Math 183
Statistical Methods

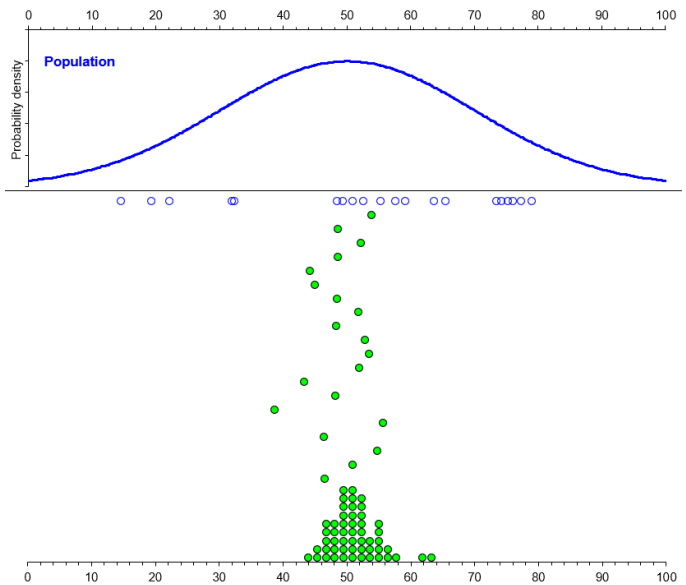
Eddie Aamari
S.E.W. Assistant Professor

eaamari@ucsd.edu
math.ucsd.edu/~eaamari/
AP&M 5880A

Today: Chapter 4 (continued)

- p-value for hypothesis testing
- Computing and interpreting p-values
- See the $\alpha = 0.05$ significance threshold as arbitrary/changeable

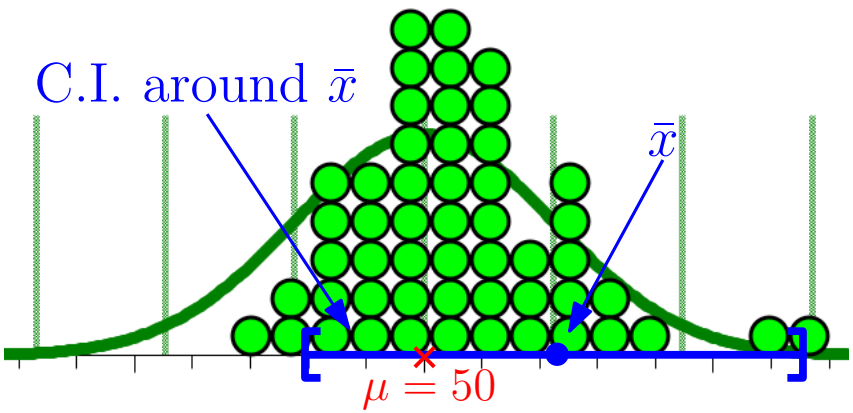
Refresher



1. Draw a sample, find $\bar{x}$ and s_x
2. Recognize that sampling variations make $\bar{x}$ an imperfect measure for μ.
3. Study the sampling distribution, which shows the variability in $\bar{x}$
4. Discover that the sampling distribution is $N\left(\mu, \sigma/\sqrt{n}\right)$
5. Use this information to build a confidence interval of μ from $\bar{x}$, that shows your uncertainty.

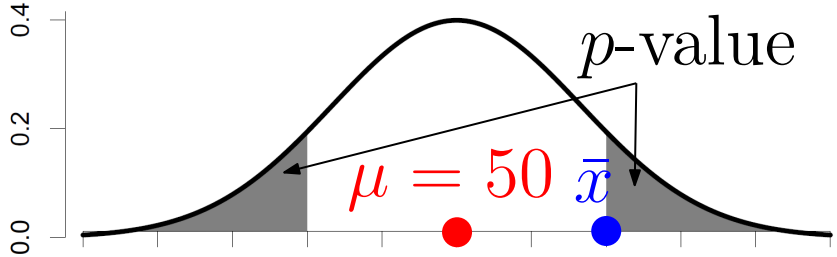
p-Values in Hypothesis Testing

Last class: We used confidence intervals around data to determine plausible parameter values for μ and, hence, if the null hypothesis was plausible.
Assume H_0: $\mu = 50$



This approach places you at $\bar{x}$, asking if $\mu = 50$ is strangely far away.

This class: We learn a new tool: the p-value.
This is commonly used across all research literature.



This approach places you at $\mu = 50$, asking if $\bar{x}$ is strangely far away.

Does Better Teaching Improve Stat Grades?

You teach the stat class in 2017 and 2018. In 2017, the average on the final was 82% with a standard deviation of 12%. Before teaching it again, you go to a seminar to improve your teaching. In a random sample of 50 students from 2018, you get an average of 85.7%. Did the seminar really improve your teaching?

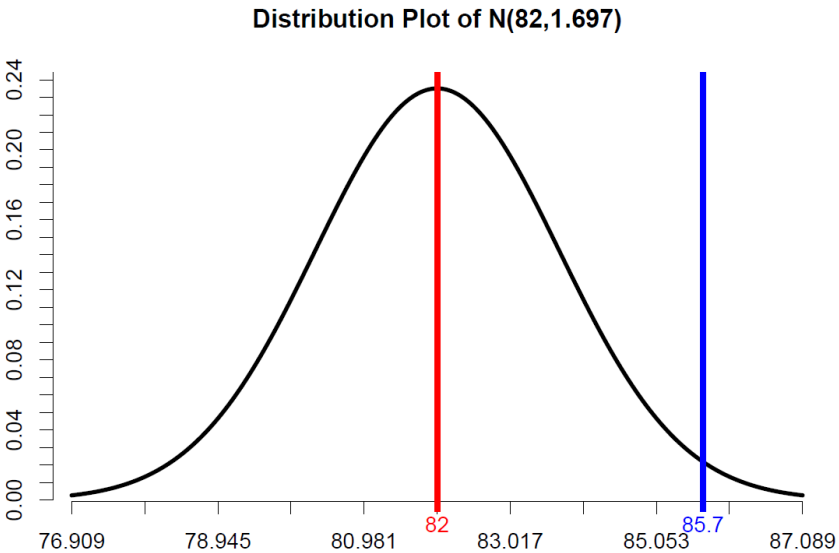
1) Set up hypotheses:

H_0: $\mu_{2018} = \mu_{2017} = 82$.
H_A: $\mu_{2018} > 82$.

2) Assume H_0, identify the sampling distribution of $\bar{x}$ under this assumption, and quantify how wacky $\bar{x}$ is in the H_0 universe.

Under H_0, the sampling distribution is

$$N\left(\mu_{2017}, \frac{\sigma_{2017}}{\sqrt{n}}\right) = N\left(82, \frac{12}{\sqrt{50}}\right)$$
$$= N\left(82, 1.697\right).$$

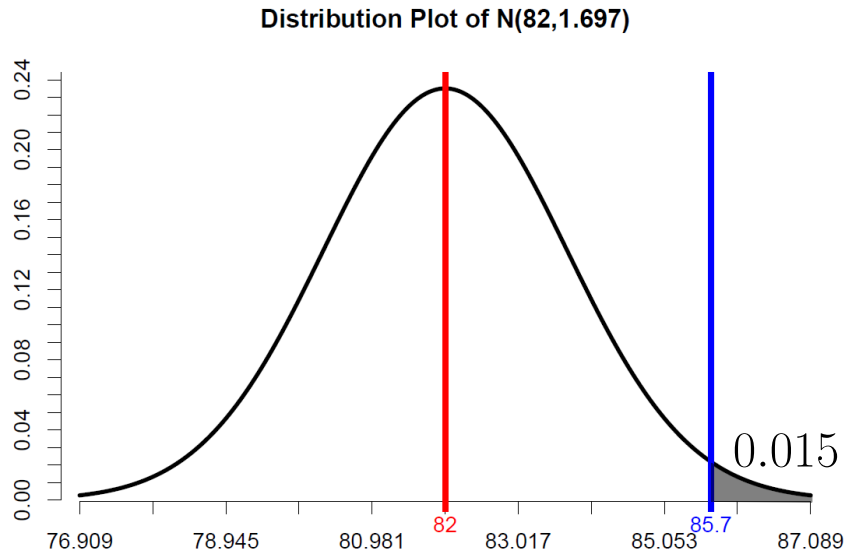


How wacky is $\bar{x} = 85.7$ compared to 82?

We'd like to ask about $P(\bar{x}|H_0)$, but this probability is 0.

Istead, we find

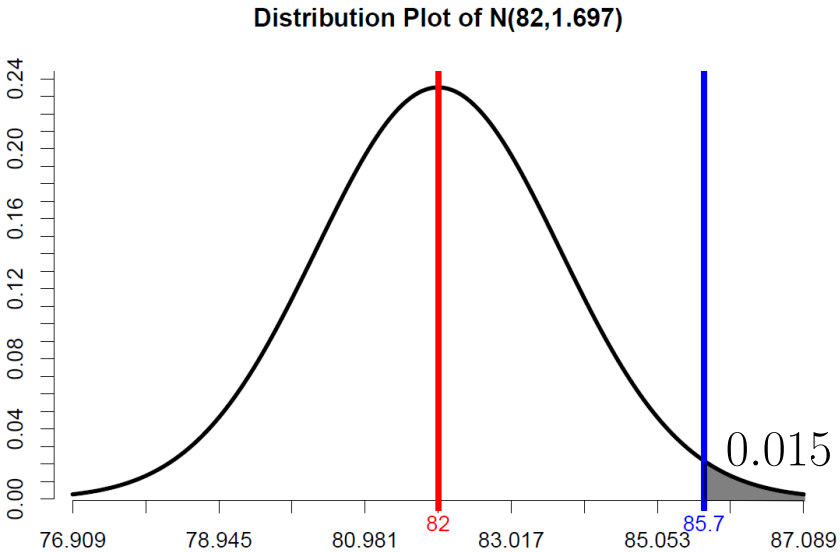
$P\left(\bar{x} \text{ or something more extreme}|H_0\right)$



By finding this area, we see that 85.7 is among the top 1.5% of averages we could ever get. This is a very strange/extreme mean to get in a world where we expect means around 82

What Do You Do With Your p-Value?

Before doing your research, you decide on a "significance level". Let's say $\alpha = 0.05$. This level signifies how strange a result would need to be for you to reject H_0 in favor of H_A.



In our example,
$p = 0.015 < \alpha = 0.05$, so we reject H_0 in favor of H_A.

General decision rule:

If $p < \alpha$, reject H_0.

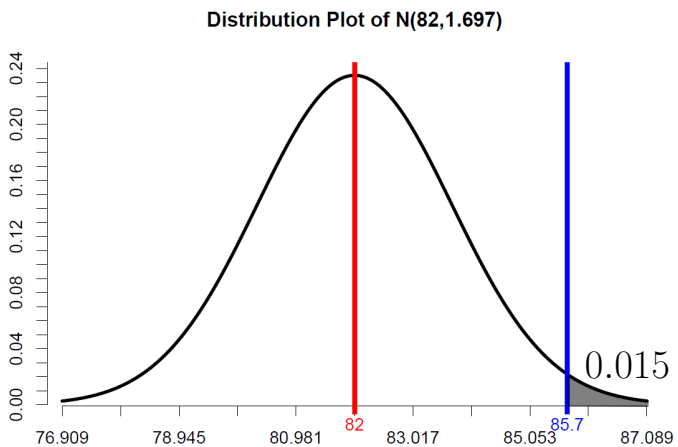
If $p \geq \alpha$, do not reject H_0.

What a p-Value Is and Is Not

The p-value is a number equal to

$$P(\text{data}^{+}|H_0 \text{ is true}).$$

As a probability, it is sharing how unlikely you would be to see such data (or something wilder) if H_0 really is true. It says that your data is among the p strangest results. Strange results cast doubt on the assumption H_0 is true and support H_A instead.



Results to the right of 85.7 support H_A: $\mu_{2018} > 82$. As we draw pictures, shade those values that cause us to favor H_A over H_0.

A p value is <u>not</u> $P(H_0$ is true).
A p value is <u>not</u> $P(H_A$ is true).

A Two-Sided Example

Let μ be the Facebook friend count average among (all) Math 183 students (ever). Your friend claims this mean is different than the average for Math 11 students. Let's find out

Suppose I tell you the Math 11 Facebook friend count average is 440 with a standard deviation of 283.

1) Set up hypotheses:

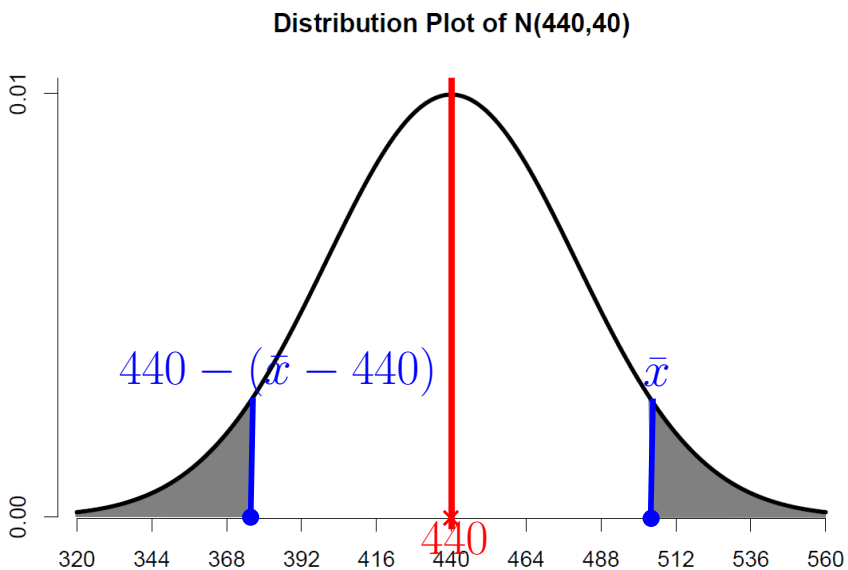
$$H_0\colon\ \mu = 440\ , \qquad H_A\colon\ \mu \neq 440.$$

(We make a two-sided alternative since it is not clear if we should predict higher or lower. Both trends are interesting information)

2) Collect some data and find the sample mean.
Say you get $n = 50$ and $\bar{x} = 500$.

3) Figure out the sampling distribution (assuming H_0), plot your sample mean on it, and shade the area(s) that contribute to the p-value.
Our sampling distribution is

$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) = N\left(440, \frac{283}{\sqrt{50}}\right) = N(440, 40).$$



With a two sided alternative, you shade both tails.

```
1 > 2*pnorm(500,mean=440,sd=40,
      lower.tail=F)
2 [1] 0.1336144
3 > 2*pnorm(380,mean=440,sd=40,
      lower.tail=T)
4 [1] 0.1336144
```

4) Get your p-value, interpret it, and make a decision.

In the case of marks at 380 and 500, we got $p = 0.134$.
This says that the mean we observed was among the 13.4% strangest means we could get.

Since $p > 0.05$, we do not reject H_0: our data are not strong enough to make us switch to the alternative hypothesis.

Practice

During exams, UCSD offers "dog therapy". You are curious if this lowers stress levels of students (scores on a 0 through 100 scale). Suppose without dogs, students have an average stress level of 81 during exams.
You sample 25 random students that were assigned to pet therapy and found average of 78.5 with a standard deviation of 7.
Does this provide strong evidence that pet therapy decreases stress levels during exams?

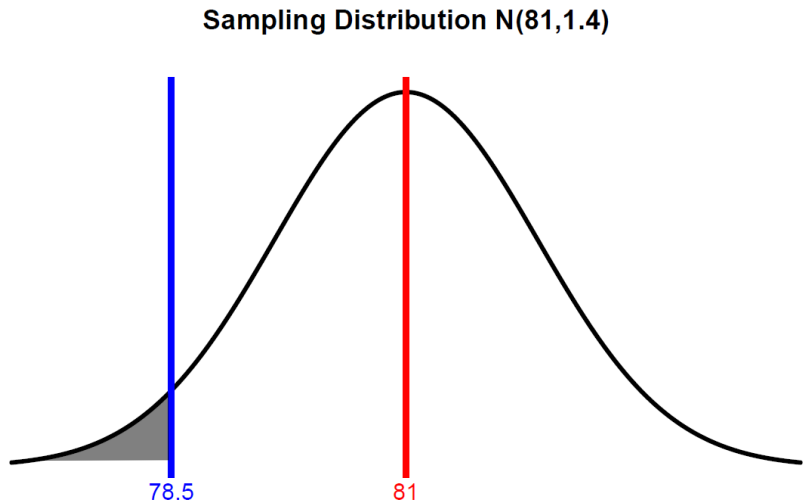
1) Define the parameter you will study.
Let μ be the average stress level of all UCSD students that have dog therapy (ever!)

2) State H_0 and H_A $\qquad H_0 : \mu = 81, \qquad H_A : \mu < 81.$

3) Find the model for the sampling distribution
Assuming H_0, it is

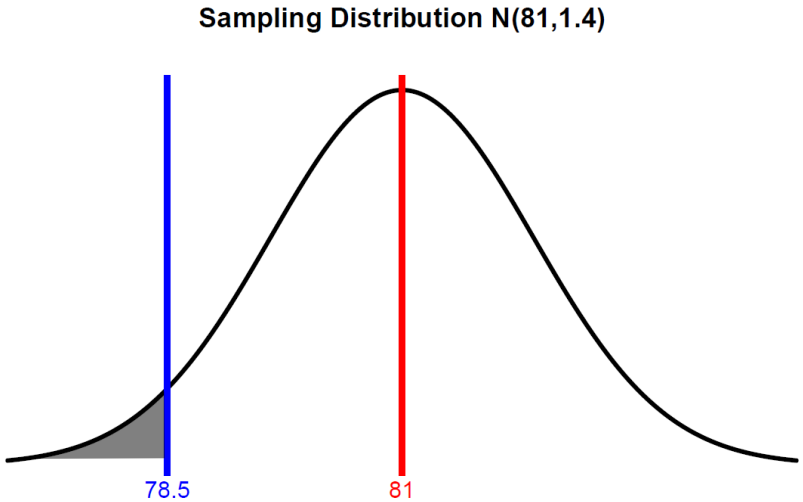
$$N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \simeq N\left(81, \frac{s_x}{\sqrt{n}}\right) = N\left(81, \frac{7}{\sqrt{25}}\right) = N\left(81, 1.4\right).$$

4) Draw a picture showing the sampling distribution, its center, and your data. Shade the area for the p-value you will find based on the nature of H_A.



Remark:
With one-sided alternatives, we shade only one tail.
With two-sided alternatives, we shade two tails.

5) Write a command in R that finds this area, or use a z-table.




```
> pnorm(78.5, mean = 81, sd =
    1.4)
[1] 0.03707277
```

6) Interpret this p-value in words and draw a conclusion
$p = 0.037$ means that, when assuming H_0 is true, the probability of getting our data (or something more extreme) is only 3.7%. Because $p = 0.037 < 0.05$, we reject H_0 in favor of H_A.
It appears that dog therapy lowers stress levels among UCSD students at exam time.

What's So Special About $\alpha = 0.05$?

$\alpha = 0.05$ is a historical artifact derived from one sentence in a 1931 book by R.A. Fisher, *The design of Experiments*. He thought that a 1 in 20 event ($= 5\%$) might be surprising enough to toss out one's belief system (H_0) in favor of something else (H_A).

Some fields have a far more demanding threshold like $\alpha = 0.0000003$. This is usually called the "5 sigma rule": you need to see an event $5SE$'s from the assumend mean in order to discard H_0 in favor of H_A Examples:

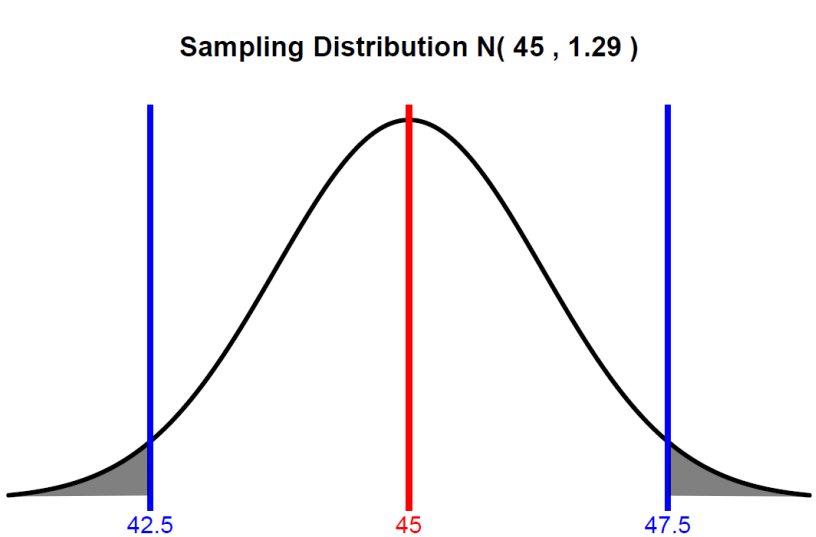
- Particule physics
- Pharmacology
- Aircraft design processes

p-Values without Technology

A researcher wonders if online ads change people's views on a movie. For those who don't view ads, the average approval score for the movie is 45 (on a 0 to 100 scale) with a standard deviation of 10. The researcher plans to show ads about the movie to 60 random people and then find their average approval score.

1) Define a parameter and hypotheses to test.
Let μ be the average approval score of the movie for people that see the ads.
We set $H_0 : \mu = 45$ and $H_A : \mu \neq 45$

2) Collect sample: the research finds $\bar{x} = 47.5$. Draw a picture of this scenario assuming H_0 and shade the area for the p-value.
Under H_0, the sampling distribution is $N\left(45, 10/\sqrt{60}\right) \simeq N(45, 1.29)$.

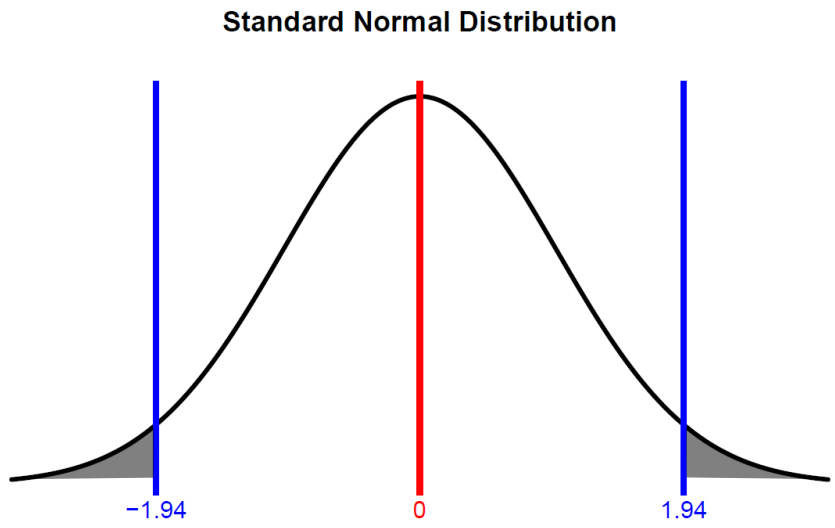


With technology: we use R

```
> 2*pnorm(42.5,mean=45,sd
    =1.29,lower.tail=T)
[1] 0.05262511
```

Without technology: we use a table. This demands we transition to the standard normal distribution using z-scores

$$z = \frac{42.5 - 45}{1.29} \simeq 1.94.$$

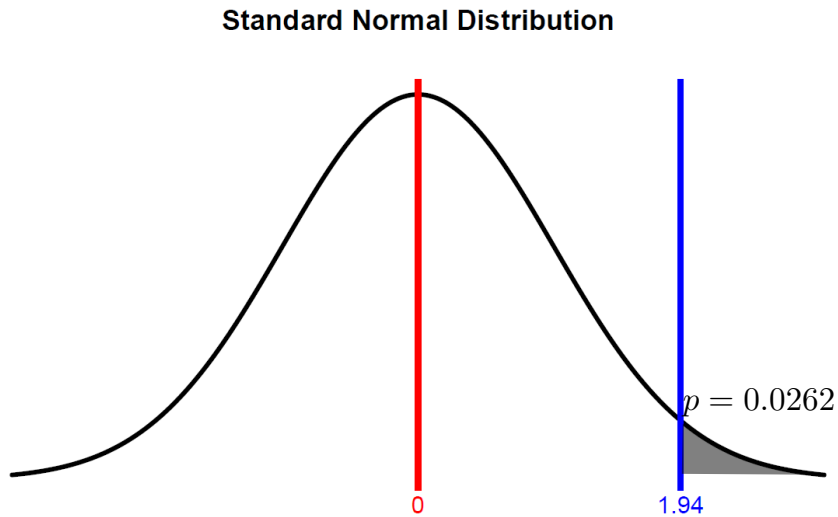


0.05	0.04	0.03	0.02	0.01	0.00	Z
0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0004	0.0004	0.0004	0.0005	0.0005	0.0005	−3.3
0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	−3.2
0.0008	0.0008	0.0009	0.0009	0.0009	0.0010	−3.1
0.0011	0.0012	0.0012	0.0013	0.0013	0.0013	−3.0
0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	−2.9
0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	−2.8
0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	−2.7
0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	−2.6
0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	−2.5
0.0071	0.0073	0.0075	0.0078	0.0080	0.0082	−2.4
0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	−2.3
0.0122	0.0125	0.0129	0.0132	0.0136	0.0139	−2.2
0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	−2.1
0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	−2.0
0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	−1.9
0.0322	0.0329	0.0336	0.0344	0.0351	0.0359	−1.8

$$p = 2 \times 0.0262 = 0.0524.$$

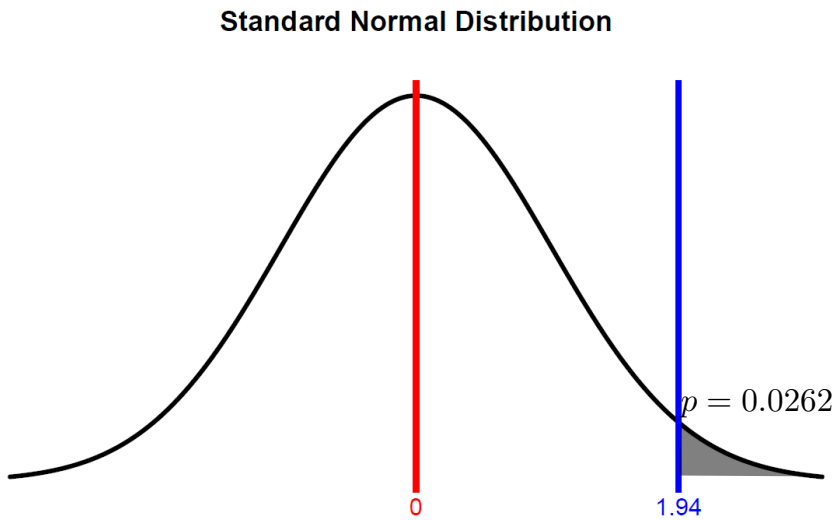
Since $p = 0.0524 > 0.05$, we do not reject H_0 at the confidence level $\alpha = 0.05$: ads do not seem to influence people's views on a movie.

After finding this out, with the same data, you decide to change the question to "Are ads increasing people's opinions about movies?" You test $H_0 : \mu = 45$ VS $H_A : \mu > 45$.



Here, the p-value is $p = 0.0262 < \alpha = 0.05$. Hence you reject H_0 and conclude that ads seem to improve people's views on movies.

Comment on this approach



This is bad practice!
Since two-sided tests always have twice the area(both tails) as one-sided tests, you can always cut your p-value in half by switching to a one-sided test.

This violates the process of science, which involves setting hypotheses and **then**, collecting data to confirm/deny those hypotheses.
If you want to update your hypotheses after seeing existing data, then you need new data to confirm/deny the updated hypotheses.