CSCI 3022 Intro to Data Science
Hypotheses and P-Values

Example:
Company A produces circuit boards, but 10% of them are defective. Company B claims that they produce fewer defective circuit boards. What are the null and alternative hypotheses?

Our data is a random sample of $n = 200$ boards from company B. What test procedure (or rule) could we devise to decide if the null hypothesis should be rejected?

Opening sol:

Example:
Company A produces circuit boards, but 10% of them are defective. Company B claims that they produce fewer defective circuit boards. What are the null and alternative hypotheses?

Our data is a random sample of $n = 200$ boards from company B. What test procedure (or rule) could we devise to decide if the null hypothesis should be rejected?

Opening sol:

Example:
Company A produces circuit boards, but 10% of them are defective. Company B claims that they produce fewer defective circuit boards. What are the null and alternative hypotheses?
Solution: Null: company B and A are equal. Alternative: company B produces fewer defective boards.
Our data is a random sample of $n = 200$ boards from company B. What test procedure (or rule) could we devise to decide if the null hypothesis should be rejected?
Solution: Reject if we observe a sample from the company B with a **very low** proportion of defects.

Announcements and Reminders

- Practicum due tonight! I'll do an office hour until 6pm.

Now what?

Using all the same math of the CLT and confidence intervals, we transition into statistical *decision* making. When do we try to actually act on or decide whether two populations are different?

With a confidence interval this is as simple as asking "is the numbers a inside my 95% confidence interval for μ?" If not, we'd decide that $\mu \neq a$.

Logic of Hypothesis Testing

More generally, we perform tests under the philosophy that we won't make interesting decisions *unless compelling evidence exists* to do so. We would *assume* that two groups are equal unless the data convincing demonstrates they are different.

We call the assumption that nothing interesting is going on the *null hypothesis*.

Analogy: Jury in a criminal trial.
The two possible conclusions:

Logic of Hypothesis Testing

More generally, we perform tests under the philosophy that we won't make interesting decisions *unless compelling evidence exists* to do so. We would *assume* that two groups are equal unless the data convincing demonstrates they are different.

We call the assumption that nothing interesting is going on the *null hypothesis*.

Analogy: Jury in a criminal trial.
The two possible conclusions:
Fail to Reject the null hypothesis if there is insufficient statistical evidence to do so.

Logic of Hypothesis Testing

More generally, we perform tests under the philosophy that we won't make interesting decisions *unless compelling evidence exists* to do so. We would *assume* that two groups are equal unless the data convincing demonstrates they are different.

We call the assumption that nothing interesting is going on the *null hypothesis*.

Analogy: Jury in a criminal trial.
The two possible conclusions:
Fail to Reject the null hypothesis if there is insufficient statistical evidence to do so.
Reject the null hypothesis in favor of the alternative if there is statistically *significant* cause to do so.

Logic of Hypothesis Testing

Notation and general process:

Logic of Hypothesis Testing

Notation and general process:

1. Assume the null hypothesis to be true, and state it: we propose that the parameter of interest θ satisfies $H_0 : \theta = \theta_0$.

Logic of Hypothesis Testing

Notation and general process:

1. Assume the null hypothesis to be true, and state it: we propose that the parameter of interest θ satisfies $H_0 : \theta = \theta_0$.

2. State the alternative to be tested: H_a :
   $\theta > \theta_0$ **OR** $\theta < \theta_0$ **OR** $\theta \neq \theta_0$

3. Draw a decision based on how improbable or probable the actual data looks if the null hypothesis is true. If the observed data is very unlikely, it might be because our hypothesis was wrong!

Why *assume* the null hypothesis?

Logic of Hypothesis Testing

Notation and general process:

1. Assume the null hypothesis to be true, and state it: we propose that the parameter of interest θ satisfies $H_0 : \theta = \theta_0$.

Why *assume* the null hypothesis?

1. Burden of proof
2. We know how to calculate probabilities when we *know* θ!

Logic of Hypothesis Testing

The alternative to the null hypothesis $H_0 : \theta = \theta_0$ will look like one of the following three assertions:

The equality sign is **always** with the null hypothesis.
The alternate hypothesis is the claim for which we are seeking statistical evidence.

Logic of Hypothesis Testing

The alternative to the null hypothesis $H_0 : \theta = \theta_0$ will look like one of the following three assertions:

1. $H_a : \quad \theta \neq \theta_0$

The equality sign is **always** with the null hypothesis.
The alternate hypothesis is the claim for which we are seeking statistical evidence.

Logic of Hypothesis Testing

The alternative to the null hypothesis $H_0 : \theta = \theta_0$ will look like one of the following three assertions:

1. $H_a : \quad \theta \neq \theta_0$
2. $H_a : \quad \theta > \theta_0$
3. $H_a : \quad \theta < \theta_0$

The equality sign is **always** with the null hypothesis.
The alternate hypothesis is the claim for which we are seeking statistical evidence.

Bring back α!

Definition: The **Significance level** α of a hypothesis test is the largest *probability* of a test statistic under the null hypothesis that would lead you to reject the null hypothesis.

Equivalently, it's the probability of the entire rejection region!

We thought of α last week during CIs as a term that widened or shrank as our tolerance for error grew, now it's very literally an *error rate*. Specifically, it's the probability of rejecting the null hypothesis when we were not supposed to do so.

Where we at?

A summary of our process:

1. State hypothesis: H_0 : fair coin, or $p = .5$.
   H_a : unfair coin, or $p \neq .5$
2. Get to flippin', collect some data
3. Compute something from our data. Maybe a sample proportion of heads $\hat{p}$?
4. Decide whether $\hat{p}$ is **too far** from $p = .5$, and make a decision accordingly.
5. α is the value that describes the probability of rejecting a null hypotheses *given* that the hypothesis was true.

So when do we reject...?

Suppose we flip our Polish Euro 10 times. How many heads does it take for us to conclude that the coin us unfair?

So when do we reject...?

Suppose we flip our Polish Euro 10 times. How many heads does it take for us to conclude that the coin us unfair?

- What would 10 heads mean?

So when do we reject...?

Suppose we flip our Polish Euro 10 times. How many heads does it take for us to conclude that the coin us unfair?

- What would 10 tails mean?

So when do we reject...?

Suppose we flip our Polish Euro 10 times. How many heads does it take for us to conclude that the coin us unfair?

- What would 6 heads mean?

So when do we reject...?

Suppose we flip our Polish Euro 10 times. How many heads does it take for us to conclude that the coin us unfair?

- Is there a difference between 60% heads if we flip 10 times and 60% heads if we flip 1000 times?

What is extreme: let's compute these!

The probability of a fixed number of H in a fixed number of trials are values of the binomial pmf.

- $P(10\, heads | fair\, coin) = \frac{1}{2^{10}}$.
- $P(10\, tails | fair\, coin) = \frac{1}{2^{10}}$.
- $P(10\, or\, more\, of\, the\, same\, type\, of\, flip | fair\, coin) = \frac{2}{2^{10}}$
- $P(9\, or\, more\, heads | fair\, coin) = \frac{1}{2^{10}} + \frac{10}{2^{10}}$
- $P(9\, or\, more\, tails | fair\, coin) = \frac{1}{2^{10}} + \frac{10}{2^{10}}$
- $P(9\, or\, more\, of\, the\, same\, type\, of\, flip | fair\, coin) = \frac{22}{2^{10}}$

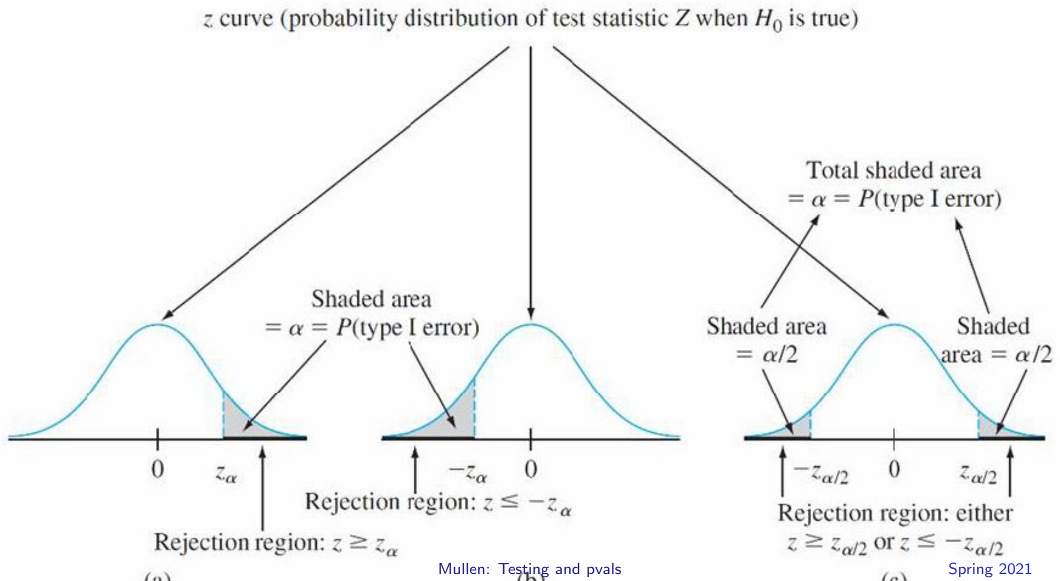
At some point, we decide that the probabilities are so unlikely (*given* the coin is fair), that if we *do* observe those outcomes, we conclude that the coin probably wasn't fair!

Rejection Regions

How would we know when the test statistic is “sufficiently rare” under the null hypothesis such that we might regard the null as false?

We could define a rejection region: a range of values that leads a researcher to reject the null hypothesis.

Rejection Regions


z curve (probability distribution of test statistic Z when H_0 is true)
Shaded area
$= \alpha = P$(type I error)
0
z_α
Rejection region: $z \geq z_\alpha$
(a)
$-z_\alpha$
0
Rejection region: $z \leq -z_\alpha$
(b)
Total shaded area
$= \alpha = P$(type I error)
Shaded area
$= \alpha/2$
Shaded
area $= \alpha/2$
$-z_{\alpha/2}$
0
$z_{\alpha/2}$
Rejection region: either
$z \geq z_{\alpha/2}$ or $z \leq -z_{\alpha/2}$
(c)

Test for Population Proportion

Example To determine if the Belgian 1 Euro coin is fair, you flip it 100 times and record the number of heads. What is the test statistic? What are the null and alternative hypotheses?

What is the distribution of the number of heads?

Test for Population Proportion

Example To determine if the Belgian 1 Euro coin is fair, you flip it 100 times and record the number of heads. What is the test statistic? What are the null and alternative hypotheses?

Test estimator: based on sample *proportion* of heads, $\hat{p}$
Test: $H_0 : p = .5$; fair coin
Against: $H_a : p \neq .5$; an unfair coin
What is the distribution of the number of heads?
$binomial(n, p)$.
We know everything about that thing! In particular, we know that

$$\hat{p} \overset{approx}{\sim} N\left(.5, \frac{.5(1-.5)}{100}\right)$$

Test for Population Proportion

Example To determine if the Belgian 1 Euro coin is fair, you flip it 100 times and record the number of heads. What is the test statistic? What are the null and alternative hypotheses?

What is the distribution of the number of heads?
$binomial(n, p)$.
We know everything about that thing! In particular, we know that

$$\hat{p} \overset{approx}{\sim} N\left(.5, \frac{.5(1-.5)}{100}\right)$$

We can use this to compute the probability given by $P(\hat{p} > z_{\alpha/2})$

Test for Population Proportion

Example (continued):
Null hypothesis:

Test statistic value:

Alternative Hypothesis   Rejection Region for α level test:

Test for Population Proportion

Example (continued):
Null hypothesis: $H_0 : p = p_0$

Test statistic value:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Alternative Hypothesis

Rejection Region for α level test:

Test for Population Proportion

Example (continued):
Null hypothesis: $H_0 : p = p_0$

Test statistic value:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Alternative Hypothesis Rejection Region for α level test:

$H_a : p > p_0$
$H_a : p < p_0$
$H_a : p \neq p_0$

Test for Population Proportion

Example (continued):
Null hypothesis: $H_0 : p = p_0$

Test statistic value:

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}}$$

Alternative Hypothesis	Rejection Region for α level test:
$H_a : p > p_0$	$z > z_\alpha$
$H_a : p < p_0$	$z < -z_\alpha$
$H_a : p \neq p_0$	$z < -z_{\alpha/2}$ **or** $z > z_{\alpha/2}$

Test for Population Proportion

Example To determine if the Belgian 1 Euro coin is fair, you flip it 250 times and it comes up heads 139 times. Do you reject the null at $\alpha = .1$ and conclude the coin is unfair?

Test for Population Proportion

Example To determine if the Belgian 1 Euro coin is fair, you flip it 250 times and it comes up heads 139 times. Do you reject the null at $\alpha = .1$ and conclude the coin is unfair?

Test statistic:

$$Z = \frac{139/250 - .5}{\sqrt{\frac{.5(1-.5)}{250}}}$$

Test for Population Proportion

Example To determine if the Belgian 1 Euro coin is fair, you flip it 250 times and it comes up heads 139 times. Do you reject the null at $\alpha = .1$ and conclude the coin is unfair?

Test statistic:

$$Z = \frac{139/250 - .5}{\sqrt{\frac{.5(1-.5)}{250}}}$$

This is approximately 1.771. Compare to $z_{\alpha/2} = 1.645$.

Test for Population Proportion

Example To determine if the Belgian 1 Euro coin is fair, you flip it 250 times and it comes up heads 139 times. Do you reject the null at $\alpha = .1$ and conclude the coin is unfair?

Test statistic:

$$Z = \frac{139/250 - .5}{\sqrt{\frac{.5(1-.5)}{250}}}$$

This is approximately 1.771. Compare to $z_{\alpha/2} = 1.645$.
Reject, because our test statistic represents a more extreme observation under the null than our the corresponding value of $z_{\alpha/2}$.

Test for Population Mean (known variance)

Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic value:

Alternative Hypothesis   Rejection Region for α level test:

Test for Population Mean (known variance)

Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic value:

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

Alternative Hypothesis Rejection Region for α level test:

Test for Population Mean (known variance)

Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic value:

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

<u>Alternative Hypothesis</u> <u>Rejection Region</u> for α level test:

$H_a : \mu > \mu_0$
$H_a : \mu < \mu_0$
$H_a : \mu \neq \mu_0$

Test for Population Mean (known variance)

Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic value:

$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

Alternative Hypothesis	Rejection Region for α level test:
$H_a : \mu > \mu_0$	$z > z_\alpha$
$H_a : \mu < \mu_0$	$z < -z_\alpha$
$H_a : \mu \neq \mu_0$	$z < -z_{\alpha/2}$ **or** $z > z_{\alpha/2}$

Test for Population Mean (known variance)

Example: Suppose a company is considering putting a new type of coating on bearings that it produces. Let denote the true mean life for the new coating. The company would not want to make any (costly) changes unless evidence **strongly** suggested that exceeds 1000 hours.

State the null and alternative hypotheses:

Now, suppose the company tests $n = 25$ randomly selected bearings with the new coating, and finds that the sample mean for life of the new coating is 1090 hours. Suppose we know that the population standard deviation is 130 hours. Conduct a test for the mean.

Test for Population Mean (known variance)

Example: Suppose a company is considering putting a new type of coating on bearings that it produces. Let denote the true mean life for the new coating. The company would not want to make any (costly) changes unless evidence **strongly** suggested that exceeds 1000 hours.

State the null and alternative hypotheses:

Null hypothesis: $H_0 : \mu = \mu_0$

Now, suppose the company tests $n = 25$ randomly selected bearings with the new coating, and finds that the sample mean for life of the new coating is 1090 hours. Suppose we know that the population standard deviation is 130 hours. Conduct a test for the mean.

Test for Population Mean (known variance)

Example: Suppose a company is considering putting a new type of coating on bearings that it produces. Let denote the true mean life for the new coating. The company would not want to make any (costly) changes unless evidence **strongly** suggested that exceeds 1000 hours.

State the null and alternative hypotheses:

Null hypothesis: $H_0 : \mu = \mu_0$
Alt hypothesis: $H_a : \mu > \mu_0$

Now, suppose the company tests $n = 25$ randomly selected bearings with the new coating, and finds that the sample mean for life of the new coating is 1090 hours. Suppose we know that the population standard deviation is 130 hours. Conduct a test for the mean.

Test for Population Mean (known variance)

Example: Suppose a company is considering putting a new type of coating on bearings that it produces. Let denote the true mean life for the new coating. The company would not want to make any (costly) changes unless evidence **strongly** suggested that exceeds 1000 hours.

State the null and alternative hypotheses:

Null hypothesis: $H_0 : \mu = \mu_0$
Alt hypothesis: $H_a : \mu > \mu_0$

Now, suppose the company tests $n = 25$ randomly selected bearings with the new coating, and finds that the sample mean for life of the new coating is 1090 hours. Suppose we know that the population standard deviation is 130 hours. Conduct a test for the mean. Compare

$Z = \frac{1090-1000}{130/\sqrt{25}} = 3.45$ to $z_{.01} = 2.32$ and **Reject**

Errors in Hypothesis Testing

Definitions: *Type One and Type Two Errors*

A *type I error* is when the null hypothesis is rejected, but it is true. (False Positive).

A *type II error* is not rejecting H_0 when H_0 is false. (False Negative)

Errors in Hypothesis Testing

Definitions: *Type One and Type Two Errors*

A *type I error* is when the null hypothesis is rejected, but it is true. (False Positive).

We commit type I errors at a rate of α.

A *type II error* is not rejecting H_0 when H_0 is false. (False Negative)

We commit type II errors at a rate of β.

Errors in Hypothesis Testing

Typically, we specify the largest value of a type I error, α, that can be tolerated, and then find a rejection region with that α.

The resulting value of is often referred to as the **significance** level of the test.

Traditional levels of significance are .10, .05, and .01, though the level in any particular problem will depend on the seriousness of a type I error. The more serious the type I error, the smaller the significance level should be.

Errors in Hypothesis Testing

We can also obtain a smaller value of α - the probability of a type I error - by decreasing the size of the rejection region. However, this results in a larger value of β - the probability of a type II error - for all parameter values consistent with H_a.

No rejection region will simultaneously make α and β small at the same time. A region must be chosen to strike a **compromise** between these errors.

Errors in Hypothesis Testing

Suppose we are evaluating a medical procedure. It has the following outcomes:

1. Procedure works: subject is cured!
2. Procedure doesn't work: subject dies.

Do we want our test - which compared this procedure as H_a to a null hypothesis of "don't get the procedure" - to focus on decreasing α or β?

Errors in Hypothesis Testing

Suppose we are evaluating a medical procedure. It has the following outcomes:

1. Procedure works: subject is cured!
2. Procedure doesn't work: subject dies.

Do we want our test - which compared this procedure as H_a to a null hypothesis of "don't get the procedure" - to focus on decreasing α or β?
$\boxed{\alpha}$: probability of endorsing the test when it doesn't actually work. Result: subjects die.
$\boxed{\beta}$: probability of *not* endorsing the test when it really works! Result: subjects don't get treated.

Errors in Hypothesis Testing

Suppose we are evaluating a medical procedure. It has the following outcomes:

1. Procedure works: subject is cured!
2. Procedure doesn't work: subject dies.

Do we want our test - which compared this procedure as H_a to a null hypothesis of "don't get the procedure" - to focus on decreasing α or β?
$\boxed{\alpha}$: probability of endorsing the test when it doesn't actually work. Result: subjects die.

Where is your risk tolerance?

Example An inventor has developed a new, energy-efficient lawn mower engine. He claims that the engine will run continuously for more than 5 hours (300 minutes) on a single gallon of regular gasoline. (The leading brand lawnmower engine runs for 300 minutes on 1 gallon of gasoline.)

From his stock of engines, the inventor selects a simple random sample of 50 engines for testing. The engines run for an average of 305 minutes. The *sample* standard deviation s is found at 30 minutes.

Test the hypothesis that the mean run time is more than 300 minutes. Use a 0.01 level of significance.

Large Sample Testing

When the sample size is large, the z tests are easily modified to yield valid test procedures without requiring either a normal population distribution or known standard deviation.

Earlier, we used the key result to justify large-sample confidence intervals:
A large $n\,(> 30)$ implies that the standardized variable

has approximately a standard normal distribution.

Large Sample Testing

When the sample size is large, the z tests are easily modified to yield valid test procedures without requiring either a normal population distribution or known standard deviation.

Earlier, we used the key result to justify large-sample confidence intervals:
A large $n\,(> 30)$ implies that the standardized variable

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

has approximately a standard normal distribution.

Test for Population Mean (Large Sample)

Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic value:

Alternative Hypothesis   Rejection Region for α level test:

Test for Population Mean (Large Sample)

Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic value:

$$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

Alternative Hypothesis Rejection Region for α level test:

Test for Population Mean (Large Sample)

Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic value:

$$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

Alternative Hypothesis   Rejection Region for α level test:

$H_a : \mu > \mu_0$
$H_a : \mu < \mu_0$
$H_a : \mu \neq \mu_0$

Test for Population Mean (Large Sample)

Null hypothesis: $H_0 : \mu = \mu_0$

Test statistic value:

$$Z = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

Alternative Hypothesis	Rejection Region for α level test:
$H_a : \mu > \mu_0$	$z > z_\alpha$
$H_a : \mu < \mu_0$	$z < -z_\alpha$
$H_a : \mu \neq \mu_0$	$z < -z_{\alpha/2}$ **or** $z > z_{\alpha/2}$

Example, cont'd: An inventor has developed a new, energy-efficient lawn mower engine. He claims that the engine will run continuously for more than 5 hours (300 minutes) on a single gallon of regular gasoline. (The leading brand lawnmower engine runs for 300 minutes on 1 gallon of gasoline.)

$n = 50$; $\bar{X} = 305, s = 30$. Is $\mu > 300$? at $\alpha = .01$?

Test the hypothesis that the mean run time is more than 300 minutes. Use a 0.01 level of significance.

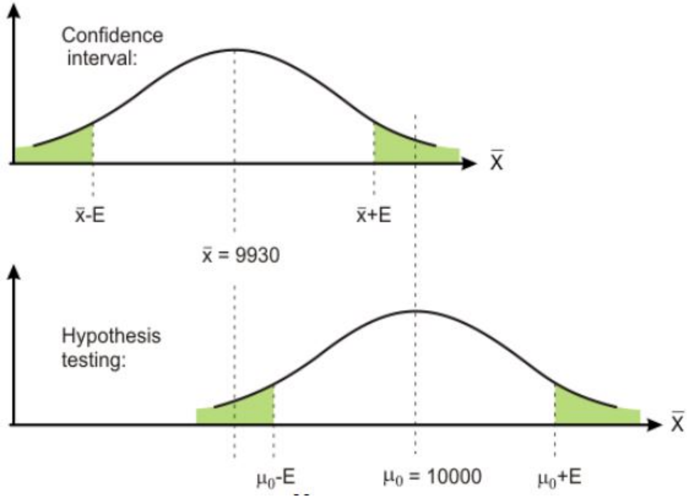
Example, cont'd: An inventor has developed a new, energy-efficient lawn mower engine. He claims that the engine will run continuously for more than 5 hours (300 minutes) on a single gallon of regular gasoline. (The leading brand lawnmower engine runs for 300 minutes on 1 gallon of gasoline.)

$n = 50$; $\bar{X} = 305, s = 30$. Is $\mu > 300$? at $\alpha = .01$?

Test the hypothesis that the mean run time is more than 300 minutes. Use a 0.01 level of significance.
Compare $Z = \frac{\bar{X}-\mu_0}{s/\sqrt{n}} = \frac{305-300}{30/\sqrt{50}} = 1.17$ to $z_{.01} = 2.32$ and **Fail to Reject**.

Rejection Regions and CIs

Rejection regions have a lot in common with confidence intervals.



p-values

The p-value measures the ”extremeness” of the test statistic.

Definition: *p-value:*

A *p-value* is the probability, under the null hypothesis, that we would get a test statistic at least as extreme as the one we calculated.

Idea: So, the smaller the p-value, the more evidence there is in the sample data against the null hypothesis (so the story goes. . .).

So what constitutes “sufficiently small” and “extreme enough” to make a decision about the null hypothesis?

p-values: a Hypothesis Testing decision

1. Select a significance level (as before, the desired type I error probability), which would define the rejection region.

p-values: a Hypothesis Testing decision

1. Select a significance level (as before, the desired type I error probability), which would define the rejection region.
2. Instead, calculate your test statistic (e.g. a z_{stat} or Z-score),

p-values: a Hypothesis Testing decision

1. Select a significance level (as before, the desired type I error probability), which would define the rejection region.

2. Instead, calculate your test statistic (e.g. a z_{stat} or Z-score),

3. ...and from it calculate the associated probability e.g. $p = P(Z > z_{stat})$ - or the less than, two sided variant - and use that.

p-values: a Hypothesis Testing decision

1. Select a significance level (as before, the desired type I error probability), which would define the rejection region.
2. Instead, calculate your test statistic (e.g. a z_{stat} or Z-score),
3. ...and from it calculate the associated probability e.g. $p = P(Z > z_{stat})$ - or the less than, two sided variant - and use that.
4. decision rule is:

$$\textbf{Reject if:} \quad p \leq \alpha$$

p-values: a Hypothesis Testing decision

1. Select a significance level (as before, the desired type I error probability), which would define the rejection region.
2. Instead, calculate your test statistic (e.g. a z_{stat} or Z-score),

4. decision rule is:

$$\textbf{Reject if:} \quad p \leq \alpha$$

Thus if the p-value exceeds the chosen significance level, the null hypothesis cannot be rejected at that level.

Note, the p-value can be thought of as the smallest significance level at which H_0 can be rejected.

p-values

The p-value measures the “extremeness” of the test statistic.

Note:

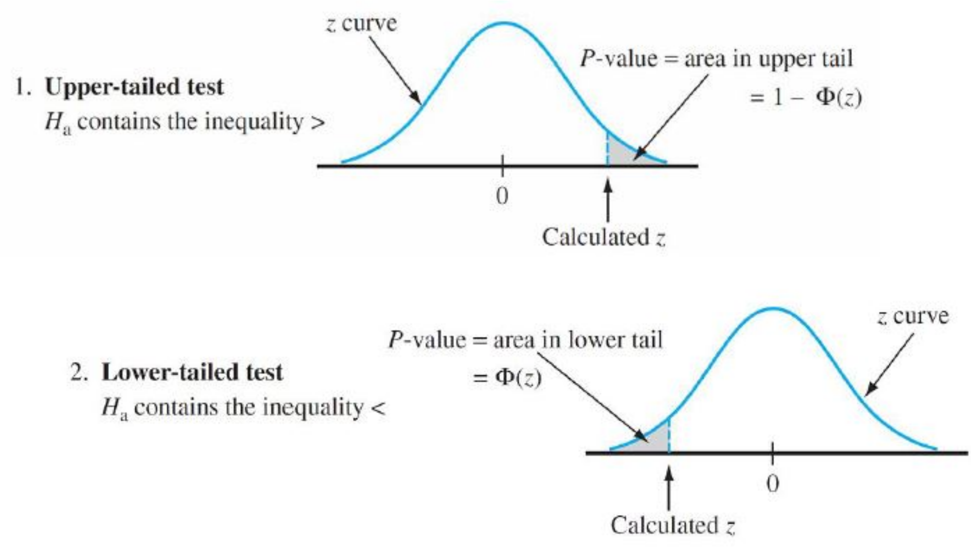
1. This probability is calculated assuming that the null hypothesis is true.
2. Beware: The p-value is not the probability that H_0 is true, nor is it an error probability!
3. The p-value is between 0 and 1.

p-values

The calculation of the p-value depends on whether the test is upper-, lower-, or two-tailed.

Each of these is the probability of getting a value at least as extreme as what was obtained (assuming H_0 true).

p-values for Z tests



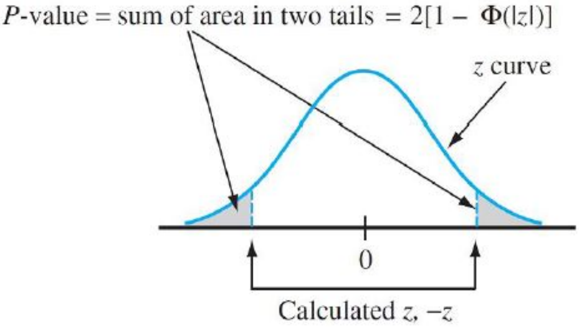
1. **Upper-tailed test**
   H_a contains the inequality >

2. **Lower-tailed test**
   H_a contains the inequality <

p-values for Z tests

3. **Two-tailed test**

   H_a contains the inequality $\neq$



P-value = sum of area in two tails = $2[1 - \Phi(|z|)]$

p-values for Z tests

Back to the lawnmower engine example: There, we had

$H_0 : \mu = 300$ vs $H_a : \mu > 300$

and $Z = 1.17$.

What is the p-value for this result? What decision would you make?

p-values for Z tests

Back to the lawnmower engine example: There, we had

$H_0 : \mu = 300$ vs $H_a : \mu > 300$

and $Z = 1.17$.

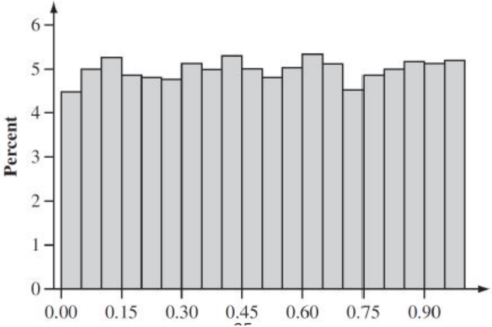
What is the p-value for this result? What decision would you make?

Hopefully the same one! In particular,
$P(Z > 1.17) =$ `1-stats.norm.cdf(1.17)` $= .12 = p > \alpha$

Distribution of p-values

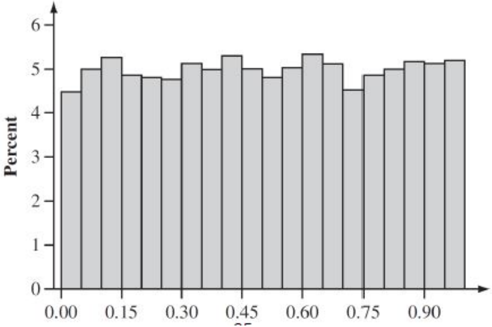
Figure below shows a histogram of the 10,000 P-values from a simulation experiment under a null $\mu = 20$ (with n $=$ 4 and $\sigma = 2$).

When H_0 is true, the probability distribution of the P-value is a uniform distribution on the interval from 0 to 1.



Distribution of p-values

Figure below shows a histogram of the 10,000 P-values from a simulation experiment under a null $\mu = 20$ (with n = 4 and σ= 2).
When H_0 is true, the probability distribution of the P-value is a uniform distribution on the interval from 0 to 1.



These data comes from a process where the null hypothesis is *TRUE*. Rejecting the null hypothesis would be an error.

Distribution of p-values

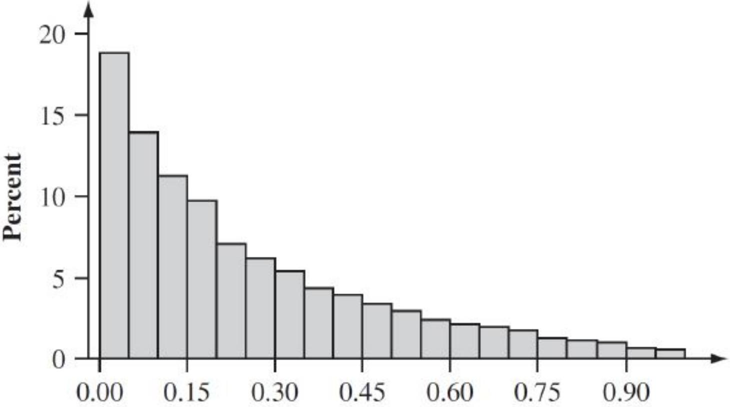
About 4.5% of these p-values are in the interval from 0 to .05.

Thus when using a significance level of .05, the null hypothesis is rejected in roughly 4.5% of these 10,000 tests.

If we continued to generate samples and carry out the test for each sample at significance level .05, in the long run, 5% of the p-values would be in the first class interval.

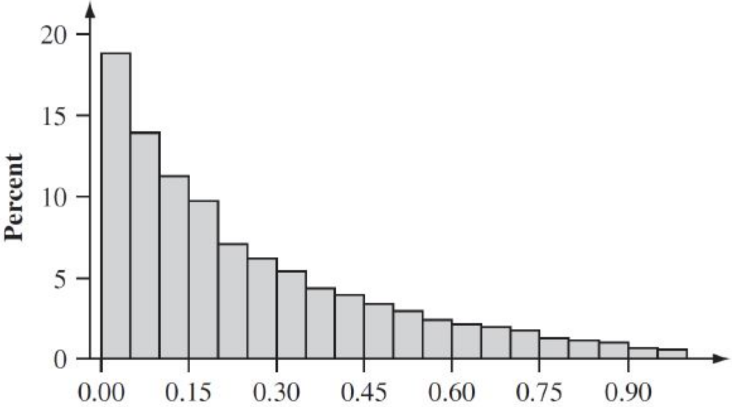
Distribution of p-values

A histogram of the p-values when we simulate under an alternative hypothesis. There is a much greater tendency for the p-value to be small (closer to 0) when $\mu = 21$ than when $\mu = 20$



Distribution of p-values

A histogram of the p-values when we simulate under an alternative hypothesis. There is a much greater tendency for the p-value to be small (closer to 0) when $\mu = 21$ than when $\mu = 20$



Distribution of p-values

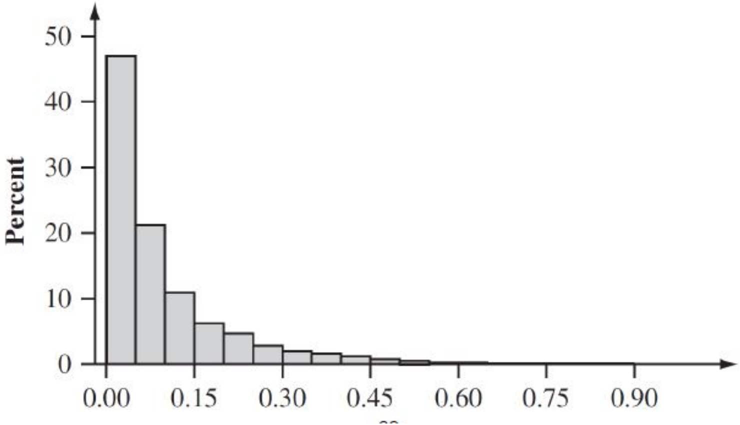
Again, H_0 is rejected at significance level .05 whenever the p-value is at most .05 (in the first bin).

Unfortunately, this is the case for only about 19% of the p-values. So only about 19% of the 10,000 tests correctly reject the null hypothesis; for the other 81%, a type II error is committed.

The difficulty is that the sample size is quite small and 21 is not very different from the value asserted by the null hypothesis (the "effect size" is small).

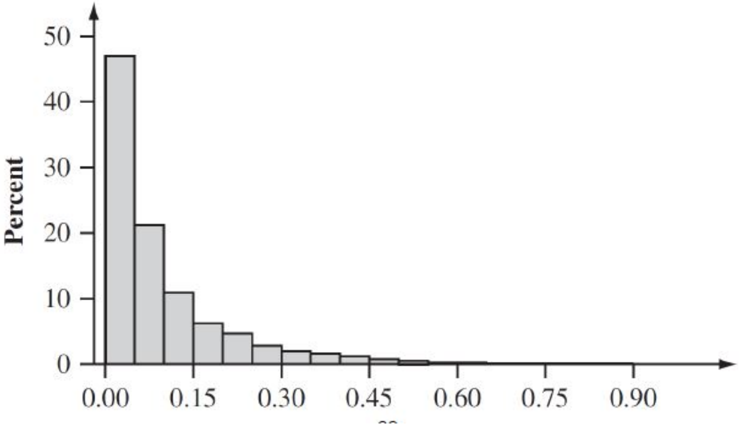
Distribution of p-values

The figure below illustrates what happens to the pvalue when H_0 is false because $\mu = 22$.



Distribution of p-values

The figure below illustrates what happens to the pvalue when H_0 is false because $\mu = 22$.



Again, these data come from H_a, but we reach the correct conclusion more often. The

Distribution of p-values

The histogram is even more concentrated toward values close to 0 than was the case when $\mu = 21$.

In general, as moves further to the right of the null value 20, the distribution of the p-value will become more and more concentrated on values close to 0.

Even here a bit fewer than 50% of the p-values are smaller than .05. So it is still slightly more likely than not that the null hypothesis is incorrectly not rejected. Only for values of much larger than 20 (e.g., at least 24 or 25) is it highly likely that the p-value will be smaller than .05 and thus give the correct conclusion.

Daily Recap

Today we learned

1. Intro and Basics of Hypothesis Tests

Moving forward:

- nb day Friday for HTs

Next time in lecture:

- More Hypotheses!