	1.71	2.06	2.49	2.80
25	1.32	1.71	2.06	2.49	2.79

Example 1. For testing $H_0 : \mu = 10$ vs. $H_a : \mu > 10$ based on a sample of size $n = 21$, if the t-statistic is 2.23,

- Comparing the t-statistic $= 2.23$ with numbers in the row for $df = n - 1 = 21 - 1 = 20$, we find that $t = 2.23$ is between 2.09 and 2.53
- one-sided p-value is between 0.025 and 0.01.

How to Use the t-Table to Find p-Values?

one tail		0.1	0.05	0.025	0.01	0.005
two tails		0.2	0.10	0.050	0.02	0.010
df	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85
	21	1.32	1.72	2.08	2.52	2.83
	22	1.32	1.72	2.07	2.51	2.82
	23	1.32	1.71	2.07	2.50	2.81
	24	1.32	1.71	2.06	2.49	2.80
	25	1.32	1.71	2.06	2.49	2.79

Example 2. For testing $H_0 : \mu = 60$ vs. $H_a : \mu \neq 60$ based on a sample of size $n = 24$, if the t-statistic is $t = 2.6$.

- Comparing the t-statistic $= 2.6$ with numbers in the row for $df = n - 1 = 24 - 1 = 23$, we find that $t = 2.6$ is between 2.50 and 2.81
- two-sided P-value is between 0.02 and 0.01.

How to Use the t-Table to Find p-Values?

one tail		0.1	0.05	0.025	0.01	0.005
two tails		0.2	0.10	0.050	0.02	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	⋮	⋮	⋮	⋮	⋮	⋮
	49	1.30	1.68	2.01	2.40	2.68
	50	1.30	1.68	2.01	2.40	2.68
	60	1.30	1.67	2.00	2.39	2.66
	70	1.29	1.67	1.99	2.38	2.65

Example 3. For testing $H_0 : \mu = 20$ vs. $H_a : \mu < 20$ based on a sample of size $n = 57$, if the t-statistic is $t = -1.55$.

- df $= 57 - 1 = 56$ is not on the table.
- The dfs above and below 56 in the table are 50 and 60.
  If df $= 50$, p-value would be between 0.1 and 0.05.
  If df $= 60$, p-value would also be between 0.1 and 0.05.
- So for df $= 56$, the p-value is also between 0.1 and 0.05.

Example: Thermal Conductivity of Glass

Thermal Conductivity is measured in terms of watts of heat power transmitted per square meter of surface per degree Celsius of temperature difference on the two sides of the material.

In these units, glass has conductivity about 1.

The National Institute of Standards and Technology provides exact data on properties of materials. Here are measurements of the thermal conductivity of 11 randomly selected pieces of a particular type of glass:

```
1.11, 1.07, 1.11, 1.07, 1.12, 1.08,
1.08, 1.18, 1.18, 1.18, 1.12
```

Example: Thermal Conductivity of Glass — Hypotheses

We want to investigate if the mean conductivity of this type of glass is greater than 1.

The hypotheses are

$$H_0 : \mu = 1, \quad H_A : \mu > 1$$

where μ is the mean conductivity of this type of glass.

Example: Thermal Conductivity of Glass — t-Statistic

The sample mean and sample SD are

$$\overline{x} \approx 1.1182, \qquad s \approx 0.04378$$

```
> conduct = c(1.11,1.07,1.11,1.07,1.12,1.08,1.08,1.18,1.18,1.18,1.12)
> mean(conduct)
[1] 1.118182
> sd(conduct)
[1] 0.04377629
```

The t-statistics is

$$t = \frac{\overline{x} - \mu_0}{s/\sqrt{n}} = \frac{1.118 - 1}{0.04378/\sqrt{11}} \approx 8.95$$

$$df = 11 - 1 = 10$$

Note $\mu_0 = 1$ because in the null hypothesis we set $\mu = 1$.

Example: Thermal Conductivity of Glass — p-value

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	2.23	2.76	3.17
	11	1.36	1.80	2.20	2.72	3.11

$$t = 8.9 > 3.17 \quad \Rightarrow \quad \text{one sided } p\text{-value} < 0.005$$

Conclusion:

The data provide convincing evidence that the mean conductivity of this type of glass is > 1

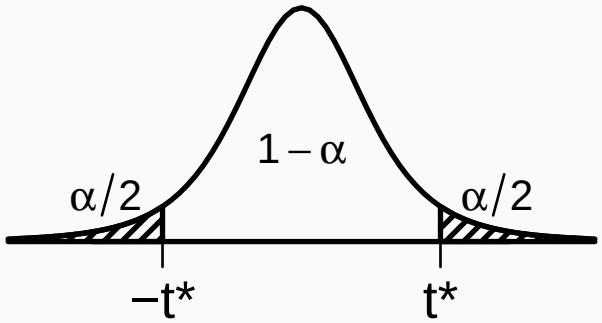
t-Confidence Interval for the Mean

A $(1-\alpha)100\%$ CI for μ is given by

$$\overline{X} \pm t^\star \times \frac{s}{\sqrt{n}}$$

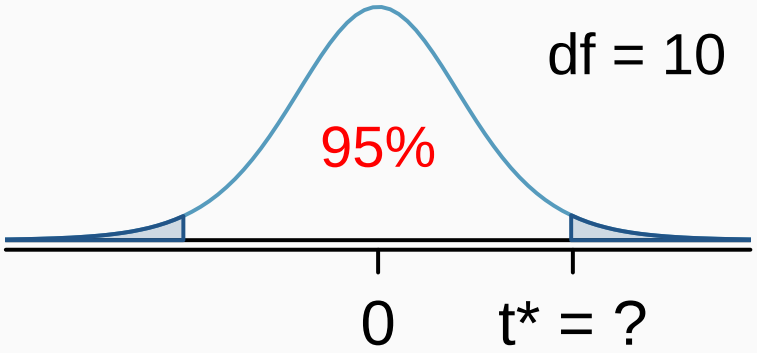
where t^* is the value of the t distribution with $n-1$ degrees of freedom such that

density curve of t_{n-1}



one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	*0.100*	*0.050*	0.020	*0.010*
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	4	1.53	2.13	2.78	3.75	4.60
	5	1.48	2.02	2.57	3.36	4.03
	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	⋮	⋮	⋮	⋮	⋮	⋮
			↑ t^* for 90% CI	↑ t^* for 95% CI		↑ t^* for 99% CI

Finding the Critical Value $t^\star$



$n = 11$, $df = 11 - 1 = 10$, $t^\star$ is at the intersection of row $df = 10$ and two tail probability 0.05.

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	*0.050*	0.020	0.010
df	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36
	9	1.38	1.83	2.26	2.82	3.25
	10	1.37	1.81	*2.23*	2.76	3.17

95% CI for the mean conductivity of this type of glass is

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}} = 1.1182 \pm 2.23 \times \frac{0.04378}{\sqrt{11}} = 1.1182 \pm 0.0294 = (1.0888, 1.1476)$$

The CI does not contain 1, which agrees with the conclusion from the hypothesis test that is μ significantly higher than 1.

T-Tests and T-Confidence Intervals in R

```
> conduct = c(1.11,1.07,1.11,1.07,1.12,1.08,1.08,1.18,1.18,1.18,1.12)
> t.test(conduct, mu = 1, alternative = "greater")

        One Sample t-test

data:  conduct
t = 8.9538, df = 10, p-value = 2.167e-06
alternative hypothesis: true mean is greater than 1
95 percent confidence interval:
 1.094259      Inf
sample estimates:
mean of x
 1.118182
```

Note that the 95% CI given (1.094259, Inf) $= (1.094259, \infty)$ is one-sided since we conducted a one-sided test.

T-Tests and T-Confidence Intervals in R

To conduct a “two-sided” test in R, change the “alternative” to “`two.sided`”

```
> t.test(conduct, mu = 1, alternative = "two.sided")

        One Sample t-test

data:  conduct
t = 8.9538, df = 10, p-value = 4.334e-06
alternative hypothesis: true mean is not equal to 1
95 percent confidence interval:
 1.088773 1.147591
sample estimates:
mean of x
 1.118182
```

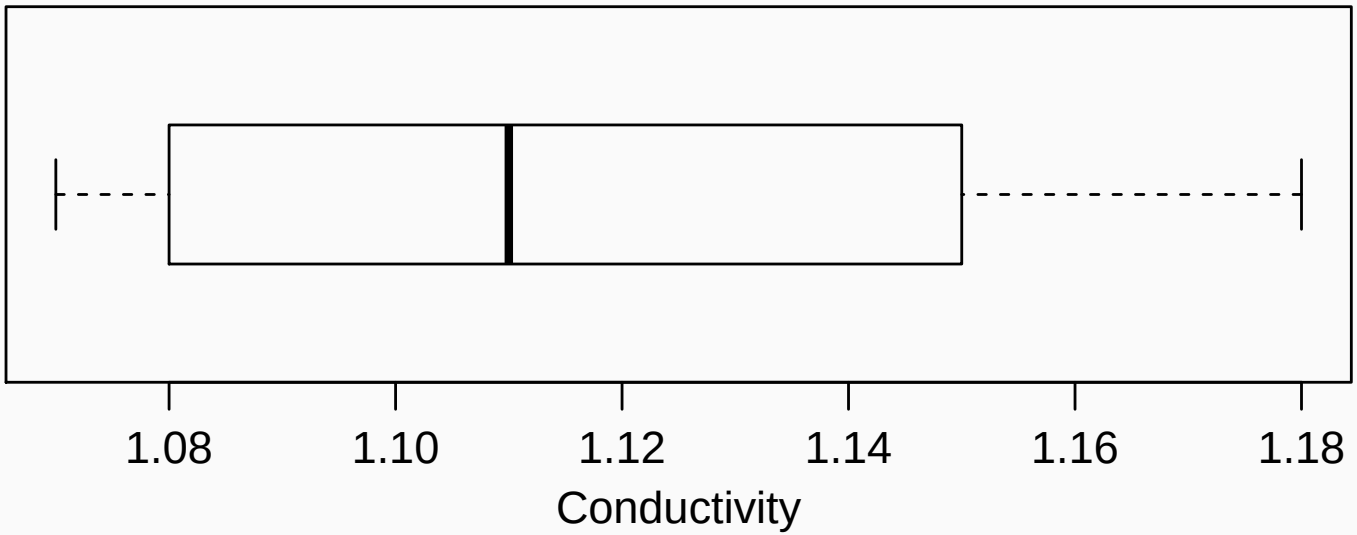
Conditions to Use t-Tests and t-Confidence Intervals

Though t-tests and t-confidence intervals don't require a big sample, they still require the following

- *Independence:* The observations should be independent
- *Normality*:
  - For the t-statistic to have a t-distribution, the population distribution has to be normal, which is rarely true.
  - In particular, it's inherently difficult to verify normality in small data sets.
  - Fortunately, the t-test and t-CI have some *robustness against non-normality* **except in the case of outliers and strong skewness**. However, their impact diminishes as the sample size gets larger.

Checking Conditions for the Thermal Conductivity Example

- *Independence:* Suppose the observations are independent.
- *Normality*: The sample distribution does not appear to be extremely skewed, but it's very difficult to assess with such a small sample size. We might want to think about whether we would expect the population distribution to be skewed or not



Example: Arsenic

Arsenic is toxic to humans and people can be exposed to it through contaminated drinking water, food, dust, and soil. Scientists have devised a non-invasive way to measure a person's level of arsenic poisoning: by examining toenail clippings. In a recent study[1], scientists measured the level of arsenic (in mg/kg) in toenail clippings of 8 people who lived near a former arsenic mine in Great Britain as follows: `0.8, 1.9, 2.7, 3.4, 3.9, 7.1, 11.9, 26.0`

Suppose the 8 people examined were randomly sampled from residents near the former arsenic mine. Is it legitimate to construct a 95% CI for the mean level of arsenic (in mg/kg) in toenail clippings for residents near the former arsenic mine using a *t*-CI?

[1] M. Button, G. R. T. Jenkin, C. F. Harrington and M. J. Watts, "Human toenails as a biomarker of exposure to elevated environment arsenic," *Journal of Environmental Monitoring*, 2009; 11(3):610-617. Data are reproduced from summary statistics and are approximate.

Example: Arsenic

Data: `0.8, 1.9, 2.7, 3.4, 3.9, 7.1, 11.9, 26.0`

Data Summary:

```
min  Q1 median  Q3 max   mean       sd n
0.8 2.5   3.65 8.3  26 7.2125 8.368041 8
```

At such a small sample size ($n = 8$), a t-CI can be used only if the population is fairly normal.

However, from the data summary we can see the sample is severely right-skewed (e.g., min/Q1 is much closer to the median than max/Q3 is), and there is an extreme outlier 26.0 that is over 3 IQRs above Q3.

It's hence not legitimate to use a t-CI.

Example: Utility Company Survey I

A utility company serves 50,000 households. As a part of a survey of customer attitudes, they take a SRS of 400 of these households. The average number of TVs in the sample households turns out to be 1.86, and the SD is 0.90.

If possible, find a 95%-confidence interval for the mean number of TVs in all 50,000 households. If this isn't possible, explain why.

Example: Utility Company Survey I (Cont'd)

- Population: the 50,000 households served by the utility company
- Parameter: mean # of TVs in the 50,000 households in the population.
- The observations are (nearly) independent since the sample is a SRS of the population and the sample size is $< 10\%$ of the population size
- At such a big sample size ($n = 400$), the t-CI is fairly robust to skewness (though it is safer to check for skewness and outliers).
- 95% confidence interval is

$$\text{sample mean} \pm t^* \times \frac{\text{sample SD}}{\sqrt{n}}$$

$$= 1.86 \pm 1.97 \times \frac{0.90}{\sqrt{400}} \approx 1.86 \pm 0.09 = (1.77, 1.95).$$

With 95% confidence, we may assert that the 50,000 households had 1.77 to 1.95 TVs on average.

Example: Utility Company Survey II

As part of the survey, all persons age 16 and over in the 400 sample households are interviewed. This makes 900 people.

On average, the sampled people watched 5.20 hours of TV the Sunday before the survey, and the SD was 4.50 hours.

True or False and explain: a 95%-confidence interval for the average number of hours spent watching TV by all persons age 16 and over in the 50,000 households on that Sunday is

$$\text{sample mean} \pm t^* \times \frac{\text{sample SD}}{\sqrt{n}}$$

$$= 5.20 \pm 1.96 \times \frac{4.50}{\sqrt{900}} \approx 5.2 \pm 0.294$$

Why $t^* = 1.96$ here?

Example: Utility Company Survey II

- Population: all persons age 16 and over in the 50,000 households served by the utility company
- Parameter: the average number of hours spent watching TV by all persons age 16 and over in the 50,000 households on that Sunday.
- The sample is NOT a **SRS** from the target population. There will be dependencies among the hours of TV watched among members of the same household.
- Hence we cannot construct the CI use the formula

$$\text{sample mean} \pm t^* \times \frac{\text{sample SD}}{\sqrt{n}}$$

  which assumes the observations in the sample were independent.

Recap: Inference Using the t-Distribution

- Conditions:
  - independence of observations (use your judgement)
  - not extremely skew, no outlier
- Hypothesis testing:

$$T_{df} = \frac{\text{sample mean} - \text{null value}}{(\text{sample SD})/\sqrt{n}}, \text{ where } df = n - 1$$

- Confidence interval:

$$\text{sample mean} \pm t^{\star}_{df} \times (\text{sample SD})/\sqrt{n}$$