STA286 Lecture 30

Neil Montgomery

Last edited: 2017-04-05 11:06

what hath 0.05 wrought?

How to misuse hypothesis testing to destroy the universe:

1. Believe that rejecting H_0 means that H_0 is false, and not rejecting H_0 means H_0 is true.

what hath 0.05 wrought?

How to misuse hypothesis testing to destroy the universe:

1. Believe that rejecting H_0 means that H_0 is false, and not rejecting H_0 means H_0 is true.
2. Don't worry about the actual effect size. Just worry about rejecting H_0.

what hath 0.05 wrought?

How to misuse hypothesis testing to destroy the universe:

1. Believe that rejecting H_0 means that H_0 is false, and not rejecting H_0 means H_0 is true.
2. Don't worry about the actual effect size. Just worry about rejecting H_0.
3. Be a journal that only accepts publications in which H_0 is rejected.

what hath 0.05 wrought?

How to misuse hypothesis testing to destroy the universe:

1. Believe that rejecting H_0 means that H_0 is false, and not rejecting H_0 means H_0 is true.
2. Don't worry about the actual effect size. Just worry about rejecting H_0.
3. Be a journal that only accepts publications in which H_0 is rejected.
4. Be a researcher who only publishes results in which H_0 is rejected.

what hath 0.05 wrought?

How to misuse hypothesis testing to destroy the universe:

1. Believe that rejecting H_0 means that H_0 is false, and not rejecting H_0 means H_0 is true.
2. Don't worry about the actual effect size. Just worry about rejecting H_0.
3. Be a journal that only accepts publications in which H_0 is rejected.
4. Be a researcher who only publishes results in which H_0 is rejected.
5. Believe in the sanctity of $\alpha = 0.05$.

what hath 0.05 wrought?

How to misuse hypothesis testing to destroy the universe:

1. Believe that rejecting H_0 means that H_0 is false, and not rejecting H_0 means H_0 is true.
2. Don't worry about the actual effect size. Just worry about rejecting H_0.
3. Be a journal that only accepts publications in which H_0 is rejected.
4. Be a researcher who only publishes results in which H_0 is rejected.
5. Believe in the sanctity of $\alpha = 0.05$.
6. Perform as many hypothesis tests as you like on the same dataset.

what hath 0.05 wrought?

How to misuse hypothesis testing to destroy the universe:

1. Believe that rejecting H_0 means that H_0 is false, and not rejecting H_0 means H_0 is true.
2. Don't worry about the actual effect size. Just worry about rejecting H_0.
3. Be a journal that only accepts publications in which H_0 is rejected.
4. Be a researcher who only publishes results in which H_0 is rejected.
5. Believe in the sanctity of $\alpha = 0.05$.
6. Perform as many hypothesis tests as you like on the same dataset.
7. Use "one-sided alternatives" because you think you really know what you are doing.

what hath 0.05 wrought?

How to misuse hypothesis testing to destroy the universe:

1. Believe that rejecting H_0 means that H_0 is false, and not rejecting H_0 means H_0 is true.
2. Don't worry about the actual effect size. Just worry about rejecting H_0.
3. Be a journal that only accepts publications in which H_0 is rejected.
4. Be a researcher who only publishes results in which H_0 is rejected.
5. Believe in the sanctity of $\alpha = 0.05$.
6. Perform as many hypothesis tests as you like on the same dataset.
7. Use "one-sided alternatives" because you think you really know what you are doing.
8. Think that you really know what you are doing.

p-values

a story that never really happened

Are the doors the right width or not?

$$H_0 : \mu = 700$$
$$H_1 : \mu \neq 700$$

a story that never really happened

Are the doors the right width or not?

$$H_0 : \mu = 700$$
$$H_1 : \mu \neq 700$$

Use population model $N(\mu, \sigma = 0.5)$. Set $\alpha = 0.05$. Plan to collect a sample of size $n = 10$. The rejection region is:

$$\left\{\overline{X} < 700 - 1.96\frac{0.5}{\sqrt{10}}\right\} \cup \left\{\overline{X} > 700 + 1.96\frac{0.5}{\sqrt{10}}\right\}$$
$$\left\{\overline{X} < 699.69\right\} \cup \left\{\overline{X} > 700.31\right\}$$

a story that never really happened

Are the doors the right width or not?

$$H_0 : \mu = 700$$
$$H_1 : \mu \neq 700$$

Use population model $N(\mu, \sigma = 0.5)$. Set $\alpha = 0.05$. Plan to collect a sample of size $n = 10$. The rejection region is:

$$\left\{\overline{X} < 700 - 1.96\frac{0.5}{\sqrt{10}}\right\} \cup \left\{\overline{X} > 700 + 1.96\frac{0.5}{\sqrt{10}}\right\}$$
$$\left\{\overline{X} < 699.69\right\} \cup \left\{\overline{X} > 700.31\right\}$$

You actually measure 10 doors. The sample average is $\overline{x} = 699.68$.

a story that never really happened

Are the doors the right width or not?

$$H_0 : \mu = 700$$
$$H_1 : \mu \neq 700$$

Use population model $N(\mu, \sigma = 0.5)$. Set $\alpha = 0.05$. Plan to collect a sample of size $n = 10$. The rejection region is:

$$\left\{\overline{X} < 700 - 1.96\frac{0.5}{\sqrt{10}}\right\} \cup \left\{\overline{X} > 700 + 1.96\frac{0.5}{\sqrt{10}}\right\}$$
$$\left\{\overline{X} < 699.69\right\} \cup \left\{\overline{X} > 700.31\right\}$$

You actually measure 10 doors. The sample average is $\overline{x} = 699.68$. We have a 2319! PUSH THE RED BUTTON!!! **REJECT THE NULL! REJECT THE NULL!**

a story that never really happened

So you cancel the contract with the supplier, who goes out of business.

a story that never really happened

So you cancel the contract with the supplier, who goes out of business.

It turns out that the summer student who compiled the data made a small error in recording one of the door widths - putting 700.2 instead of 700.4 for that one record.

a story that never really happened

So you cancel the contract with the supplier, who goes out of business.

It turns out that the summer student who compiled the data made a small error in recording one of the door widths - putting 700.2 instead of 700.4 for that one record.

So actually $\overline{x}$ is 699.70.

a story that never really happened

So you cancel the contract with the supplier, who goes out of business.

It turns out that the summer student who compiled the data made a small error in recording one of the door widths - putting 700.2 instead of 700.4 for that one record.

So actually $\overline{x}$ is 699.70.

Everthing has ***completely changed***. H_0 is not rejected. FAIL TO REJECT! FAIL TO REJECT! Situation is niner-niner-zero.

But it's too late. The market has decided. Lives are destroyed. Demagogues rise to power in the wake of mass disillusionment.

a story that never really happened

So you cancel the contract with the supplier, who goes out of business.

It turns out that the summer student who compiled the data made a small error in recording one of the door widths - putting 700.2 instead of 700.4 for that one record.

So actually $\overline{x}$ is 699.70.

Everthing has ***completely changed***. H_0 is not rejected. FAIL TO REJECT! FAIL TO REJECT! Situation is niner-niner-zero.

But it's too late. The market has decided. Lives are destroyed. Demagogues rise to power in the wake of mass disillusionment.

Another option is to use something called a *p-value*.

p-value

A p-value is the probability (calculated using the H_0 parameter value) of observing a value of the test statistic "more extreme" that what was actually observed.

"More extreme" just means further away (in absolute value) than the H_0 parameter value.

p-value

A p-value is the probability (calculated using the H_0 parameter value) of observing a value of the test statistic "more extreme" that what was actually observed.

"More extreme" just means further away (in absolute value) than the H_0 parameter value.

In the Doors example, $\overline{x}$ was thought at first to be 699.68, which is 0.32mm away from 700. The probability of being *more than* 0.32mm away from 700 is:

$$P(\overline{X} < 699.68) + P(\overline{X} > 700.32) = 0.021 + 0.021 = 0.043$$

p-value

A p-value is the probability (calculated using the H_0 parameter value) of observing a value of the test statistic "more extreme" that what was actually observed.

"More extreme" just means further away (in absolute value) than the H_0 parameter value.

In the Doors example, $\overline{x}$ was thought at first to be 699.68, which is 0.32mm away from 700. The probability of being *more than* 0.32mm away from 700 is:

$$P(\overline{X} < 699.68) + P(\overline{X} > 700.32) = 0.021 + 0.021 = 0.043$$

After the correction, $\overline{x}$ is now 699.70. The p-value is now:

$$P(\overline{X} < 699.70) + P(\overline{X} > 700.30) = 0.029 + 0.029 = 0.058$$

p-value

A p-value is the probability (calculated using the H_0 parameter value) of observing a value of the test statistic "more extreme" that what was actually observed.

"More extreme" just means further away (in absolute value) than the H_0 parameter value.

In the Doors example, $\overline{x}$ was thought at first to be 699.68, which is 0.32mm away from 700. The probability of being *more than* 0.32mm away from 700 is:

$$P(\overline{X} < 699.68) + P(\overline{X} > 700.32) = 0.021 + 0.021 = 0.043$$

After the correction, $\overline{x}$ is now 699.70. The p-value is now:

$$P(\overline{X} < 699.70) + P(\overline{X} > 700.30) = 0.029 + 0.029 = 0.058$$

Too bad you didn't know about p-values before causing WWIII.

the use and interpretation of p-values

P-values should be used to evaluate the evidence against H_0.

The smaller the p-value, the stronger the evidence.

the use and interpretation of p-values

P-values should be used to evaluate the evidence against H_0.

The smaller the p-value, the stronger the evidence.

There is no magic threshold for how small is "small enough."

the use and interpretation of p-values

P-values should be used to evaluate the evidence against H_0.

The smaller the p-value, the stronger the evidence.

There is no magic threshold for how small is "small enough." So don't ask.

the use and interpretation of p-values

P-values should be used to evaluate the evidence against H_0.

The smaller the p-value, the stronger the evidence.

There is no magic threshold for how small is "small enough." So don't ask. Please stop asking.

the use and interpretation of p-values

P-values should be used to evaluate the evidence against H_0.

The smaller the p-value, the stronger the evidence.

There is no magic threshold for how small is "small enough." So don't ask. Please stop asking.

Some language you might overhear me using:

- 0.126 might be "no evidence"

the use and interpretation of p-values

P-values should be used to evaluate the evidence against H_0.

The smaller the p-value, the stronger the evidence.

There is no magic threshold for how small is "small enough." So don't ask. Please stop asking.

Some language you might overhear me using:

- 0.126 might be "no evidence"
- 0.063 might be "weak evidence"

the use and interpretation of p-values

P-values should be used to evaluate the evidence against H_0.

The smaller the p-value, the stronger the evidence.

There is no magic threshold for how small is "small enough." So don't ask. Please stop asking.

Some language you might overhear me using:

- 0.126 might be "no evidence"
- 0.063 might be "weak evidence"
- 0.0031 might be "evidence"

the use and interpretation of p-values

P-values should be used to evaluate the evidence against H_0.

The smaller the p-value, the stronger the evidence.

There is no magic threshold for how small is "small enough." So don't ask. Please stop asking.

Some language you might overhear me using:

- 0.126 might be "no evidence"
- 0.063 might be "weak evidence"
- 0.0031 might be "evidence"
- 0.0000014 might be "strong evidence"

the use and interpretation of p-values

P-values should be used to evaluate the evidence against H_0.

The smaller the p-value, the stronger the evidence.

There is no magic threshold for how small is "small enough." So don't ask. Please stop asking.

Some language you might overhear me using:

- 0.126 might be "no evidence"
- 0.063 might be "weak evidence"
- 0.0031 might be "evidence"
- 0.0000014 might be "strong evidence"
- 3×10^{-15} might be "overwhelming evidence"

the use and interpretation of p-values

P-values should be used to evaluate the evidence against H_0.

The smaller the p-value, the stronger the evidence.

There is no magic threshold for how small is "small enough." So don't ask. Please stop asking.

Some language you might overhear me using:

- 0.126 might be "no evidence"
- 0.063 might be "weak evidence"
- 0.0031 might be "evidence"
- 0.0000014 might be "strong evidence"
- 3×10^{-15} might be "overwhelming evidence"

the use and interpretation of p-values

P-values should be used to evaluate the evidence against H_0.

The smaller the p-value, the stronger the evidence.

There is no magic threshold for how small is "small enough." So don't ask. Please stop asking.

Some language you might overhear me using:

- 0.126 might be "no evidence"
- 0.063 might be "weak evidence"
- 0.0031 might be "evidence"
- 0.0000014 might be "strong evidence"
- 3×10^{-15} might be "overwhelming evidence"

Think in terms of orders of magnitude.

things that annoy me

When people ask me how small a p-value “has” to be.

things that annoy me

When people ask me how small a p-value "has" to be.

When people compute a p-value, and then say "The p-value is smaller than 0.05, so I reject the null hypothesis."

$100 \cdot (1 - \alpha)\%$ C.I. versus hypothesis test with size α

A C.I. formula and a rejection region formula for $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ will be based on (something like):

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

$100 \cdot (1 - \alpha)\%$ C.I. versus hypothesis test with size α

A C.I. formula and a rejection region formula for $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ will be based on (something like):

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

For the C.I., unwrap to isolate μ in the middle. For the R.R., put $\mu = \mu_0$ and unwrap to isolate $\overline{X}$ in the middle.

$100 \cdot (1-\alpha)\%$ C.I. versus hypothesis test with size α

A C.I. formula and a rejection region formula for $H_0 : \mu = \mu_0$ versus $H_1 : \mu \neq \mu_0$ will be based on (something like):

$$P\left(-z_{\alpha/2} < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}\right) = 1 - \alpha$$

For the C.I., unwrap to isolate μ in the middle. For the R.R., put $\mu = \mu_0$ and unwrap to isolate $\overline{X}$ in the middle.

The following is true:

Reject H_0 at level α $\iff$ $100 \cdot (1-\alpha)\%$ C.I. does not containμ_0

the "one-sample t test"

You'll never know σ, so use the data to estimate σ with s, as usual.

From the Doors example where $n = 10$ and $\overline{x} = 699.70$, suppose also that that $s = 0.599$.

the "one-sample t test"

You'll never know σ, so use the data to estimate σ with s, as usual.

From the Doors example where $n = 10$ and $\overline{x} = 699.70$, suppose also that that $s = 0.599$.

The p-value is now calculated based on:

$$\frac{\overline{X} - 700}{s/\sqrt{n}} \sim t_9$$

$$P(\overline{X} < 699.70) + P(\overline{X} > 700.30) = P(t_9 < -1.585) + P(t_9 > 1.585) = 0.074 + 0.074 = 0.148$$

the "two-sample t-test"

A more realistic hypothesis testing scenario.

Two populations: $N(\mu_1, \sigma_1)$ and $N(\mu_2, \sigma_2)$. The obvious hypotheses will always be:

$$H_0 : \mu_1 = \mu_2$$
$$H_1 : \mu_1 \neq \mu_2$$

The "parameter" is $\theta = \mu_1 - \mu_2$, estimated (as usual) by $\overline{X_1} - \overline{X_2}$ from samples of sizes n_1 and n_2.

Two possibilities:

$$\frac{\overline{X_1} - \overline{X_2}}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} \sim t_{n_1+n_2-2} \qquad \text{or} \qquad \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} \sim t_\nu$$

two-sample t-test example

Modified from 10.106. Can nutritional counselling change blood cholesterol level? A group of 15 people received counseling for 8 weeks. A group of 18 people did not.

The readings are made available by the textbook in the following terrible manner:

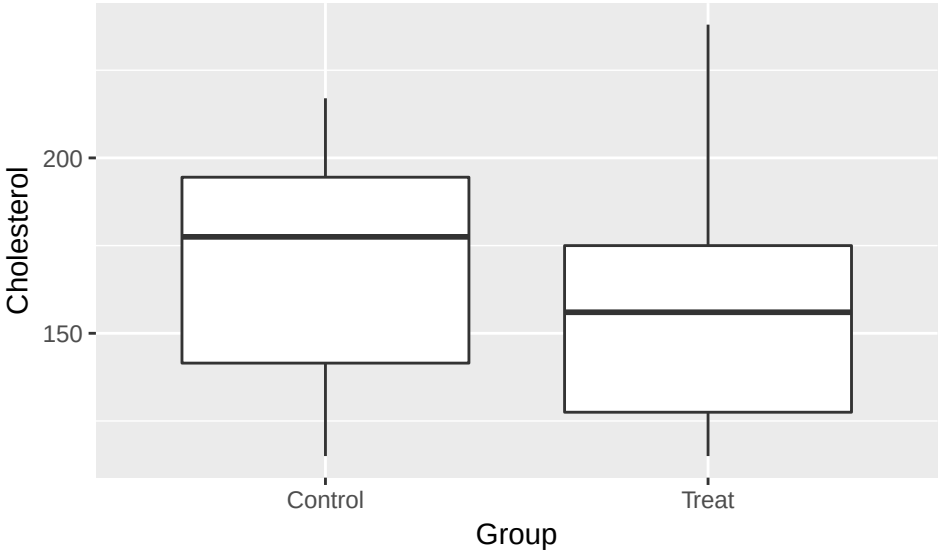
Treat	Control
129	151
131	132
154	196
172	195
115	188
126	198
175	187
191	168
122	115
238	165
159	137
156	208
176	133
175	217
126	191
	193
	140
	146

two-sample t-test example

A Real Dataset:

```
##       Group Cholesterol
## 1     Treat         129
## 2     Treat         131
## 3     Treat         154
## 4     Treat         172
## 5     Treat         115
## 6     Treat         126
## 7     Treat         175
## 8     Treat         191
## 9     Treat         122
## 10    Treat         238
## 11    Treat         159
## 12    Treat         156
## 13    Treat         176
## 14    Treat         175
```

two-sample t-test example - plot



two-sample t-test example - equal variance version

```
## # A tibble: 2 × 4
##     Group     n  X_bar      S
##    <fctr> <int>  <dbl>  <dbl>
## 1 Control    18 170.00 30.788
## 2   Treat    15 156.33 33.090

##
##  Two Sample t-test
##
## data:  Cholesterol by Group
## t = 1.23, df = 31, p-value = 0.23
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.0417 36.3750
## sample estimates:
## mean in group Control   mean in group Treat
##                170.00                156.33
```

two-sample t-test example - no variance assumption version

```
##
##  Welch Two Sample t-test
##
## data:  Cholesterol by Group
## t = 1.22, df = 29, p-value = 0.23
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -9.2584 36.5917
## sample estimates:
## mean in group Control   mean in group Treat
##                170.00                156.33
```