

CHAPTER 18 & 23.3

p-Values

Quick review

A statistical hypothesis is a claim about the value of a parameter of a population characteristic.

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

A test statistic is calculated from the sample data, assuming that the null hypothesis is true. It is used in the decision about whether or not to reject the null hypothesis.

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

p-values

What is a p-value?

A p-value is a *probability* value.

Therefore, it is any value between 0 and 1.

When will we be using p-values?

Use them with hypothesis tests. A smaller p-value means that there is stronger evidence in favor of the alternative hypothesis.

What does a p-value represent?

”A p-value is the probability of obtaining the observed difference (or a larger one) in the outcome measure, given that no difference exists between treatments in the population”

How do we get/find/create p-values?

We use statistical hypothesis tests to discover the p-value.

$$\text{p-value} = p(Z \geq z)$$

The p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct.

The p-value is used as an alternative to rejection points to provide the smallest level of significance at which the null hypothesis would be rejected.

Example

A chemist creates a new drug to promote weight loss.

Does the drug actually work? That is, does drug X reduce weight in people?

You gather two samples of volunteers. Sample A and Sample B.

Group A gets a placebo while group B gets drug X.

Weigh each person at the beginning of the study and again after 1 month.

After 1 month the average weight loss in group A is 0 pounds.

After 1 month the average weight loss in group B is 2.5 pounds.

So, did the drug work, or is this happenstance? A 2.5-pound loss could easily happen just by coincidence... would you be convinced at 3.5 pounds or 5 pounds or 10 pounds...

You can find the probability of a 2.5-pound weight loss happening, assuming that the null hypothesis is true, i.e., a p-value!

H_0 : There is no difference between a placebo and Drug X.

Ask yourself, “What would happen in a world where the weight difference in volunteers who received drug X is the same as the weight difference in those who received the placebo?”

H_0 : The weight difference in those who receive drug X is the same as the weight difference in those who receive the placebo. i.e., the status quo.

If the null were true:

Then what is the probability of discovering a 2.5-pound reduction or more in body weight in those treated with drug X from our sample (group B) compared with the placebo (Group A) ?

We use statistical hypothesis tests to discover the p-value.

Get data – run test – look at p-value.

Suppose we follow the steps above and get a p-value of 0.02 or 2%.

What does this mean?

If the null hypothesis were true (i.e. the two population means were identical), then there is only a 2% chance of observing a difference as large, or larger, than what we observed in our sample.

So, in a world where the weight difference in those who receive drug X is the same as the weight difference in those who receive the placebo, then there is only a 2% chance of observing a weight loss of 2.5 pounds, or more, between our sample groups.

With only a 2% chance, should we believe this 'world' is real or perhaps not?

What is accounting for this 2%?

Random noise or random chance affects the p-value.

Things like variation, genetics (in humans), sampling method,...

But not necessarily drug X.

This probability, or p-value, measures the strength of evidence against the null hypothesis.

By analogy, in a court of law the defendant is the null hypothesis, that is to say they are innocent until proven guilty. Likewise, we do not reject the null unless there is sufficient evidence to do so.

What kind of evidence?

The smaller the p-value, the stronger the evidence against the null hypothesis.

Suppose you read an article that claims 26% of U.S. college students are bilingual. You wonder if CU students have a higher percentage (one-tail test) than 26%.

So, you decide to figure out an answer.

First step is to **design** your study, then: **Get data** – **run test** – **look at p-value**.

In your **design**, you must state the hypotheses and decide your significance level.

Let p represent the proportion of CU students that are bilingual.

$H_0: p = 0.26$ That is to say, 26% of CU students are bilingual.

$H_1: p > 0.26$

Significance level: $\alpha = .05$ or 5%. This is your (prior) decision; part of your design.

Get data

You then sample 120 students from CU and discover that 40 of them are bilingual.

You analyze these numbers and find that the test statistic for these results produces a $Z \approx 1.83$. (BTW where did 1.83 come from...)

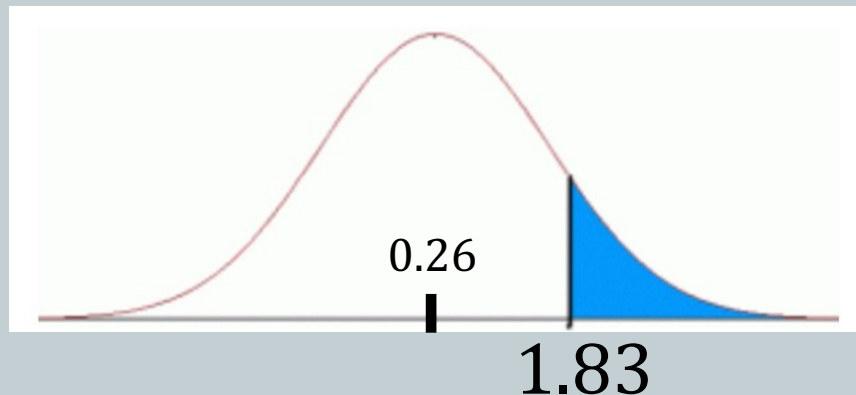
How did we get the $Z = 1.83$?

Scratch-paper work:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{.33 - .26}{\sqrt{\frac{.26(1-.26)}{120}}} \approx 1.83$$

The 1.83 answers the question:

“How many SD’s above the assumed $p=0.26$ is $\hat{p} = 0.33 = \frac{40}{120}$?”



The assumed distribution of sample proportions from the CU student body.

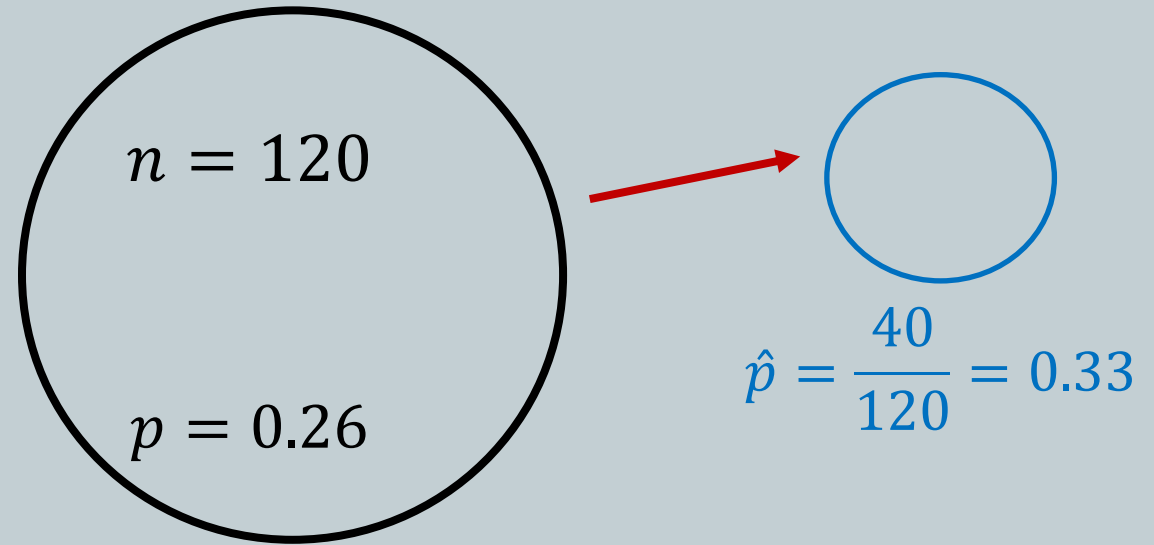
And yet our sample was $\hat{p} = .33$
How abnormal is that?

Assuming the **necessary conditions** are met, what is the p-value for this test?

necessary conditions

- 1] Normal
- 2] Random
- 3] Independent

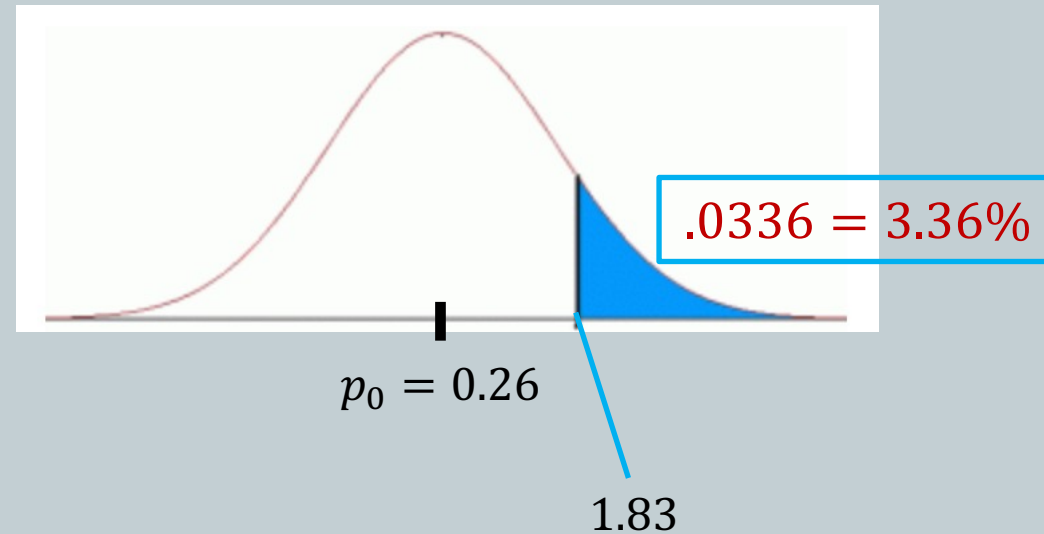
These allow us to assume the sampling distribution of the sample proportions is roughly normal.



$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} = \frac{.33 - .26}{\sqrt{\frac{.26(1 - .26)}{120}}} \approx 1.83$$

$$\text{p-value} = p(Z \geq 1.83) = 1 - p(Z \leq 1.83) = 1 - 0.9664 = 0.0336$$

$$\text{p-value} = p(Z \geq 1.83) = 0.0336$$



So, at the 5% significance level we **reject** H_0 in favor of $H_1: p > 0.26$.

$H_0: p = 0.26$ This hypothesis states that 26% of CU students are bilingual.

Notice that at the 1% significance level we **fail to reject** H_0

However, it is unethical to decide the significance level **AFTER** completing the experiment. Similar to p-hacking.

Notes to keep in mind:

Remember to set-up the null hypothesis as the status quo.

We then either accept H_1 or default to H_0 (reject H_0 or fail to reject H_0)

Assume H_0 is true and see if we get result at least as extreme

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \approx \frac{s}{\sqrt{n}} = \hat{\sigma}_{\bar{X}}$$

How far away (how many standard deviations away) from the mean is the sample statistic ? What is the probability of getting a result that many SD's away from the mean ?

A Z-score is defacto the distance (measured in SD's) from the mean.

Divide by sampling distributions SD (that is, our estimate of it)

- Suppose you want your buffalos to gain more weight for health reasons.
- Current mean weight gain is 20 pounds per 10-days and you want more gain.
- So, you switched from food A to food B.
- You want to know if this has had any affect



1] State your hypotheses and significance level

H_0 : $\mu = 20$ after the food change. The new food has no affect.

H_1 : $\mu > 20$ after the change in food. Significant weight gain occurs.

Where "significant" is defined by you to be $\alpha = .05$, or 5%.

2] Gather stats after the introduction of the new food product. This will get you \bar{x} and s .

Let's say $\bar{x} = 25$.

3] Calculate the p-value. that is, calculate $P(\bar{x} \geq 25 | H_0 \text{ is true})$

4] p-value $< \alpha$ implies we reject H_0 as we have meaningful evidence for H_1 .

p-value $\geq \alpha$ implies we do not reject H_0 .






Assuming the null is true, if probability is lower than α then we reject H_0
(i.e., we have evidence for H_1)

So, if p-value = 0.03, then we reject the null because $0.03 < 0.05$.

If probability is at, or higher, than α , then do not reject H_0 .

Alcohol is known to slow a person's reaction time. However, a new drug has been developed to counter the effects of alcohol.

100 randomly chosen volunteers were injected with the new drug and tested for reaction time on an electronic board. Their mean response time was 1.05 seconds with a standard deviation of 0.5 seconds.

BLOOD ALCOHOL CONCENTRATION	NUMBER OF DRINKS	EFFECTS ON DRIVING
0.02% BAC		<ul style="list-style-type: none">• Decline in visual functions• Inability to perform two tasks at the same time• Loss of judgment• Altered mood
0.05% BAC		<ul style="list-style-type: none">• Reduced coordination• Reduced ability to track moving objects• Difficulty steering• Slower response to emergency driving situations
0.08% BAC		<ul style="list-style-type: none">• Reduced ability to concentrate• Short-term memory loss• Lack of speed control• Impaired perception and self-control
0.10% BAC		<ul style="list-style-type: none">• Clear deterioration of reaction time• Reduced ability to maintain lane position• Reduced ability to brake appropriately• Slurred speech
0.15% BAC		<ul style="list-style-type: none">• Substantial impairment in vehicle control• Loss of auditory information processing• Major loss of balance• Vomiting may occur

Source: Centers for Disease Control and Prevention

The mean response time for 100 randomly chosen people without the drug was 1.2 seconds.

Do you think the drug has an effect on reaction time? How sure are you?

Set up the null hypothesis.

This is the status quo, the drug has no affect, nothing unusual is going on, the mean response time of those people taking the drug is also 1.2 seconds just like the population of people not taking the drug...

So, if μ is the mean response time of drug taking people, we have

$H_0: \mu = 1.2$ seconds, even though these people are taking the drug. i.e., the drug has no effect.

The alternative hypothesis states, “No, something IS going on, the drug does have an effect.”

$H_1: \mu \neq 1.2$ seconds when the drug is given (two-tail; either way).

Do we accept the alternative hypothesis, or do we default to the null hypothesis because the data just isn't convincing?

We start by assuming the null hypothesis is true,

So, let's assume the null hypothesis is true.

If the null hypothesis was true, then what is the probability that we would have gotten these results (1.05 seconds) with the sample?

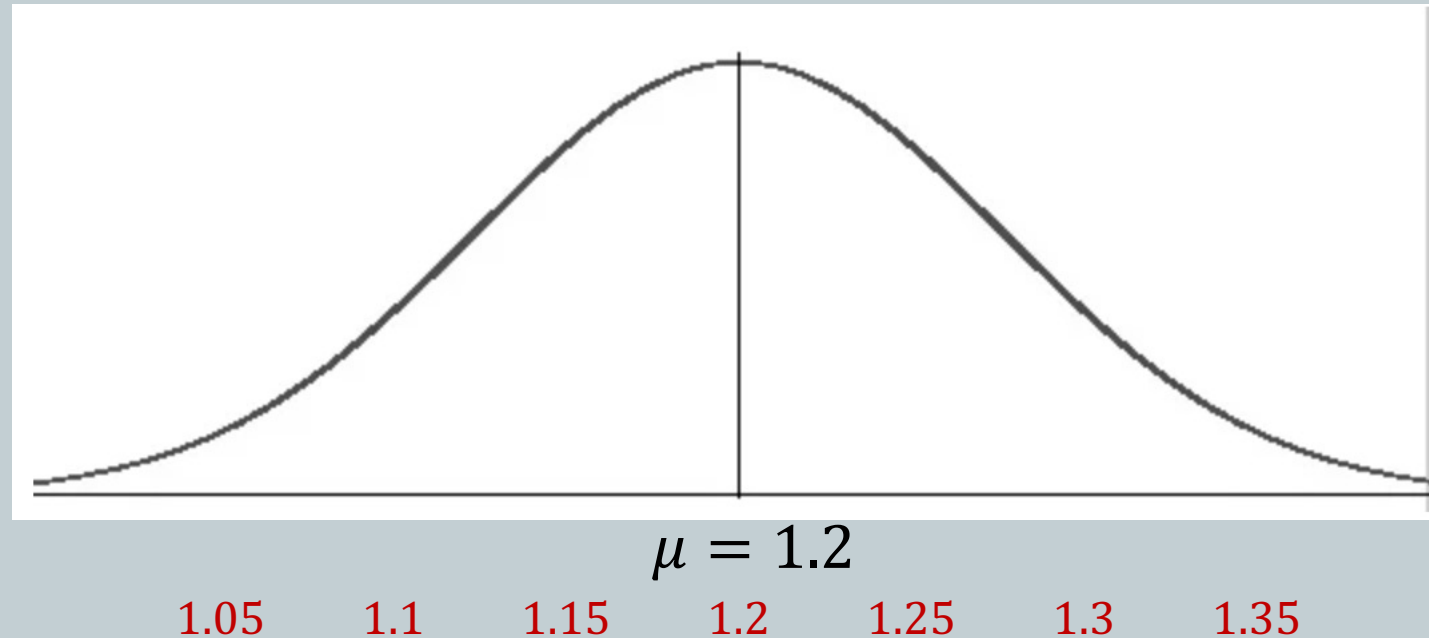
And, if that probability is really small, then the null hypothesis probably isn't true. In which case you would reject the null in favor of the alternate hypothesis.

Let's assume the null is true and figure out the probability of getting this result (a sample mean of 1.05 seconds with a standard deviation of .5 seconds) or a result more extreme.

Consider the sampling distribution if we assume the null hypothesis.

It would be normal with a mean of 1.2 seconds since we are assuming the null.

So, $\mu_{\bar{X}} = \mu = 1.2$.



What is the standard deviation of our sampling distribution?

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{\sigma}{\sqrt{100}} \approx \frac{s}{\sqrt{100}} = \frac{0.5}{10} = \frac{1}{20} = 0.05$$

Since we approximated, we can write $\hat{\sigma}_{\bar{X}} = 0.05$.

Now, what is the probability of getting 1.05 as we did in our sample?
AKA, how many standard deviations away from 1.2 is our value of 1.05?

More importantly, what is the probability of getting that many standard deviations away from the mean?

This is answered with a Z -score and we divide by our best estimate of the sampling distributions standard deviation: $\hat{\sigma}_{\bar{X}} = 0.05$.

$$Z = \frac{1.2 - 1.05}{0.05} = 3 \quad \text{which implies 1.05 is 3 standard deviations away from the mean.}$$

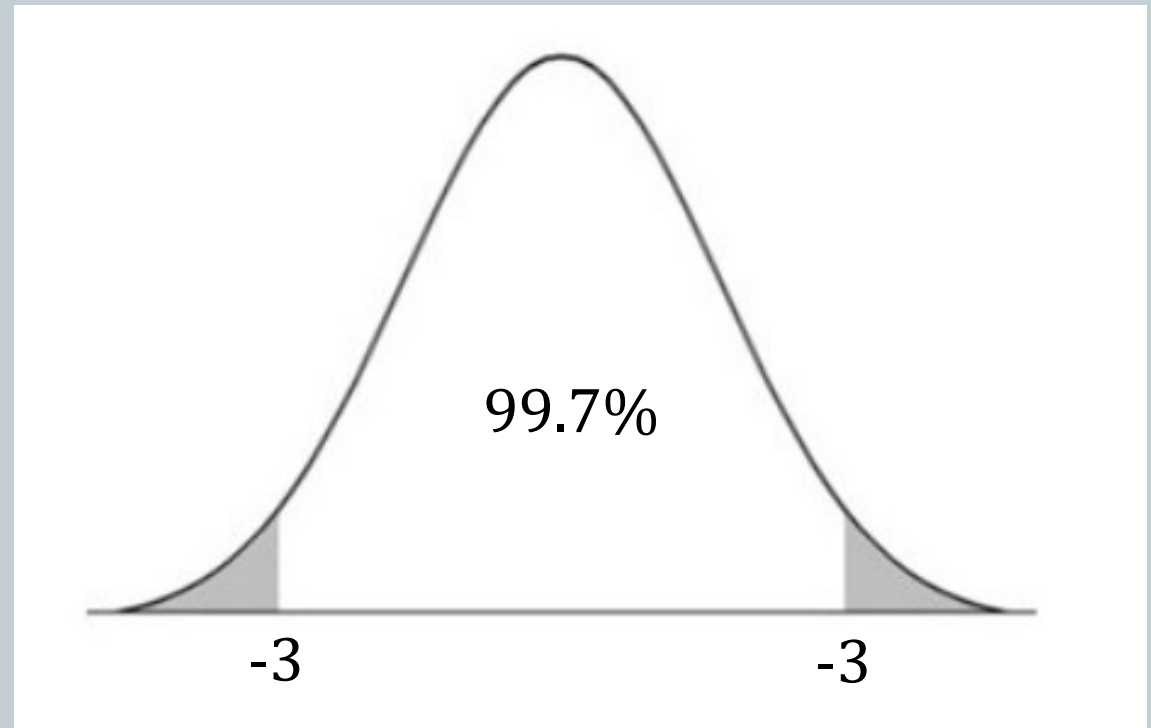
`norm.ppf(.0015) ≈ 3`

or 68-95-99.7 rule

The tails contain just $0.3\% = .003$

If we assume the drug has no effect, then the probability of getting a sample this extreme, or more extreme, is only 0.3%.

This result seems to favor the alternative hypothesis.



Therefore, we **reject** the null, that is we reject the following idea:

$H_0: \mu = 1.2$ seconds, even though these people are taking the drug. i.e., the drug has no effect.

In favor of the alternate hypothesis, $H_1: \mu \neq 1.2$.

The probability of getting a result more extreme than this, given the null is true, is called a **p-value**.

Our p-value, our probability value, is 0.003.

There is a very small probability that we could have gotten this result, **IF** the null hypothesis were true. So, we will reject it.

We often create a threshold, **prior** to carrying out the experiment, beyond which we will reject the null. Typically, 5%.

i.e., if there is less than a 1 in 20 chance, then I am going to reject the null.

Conclusion: We think the drug has some effect.

Raisin Bran advertises that they put “Two scoops of raisins in every box.”
Suppose a ‘scoop’ of raisins weighs 35 grams.

Can you sue Kellogg's for false advertising if you sample 40 boxes of Raisin Bran and find a mean raisin weight of 68.7 grams with a standard deviation of 6.35 grams?



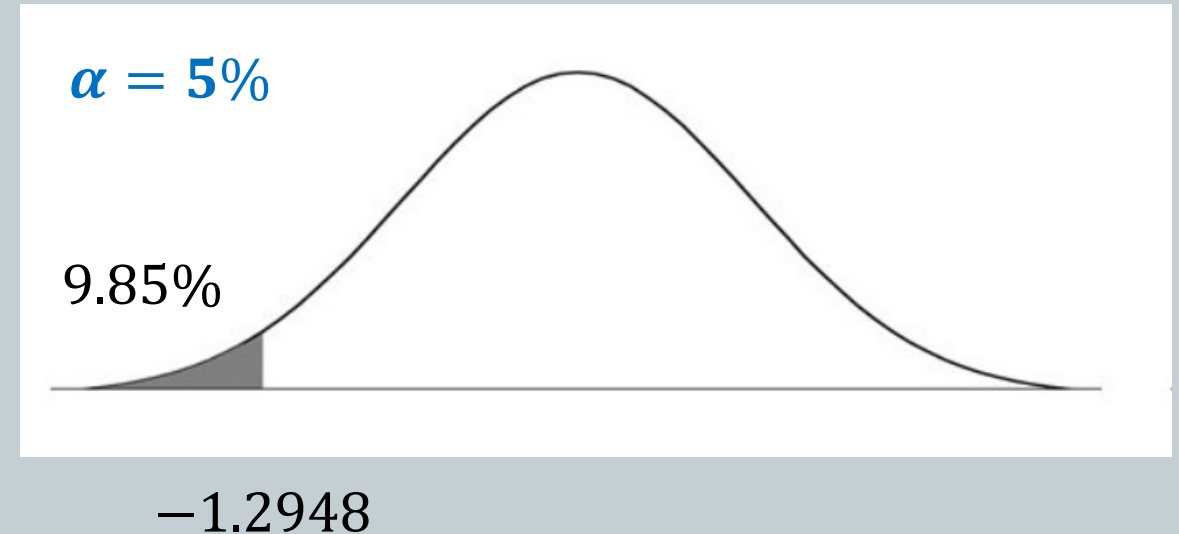
Not all of the 40 boxes sampled were less than 70. Too easy.
Not all of the 40 boxes sampled were more than 70. Too easy.
Suppose they ranged in values at least from 55 to 85.

$$H_0: \mu = 70 \quad H_1: \mu < 70 \quad \alpha = 0.05$$

$$Z = \frac{68.7 - 70}{6.35 / \sqrt{40}} = -1.2948$$

$$\Phi(-1.2948) = 0.0985$$

$$\text{p-value} = p(Z \leq -1.2948) = 0.0985$$



This is a p-value of .0985, and $.0985 > .05$, therefore we cannot reject the null.

That is to say, a 1 in 10 chance is not evidence enough to go after Kelloggs.

Bare bones

A statistician named Jamie wants to test the hypothesis:

$H_0: \mu = 120$ against the alternative hypothesis

$H_1: \mu < 120$

assuming $\alpha = 0.05$.

For this study, Jamie took a sample and gathered values:

$n = 40$, $\sigma = 32.17$ and $\bar{x} = 105.37$.

What can be concluded for this hypothesis?

Solution:

We know that:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{32.17}{\sqrt{40}} = 5.0865$$

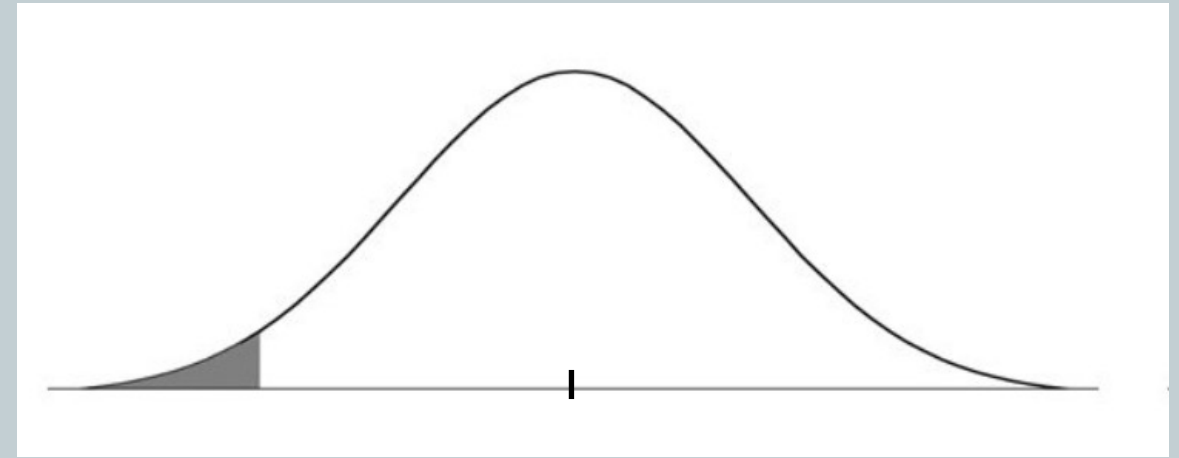
Now, using the test static formula, we get:

$$Z = \frac{105.37 - 120}{5.0865} = -2.8762$$

$$P(Z < -2.8762) = 0.0021$$

$$\text{p-value} = 0.0021 < 0.05$$

Therefore, we reject the null.



$$\mu = 120$$

Alpha levels are controlled by the researcher and are related to confidence levels. You get an alpha level by subtracting your confidence level from 100%.

For example, if you want to be 98 percent confident in your research, the alpha level would be 2% ($100\% - 98\%$).

When you run the hypothesis test, the test will give you a value for p . Compare that value to your chosen alpha level.

For example, let's say you chose an alpha level of 5% (0.05).

If the results from the test give you:

- A small p (≤ 0.05), reject the null hypothesis.

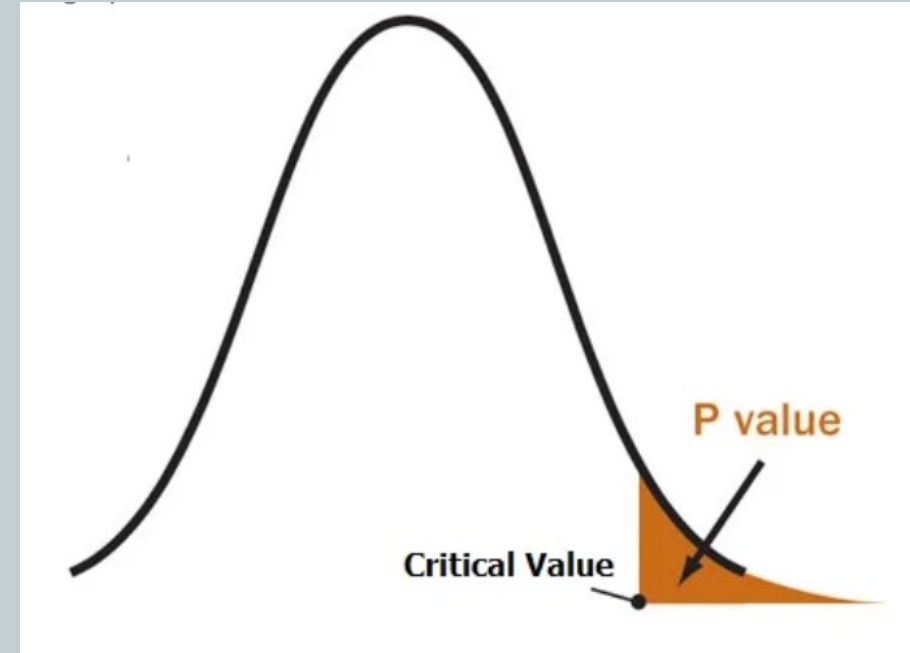
- This is strong evidence that the null hypothesis is invalid.

- A large p (> 0.05) means the alternate hypothesis is weak.

- So, you do not reject the null.

In general:

$p \text{ value} > .10$ “Not significant”
 $p \text{ value} \leq .10$ “Marginally significant”
 $p \text{ value} \leq .05$ “Significant”
 $p \text{ value} \leq .01$ “Highly Significant”



Concluding thoughts:

The p -value is a number (between 0 and 1), calculated from a statistical test, that describes how likely you are to have found a particular set of observations if the null hypothesis were true.

p -values are used in hypothesis testing to help decide whether to reject the null hypothesis.

The formula for the calculation for p -value is

- 1] Find the test statistic and the Z-score that it creates
- 2] Find the corresponding level of p from the Z-value obtained
- 3] Make a conclusion based on p and α .

The smaller the p -value, the more likely you are to reject the null hypothesis.

The level of significance, α is a predefined threshold that should be set by the researcher. It is generally fixed as 0.05.

Examples:

We are testing the hypothesis that the average milk consumption per month in Eckley, Colorado is greater than 7 gallons per household per week; and we want 95% confidence.

Sample 30 homes. The average is 8.4, and the sample standard deviation is 4.29.

$$H_0: \mu \leq 7$$

- 1] What is the Z -value for a 1-tailed test at 95%?
- 2] What is the Z -value for our sample mean of 8.4?
- 3] What is the p -value for our sample mean of 8.4?
- 4] Do we reject the null hypothesis?

1] What is the Z-value for a 1-tailed test at 95%?

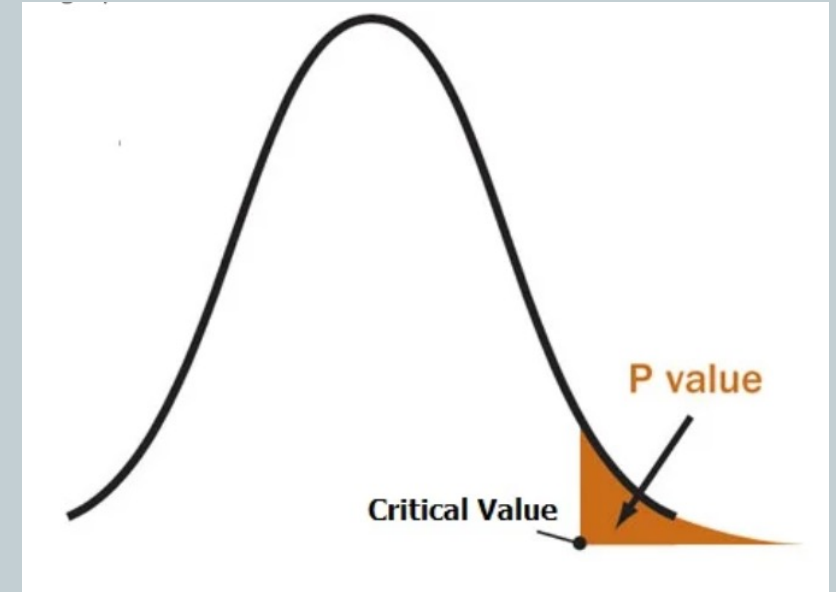
$$P(Z > 1.645) = .95 \quad Z\text{-value} = 1.645$$

2] What is the Z-value for our sample mean of 8.4?

$$Z_{8.4} = \frac{8.4 - 7}{0.78} = 1.8 \quad \left(4.29 / \sqrt{30} = 0.78 \right)$$

3] What is the p-value for our sample mean of 8.4?

$$P(Z > 1.8) = 1 - \Phi(1.8) = 0.0359$$



4] Do we reject the null hypothesis?

We **reject the null** hypothesis since $1.8 > 1.645$ and $.0359 = 3.59\% < 5\%$

We are 95% confident that we will not get a sample mean of 8.4 when the true population mean is 7.

The p-value process:

State your null hypothesis.

The null hypothesis is a commonly accepted fact. It's the default, or what we'd believe if the experiment was never conducted. It's the least exciting result, showing no significant difference between two or more groups.

Researchers work to nullify or disprove null hypotheses.

State an alternative hypothesis.

You'll want to prove an alternative hypothesis. This is the opposite of the null hypothesis, demonstrating or supporting a statistically significant result. By rejecting the null hypothesis, you accept the alternative hypothesis.

Determine a significance level.

This is the determiner, also known as the alpha (α). It defines the probability that the null hypothesis will be rejected. A typical significance level is set at 0.05 (or 5%). You may also see 0.1 or 0.01, depending on the area of study.

If you set the alpha at 0.05, then there is a 5% chance you'll find support for the alternative hypothesis (thus rejecting the null hypothesis) when, in truth, the null hypothesis is actually true, and you were wrong to reject it.

In other words, the significance level is a statistical way of demonstrating how confident you are in your conclusion. If you set a high alpha (0.25), then you'll have a better shot at supporting your alternative hypothesis, since you don't need to find as big a difference between your test groups. However, you'll also have a bigger chance at being wrong about your conclusion.

Calculate the p-value.

The p-value, or calculated probability, indicates the probability of achieving the results of the null hypothesis. While the alpha is the significance level you're trying to achieve, the p-level is what your actual data is showing when you calculate it. A low p-value offers stronger support for your alternative hypothesis.

Draw a conclusion.

If your p-value meets your significance level requirements, then your alternative hypothesis may be valid and you may reject the null hypothesis. In other words, if your p-value is less than your significance level (e.g., if your calculated p-value is 0.02 and your significance level is 0.05), then you can reject the null hypothesis and accept your alternative hypothesis.

Suppose someone wants to show that acupuncture will cure anxiety

H_0 : Acupuncture has no effect on anxiety

H_1 : Acupuncture alleviates anxiety

Significance level: $\alpha = 0.25$ (this level allows for a better shot at proving H_1)

p-value: We'll say the p-value is calculated as 0.05.

Conclusion: After providing one group with acupuncture and the other with a placebo, the difference between the two groups is statistically significant with a p-value of 0.05, well below the alpha of 0.25.

Therefore, this study supports the alternative hypothesis that acupuncture alleviates anxiety.

Does vitamin C prevent the common cold?

H_0 : Children who take vitamin C are no less likely to become ill during flu season.

H_1 : Children who take vitamin C are less likely to become ill during flu season.

significance level: $\alpha = 0.05$

p-value: The p-value is calculated to be 0.20.

conclusion: After providing one group with vitamin C during flu season and the other with a placebo, it is recorded whether or not participants got sick. Analysis of the results reveals a p-value of 0.2. That is above the desired significance level of 0.05, we therefore fail to reject the null hypothesis.

Based on this experiment there is no support for the alternative hypothesis that vitamin C can prevent colds.

Next time: Bootstrapping