Hypothesis Testing for a Proportion

Remember when...

Gender discrimination experiment:

		Promotion		
		Promoted	Not Promoted	Total
Gender	Male	21	3	24
	Female	14	10	24
	Total	35	13	48

Remember when...

Gender discrimination experiment:

		Promotion		
		Promoted	Not Promoted	Total
Gender	Male	21	3	24
	Female	14	10	24
	Total	35	13	48

$\hat{p}_{males}$ = 21 / 24 = 0.88

$\hat{p}_{females}$ = 14 / 24 = 0.58

Remember when...

Gender discrimination experiment:

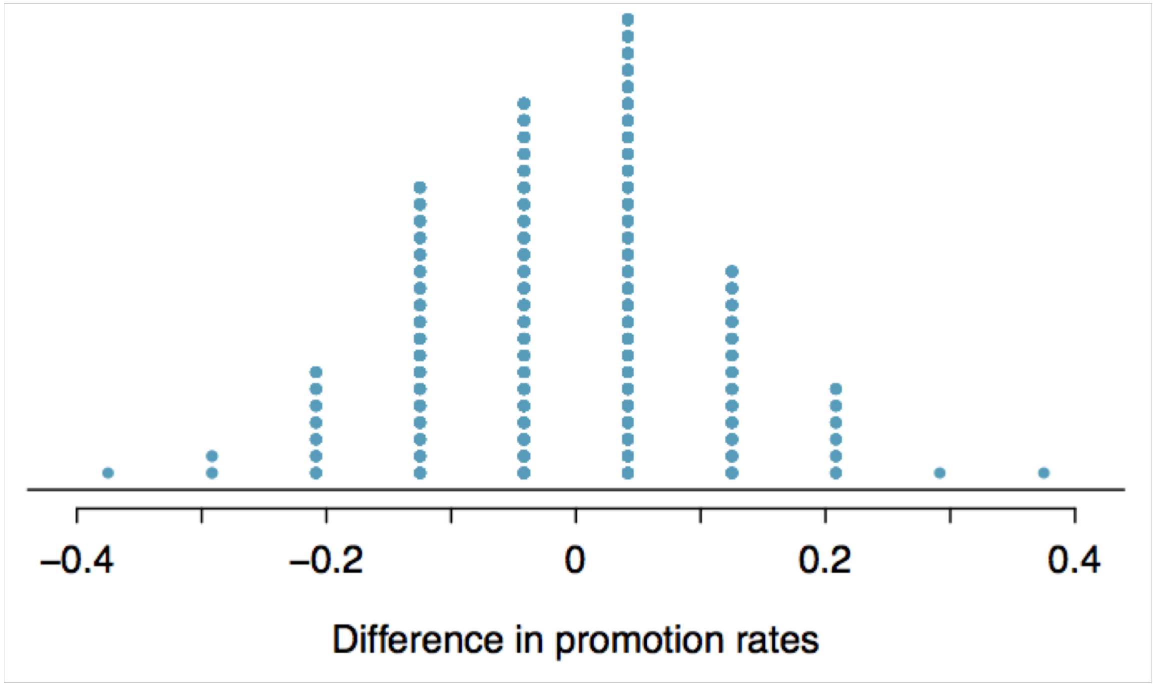
		Promotion		
		Promoted	Not Promoted	Total
Gender	Male	21	3	24
	Female	14	10	24
	Total	35	13	48

$$\hat{p}_{males} = 21 / 24 = 0.88$$
$$\hat{p}_{females} = 14 / 24 = 0.58$$

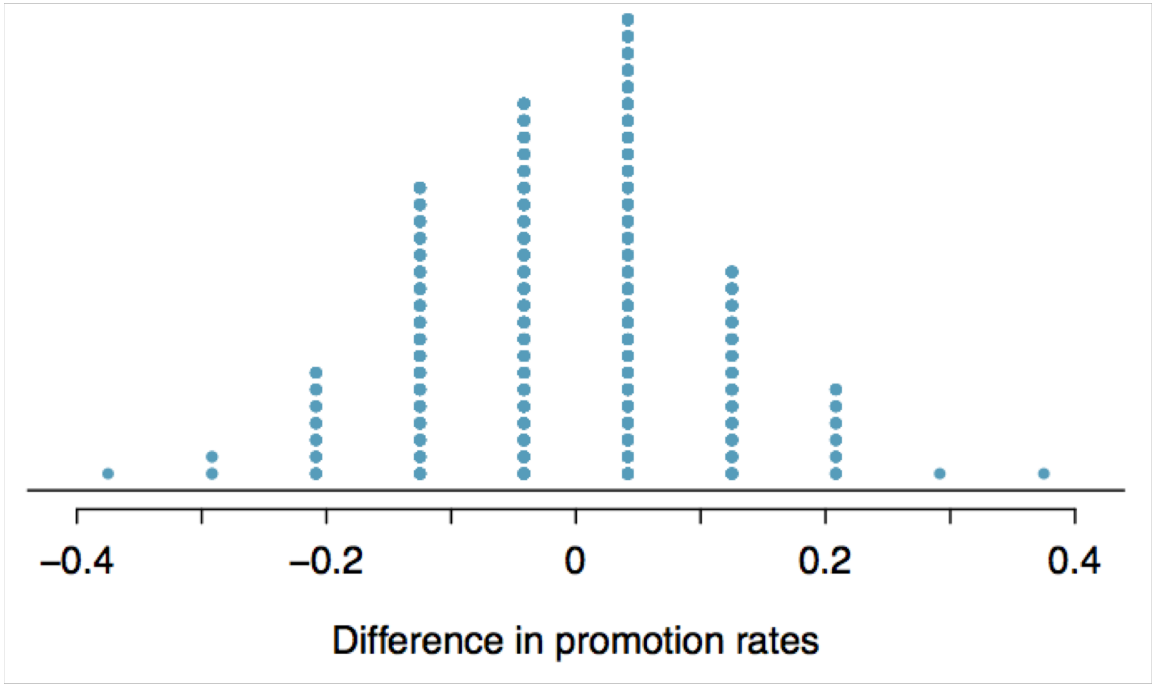
Possible explanations:

- Promotion and gender are *independent*, no gender discrimination, observed difference in proportions is simply due to chance.
  → null (nothing is going on)
- Promotion and gender are *dependent*, there is gender discrimination, observed difference in proportions is not due to chance.
  → alternative (something is going on)

Result



Result



Since it was quite unlikely to obtain results like the actual data or something more extreme in the simulations (male promotions being 30% or more higher than female promotions), we decided to reject the null hypothesis in favor of the alternative.

Recap: hypothesis testing framework

- We start with a *null hypothesis* (H_0) that represents the status quo.

Recap: hypothesis testing framework

- We start with a *null hypothesis* (H_0) that represents the status quo.
- We also have an *alternative hypothesis* (H_A) that represents our research question, i.e. what we're testing for.

Recap: hypothesis testing framework

- We start with a *null hypothesis* (H_0) that represents the status quo.
- We also have an *alternative hypothesis* (H_A) that represents our research question, i.e. what we're testing for.
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or traditional methods based on the central limit theorem (coming up next...).

Recap: hypothesis testing framework

- We start with a *null hypothesis* (H_0) that represents the status quo.
- We also have an *alternative hypothesis* (H_A) that represents our research question, i.e. what we're testing for.
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or traditional methods based on the central limit theorem (coming up next...).
- If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

Recap: hypothesis testing framework

- We start with a *null hypothesis* (H_0) that represents the status quo.
- We also have an *alternative hypothesis* (H_A) that represents our research question, i.e. what we're testing for.
- We conduct a hypothesis test under the assumption that the null hypothesis is true, either via simulation or traditional methods based on the central limit theorem (coming up next...).
- If the test results suggest that the data do not provide convincing evidence for the alternative hypothesis, we stick with the null hypothesis. If they do, then we reject the null hypothesis in favor of the alternative.

We'll formally introduce the hypothesis testing framework using an example on testing a claim about a population mean.

Testing hypotheses using confidence intervals

Earlier we calculated a 95% confidence interval for the proporton of American Facebook users who think Facebook categorizes their interests accurately as 64% to 67%. Based on this confidence interval, do the data support the hypothesis that majority of American Facebook users think Facebook categorizes their interests accurately.

The associated hypotheses are:

H_0: $p = 0.50$: 50% of American Facebook users think Facebook categorizes their interests accurately

H_A: $p > 0.50$: More than 50% of American Facebook users think Facebook categorizes their interests accurately

Null value is not included in the interval → reject the null hypothesis.

This is a quick-and-dirty approach for hypothesis testing, but it doesn’t tell us the likelihood of certain outcomes under the null hypothesis (p-value)

Decision errors

- Hypothesis tests are not flawless.
- In the court system innocent people are sometimes wrongly convicted, and the guilty sometimes walk free.
- Similarly, we can make a wrong decision in statistical hypothesis tests as well.
- The difference is that we have the tools necessary to quantify how often we make errors in statistics.

Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true		
	H_A true		

Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	
	H_A true		

Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	
	H_A true		✓

Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true		✓

- A *Type 1 Error* is rejecting the null hypothesis when H_0 is true.

Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true	*Type 2 Error*	✓

- A *Type 1 Error* is rejecting the null hypothesis when H_0 is true.
- A *Type 2 Error* is failing to reject the null hypothesis when H_A is true.

Decision errors (cont.)

There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.

		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true	*Type 2 Error*	✓

- A *Type 1 Error* is rejecting the null hypothesis when H_0 is true.
- A *Type 2 Error* is failing to reject the null hypothesis when H_A is true.

We (almost) never know if H_0 or H_A is true, but we need to consider all possibilities.

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

H_0: Defendant is innocent

H_A: Defendant is guilty

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

- Declaring the defendant guilty when they are actually innocent

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

H_0: Defendant is innocent

H_A: Defendant is guilty

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

Type 2 error

- Declaring the defendant guilty when they are actually innocent

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

H_0: Defendant is innocent

H_A: Defendant is guilty

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

Type 2 error

- Declaring the defendant guilty when they are actually innocent

Type 1 error

Which error do you think is the worse error to make?

Hypothesis Test as a trial

If we again think of a hypothesis test as a criminal trial then it makes sense to frame the verdict in terms of the null and alternative hypotheses:

H_0: Defendant is innocent

H_A: Defendant is guilty

Which type of error is being committed in the following circumstances?

- Declaring the defendant innocent when they are actually guilty

Type 2 error

- Declaring the defendant guilty when they are actually innocent

Type 1 error

Which error do you think is the worse error to make?

“better that ten guilty persons escape than that one innocent suffer”
- William Blackstone

Type 1 error rate

- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a *significance level* of 0.05, $\alpha = 0.05$.

Type 1 error rate

- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a *significance level* of 0.05, $\alpha = 0.05$.
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.

Type 1 error rate

- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a *significance level* of 0.05, $\alpha = 0.05$.
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.
- In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error if the null hypothesis is true.

$$P(\text{Type 1 error} \mid H_0 \text{ true}) = \alpha$$

Type 1 error rate

- As a general rule we reject H_0 when the p-value is less than 0.05, i.e. we use a *significance level* of 0.05, $\alpha = 0.05$.
- This means that, for those cases where H_0 is actually true, we do not want to incorrectly reject it more than 5% of those times.
- In other words, when using a 5% significance level there is about 5% chance of making a Type 1 error if the null hypothesis is true.

$$P(\text{Type 1 error} \mid H_0 \text{ true}) = \alpha$$

- This is why we prefer small values of α -- increasing α increases the Type 1 error rate.

Facebook interest categories

The same survey asked the 850 respondents how comfortable they are with Facebook creating a list of categories for them. 41% of the respondents said they are comfortable. Do these data provide convincing evidence that the proportion of American Facebook users are comfortable with Facebook creating a list of interest categories for them is different than 50%?

https://www.pewinternet.org/2019/01/16/facebook-algorithms-and-personal-data/

Facebook interest categories

The same survey asked the 850 respondents how comfortable they are with Facebook creating a list of categories for them. 41% of the respondents said they are comfortable. Do these data provide convincing evidence that the proportion of American Facebook users are comfortable with Facebook creating a list of interest categories for them is different than 50%?

Setting the hypotheses

The *parameter of interest* is the proportion of <u>all</u> American Facebook users who are comfortable with Facebook creating categories of interests for them.

There may be two explanations why our sample proportion is lower than 0.50 (minority).

- The true population proportion is different than 0.50.
- The true population mean is 0.50, and the difference between the true population proportion and the sample proportion is simply due to natural sampling variability.

Facebook interest categories

The same survey asked the 850 respondents how comfortable they are with Facebook creating a list of categories for them. 41% of the respondents said they are comfortable. Do these data provide convincing evidence that the proportion of American Facebook users are comfortable with Facebook creating a list of interest categories for them is different than 50%?

Setting the hypotheses

We start with the assumption that 50% of American Facebook users are comfortable with Facebook creating categories of interests for them

H_0: $p = 0.50$

We test the claim that the proportion of American Facebook users who are comfortable with Facebook creating categories of interests for them is different than 50%.

H_A: $p \neq 0.50$

Facebook interest categories - conditions

Which of the following is not a condition that needs to be met to proceed with this hypothesis test?

(a) Respondents in the sample should be independent of each other with respect to whether or not they feel comfortable with their interests being categorized by Facebook.

(b) Sampling should have been done randomly.

(c) The sample size should be less than 10% of the population of all American Facebook users.

(d) There should be at least 30 respondents in the sample.

(e) There should be at least 10 expected successes and 10 expected failure.

Facebook interest categories - conditions

Which of the following is not a condition that needs to be met to proceed with this hypothesis test?

(a) Respondents in the sample should be independent of each other with respect to whether or not they feel comfortable with their interests being categorized by Facebook.

(b) Sampling should have been done randomly.

(c) The sample size should be less than 10% of the population of all American Facebook users.

(d) There should be at least 30 respondents in the sample.

(e) There should be at least 10 expected successes and 10 expected failure.

Test statistic

In order to evaluate if the observed sample mean is unusual for the hypothesized sampling distribution, we determine how many standard errors away from the null it is, which is also called the *test statistic*.

$$\hat{p} \sim N\left(\mu = 0.50, SE = \sqrt{\frac{0.50 \times 0.50}{850}}\right)$$

$$Z = \frac{0.41 - 0.50}{0.0171} = -5.26$$

The sample proportion is 5.26 standard errors away from the hypothesized value. Is this considered unusually low? That is, is the result *statistically significant*?

Yes, and we can quantify how unusual it is using a p-value.

p-values

We then use this test statistic to calculate the *p-value*, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.

If the p-value is *low* (lower than the significance level, α, which is usually 5%) we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence *reject* H_0.

If the p-value is *high* (higher than α) we say that it is likely to observe the data even if the null hypothesis were true, and hence *do not reject* H_0.

Facebook interest categories - p-value

p-value: probability of observing data at least as favorable to H_A as our current data set (a sample proportion lower than 0.41), if in fact H_0 were true (the true population proportion was 0.50).

$$P(\hat{p} < 0.41 \; or \; \hat{p} > 0.59 \mid p = 0.50) = P(|Z| > 5.26) < 0.0001$$

Facebook interest categories - Making a decision

p-value < 0.0001

- If 50% of all American Facebook users are comfortable with Facebook creating these interest categories, there is less than a 0.01% chance of observing a random sample of 850 American Facebook users where 41% or fewer or 59% of higher feel comfortable with it.
- This is a pretty low probability for us to think that the observed sample proportion, or something more extreme, is likely to happen simply by chance.

Since p-value is *low* (lower than 5%) we *reject H_0*.

The data provide convincing evidence that the proportion of American Facebook users who are comfortable with Facebook creating a list of interest categories for them is different than 50%.

The difference between the null value of 0.50 and observed sample proportion of 0.41 is *not due to chance* or sampling variability.

Choosing a significance level

Choosing a significance level for a test is important in many contexts, and the traditional level is 0.05. However, it is often helpful to adjust the significance level based on the application.

We may select a level that is smaller or larger than 0.05 depending on the consequences of any conclusions reached from the test.

If making a Type 1 Error is dangerous or especially costly, we should choose a small significance level (e.g. 0.01). Under this scenario we want to be very cautious about rejecting the null hypothesis, so we demand very strong evidence favoring H_A before we would reject H_0.

If a Type 2 Error is relatively more dangerous or much more costly than a Type 1 Error, then we should choose a higher significance level (e.g. 0.10). Here we want to be cautious about failing to reject H_0 when the null is actually false.

One vs. two sided hypothesis tests

In two sided hypothesis tests we are interested in whether p is either above or below some null value p_0: $H_A : p \neq p_0$.

In one sided hypothesis tests we are interested in p differing from the null value p_0 in one direction (and not the other):

If there is only value in detecting if population parameter is less than p_0, then H_A: $p < p_0$.

If there is only value in detecting if population parameter is greater than p_0, then $H_A : p > p_0$.

Two-sided tests are often more appropriate as we often want to detect if the data goes clearly in the opposite direction of a hypothesis direction as well.