

Lecture 18: t-distribution

Chapter 5.3

Goals for Today

- ▶ Continue difference of means example: Hypothesis testing
- ▶ Paired Differences
- ▶ What do we do when n is small?

Differences in Means: Hypothesis Tests

Recall the Cherry Run race data. We test

- ▶ $H_0 : \mu_w - \mu_m = 0$
- ▶ $H_A : \mu_w - \mu_m \neq 0$

at the $\alpha = 0.01$ significance level.

Differences in Means: Hypothesis Tests

Recall the Cherry Run race data. We test

- ▶ $H_0 : \mu_w - \mu_m = 0$
- ▶ $H_A : \mu_w - \mu_m \neq 0$

at the $\alpha = 0.01$ significance level.

We note

- ▶ The null states no difference in race times

Differences in Means: Hypothesis Tests

Recall the Cherry Run race data. We test

- ▶ $H_0 : \mu_w - \mu_m = 0$
- ▶ $H_A : \mu_w - \mu_m \neq 0$

at the $\alpha = 0.01$ significance level.

We note

- ▶ The null states no difference in race times
- ▶ The two sided alternative is more **conservative** than the one-sided alternative $\mu_w - \mu_m > 0$.

Hypothesis Testing Procedure

We verified the conditions when looking at confidence intervals

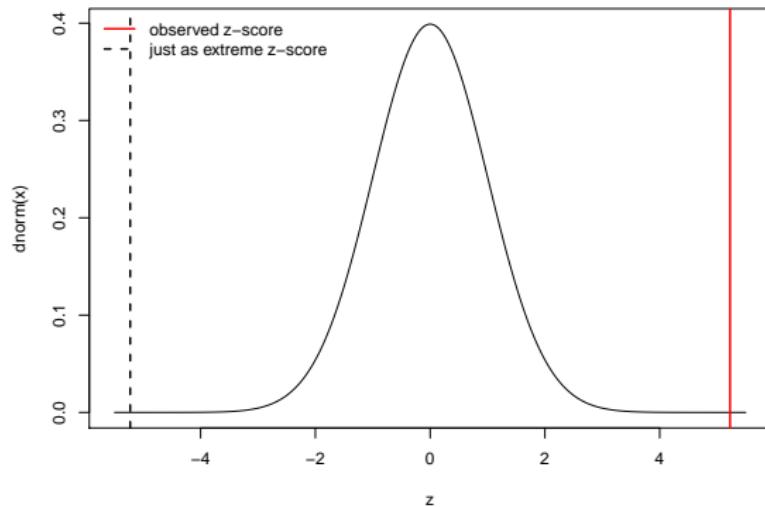
Hypothesis Testing Procedure

We verified the conditions when looking at confidence intervals

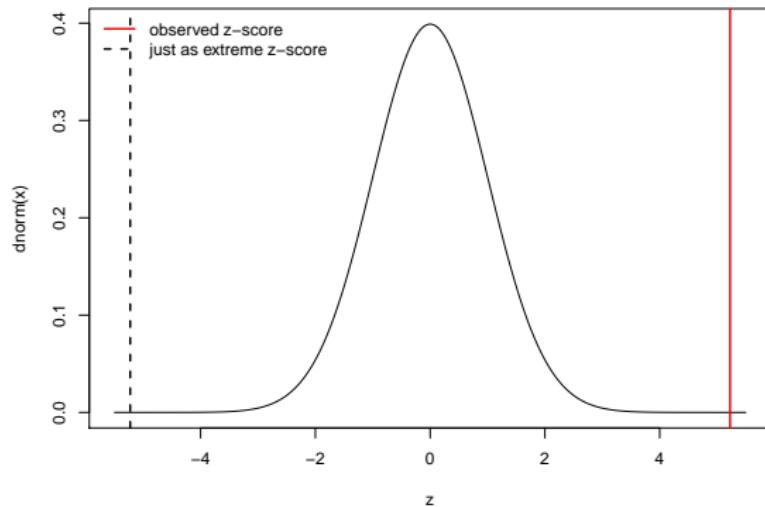
The **test statistic** is the z-score of the point estimate $\bar{x}_1 - \bar{x}_2$ of $\mu_w - \mu_m$ under the null hypothesis:

$$\begin{aligned} z &= \frac{\text{point estimate} - \text{null value}}{SE} = \frac{\bar{x}_1 - \bar{x}_2 - 0}{SE_{\bar{x}_w - \bar{x}_m}} = \frac{14.48}{2.77} \\ &= 5.227 \end{aligned}$$

Hypothesis Tests



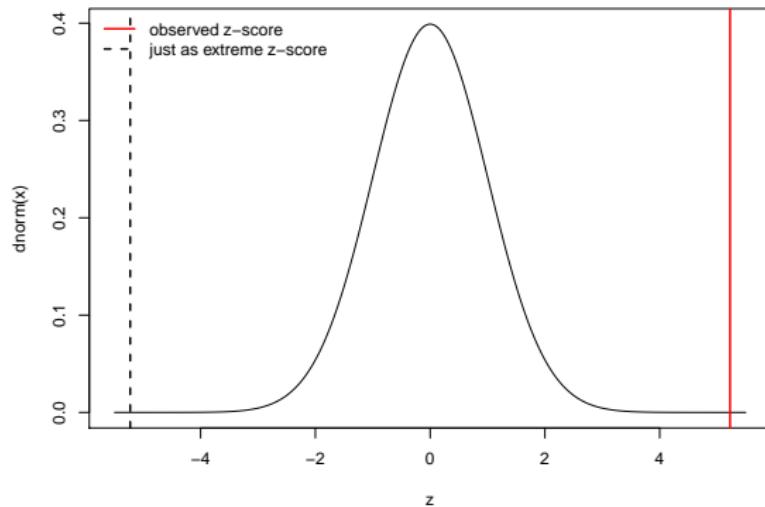
Hypothesis Tests



4. Compute the p-value

- ▶ Since the sampling distribution is normal, we use the z-table.
- ▶ Since $H_A : \mu_w - \mu_m \neq 0$, we have a two-sided p-value $2 \times 0 = 0$

Hypothesis Tests



4. Compute the p-value
 - ▶ Since the sampling distribution is normal, we use the z-table.
 - ▶ Since $H_A : \mu_w - \mu_m \neq 0$, we have a two-sided p-value $2 \times 0 = 0$
5. Since the p-value $0 < \alpha = 0.01$, we reject H_0 and declare that men and women did not have equal mean finish times.

Chapter 5.1: Paired Data

Two sets of observations are **paired** if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

Chapter 5.1: Paired Data

Two sets of observations are **paired** if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

Examples:

- ▶ Cholesterol levels before and after some intervention

Chapter 5.1: Paired Data

Two sets of observations are **paired** if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

Examples:

- ▶ Cholesterol levels before and after some intervention
- ▶ Disease rates amongst pairs of twins

Chapter 5.1: Paired Data

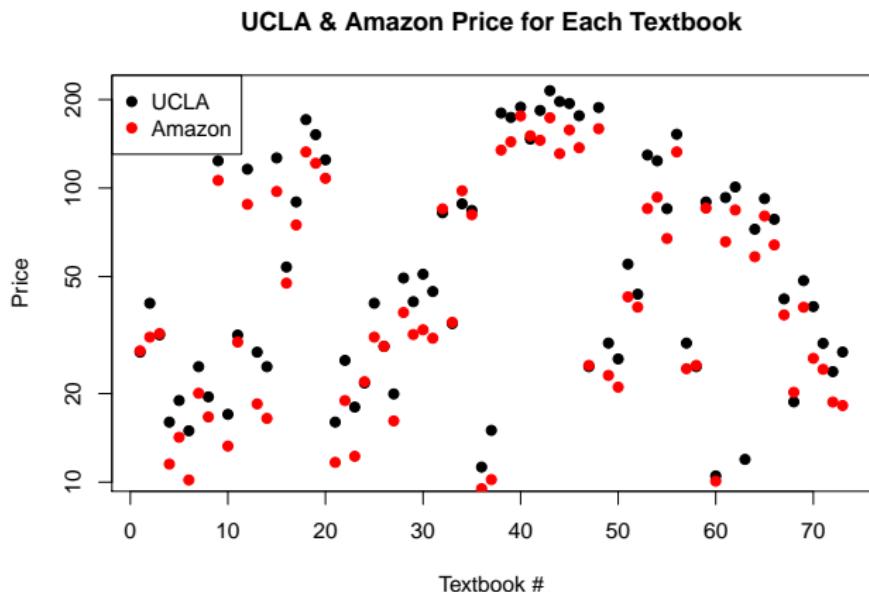
Two sets of observations are **paired** if each observation in one set has a special correspondence or connection with exactly one observation in the other data set.

Examples:

- ▶ Cholesterol levels before and after some intervention
- ▶ Disease rates amongst pairs of twins
- ▶ In the text: price of the same textbook at the UCLA bookstore vs Amazon

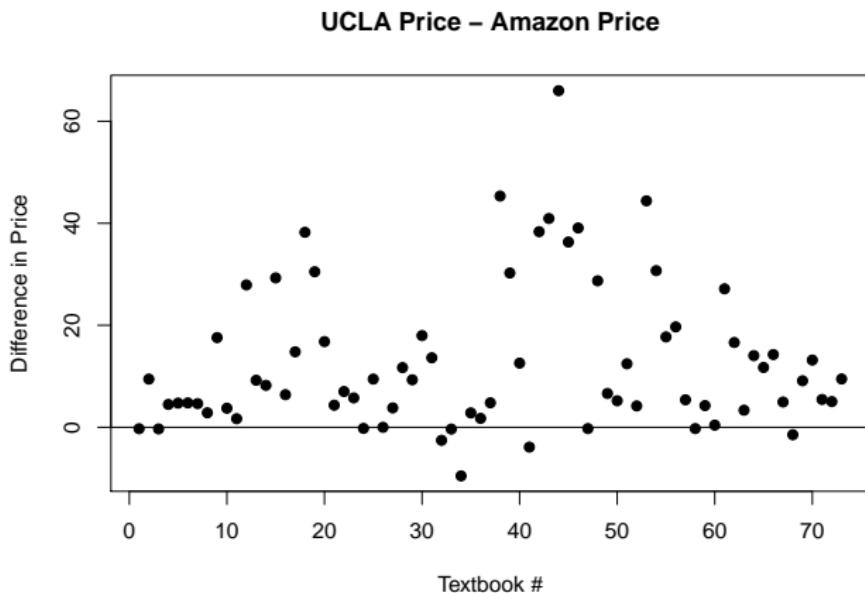
Paired Differences

The methodology for paired data remains the same, except our **observations** are the difference in pairs. Example, for the UCLA Bookstore vs Amazon book price example in the text



Paired Differences

The methodology for paired data remains the same, except our **observations** are the difference in pairs. Example, for the UCLA Bookstore vs Amazon book price example in the text



Paired Differences

We have

Paired Differences

We have

- ▶ population parameter is μ_{diff}

Paired Differences

We have

- ▶ population parameter is μ_{diff}
- ▶ point estimate \bar{x}_{diff} of μ_{diff}

Paired Differences

We have

- ▶ population parameter is μ_{diff}
- ▶ point estimate \bar{x}_{diff} of μ_{diff}
- ▶ Conditions: not on the original observations, but rather the differences: 10% rule, sample size n , and not too skewed differences.

Paired Differences

We have

- ▶ population parameter is μ_{diff}
- ▶ point estimate \bar{x}_{diff} of μ_{diff}
- ▶ Conditions: not on the original observations, but rather the differences: 10% rule, sample size n , and not too skewed differences.
- ▶ If met, \bar{x}_{diff} has a normal sampling distribution with mean μ_{diff} and $SE_{diff} = \frac{s_{diff}}{\sqrt{n_{diff}}}$.

Sample Size n

We need a **large sample size n** for two reasons:

1. Ensure the sampling distribution of \bar{x} is normal regardless of the true distribution by the Central Limit Theorem.
2. ensure s is a good estimate of σ , which is used in $SE = \frac{\sigma}{\sqrt{n}}$

Sample Size n

We need a **large sample size n** for two reasons:

1. Ensure the sampling distribution of \bar{x} is normal regardless of the true distribution by the Central Limit Theorem.
2. ensure s is a good estimate of σ , which is used in $SE = \frac{\sigma}{\sqrt{n}}$

What do we do when the sample size n is small? We're stuck, except when

- ▶ the observations are normal
- ▶ the sample observations are independent

the sampling distribution of \bar{x} is nearly normal **regardless** of sample size n .

Verifying Normality

Be cautious when verifying the normality condition for small n . It is important to not only examine the data but also think about where the data come from. For example, ask:

- ▶ Would I expect this distribution to be symmetric?
- ▶ Am I confident that outliers are rare?

t Distribution

Let x_1, \dots, x_n be a random sample from a normal distribution.

t Distribution

Let x_1, \dots, x_n be a random sample from a normal distribution.
Then the standardized variable

$$t = \frac{\bar{x} - \mu}{SE} = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

has probability distribution called a *t* distribution with $n - 1$ degrees of freedom (*df*).

t Distribution

Properties of the *t* distribution:

- ▶ a *t*-distribution has only one parameter: the degrees of freedom df .

t Distribution

Properties of the *t* distribution:

- ▶ a *t*-distribution has only one parameter: the degrees of freedom df .
- ▶ It is bell-shaped and centered at 0, like a *z*-curve

t Distribution

Properties of the *t* distribution:

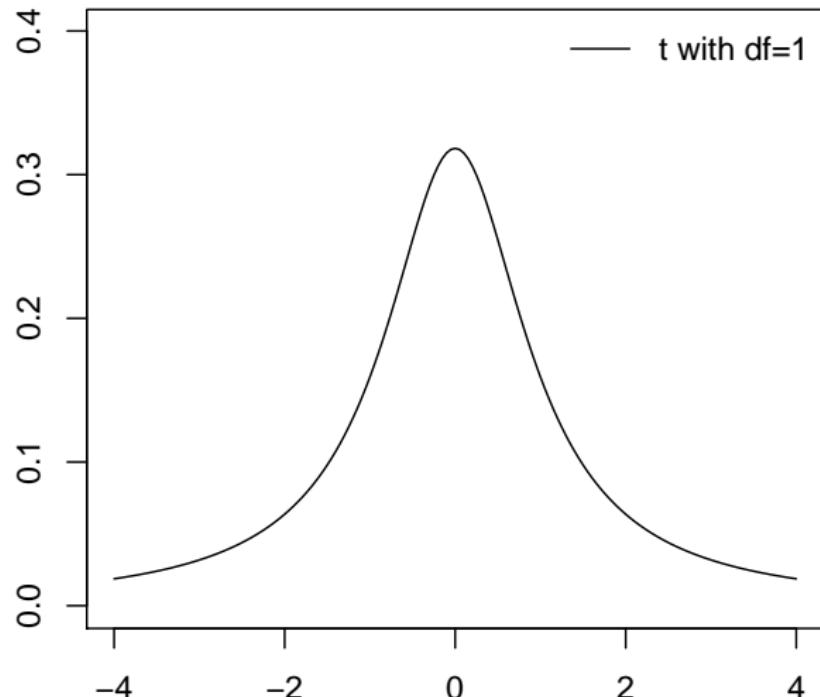
- ▶ a *t*-distribution has only one parameter: the degrees of freedom *df*.
- ▶ It is bell-shaped and centered at 0, like a *z*-curve
- ▶ Any *t* curve is more spread out than a *z* curve.
i.e. it has fatter tails

t Distribution

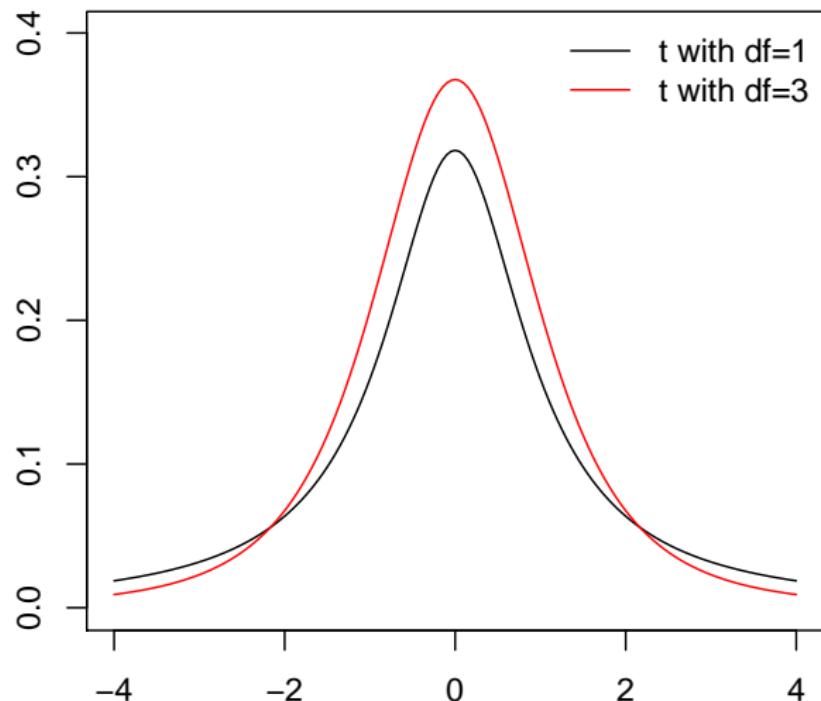
Properties of the *t* distribution:

- ▶ a *t*-distribution has only one parameter: the degrees of freedom df .
- ▶ It is bell-shaped and centered at 0, like a *z*-curve
- ▶ Any *t* curve is more spread out than a *z* curve.
i.e. it has fatter tails
- ▶ As the df goes to ∞ , the *t* curve approaches the *z* curve.

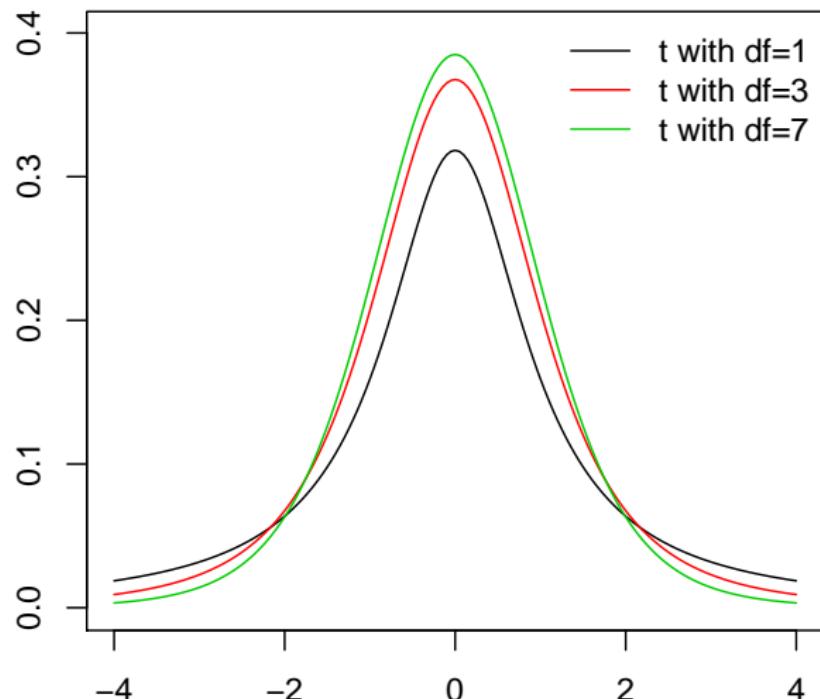
t Distribution Examples



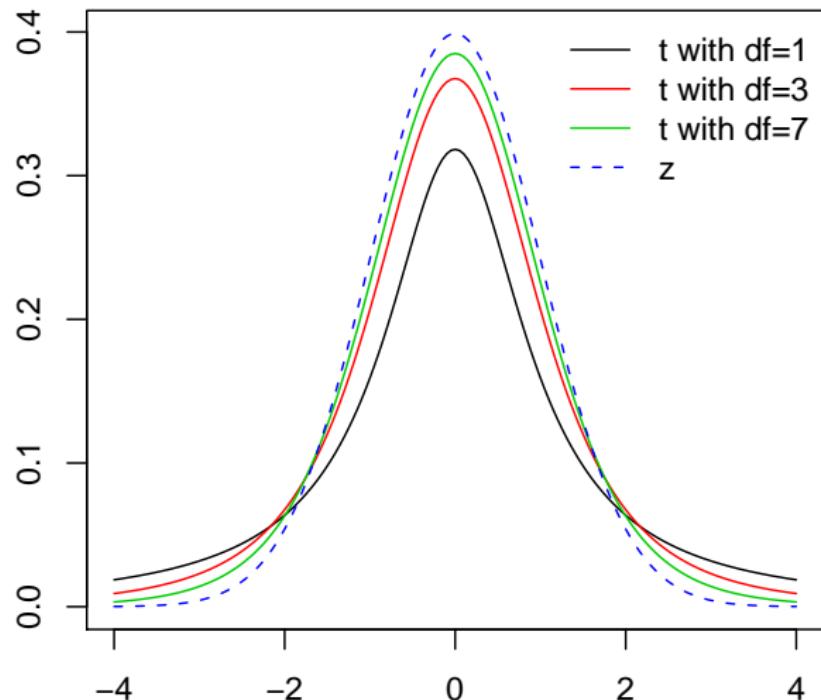
t Distribution Examples



t Distribution Examples



t Distribution Examples



Conditions for Using t Distribution

We use the t distribution when you have a [small sample](#) and

Conditions for Using t Distribution

We use the t distribution when you have a **small sample** and

- ▶ **Independence of observations:**
 - ▶ 10% rule on sample vs population size
 - ▶ or if we have an experiment or random process we check that each observation were independent

Conditions for Using t Distribution

We use the t distribution when you have a **small sample** and

- ▶ **Independence of observations:**
 - ▶ 10% rule on sample vs population size
 - ▶ or if we have an experiment or random process we check that each observation were independent
- ▶ **Observations come from a nearly normal distribution:** Difficult to verify with small data sets:
 - ▶ take a look at a histogram of the data
 - ▶ consider whether any previous experiences alert us that the data may not be nearly normal

t-Tables

If $n = 19$, we use $df = 19 - 1 = 18$ and do a look up on the *t*-table on page 410:

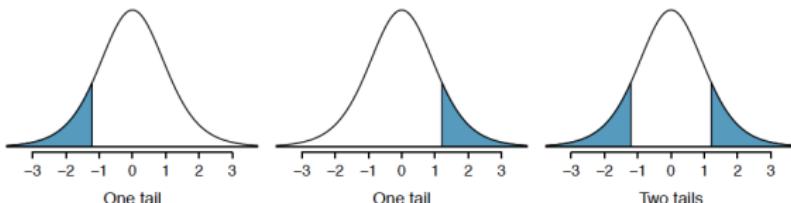


Figure B.1: Three *t* distributions.

	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
	10	1.37	1.81	2.23	2.76	3.17
	11	1.36	1.80	2.20	2.72	3.11

Confidence Intervals

Confidence intervals: Use t_{df}^* instead of z^*

$$[\bar{x} - t_{df}^* SE, \bar{x} + t_{df}^* SE] = \left[\bar{x} - t_{df}^* \times \frac{s}{\sqrt{n}}, \bar{x} + t_{df}^* \times \frac{s}{\sqrt{n}} \right]$$

So for example, to get a 95% C.I. based on $n = 18 \Rightarrow df = 17$,
 $t_{17}^* = 2.11$.

t-Test

The *t*-test is the same as the previous test, where instead of finding the *z*-score of \bar{x} , we find the *t*-statistic/score.

t-Test

The *t*-test is the same as the previous test, where instead of finding the *z*-score of \bar{x} , we find the *t*-statistic/score.

Example 5.19 on page 252: NYC is known as the city that never sleeps. A random sample of 25 New Yorkers were asked how much sleep they get per night. Does the data below provide strong evidence that New Yorkers sleep less than 8 hours a night on average? Set $\alpha = 0.05$

n	\bar{x}	s	min	max
25	7.73	0.77	6.17	9.78

t-Test

2. Conditions:

- ▶ Independence: 25 is obviously less than 10% of the study population of NYC
- ▶ What about normality? Not an exact science. The halfway point of the min and max is 7.975, which is fairly close to $\bar{x} = 7.73$. So symmetric enough.

t-Test

2. Conditions:

- ▶ **Independence:** 25 is obviously less than 10% of the study population of NYC
- ▶ What about normality? Not an exact science. The halfway point of the min and max is 7.975, which is fairly close to $\bar{x} = 7.73$. So symmetric enough.

3. The test statistic is the *t*-statistic:

$$t = \frac{\bar{x} - \text{null value}}{SE} = \frac{\bar{x} - \text{null value}}{\frac{s}{\sqrt{n}}} = \frac{7.73 - 8}{\frac{0.77}{\sqrt{25}}} = -1.75$$

Since $n = 25$, $df = 25 - 1 = 24$.

t-Test

4. *p*-Value. We use the *t* distribution i.e. the *t*-table on page 410:

one-tail	0.100	0.050	0.025	0.010	0.005
two-tail	0.200	0.100	0.050	0.020	0.010
df = 24	1.32	1.71	2.06	2.49	2.80

Since 1.75 is in between one-tail values of 1.71 and 2.06 and by symmetry, the p-value is somewhere between 0.05 and 0.025.

t-Test

4. *p*-Value. We use the *t* distribution i.e. the *t*-table on page 410:

one-tail	0.100	0.050	0.025	0.010	0.005
two-tail	0.200	0.100	0.050	0.020	0.010
df = 24	1.32	1.71	2.06	2.49	2.80

Since 1.75 is in between one-tail values of 1.71 and 2.06 and by symmetry, the *p*-value is somewhere between 0.05 and 0.025.

5. Decision: Since the *p*-value < $\alpha = 0.05$, we reject the null hypothesis that NY'ers sleep 8 hours a night at the $\alpha = 0.05$ significance level.

History of t Distribution

The t distribution was derived by William Sealy Gosset in 1908, a chemist/statistician at the Guinness Brewery in Dublin, Ireland.



History of t Distribution

Gosset was concerned with small-sample statistics about barley given that brewers are limited in the number of batches of beer they can brew.

History of t Distribution

Gosset was concerned with small-sample statistics about barley given that brewers are limited in the number of batches of beer they can brew.

Guinness prohibited its employees from publishing. So Gosset had to use the pseudonym “Student” to conceal his identity.

History of t Distribution

Gosset was concerned with [small-sample statistics](#) about barley given that brewers are limited in the number of batches of beer they can brew.

Guinness prohibited its employees from publishing. So Gosset had to use the pseudonym “Student” to conceal his identity.

In particular, the [\(Student's\) t-test](#) is one of the most widely used statistical tests in the world.

History of t Distribution

In fact if you go to the Guinness Brewery at St James's Gate in Dublin, Ireland...



History of t Distribution

