

# **STAT 22000 Lecture Slides**

## **Inference on Population Means**

### **Using $t$ -Distributions**

---

Yibi Huang  
Department of Statistics  
University of Chicago

This set of slides covers Section 5.1 in the 3rd edition of *OpenIntro Statistics* (or Section 7.1 in the 4th edition).

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

## What if $\sigma$ is Unknown?

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

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

We use this fact to construct confidence intervals and do hypothesis tests about  $\mu$  in Section 4.2 and 4.3.

As the population SD  $\sigma$  is UNKNOWN, we replace it with the **sample SD  $s$** , with the restrictions that

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

This lecture we will introduce a way to remove the restriction on sample size.

## Student's $t$ -Distributions

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

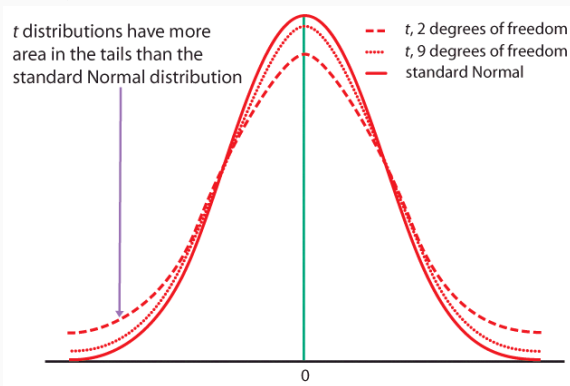
$$t = \frac{\bar{X} - \mu}{s / \sqrt{n}}$$

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

$$t = \frac{\bar{X} - \mu}{s / \sqrt{n}} \sim t_{n-1} \quad \text{v.s.} \quad z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}} \sim N(0, 1)$$

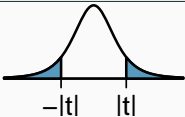
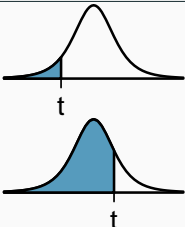
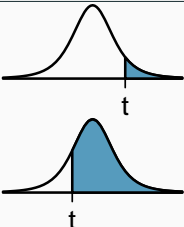
- The  $t$ -statistic has greater variability than the  $z$ -statistic because of the extra variability in the sample SD  $s$ .  
So the  $t$ -distribution has heavier tail than  $N(0, 1)$ .
- As the sample size increases, the sample SD  $s$  becomes a more accurate estimate of the population SD  $\sigma$ .  
So as df increase,  $t$ -curve approaches  $N(0, 1)$  curve

# One-sample $t$ Test of a Population Mean

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

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

The  $p$ -value depends on  $H_a$ .

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

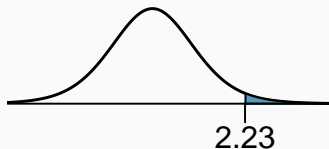
Then we reject  $H_0$  when  $P$ -value  $< \alpha$ .

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

## How to Use R to Find Upper One-Sided $P$ -Values?

Example 1. Testing  $H_0 : \mu = 10$  vs.  $H_a : \mu > 10$ , sample size  $n = 21$ , the  $t$ -statistic is 2.23.

Upper One-sided  $P$ -value =



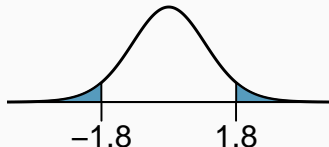
```
> pt(2.23,df=20,lower.tail=F)
[1] 0.01868333
```



## How to Use R to Find Two-Sided $p$ -Values?

Example 2. Testing  $H_0 : \mu = 60$  vs.  $H_a : \mu \neq 60$ , sample size  $n = 24$ , the  $t$ -statistic is  $t = 1.8$ .

Two-sided  $P$ -value =

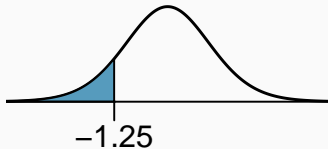


```
> 2*pt(1.8,df=24-1,lower.tail=F)
[1] 0.08499266
```

## How to Use R to Find Lower One-Sided $P$ -Values?

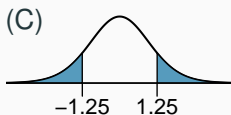
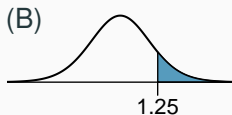
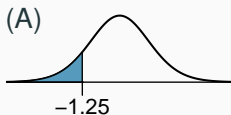
Example 3. Testing  $H_0 : \mu = 20$  vs.  $H_a : \mu < 20$ , sample size  $n = 17$ ,  $t$ -statistic  $= -1.25$ .

Lower one-sided  $P$ -value =

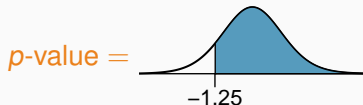


```
> pt(-1.25, df=17-1)  
[1] 0.1146372
```

Example 4. For testing  $H_0 : \mu = 20$  vs.  $H_a : \mu > 20$  with a sample of size  $n = 17$  and a  $t$ -statistic  $= -1.25$ , which of the following areas represents the corresponding  $p$ -value?



Ans: (D). Note  $H_a : \mu > 20$  is upper one-sided. The  $p$ -value is the upper tail area above the  $t$ -statistic,



```
> pt(-1.25, df=17-1, lower.tail=F)  
[1] 0.8853628
```

## 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  $\mu$  is the mean conductivity of this type of glass.

## Example: Thermal Conductivity of Glass — $t$ -Statistic & $P$ -value

The sample mean and sample SD are

$$\bar{x} \approx 1.1182, \quad \text{and} \quad s \approx 0.04378.$$

```
> 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

$$t = \frac{\bar{x} - \mu_0}{s / \sqrt{n}} = \frac{1.118 - 1}{0.04378 / \sqrt{11}} \approx 8.95, \quad df = 11 - 1 = 10$$

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

```
> pt(8.95, df=10, lower.tail=F)
[1] 2.175242e-06
```

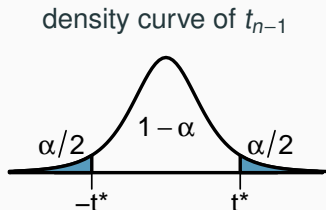
**Conclusion:** The tiny  $P$ -value  $2.175 \times 10^{-6}$  provides convincing evidence that the mean conductivity of this type of glass is  $> 1$

## $t$ -Confidence Interval for the Mean

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

$$\bar{X} \pm t^* \times \frac{s}{\sqrt{n}}$$

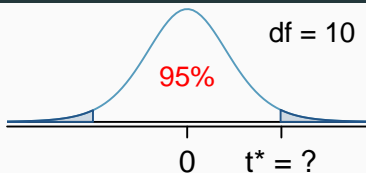
where  $t^*$  is the value of the  $t$  distribution with  $n - 1$  degrees of freedom such that



which can be found in R using the `qt()` command.

```
qt(alpha/2, df, lower.tail=F)
```

## Finding the Critical Value $t^*$



$n = 11$ ,  $df = 11 - 1 = 10$ ,  $t^*$  can be found in R as follows to be  $\approx 2.228$

```
> qt(0.05/2, df=10, lower.tail=F)
[1] 2.228139
```

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

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}} = 1.1182 \pm 2.228 \times \frac{0.04378}{\sqrt{11}} = 1.1182 \pm 0.0294 \\ = (1.0888, 1.1476)$$

The CI does not contain 1, which agrees with the conclusion from the hypothesis test that is  $\mu$  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  $(1.094259, \text{Inf}) = (1.094259, \infty)$  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
```

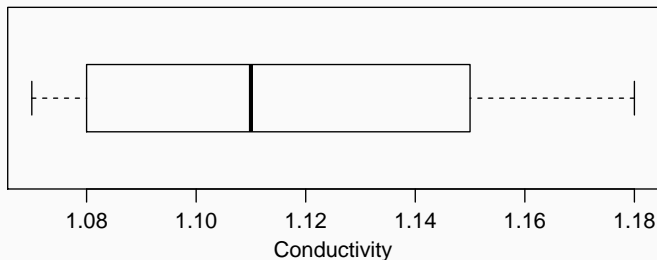
## Conditions to Use $t$ -Tests and $t$ -Confidence Intervals

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<sup>1</sup>, 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?

---

<sup>1</sup> 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

$$\begin{aligned} & \text{sample mean} \pm t^* \times \frac{\text{sample SD}}{\sqrt{n}} \\ &= 1.86 \pm 1.97 \times \frac{0.90}{\sqrt{400}} \approx 1.86 \pm 0.09 = (1.77, 1.95). \end{aligned}$$

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

$$\begin{aligned} & \text{sample mean} \pm t^* \times \frac{\text{sample SD}}{\sqrt{n}} \\ &= 5.20 \pm 1.96 \times \frac{4.50}{\sqrt{900}} \approx 5.2 \pm 0.294 \end{aligned}$$

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:

$$\text{sample mean} \pm t_{df}^* \times (\text{sample SD}) / \sqrt{n}$$