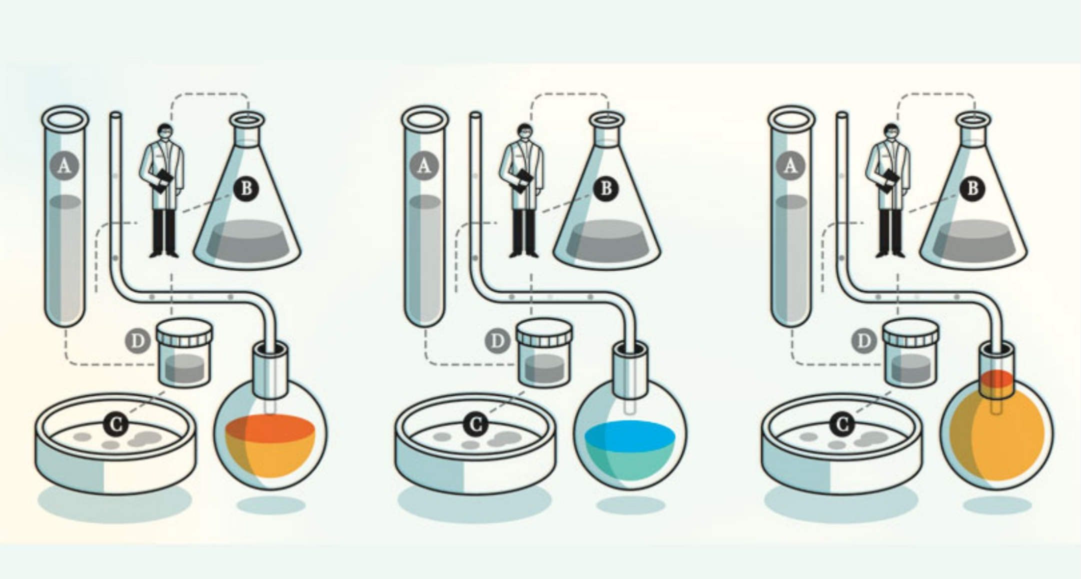
CHAPTER 25

Hypothesis Testing



The main purpose of statistics is to test a hypothesis.

You might run an experiment and find that a certain drug is effective at treating blood pressure.

But if you can’t repeat the experiment, then no one will take your results seriously.

What is a hypothesis?

- A hypothesis is simply an educated guess about something you are interested in.
- A hypothesis should be testable;

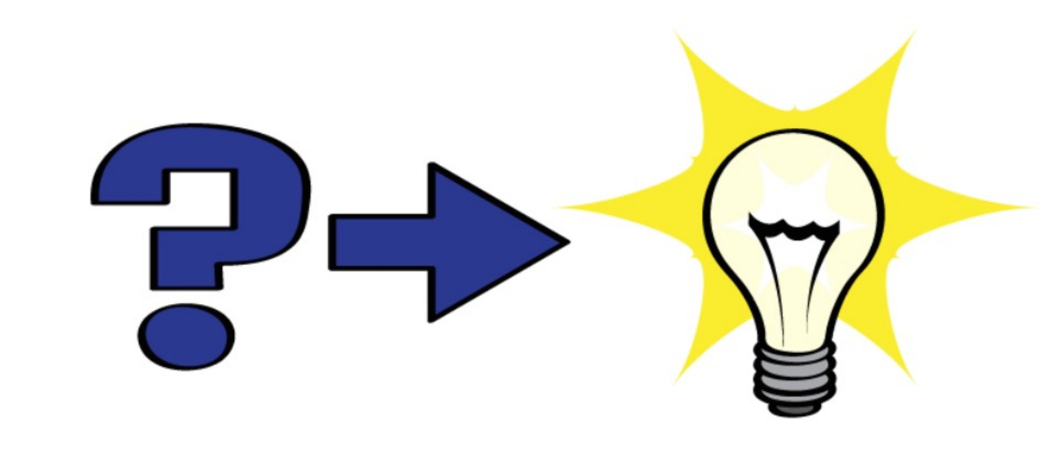
  either by experiment or observation.

Examples of Hypotheses:

- This new medication might work.
- This new way of teaching will work better.
- Traveling this other route is faster.

A hypothesis just needs to be able to be tested. Then we can analyze the data from the experiment to determine the results.

How do I make a hypothesis?



When you propose a hypothesis,
write it as a statement.
”If I do this thing to an independent variable,
then this other thing will happen to the dependent variable.”

Examples of hypothesis statements:

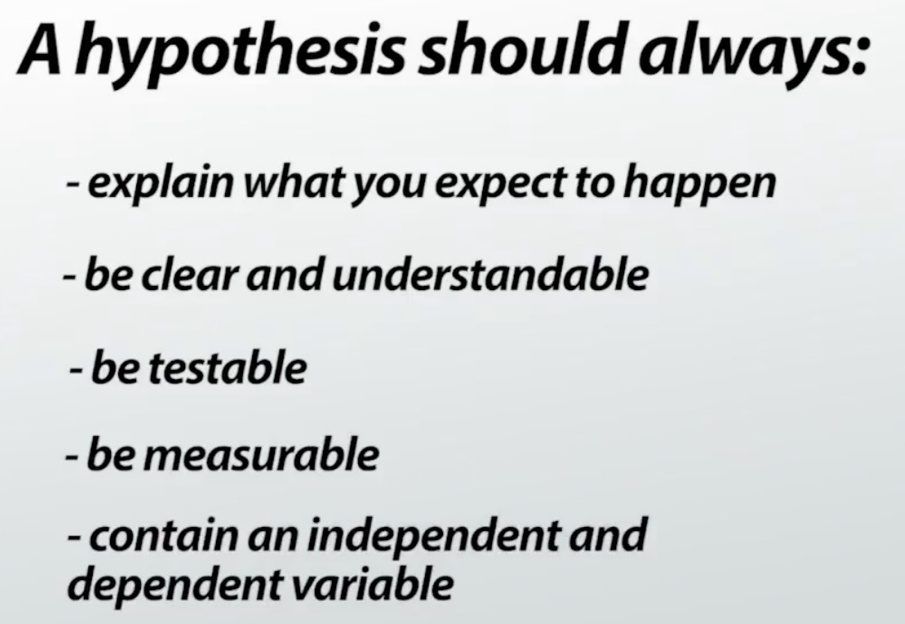
If I put more fertilizer on the garden,
then the corn will increase in size.

If a doctor has a kindly beside manner,
then the patient’s overall depression level is lower.

If statistics exams are opened for 24 hours,
then the student test scores will improve.

A good hypothesis statement should:

1] Include an ”if” and “then” statement.
2] Include both the independent and dependent variable
3] Be testable by experiment, survey, or other scientifically sound technique
4] Be based on information in prior research
5] Have design criteria, i.e. programming projects



A hypothesis should always:

- *explain what you expect to happen*
- *be clear and understandable*
- *be testable*
- *be measurable*
- *contain an independent and dependent variable*

The blue and orange statements are paramount to understanding Hypothesis testing

Hypothesis testing in statistics is a way for you to test the results of an experiment (or survey) to see if you have meaningful results.

You are basically testing whether your results are valid by figuring out the odds that your results have happened by chance.

If the odds that something happened are very low, say 1%, then we reject the idea that the change was simply coincidence.

If your results may have happened by chance, then the experiment won't be repeatable and so has little use.

The first step in creating a hypothesis test is to come up with the null hypothesis

If you want to demonstrate something that goes against ‘common knowledge’ then you find evidence against what most people believe. What most people believe is the null hypothesis.

The null hypothesis is a statement concerning the accepted population parameter, μ. Your goal is to find evidence against the null hypothesis.

If the FDA claims that a 1 kg container of cooking oil is 0.15% cholesterol at most, (and you want to show this is false) then the null hypothesis would be “Each 1 Kg container of cooking oil is 0.15% cholesterol”

Assuming the null is true, you then calculate a test statistic from a sample showing, with high confidence, that we should reject the null hypothesis in favor of an alternate hypothesis.

How to run a hypothesis test:

1] Figure out your null hypothesis

2] State your null hypothesis

3] Choose what kind of test you need to perform

4] Run the test, look at the results:

There are only two possible outcomes:

i. You reject the null hypothesis

ii. You fail to reject the null hypothesis

(accept)
(no change in our status)
I Fail to Reject the Null Hypothesis.
Will you marry me?

We do not ever "accept the null hypothesis"

semantics: fail to reject the null… ~ support the null…

What is a null hypothesis?

Think of the null hypothesis as the *dull hypothesis*.

It is the status quo. It is the hypothesis that has always existed. It is the hypothesis that says there is no change. It is the hypothesis that everyone accepts as the way it is.

You are looking for a fact that is nullifiable.

i.e., you want a hypothesis that you can reject in favor of some new exciting thing.

Cooking oil actually contains 0.30% cholesterol!

Generally accepted facts that you might want to nullify in favor of a new idea:

There are 8 planets in the solar system

Bears hibernate for the winter

Eating salt increases blood pressure

Ibuprofen is more effective than aspirin in helping a person who has had a heart attack.

Contrary to popular belief, people can see through walls.

Redheads are insecure about their hair color.

Young boys are prone to more behavioral problems than young girls.

Children of obese parents are more likely to become obese themselves.

Dr. Stuart has telekinetic abilities and can read minds.

People are more susceptible to colds in the fall than the winter.

Teen-age males are more reckless drivers than teen-age females.

Raisin Bran contains two-scoops or raisins.

These are examples of what we might call the null hypothesis.

In any hypothesis testing problem, there are always two competing hypothesis under consideration:

1] H_0 is the null hypothesis
2] H_1 is the alternative hypothesis (also notated H_a)

The objective of hypothesis testing is to choose, based on sampled data, between two competing hypothesis about the value of a population parameter.

By analogy, in a court of law, we might consider the null hypothesis:

H_0: The defendant is not guilty

Then the jury is presented with evidence that supports the alternate hypothesis:

H_1: The defendant is guilty.

If the evidence is strong enough, then the jury rejects H_0 (rejects the null hypothesis)

Is there evidence for the alternate hypothesis?

The burden of proof is placed on those that believe the alternate claim.

The initially favored claim H_0 will not be rejected in favor of the alternate claim H_1, unless the sample evidence provides enough support for the alternative.

Remember, there are only two outcomes:

1] Reject H_0 (in favor of H_1)

2] Fail to reject H_0

We start by assuming the null because we typically do not want to accept another assertion unless data can be shown to strongly support it.

The reluctance to change is due to time, money, effort,…

For instance, if your current method of business is providing $\$10{,}000$ per month, then change is not desired unless there is evidence that the change will provide much more than $\$10{,}000$ per month.

The null hypothesis versus the alternative hypothesis (in notation)

The null hypothesis is usually denoted with something like:

$$H_0\colon \theta = \theta_0$$

i.e., the population parameter is the same as… they are equal…

While the alternative hypothesis is denoted as one of the following:

$$H_1\colon \theta \neq \theta_0 \qquad H_1\colon \theta < \theta_0 \qquad H_1\colon \theta > \theta_0$$

i.e. things are not as you suspect;
the parameter is actually different… smaller… bigger…

The alternative hypothesis is the one for which we are seeking statistical evidence.

Test statistic

The test statistic is a quantity derived from the sample data.
The test statistic is calculated assuming that the null hypothesis is true.
The test statistic is used in the decision about whether to reject the null, H_0.
The test statistic is your evidence concerning the competing hypothesis, H_1.

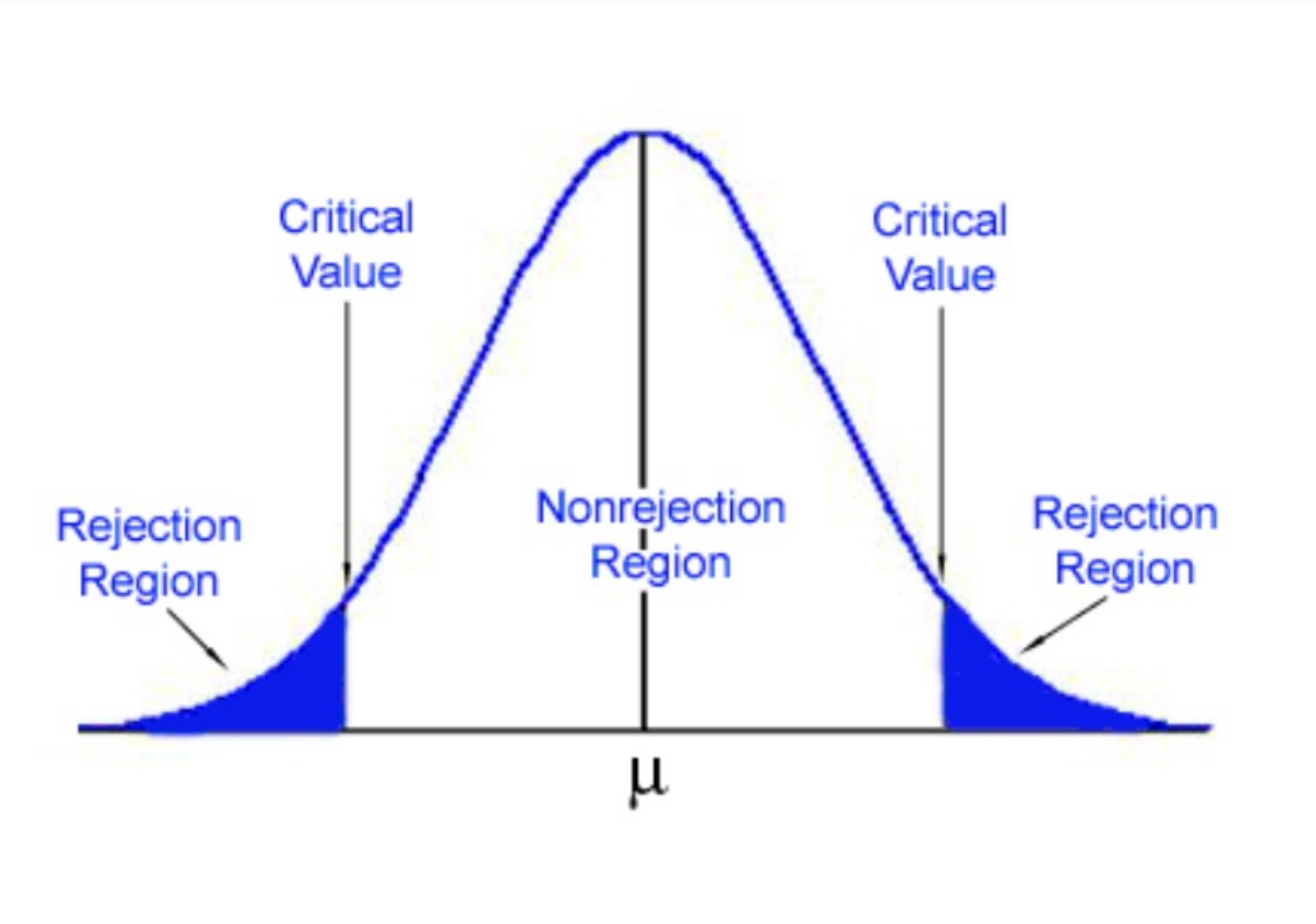
Consider the test statistic under the assumption that the null hypothesis is true by asking questions like:

How likely would we be to obtain this evidence
if the null hypothesis were true?

Standard procedure

- write down H_0 and H_1
- Compute the test statistic
- Compute $Z_{critical}$ based on α.

- critical values on the edge of the rejection region.



Rejection region and significance level

The rejection region is a range of values of the test statistic that would lead you to reject the null hypothesis.

The significance level α indicates the largest probability of the test statistic occurring under the null hypothesis that would lead you to reject the null hypothesis

IS THE BELGIUM 1-EURO COIN BIASED?

To determine if the new Belgium 1-Euro coin is ‘fair’, you flip it a large number of times and record the proportion of heads, call this $\hat{p}$.

- What is the test statistic?

- What is the null hypothesis?

- What is the alternate hypothesis?

- What would it take to convince you that the coin is not fair?

To determine if the new Belgium 1 euro coin is 'fair', you flip it a large number of times and record the proportion of heads, call this $\hat{p}$.

- What is the test statistic?

  If $X =$ the number of flips that come up heads,

  then, $\hat{p} = \frac{X}{n}$

- What is the null hypothesis?

$$H_0: p = 0.5$$

- What is the alternate hypothesis?

$$H_1: p \neq 0.5$$

- What would it take *to convince* ***you*** that the coin is not fair?

$$\hat{p} = \frac{51}{100} \text{ or } \frac{59}{100} \text{ or } \frac{65}{100} \text{ or } \frac{75}{100} \text{ or } \ldots$$

Determining if the Belgium 1-euro coin is fair.

We know the proportion of heads in a **fair** coin is 50%

So, we stated the null hypothesis, $H_0: p = 0.5$

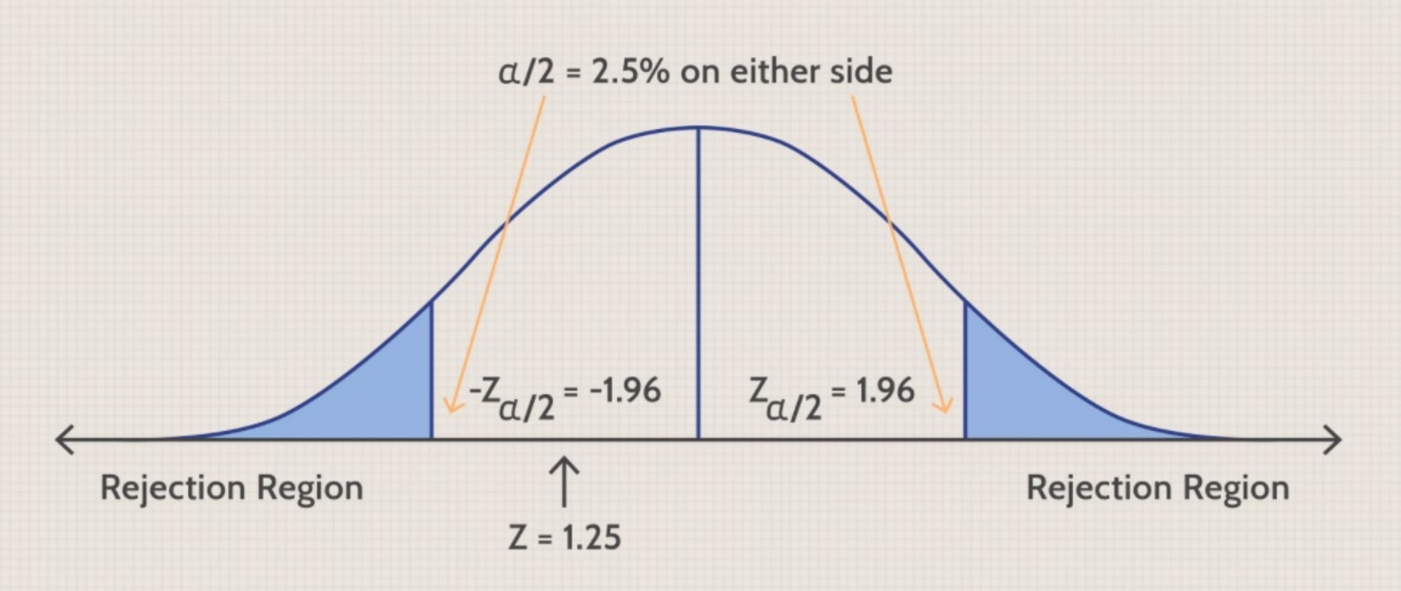
And the alternate hypothesis $H_1: p \neq 0.5$

Let's decide we would be convinced with a 95% CI in a single sample test of the population proportion.

95% CI implies 5% in the tails.

Therefore,

$$Z_{\alpha/2} = \text{norm.ppf}(0.975) = 1.96$$



Determining if the Belgium 1-euro coin is fair.

With lots of flips $\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$ and by the CLT: $Z = \frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$.

Assuming H_0 is true: $\frac{\hat{p}-0.5}{\sqrt{\frac{0.5(1-0.5)}{100}}} = Z$

Does this make sense? If the coin were fair (i.e., we are assuming H_0 is true), then what would you expect the value of Z to be after collecting a sample proportion, $\hat{p}$?

Z measures our "test statistic". If we assume H_0 is true, then Z should be centrally located.

The rejection region at significance level α, that is $\alpha = 1\ -$ confidence level, is pictured on the previous slide. The non-centrally located regions.

So, lets collect some data to determine an actual value for $\hat{p}$.

To determine if the Belgian 1-Euro coin is fair,

you flip it 250 times and find that it comes up heads 139 times.

Do you reject the null at the 0.1 significance level or not?

$H_0: p = 0.5$ $H_1: p \neq 0.5$ significance level $\alpha = 0.1$.

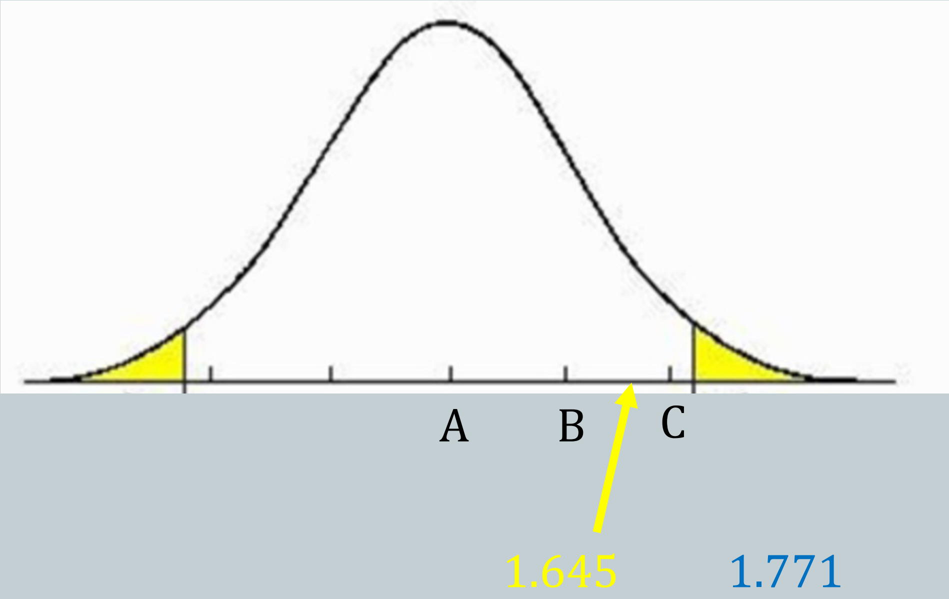
Calculate Z – assuming H_0 is true.

$$\hat{p} = \frac{139}{250} \sim N(p, \sigma^2) = N\left(p, \frac{p(1-p)}{n}\right) = N\left(0.5, \frac{1/_2 \cdot 1/_2}{250}\right) = N(.5, .001)$$ (by the CLT)

$\mathrm{A} = 0.5$

$\mathrm{B} = 0.5 + 0.032 = 0.532$

$C = 0.5 + 2 \cdot 0.032$



Calculate Z assuming H_0

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\frac{139}{250} - 0.5}{\sqrt{\frac{0.5(1-0.5)}{250}}} = 1.771$$

What Z puts α probability in the tails: $CDF\left(Z\alpha_{/_2}\right) = 1 - {}^{\alpha}/_2 = 0.95$

$Z_{crit} =$ `stats.norm.ppf(.95)` $= 1.645$

$1.771 > 1.645$ implies reject H_0 at the 10% significance level: the coin is bias.

To determine if the Belgian 1-Euro coin is fair at the 0.05 significance level,
you flip it 250 times and find that it comes up heads 139 times.
Do you reject the null at the 0.05 significance level or not?

$H_0 : p = 0.5$ $\quad\quad$ $H_1 : p \neq 0.5$ $\quad\quad$ significance level $\alpha = 0.05$.

Calculate Z – assuming H_0 is true.

Note for understanding, but not imperative to solving…

$$\hat{p} = \frac{139}{250} \sim N(p, \sigma^2) = N\left(p, \frac{p(1-p)}{n}\right) = N\left(0.5, \frac{{}^1/_2 \cdot {}^1/_2}{250}\right) = N(.5, .001)$$ (by the CLT)

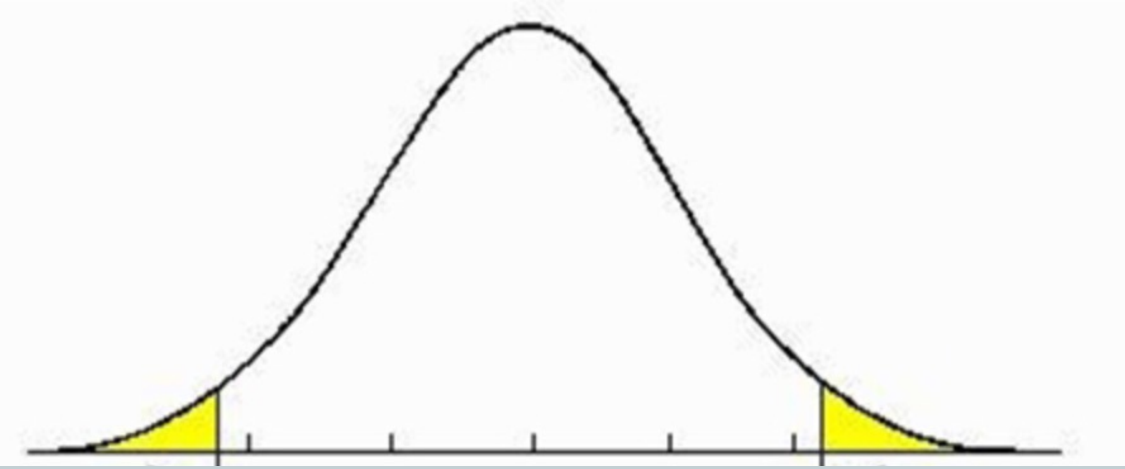
$$\sqrt{.001} \approx 0.0316$$

To determine if the Belgian 1-Euro coin is fair at the 0.05 significance level

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{\frac{139}{250} - 0.5}{\sqrt{\frac{0.5(1-0.5)}{250}}} = 1.771$$

What Z puts α probability in the tails: $CDF\left(Z_{\alpha/2} \right) = 1 - \alpha/2 = 0.975$

$Z_{crit} =$ `stats.norm.ppf(.975) =` 1.96



$1.771 \ngtr 1.96$ implies fail to reject H_0.

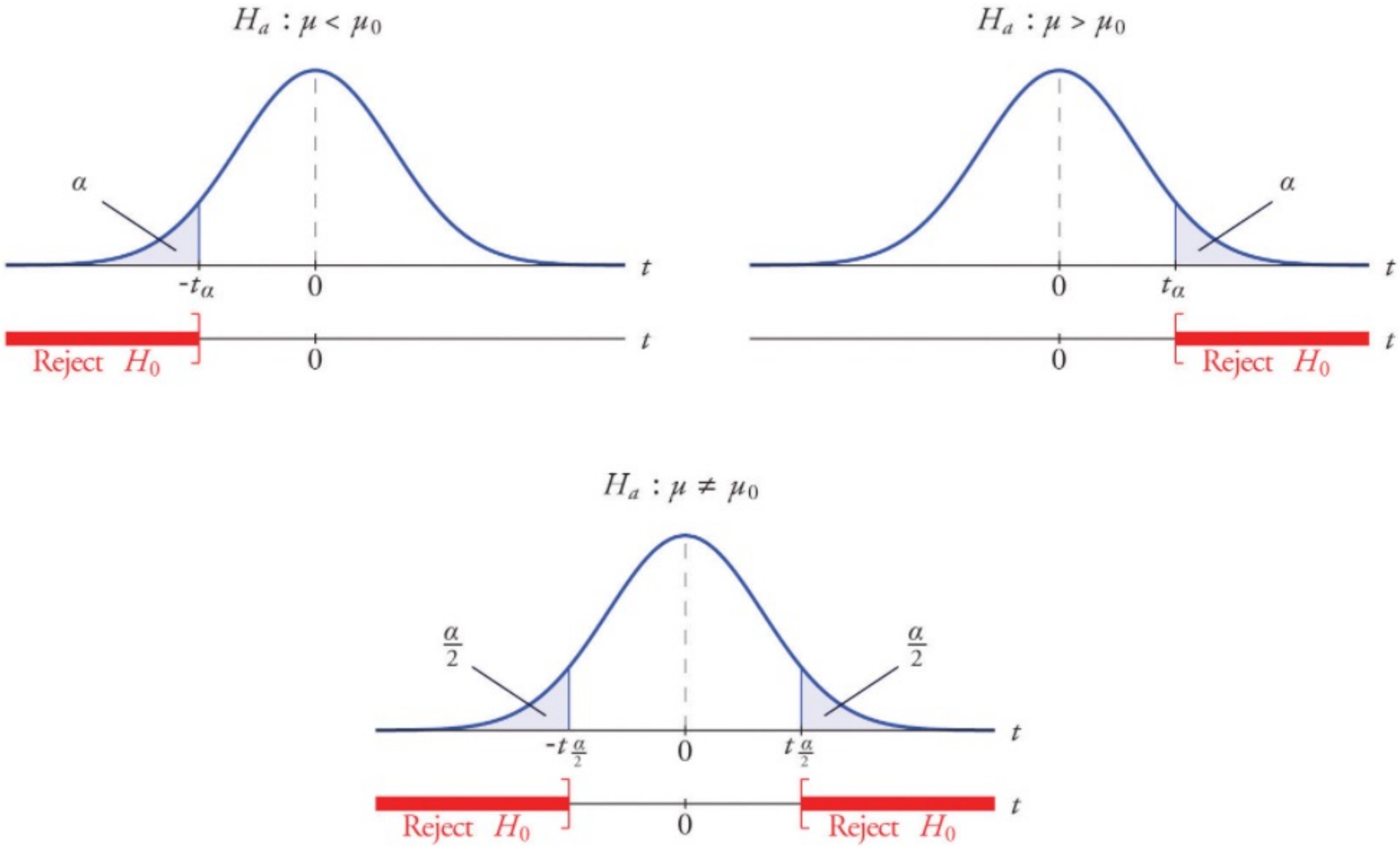
At the 5% significance level, we do not have enough evidence to claim bias.
The coin is not bias.

We have been doing two-tail confidence intervals.

There are also one-tail confidence intervals.

What is the difference?

When is a certain type called for?



There are claims that students in CSCI 3022 are smarter than the typical CU student.

A random sample of 30 entrance exam scores from the students in CSCI 3022 has a mean score of 112.5.

The CU population mean entrance exam score is 100 with a standard deviation of 15.

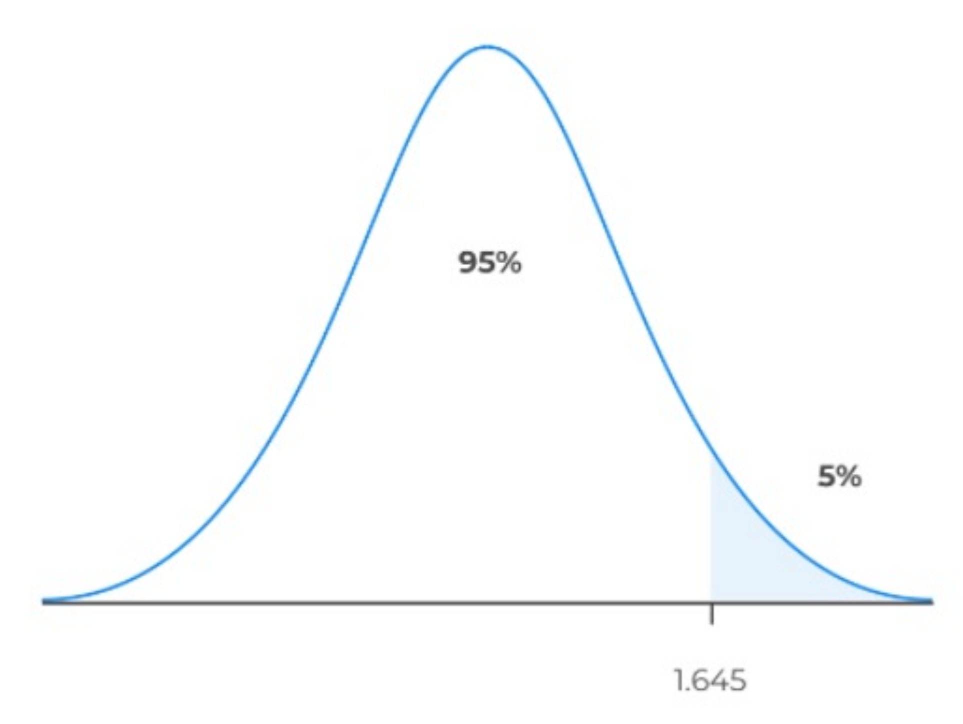
Is there sufficient evidence to support this view?

There are claims that students in CSCI 3022 are smarter than the typical CU student.

Is there sufficient evidence to support this view?

Say μ is the CSCI 3022 population mean.

H_0: $\mu = 100$
H_1: $\mu > 100$
$\alpha = 0.05$



$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} = \frac{112.5 - 100}{15/\sqrt{30}} = 4.56 > 1.645 \implies \text{reject } H_0.$$

We reject the null hypothesis (that the mean CSCI 3022 entrance exam score is 100 just like everyone else at CU) in support of the alternate hypothesis.
CSCI 3022 students are smarter than the typical CU student.

You have a business website that is currently getting about 200,000 hits per day, with a SD of 50,000 hits.

A new advertising agency comes along and says, "You should change over to us, we are awesome, we will get you more website hits, give us money,…"

You are reluctant so you try a 30-day trial with the agency.

During the 30-day trial, the new agency gets you 210,000 hits per day.

You think to yourself, "That really is no different than before", others in your company say, "210,000 is more than before. We should change."

Perform a statistical hypothesis test to determine if the new ad campaign really outperforms the old one at the 0.05 significance level.

Step one: what is your null hypothesis?

“nothing has changed, everything is the same, we have the same average as before…”

H_0: $\mu = 200$
H_1: $\mu > 200$

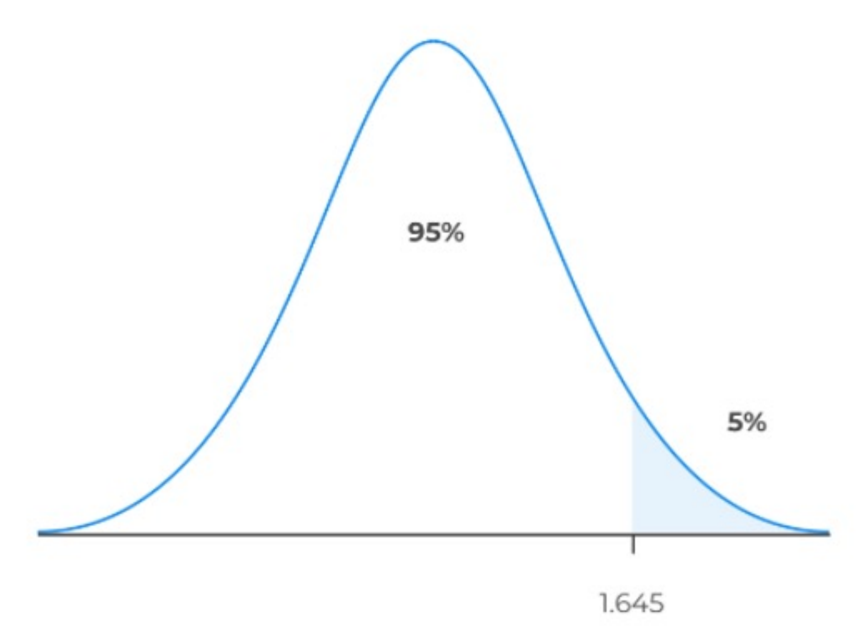
Now generate Z, assuming H_0 is true:

$$Z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{210 - 200}{50 / \sqrt{30}} = 1.095$$

$\Phi(Z_{.05}) = 0.95$ or `stats.norm.ppf(.95) = 1.645`

since $1.095 < 1.645$ we fail to reject H_0

The new advertising does not significantly change what is already going on.



Of course, we are taking samples and playing with probability, soooo….
there might be occasions when we arrive at the wrong conclusion.

Errors in hypothesis testing

	H_0 is true	H_0 is false	
Reject H_0	Type I error	OK	False Positive
Fail to reject H_0	OK	Type II error	False Negative

Q: What is the probability of making a Type I error?

A: α

Because $\alpha = p$, and $\hat{p}$ just happens by chance to land in the rejection region

Of course, we are taking samples and playing with probability, soooo…. there might be occasions when we stated something incorrectly.

What is the probability of making a Type II error?

https://www.ssc.wisc.edu/~gwallace/PA_818/Resources/Type%20II%20Error%20and%20Power%20Calculations.pdf

The probability of committing a type II error is equal to one minus the power of the test, also known as beta. The power of the test could be increased by increasing the sample size, which decreases the risk of committing a type II error.

$$1 - \beta$$

Errors in hypothesis testing

	H_0 is true	H_0 is false	
Reject H_0	Type I error	OK	False Positive
Fail to reject H_0	OK	Type II error	False Negative

So, the probability of making a Type I error is α.

Why is this relevant?

When you run a hypothesis test, how do you decide upon α ?

Choose α by considering how willing you are to risk a Type I error.

Type II errors are also important to avoid!

H_0: The airplane is safe

H_1: The airplane is less than safe

Where should we set α ?

Test on means

McDowell's hamburgers have an ice cream machine that is supposed to dispense 8 oz of ice cream when working properly.

Suppose that the average amount dispensed in a particular sample of 35 cones is 7.91 ounces with a variance of 0.03 ounces squared, s^2.

Is there evidence that the machine should be stopped, and production wait for repairs?
The lost production from a shutdown is potentially so great that management feels that the level of significance in the analysis should be 99%.

Again, we will follow the steps in our analysis of this problem.

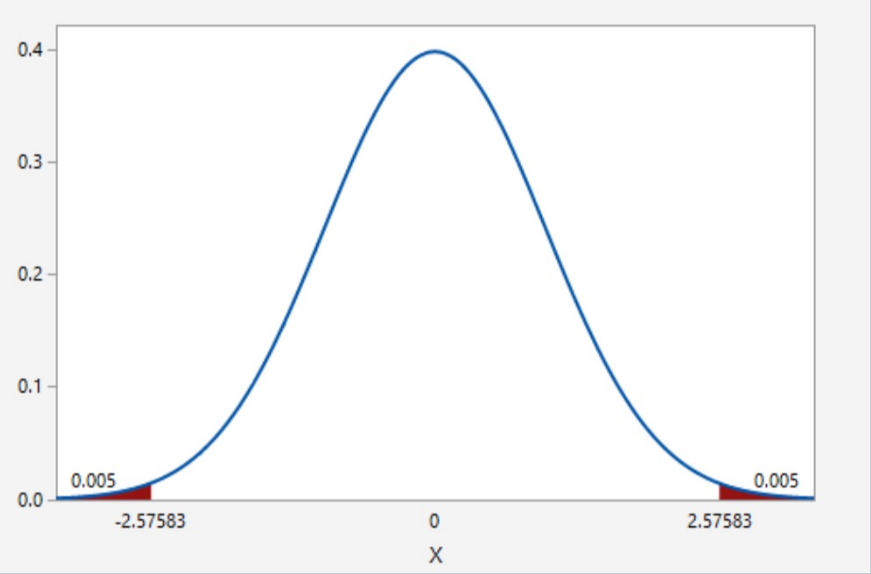
Set the Null and Alternative Hypothesis.
The random variable is the quantity of ice cream dispensed. This is a continuous random variable and the parameter we are interested in is the mean. Our hypothesis therefore is about the mean.

In this case we are concerned that the machine is not filling properly. From what we are told it does not matter if the machine is over-filling or under-filling, both seem to be an equally bad error. This tells us that this is a two-tailed test: if the machine is malfunctioning it will be shutdown regardless if it is from over-filling or under-filling.

The null and alternative hypotheses are thus:

$$H_0\colon \mu = 8 \quad \text{and} \quad H_1\colon \mu \neq 8$$

Decide the level of significance and draw the graph showing the critical value. This problem has already set the level of significance at 99%.



The decision seems an appropriate one and shows the thought process when setting the significance level. Management wants to be very certain, as certain as probability will allow, that they are not shutting down a machine that is not in need of repair.

Because this is a continuous random variable and we are interested in the mean, and the sample size is greater than 30, the appropriate distribution is the normal distribution, and the relevant critical value is 2.575.

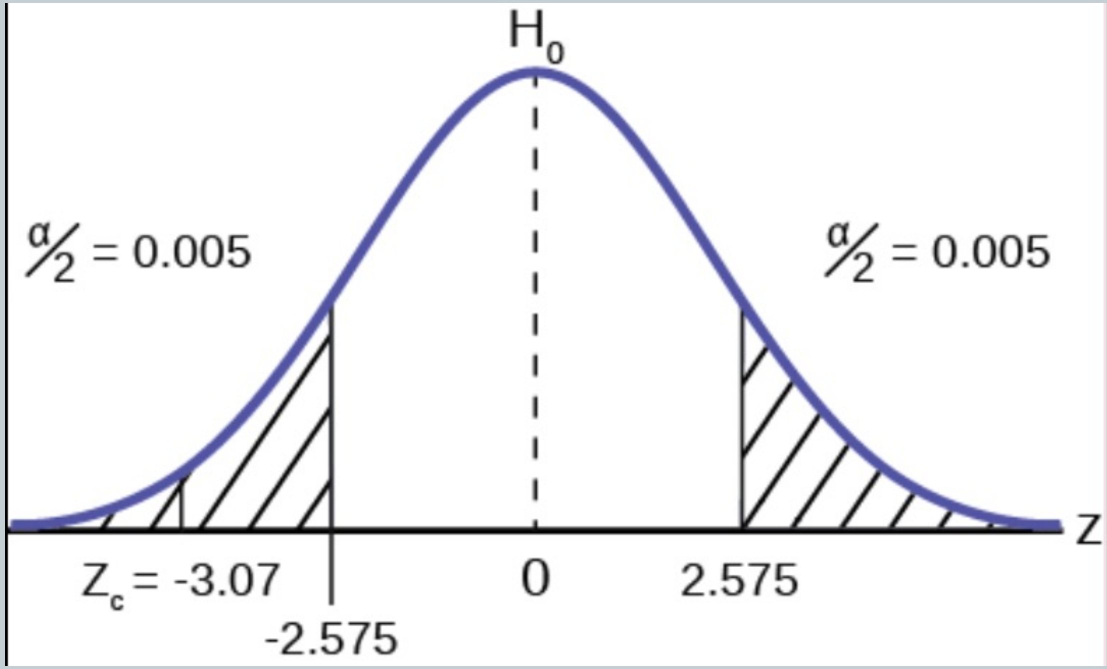
$$\Phi(0.995) = 2.575$$

Calculate sample parameters and the test statistic. The sample parameters are provided, the sample mean is 7.91 and the sample variance is .03 and the sample size is 35.

We need to note that the sample variance was provided not the sample standard deviation, which is what we need for the formula. Remembering that the standard deviation is simply the square root of the variance, we therefore know the sample standard deviation, s, is 0.173.

With this information we calculate the test statistic as -3.07 and mark it on the graph.

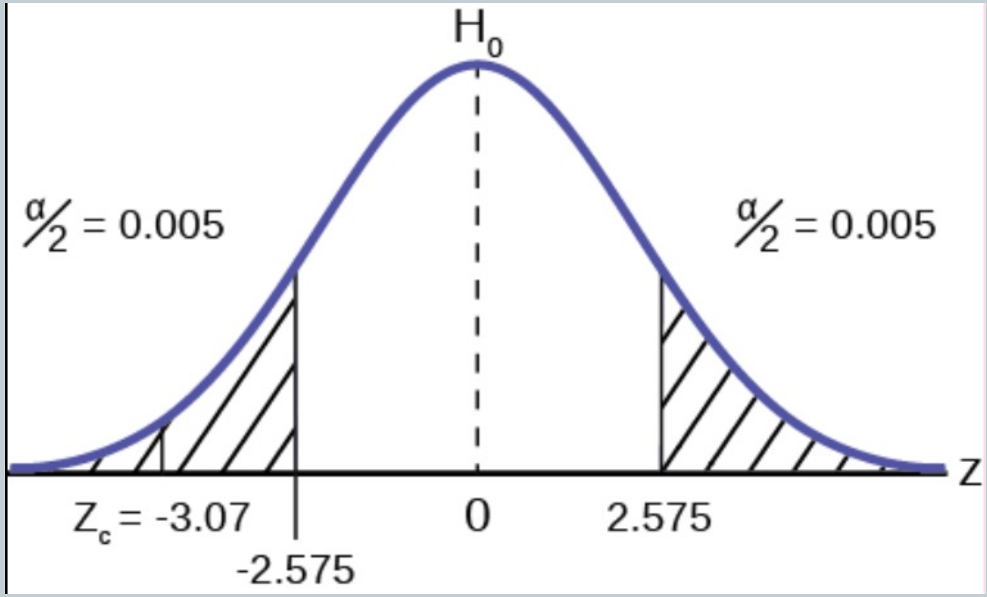
$$Z_c = \frac{\bar{X}-\mu_0}{s/\sqrt{n}} = \frac{7.91-8}{.173/\sqrt{35}} = -3.07$$



Compare test statistic and the critical values

Now we compare the test statistic and the critical value by placing the test statistic on the graph. We see that the test statistic is in the tail, decidedly greater than the critical value of 2.575.

We note that even the very small difference between the hypothesized value and the sample value is still a large number of standard deviations. The sample mean is only 0.08 ounces different from the required level of 8 ounces, but it is 3 plus standard deviations away and thus we cannot accept the null hypothesis.



Reach a Conclusion

Three standard deviations of a test statistic will guarantee that the test will fail. The probability that anything is within three standard deviations is almost zero.

Actually, it is 0.0026 on the normal distribution, which is certainly almost zero in a practical sense. Our formal conclusion would be " At a 99% level of significance we cannot accept the hypothesis that the sample mean came from a distribution with a mean of 8 ounces"

Or less formally, and getting to the point:
"At a 99% level of significance we conclude that
the machine is under filling and is in need of repair".

Hypothesis test for proportions

When you perform a hypothesis test of a population proportion *p*, you take a simple random sample from the population.

The conditions for a binomial distribution must be met, which are:

There are a certain number *n* of independent trials meaning random sampling.

The outcomes of any trial are binary, success or failure, and each trial has the same probability of a success *p*.

The shape of the binomial distribution needs to be similar to the shape of the normal distribution. To ensure this, there should be at least 5 successes and 5 failures.

Suppose a consumer group suspects that the proportion of households that have three or more cell phones is 30%.

A cell phone company has reason to believe that the proportion is not 30%.

Before they start a big advertising campaign, they conduct a hypothesis test. Their marketing people survey 150 households with the result that 43 of the households have three or more cell phones.

$H_0: p = 0.3$ $H_1: p \neq 0.3$ $n = 150$ $\hat{p} = \frac{x}{n} = \frac{43}{150} = 0.287$

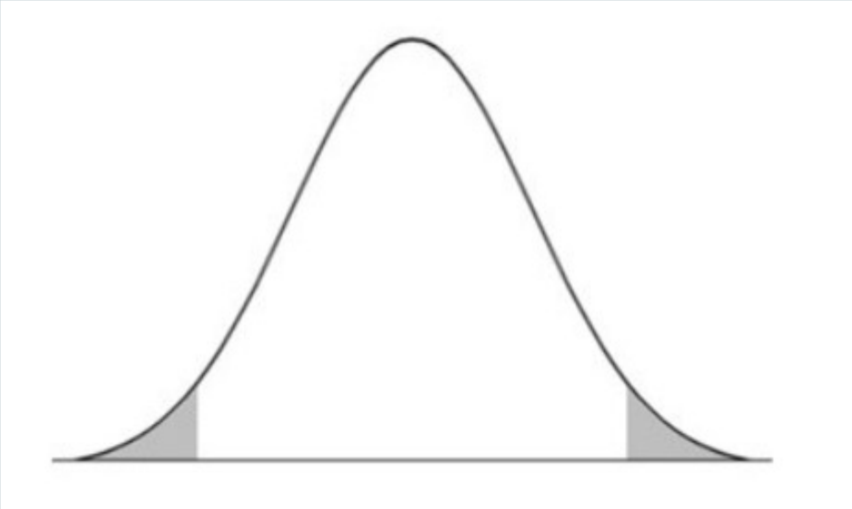
$H_0: p = 0.3$ $H_1: p \neq 0.3$ $n = 150$ $\hat{p} = \frac{x}{n} = \frac{43}{150} = 0.287$

$$Z_c = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} = \frac{0.287 - 0.3}{\sqrt{\frac{\frac{43}{150} \cdot \frac{107}{150}}{150}}} = -0.3521$$

$\Phi(0.05) = 1.645$ or `norm.ppf(.95) = 1.645`

At a 90% significance level we cannot reject the null. The consumer group is correct.

the proportion of households that
have three or more cell phones is 30%.



Nest time: p-values