Week 7

Paired *t* Procedures and *t* Procedures for Two Independent Populations

ST 314
Introduction to Statistics for Engineers

Oregon State University

Independence & Dependence

- Populations are assumed to be independent__________ if sampled or experimental units are from different, unrelated populations.
- Populations are considered to be dependent or paired_ if more than one measurement is taken on a single experimental unit or if experimental units are *paired* by a common factor.
    - Common examples
        - Before + after experiments or studies
        - Testing twins or genetically identical subjects
        - Splitting a sample to impose more than one treatment on a single unit

Independence & Dependence

Which study has independent populations and which has a dependent populations?

Study 1 - To study the change in attitude towards statistics over the course of the semester, a professor selects a simple random sample of students. She administers a questionnaire to the students at the beginning of the semester and then again at the end of the semester.

Dependent

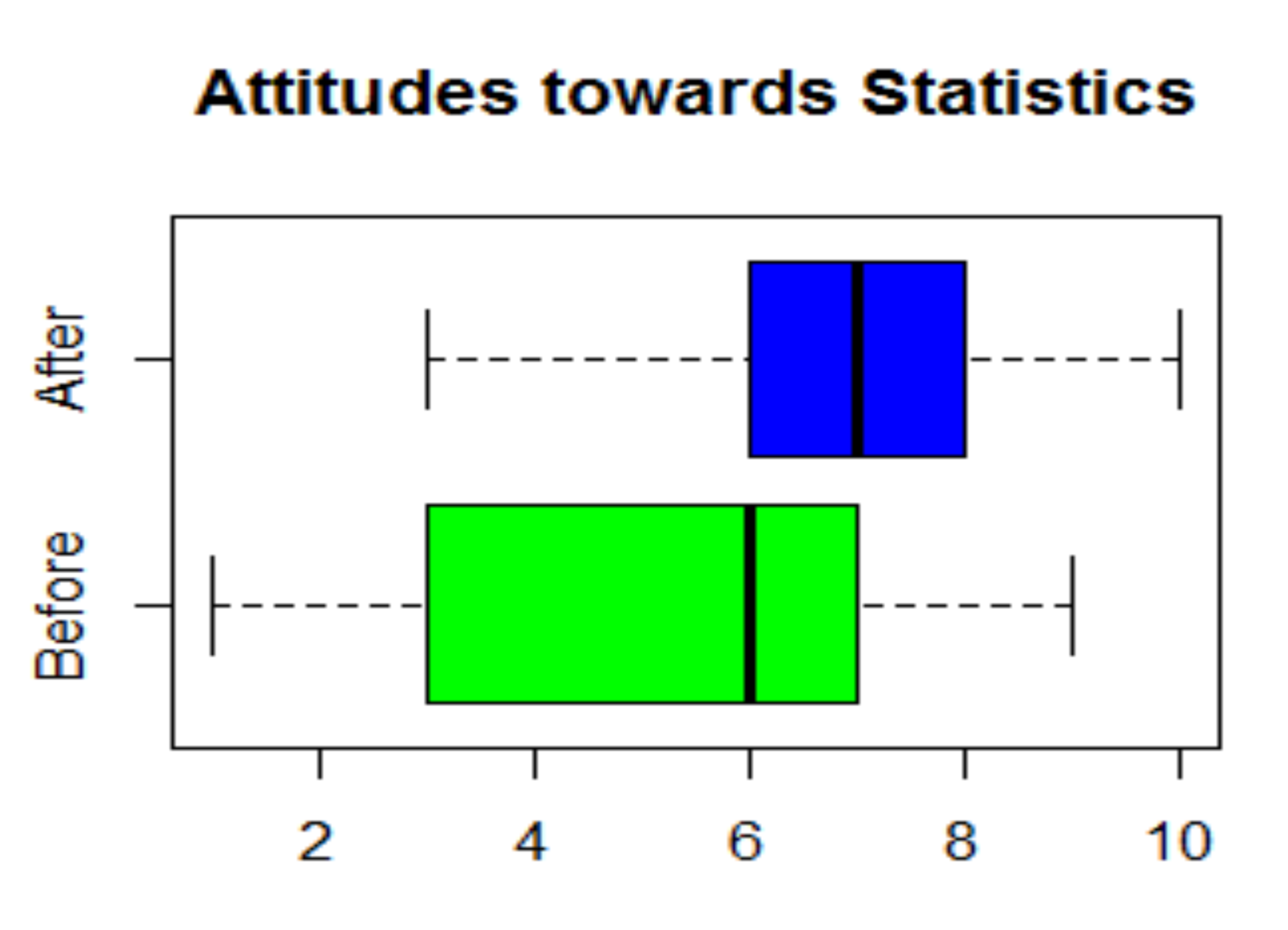
Study 2 - At the end of a term a statistics professor would like to compare the general attitudes towards statistics of students that are science based majors versus non-science based majors. The professor selects a simple random sample of students from each population and administers a questionnaire to each sample of students.

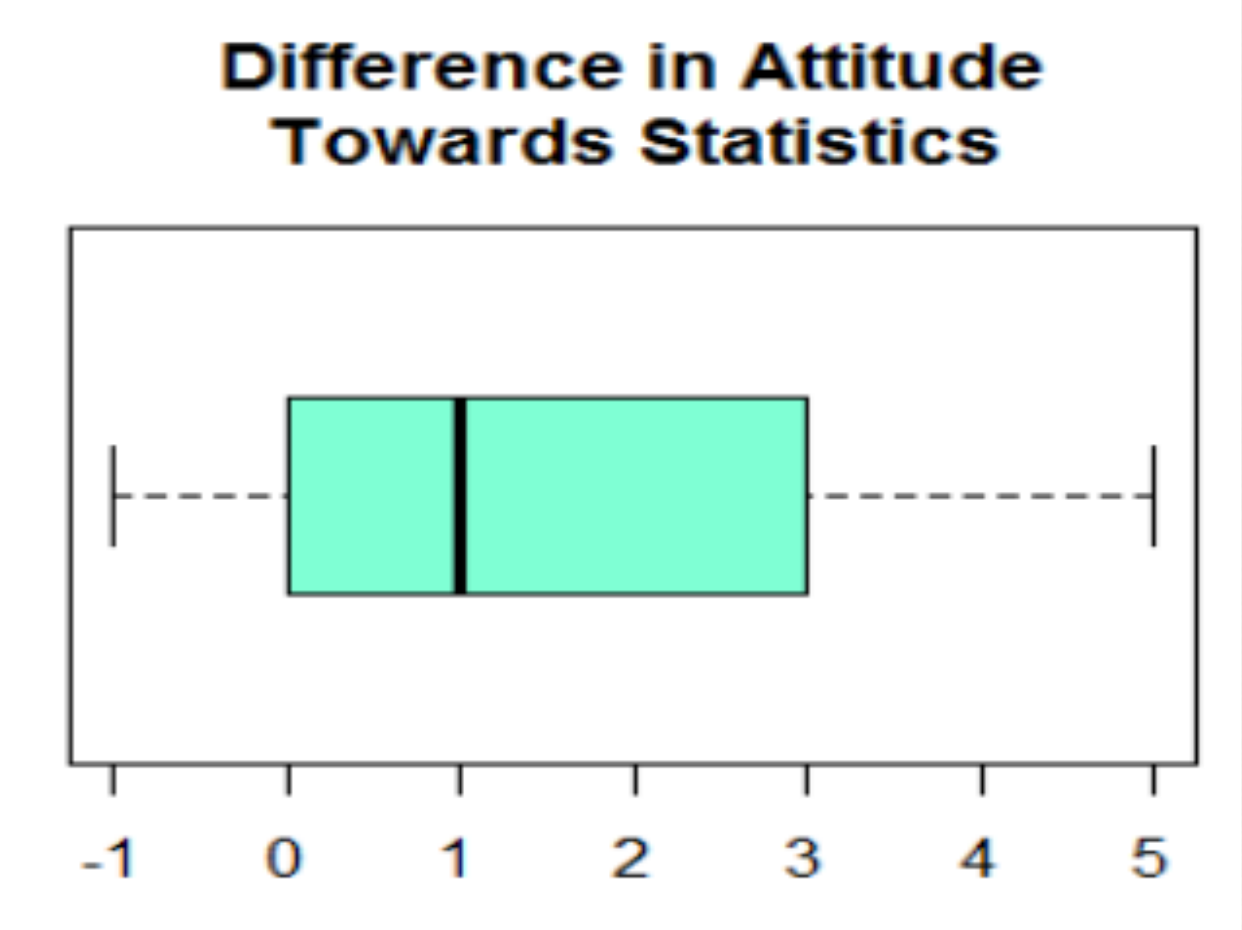
Independent

Independence & Dependence

The following are the before and after term attitude towards statistics scores for 13 randomly sampled students.

Student	A	B	C	D	E	F	G	H	I	J	K	L	M
before	1	3	7	8	6	9	5	7	6	7	1	3	7
after	5	8	6	8	9	10	6	8	8	7	5	3	6
difference (after-before)	4	5	-1	0	3	1	1	1	2	0	4	0	-1





CAUTION

Performing an independent test on dependent data will increase the chance of Type II error and will overall decrease in information.

Matched Pairs t Procedures

Four random samples of a ferrous-type substance are used to determine if there is a difference between a laboratory chemical analysis and an X-ray fluorescence analysis in measuring the iron content of the substance. Each sample is split and analyzed using both methods.

The iron content measured by the two methods for 4 random samples

Sample	1	2	3	4
Chemical	2.1	2.5	2.3	2.5
X-ray	2.0	2.3	2.1	2.4
Difference	0.1	0.2	0.2	0.1

When data are paired we can make comparisons by analyzing the differences between in each pair.

- Observed differences: diff ← c(0.1 , 0.2, 0.2, 0.1)
- Average of the sampled differences: $\bar{x}_{diff} = \frac{0.1+0.2+0.2+0.1}{4} = 0.15$
- Sample standard deviation of the observed differences: $s_{diff} = sd(diff) = 0.058$
- Number of pairs or differences: $n_{diff} = 4$

Matched Pairs *t* Confidence Interval

- When to use: want to estimate the average difference between two paired populations
- Conditions for inference:

  -Representative sample

  -sufficiently large sample (CLT) ← # of pairs

  -Paired observations
- The confidence interval for estimating the difference between population means is:

$$\bar{x}_{diff} \pm \underbrace{t^*_{n_{diff}-1}}_{\text{degrees of freedom}} \left(\frac{s_{diff}}{\sqrt{n_{diff}}} \right)$$

Matched Pairs Confidence Interval Example

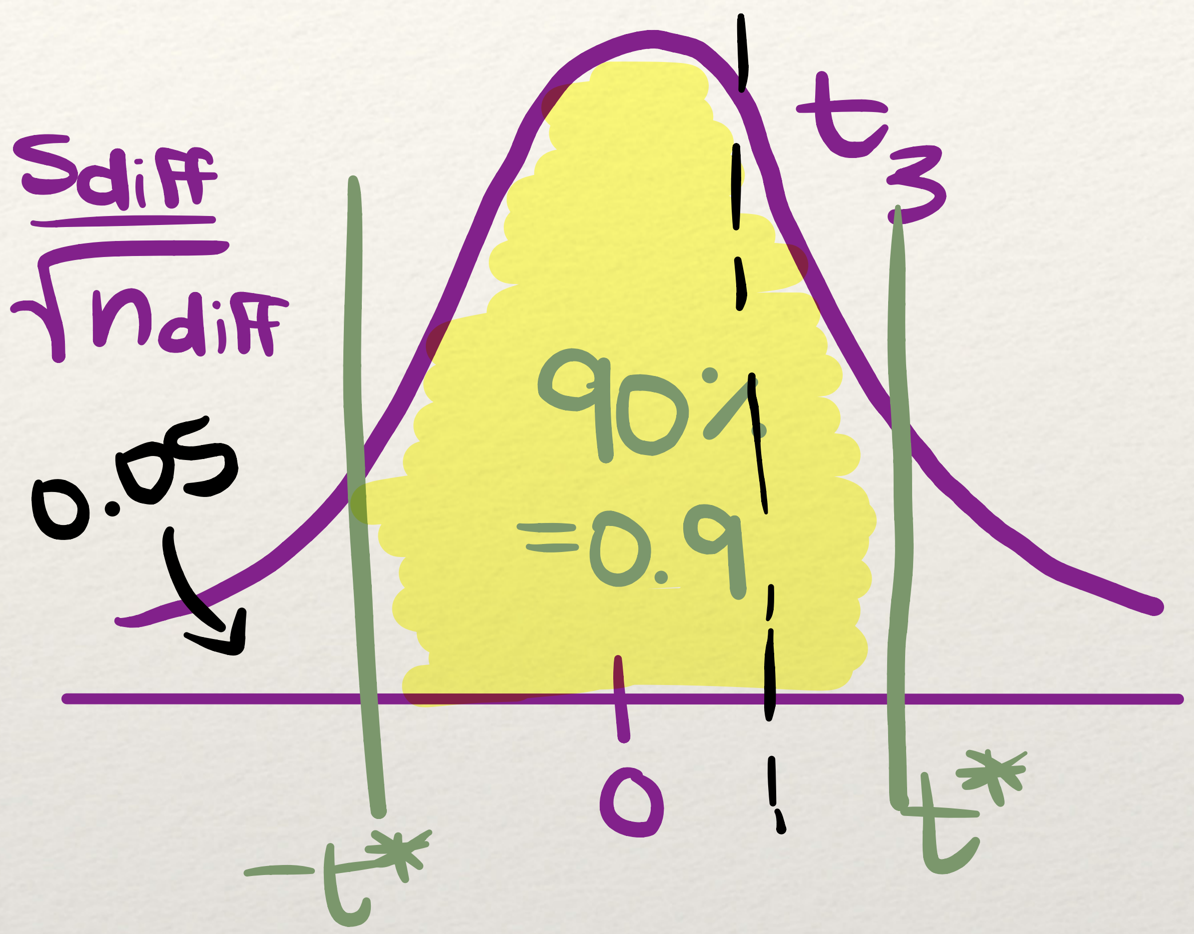
The iron content measured by the two methods for 4 random samples. Assume the differences between the Chemical and X-ray analyses come from a Normal distribution

Sample	1	2	3	4
Chemical	2.1	2.5	2.3	2.5
X-ray	2.0	2.3	2.1	2.4
Difference	0.1	0.2	0.2	0.1

Estimate the average difference between iron content of the different measuring methods with a 90% confidence interval.

$\bar{x}_{diff} = 0.15$ $\quad s_{diff} = 0.058$ $\quad n_{diff} = 4$

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



$$t^* = qt(0.95, 3) = 2.353$$

$$0.15 \pm 2.353\left(\frac{0.058}{\sqrt{4}}\right) = (0.082, 0.218)$$

We are 90% confident that the average difference between the two methods is between 0.082 and 0.218 with a point estimate of 0.15.

Matched Pairs *t* Test

- When to use: test the average difference between paired data
- Conditions required for inference:

  Same as confidence interval for paired data

- Null & Alternative hypotheses: $H_0: \mu_{diff} = 0$

  $H_A: \mu_{diff} < 0$ OR $H_A: \mu_{diff} > 0$ OR $H_A: \mu_{diff} \neq 0$

- Test statistic:

$$t = \frac{\bar{x}_{diff} - 0}{s_{diff}/\sqrt{n_{diff}}} \sim t_{\underbrace{n_{diff}-1}_{\text{degrees of freedom}}}$$

Matched Pairs *t* Test Example

The iron content measured by the two methods for 4 random samples. Assume the differences between the Chemical and X-ray analyses come from a Normal distribution

Sample	1	2	3	4
Chemical	2.1	2.5	2.3	2.5
X-ray	2.0	2.3	2.1	2.4
Difference	0.1	0.2	0.2	0.1

Test whether there is evidence the two methods measure a different amount of iron in the same substance. Use a significance level of 0.1.

$H_0: \mu_{diff} = 0$ $H_A: \mu_{diff} \neq 0$

$\bar{x}_{diff} = 0.15$ $s_{diff} = 0.058$ $n_{diff} = 4$

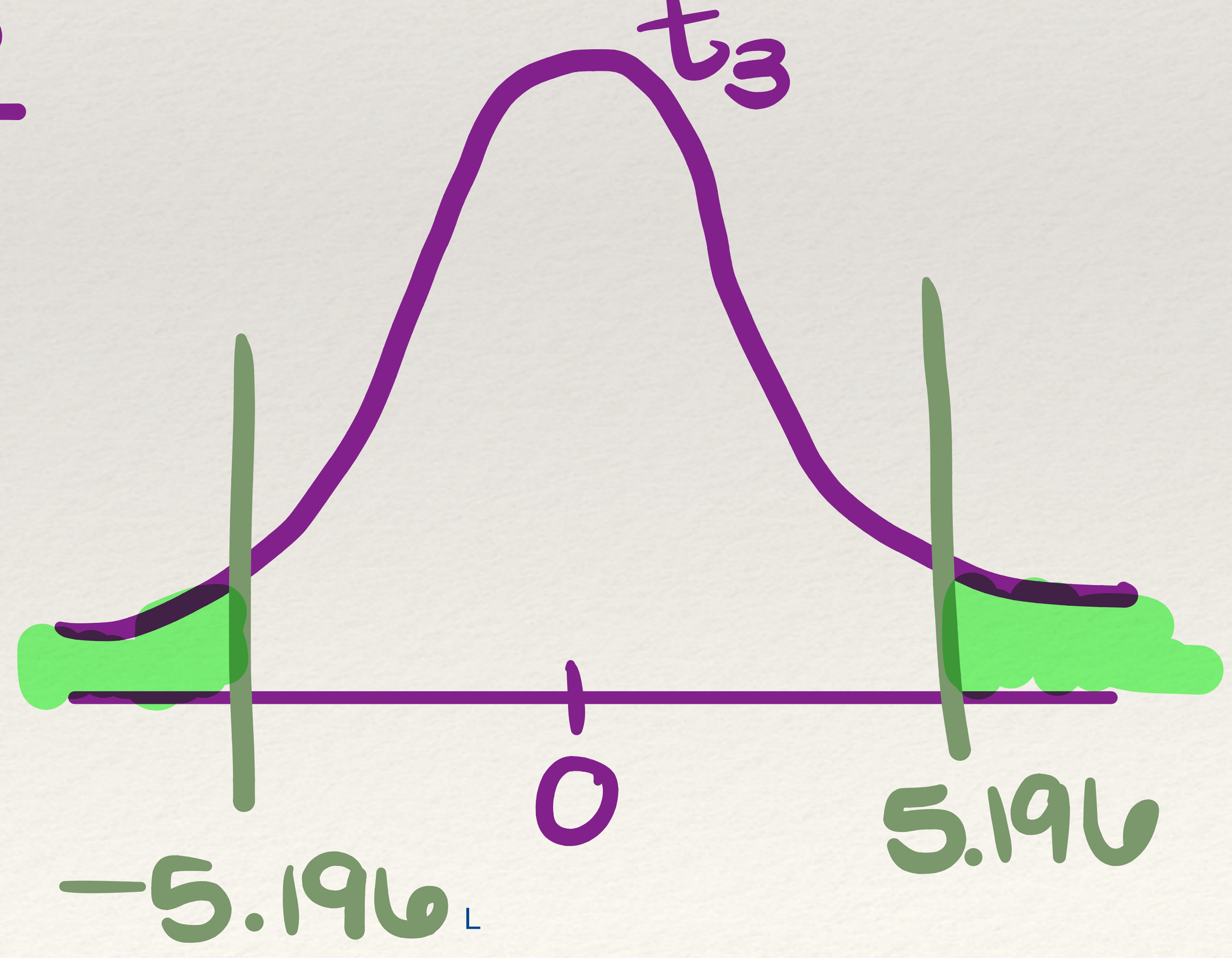
$$t = \frac{0.15 - 0}{\frac{0.058}{\sqrt{4}}} = 5.196$$

$2 * pt(-5.196, 3)$

$= 0.0138$

↑

p-value



Matched Pairs *t* Test Example

The iron content measured by the two methods for 4 random samples. Assume the differences between the Chemical and X-ray analyses come from a Normal distribution

Sample	1	2	3	4
Chemical	2.1	2.5	2.3	2.5
X-ray	2.0	2.3	2.1	2.4
Difference	0.1	0.2	0.2	0.1

Test whether there is evidence the two methods measure a different amount of iron in the same substance. Use a significance level of 0.1.

P-value = 0.0138

Reject the null hypothesis that the average difference between the two methods is 0.

There is convincing/suggestive evidence that the average difference between the two methods is not 0.

Final Exam

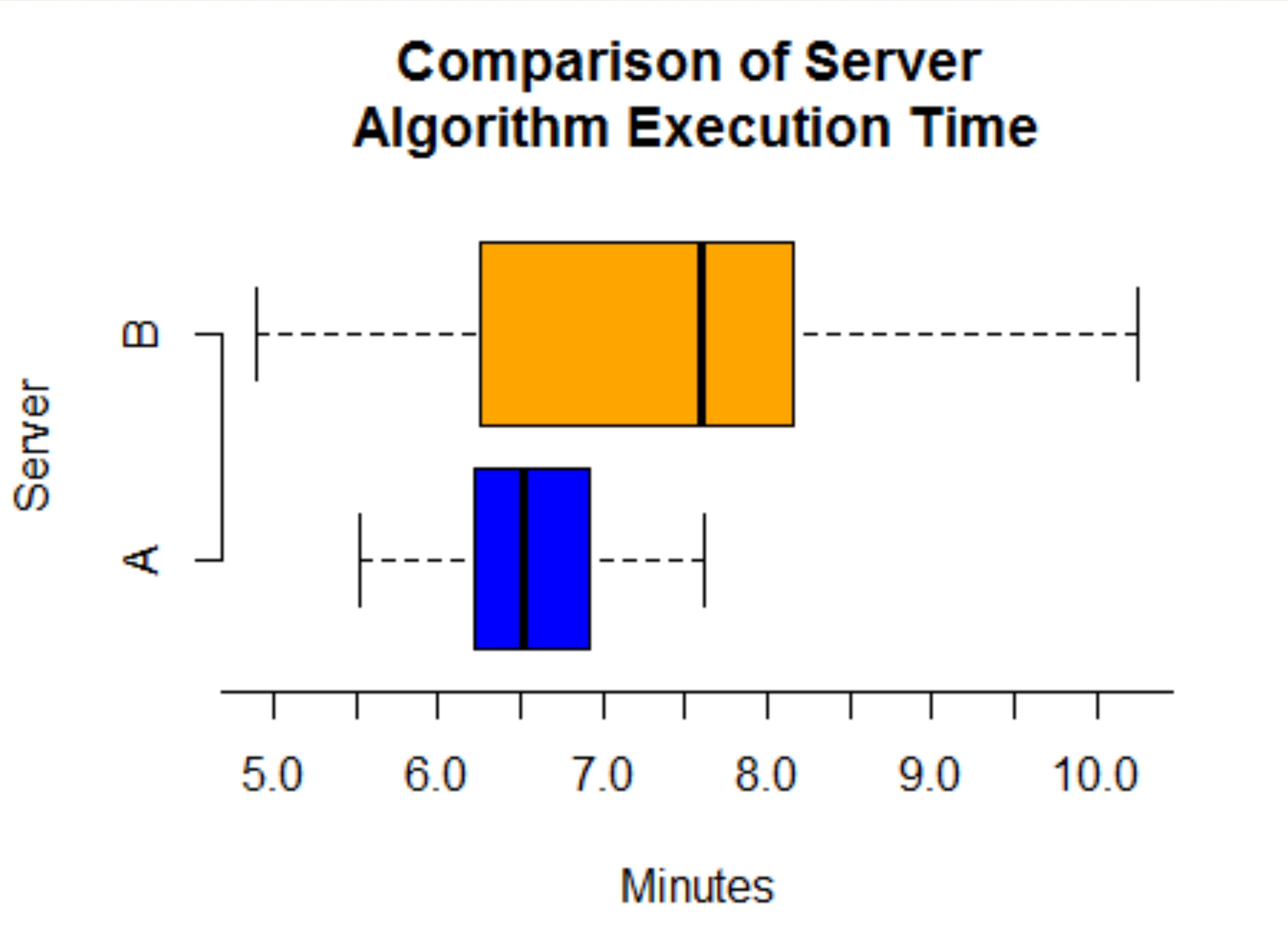
June 8, 12 am - June 9, 11:59 pm

110 minutes to complete exam

No class May 26 - I have jury duty :(

Comparing Population Means

Suppose a business manager needs access to a server. Server A is supposedly faster than server B, but an account on A is more expensive. She would like to compare the speeds of the two servers to see if server A is worth the extra cost.



Two Sample t Confidence Interval

- When to use: want to estimate the difference between 2 population means from 2 independent populations
- Conditions for inference:

-Independent data within and between groups (random sampling)

-Normality – need each sample to be sufficiently large (CLT)

- The confidence interval for the difference in population means is: want to est: $\mu_1 - \mu_2$

$$\underbrace{(\bar{x}_1 - \bar{x}_2)}_{\text{point estimate for } \mu_1-\mu_2} \pm t^*_{\nu} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

↑ degrees of freedom

Satterthwaite Approximate *t* Distribution

- Satterthwaite degrees of freedom:
  - Used in an "unpooled" *t* procedure
  - Used when we do not want to assume the population standard deviations of the two populations of interest are equal (most of the time)

$$\nu = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

- Conservative degrees of freedom:
  - Satterthwhaite can be tedious by hand. Sometimes "conservative" degrees of freedom is used.

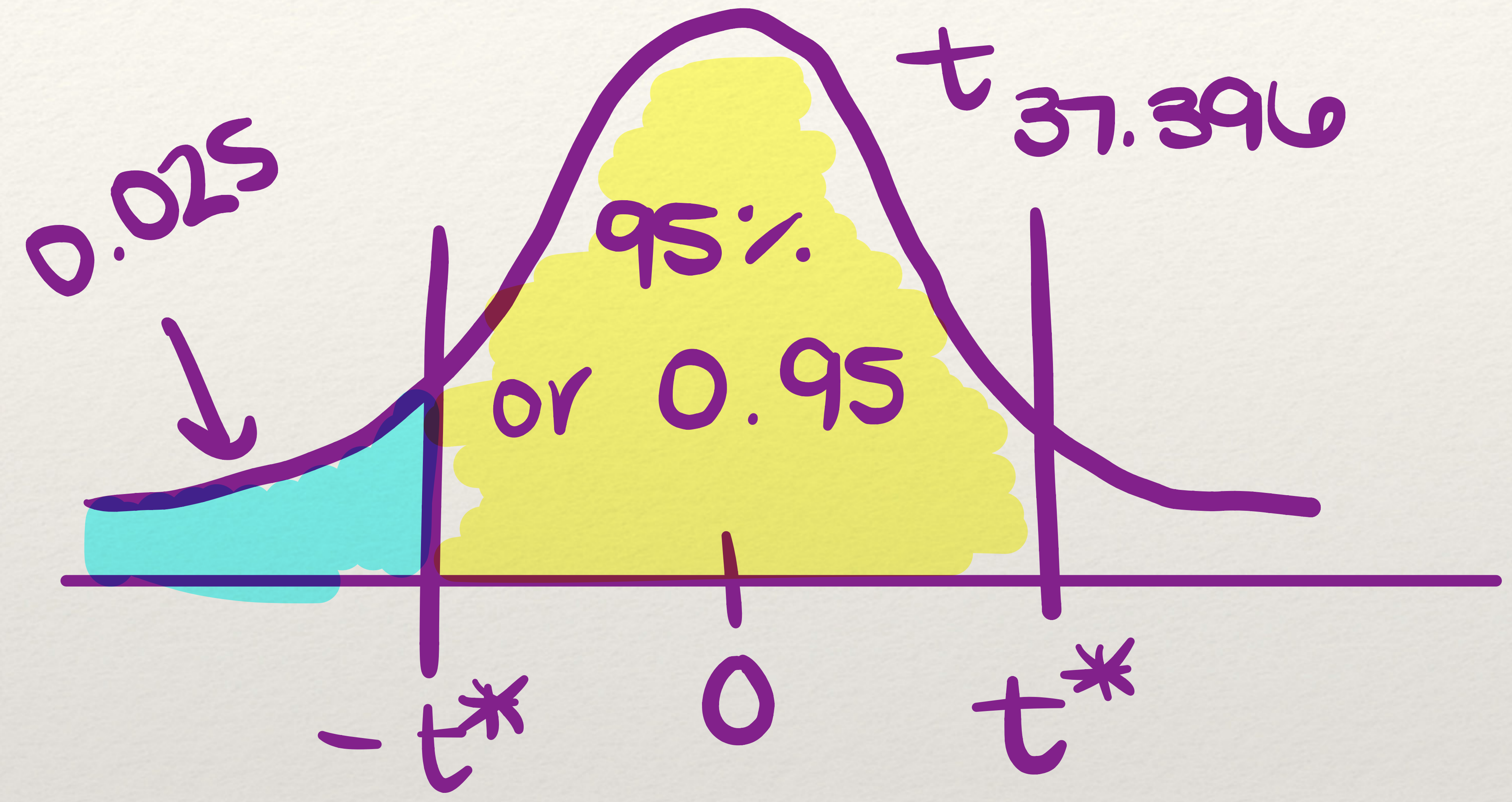
$$\min(n_1 - 1, n_2 - 1)$$

We don't need to use this b/c we have computers!

Confidence Interval Example

Suppose a business manager needs access to a server. To compare the speed in minutes between two servers A and B, a computer algorithm is executed 30 times on server A and 30 times on server B.

Degrees of Freedom $\nu = 37.396$



Random samples from Server A and B

	$\bar{x}$	s	n
Server A	6.5	0.5	30
Server B	7.3	1.3	30

R: qt(0.975, 37.396) = 2.025

$$(\bar{X}_A - \bar{X}_B) \pm t^* \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$$

Calculate a 95% confidence interval for $\mu_A - \mu_B$.

$$= (6.5 - 7.3) \pm 2.025 \sqrt{\frac{0.5^2}{30} + \frac{1.3^2}{30}} = (-1.315, -0.285)$$

Confidence Interval Example

Suppose a business manager needs access to a server. To compare the speed in minutes between two servers A and B, a computer algorithm is executed 30 times on server A and 30 times on server B.

Random samples from Server A and B

	$\bar{x}$	s	n
Server A	6.5	0.5	30
Server B	7.3	1.3	30

Calculate a 95% confidence interval for $\mu_A - \mu_B$.

(-1.315, -0.285)

We are 95% confident that server A is 0.285 minutes to 1.315 minutes faster than server B on average with a point estimate of 0.8 minutes faster.

Two Sample *t* Test

- When to use: want to test the difference between the population means from 2 independent populations
- Conditions required for inference:

Same as two sample t conf. int.

- Null & Alternative hypotheses: $H_0: \mu_1 - \mu_2 = 0$ equivalently $H_0: \mu_1 = \mu_2$

$H_A: \mu_1 - \mu_2 < 0$ OR $H_A: \mu_1 - \mu_2 > 0$ OR $H_A: \mu_1 - \mu_2 \neq 0$

- Test statistic:

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

Distribution of the Two Sample Test Statistic

The two sample t test statistic is:

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

← Represents the hypothesized difference
in most cases $\delta_0 = 0$

The distribution of this statistic follows:

$$t \sim t_\nu$$

↑ satterthwaite df

Two Sample t Test Example

Suppose a business manager needs access to a server. Server A is supposedly faster than server B, but an account on A is more expensive. She would like to compare the speeds of the two servers to see if server A is worth the extra cost.

$H_0: \mu_A - \mu_B = 0$

$H_A: \mu_A - \mu_B < 0$

Random samples from Server A and B

	$\bar{x}$	s	n
Server A	6.5	0.5	30
Server B	7.3	1.3	30

Random samples → Independence ✓

Sufficiently large samples ✓

Determine the hypotheses needed for this test and check the conditions.

$H_0: \mu_1 - \mu_2 = 0 \qquad H_a: \mu_1 - \mu_2 < 0$

Two Sample *t* Test Example

Suppose a business manager needs access to a server. Server A is supposedly faster than server B, but an account on A is more expensive. She would like to compare the speeds of the two servers to see if server A is worth the extra cost.

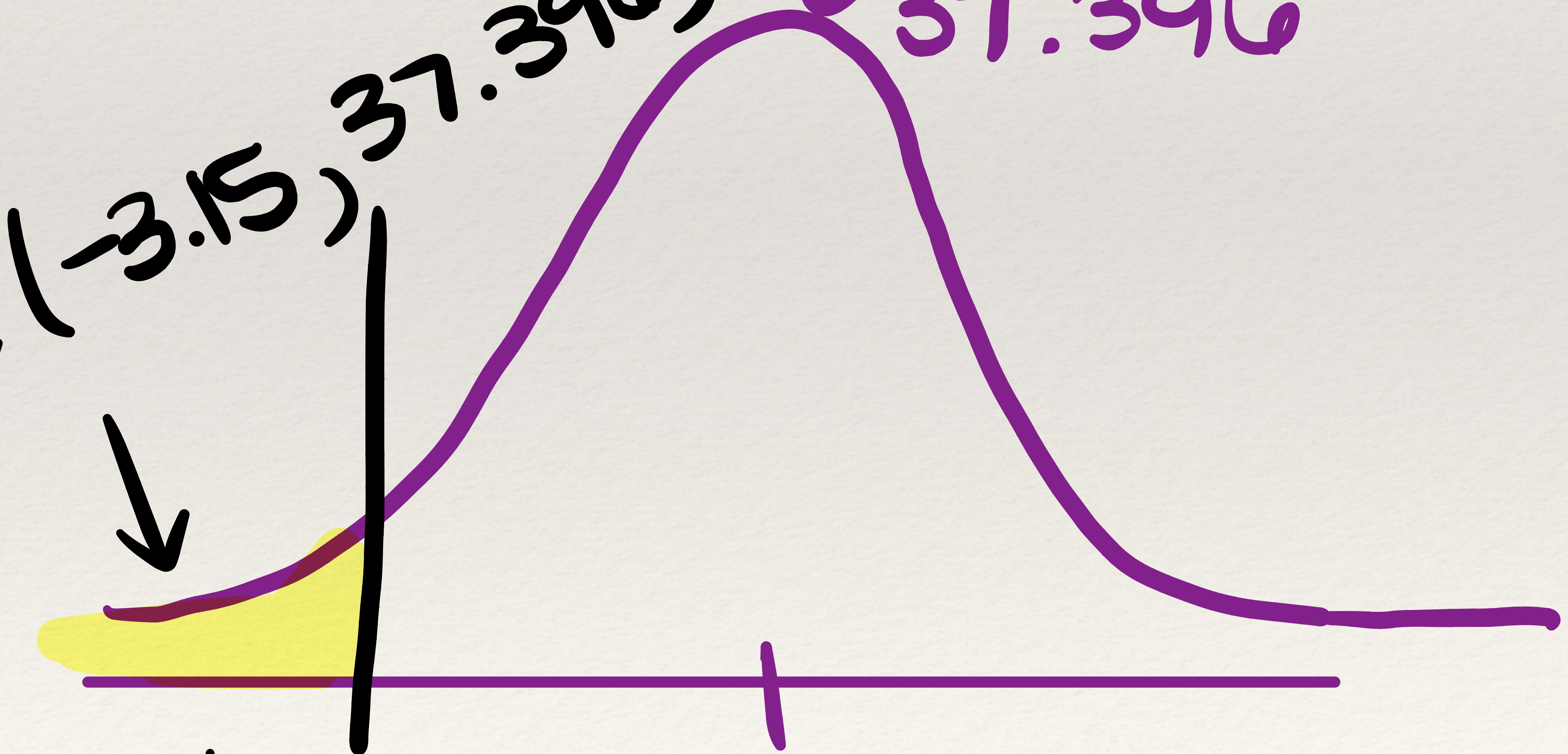
$$t = \frac{(6.5 - 7.3)}{\sqrt{\frac{0.5^2}{30} + \frac{1.3^2}{30}}} = -3.15$$

Random samples from Server A and B

	$\bar{x}$	s	n
Server A	6.5	0.5	30
Server B	7.3	1.3	30

$t_{37.396}$

R: pt(-3.15, 37.396)



$t = -3.15$ $\quad 0$

Using a significance level of 0.05, calculate the test statistic, the degrees of freedom, and p-value.

Make a conclusion.

p-value = 0.0016

Two Sample t Test Example

Suppose a business manager needs access to a server. Server A is supposedly faster than server B, but an account on A is more expensive. She would like to compare the speeds of the two servers to see if server A is worth the extra cost.

Random samples from Server A and B

	$\bar{x}$	s	n
Server A	6.5	0.5	30
Server B	7.3	1.3	30

Using a significance level of 0.05, calculate the test statistic, the degrees of freedom, and p-value.
Make a conclusion.

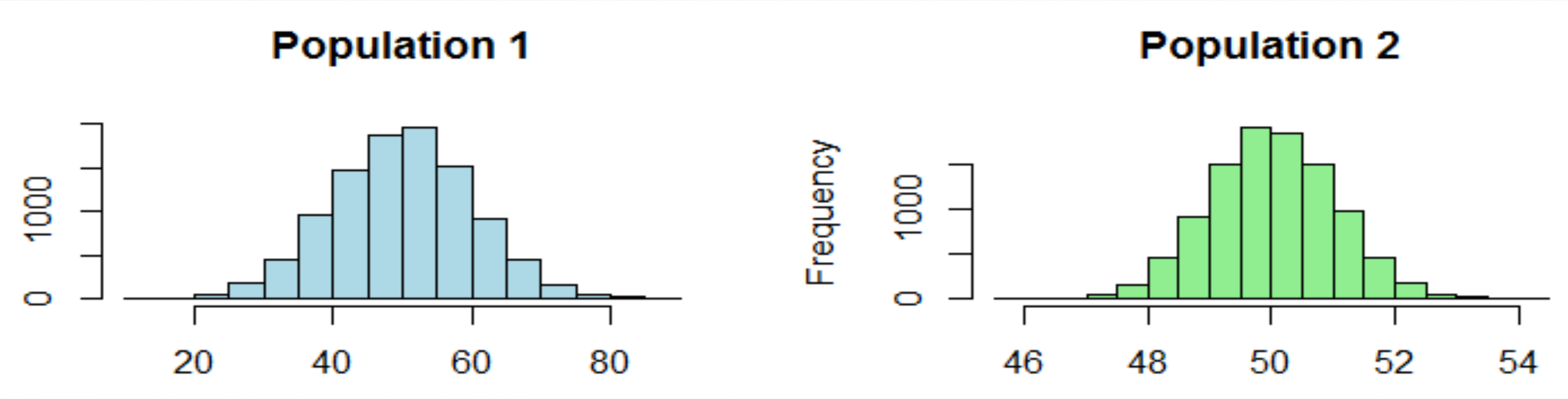
Interpreting the p-value: The probability of observing a difference in sample means of -0.8 or something less than that when the null hypothesis true is 0.0016.

Conclusion: There is convincing evidence to suggest that server A is faster than server B. We reject the null hypothesis that the average run times of the two servers are equal.

Keep in mind, it's possible we've made a Type I error.

Avoid Pooled *t* Procedures

- You may come across resources that use "pooled" procedures.
- A pooled test assumes $\sigma_1 = \sigma_2$.
- Consider the two populations below, both have the same mean but different variances. If we use an $\alpha = 0.05$, given the means are the same we should reject the null, just by chance, about 5% of the time.



$$\mu_1 = \mu_2 = 50$$
$$\sigma_1 = 10 \neq \sigma_2 = 1$$

Because the assumption of equal variance is violated, the test breaks down and rejects the null hypothesis more frequently than it should. This increases type I error.