Reminder: Confidence interval

Confidence interval, a general formula

point estimate ± z x SE*

Conditions when the point estimate = $\bar{x}$

1. *Independence*: Observations in the sample must be independent

2. *Sample size / skew*: $n \geq 30$ and population distribution should not be extremely skewed

Case Study:
Gender Discrimination

Gender Discrimination

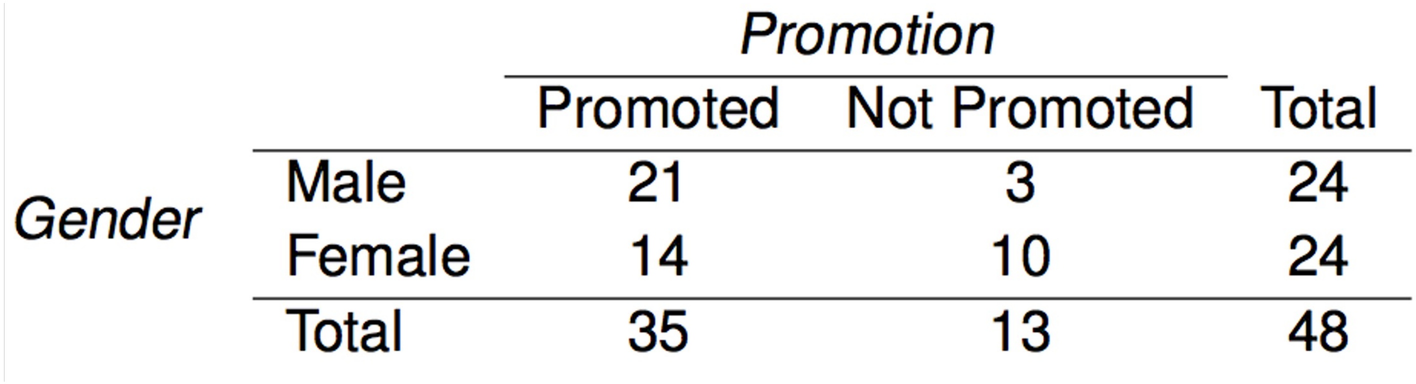
- In 1972, as a part of a study on gender discrimination, 48 male bank supervisors were each given the same personnel file and asked to judge whether the person should be promoted to a branch manager job that was described as “routine”.
- The files were identical except that half of the supervisors had files showing the person was male while the other half had files showing the person was female.
- It was randomly determined which supervisors got “male” applications and which got “female” applications.
- Of the 48 files reviewed, 35 were promoted.
- The study is testing whether females are unfairly discriminated against.

Is this an observational study or an experiment?

B.Rosen and T. Jerdee (1974), ``Influence of sex role stereotypes on personnel decisions", J.Applied Psychology, 59:9-14.

Data

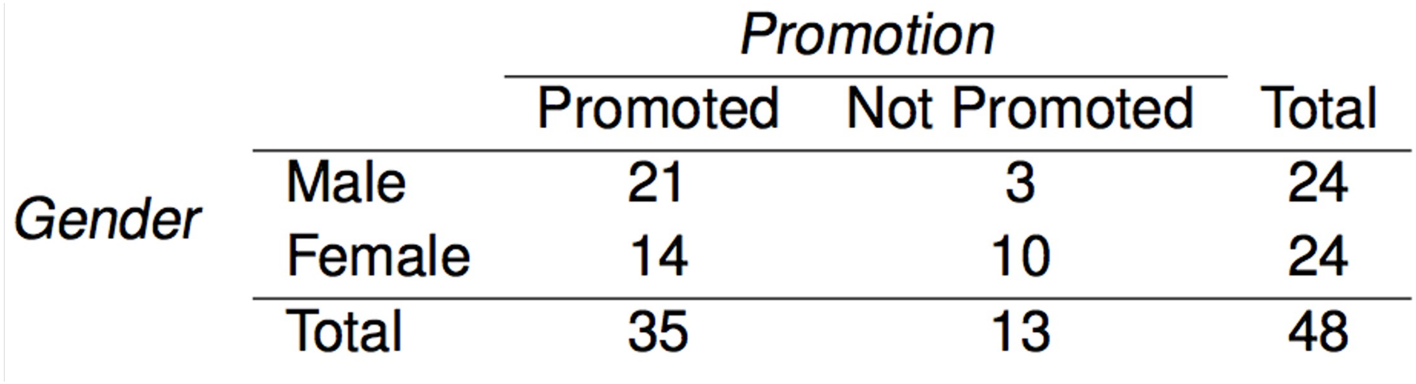
At a first glance, does there appear to be a relationship between promotion and gender?



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

Data

At a first glance, does there appear to be a relationship between promotion and gender?



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

% promoted: $\widehat{\boldsymbol{p_t}}$: 35/ 48 = 0.73

% of males promoted: $\widehat{\boldsymbol{p_m}}$: 21 / 24 = 0.88

% of females promoted: $\widehat{\boldsymbol{p_f}}$: 14 / 24 = 0.58

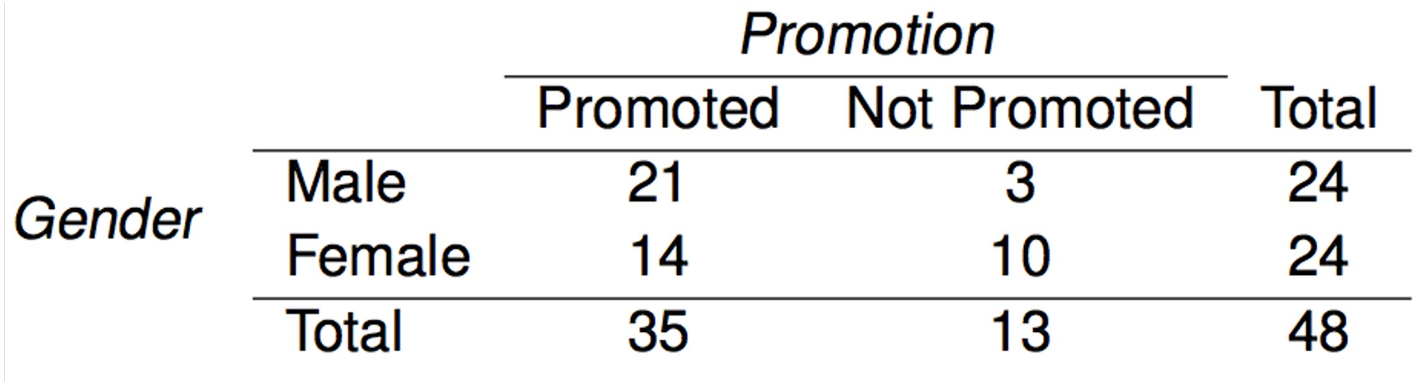
Confidence?

- Compute 95% Confidence Intervals for

- $\widehat{\boldsymbol{p_m}}$
  - $0.88 \ \pm \ 1.96\times\sqrt{\frac{0.88\times0.12}{24}} = 0.88 \pm 1.96\times0.07$
  - 0.88 ± 0.13
- $\widehat{\boldsymbol{p_f}}$
  - $0.58 \ \pm \ 1.96\times\sqrt{\frac{0.58\times0.42}{24}} = 0.58 \pm \ 1.96\times0.1$
  - 0.58 ± 0.18

Data

At a first glance, does there appear to be a relationship between promotion and gender?



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

% promoted: $\widehat{p_t}$: 35/ 48 = 0.73

% of males promoted: $\widehat{p_m}$: 21 / 24 = 0.88

% of females promoted: $\widehat{p_f}$: 14 / 24 = 0.58

If Gender and Promotion are independent, $\widehat{p_m}$ - $\widehat{p_f}$ $= \mathbf{0}\%$

Practice

We saw a difference of almost 30% (29.2% to be exact) between the proportion of male and female files that are promoted. Based on this information, which of the below is true?

A. If we were to repeat the experiment we will definitely see that more female files get promoted. This was a fluke.
B. Promotion is dependent on gender, males are more likely to be promoted, and hence there is gender discrimination against women in promotion decisions.
C. The difference in the proportions of promoted male and female files is due to chance, this is not evidence of gender discrimination against women in promotion decisions.
D. Women are less qualified than men, and this is why fewer females get promoted.

Practice

We saw a difference of almost 30% (29.2% to be exact) between the proportion of male and female files that are promoted. Based on this information, which of the below is true?

A. If we were to repeat the experiment we will definitely see that more female files get promoted. This was a fluke.

B. Promotion is dependent on gender, males are more likely to be promoted, and hence there is gender discrimination against women in promotion decisions. Maybe

C. The difference in the proportions of promoted male and female files is due to chance, this is not evidence of gender discrimination against women in promotion decisions. Maybe

D. Women are less qualified than men, and this is why fewer females get promoted.

Two Competing Claims

1. “There is nothing going on.”

   Promotion and gender are *independent*, no gender discrimination, observed difference in proportions is simply due to chance.

   → Null Hypothesis

Two Competing Claims

1. "There is nothing going on."

   Promotion and gender are *independent*, no gender discrimination, observed difference in proportions is simply due to chance.

   → Null Hypothesis

2. "There is something going on."

   Promotion and gender are *dependent*, there is gender discrimination, observed difference in proportions is not due to chance.

   → Alternative Hypothesis

A Trial as a Hypothesis Test

Hypothesis testing is very much like a court trial.

- H_0 : Defendant is innocent
  H_A : Defendant is guilty
- We then present the evidence - collect data.
- Then we judge the evidence - “Could these data plausibly have happened by chance if the null hypothesis were true?"
  - If they were very unlikely to have occurred, then the evidence raises more than a reasonable doubt in our minds about the null hypothesis.
- Ultimately we must make a decision. How unlikely is unlikely?

Image from http://www.nwherald.com/_internal/cimg!0/oo1il4sf8zzaqbboq25oevvbg99wpot

A Trial as a Hypothesis Test (cont.)

- If the evidence is not strong enough to reject the assumption of innocence, the jury returns with a verdict of “not guilty".
  - The jury does not say that the defendant is innocent, just that there is not enough evidence to convict.
  - The defendant may, in fact, be innocent, but the jury has no way of being sure.
- Said statistically, we fail to reject the null hypothesis.
  - We never declare the null hypothesis to be true, because we simply do not know whether it's true or not.
  - Therefore we never ``accept the null hypothesis".

A Trial as a Hypothesis Test (cont.)

- In a trial, the burden of proof is on the prosecution.
- In a hypothesis test, the burden of proof is on the unusual claim.
- The null hypothesis is the ordinary state of affairs (the status quo), so it's the alternative hypothesis that we consider unusual and for which we must gather evidence.

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 (today) or theoretical methods (later in the course).
- 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.

Simulating the experiment...

... under the assumption of independence, i.e. leave things up to chance.

If results from the simulations based on the *chance model* look like the data, then we can determine that the difference between the proportions of promoted files between males and females was simply *due to chance* (promotion and gender are independent).

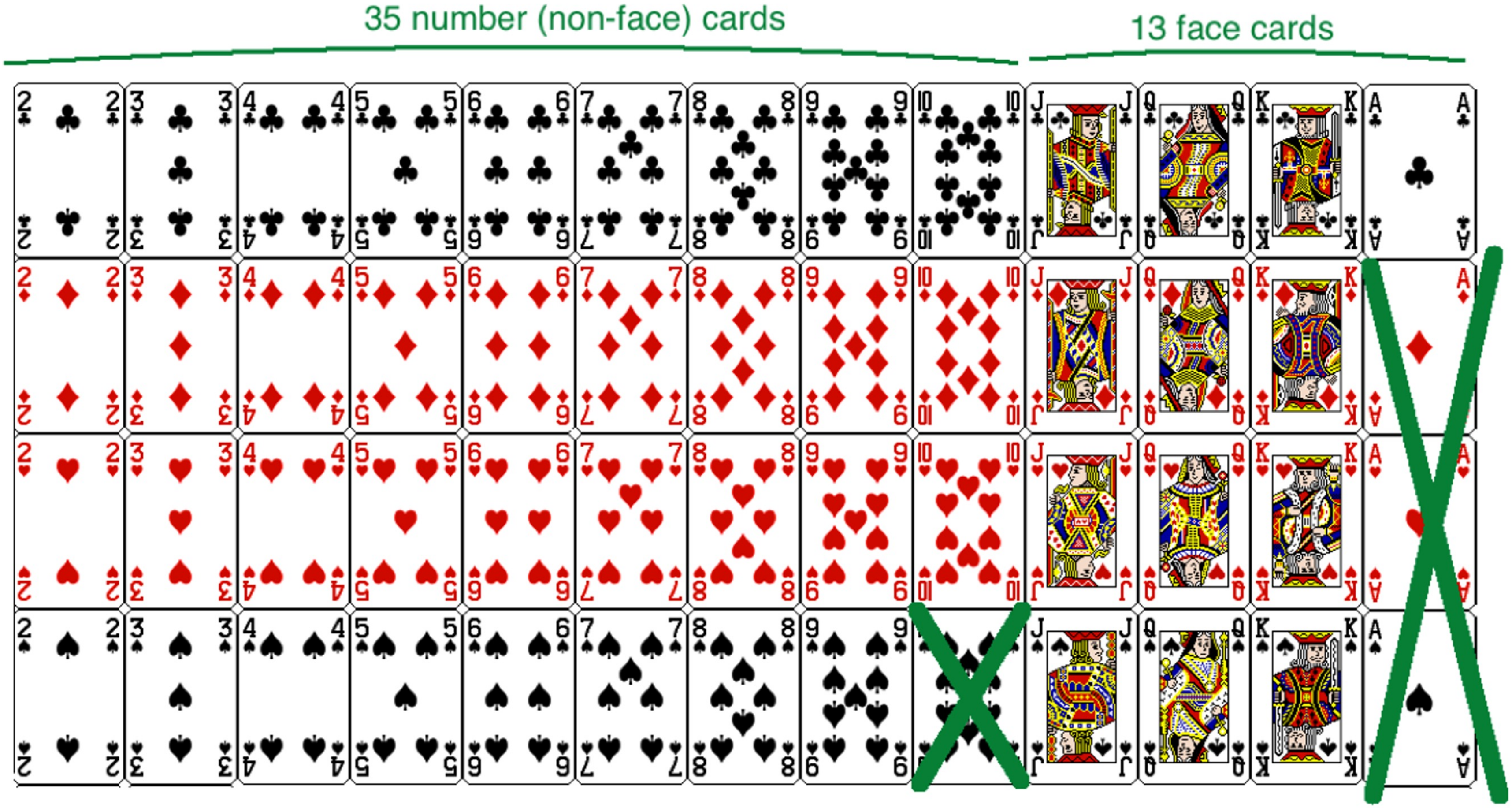
If the results from the simulations based on the chance model do not look like the data, then we can determine that the difference between the proportions of promoted files between males and females was not due to chance, but *due to an actual effect of gender* (promotion and gender are dependent).

Application Activity: Simulating the Experiment

Use a deck of playing cards to simulate this experiment.

1. Let a face card represent *not promoted* and a non-face card represent a *promoted*. Consider aces as face cards.
   - Set aside the jokers.
   - Take out 3 aces >> there are exactly 13 face cards left in the deck (face cards: A, K, Q, J).
   - Take out a number card >> there are exactly 35 number (non-face) cards left in the deck (number cards: 2-10).
2. Shuffle the cards and deal them intro two groups of size 24, representing males and females.
3. Count and record how many files in each group are promoted (number cards).
4. Calculate the proportion of promoted files in each group and take the difference (male - female), and record this value.
5. Repeat steps 2 - 4 many times.

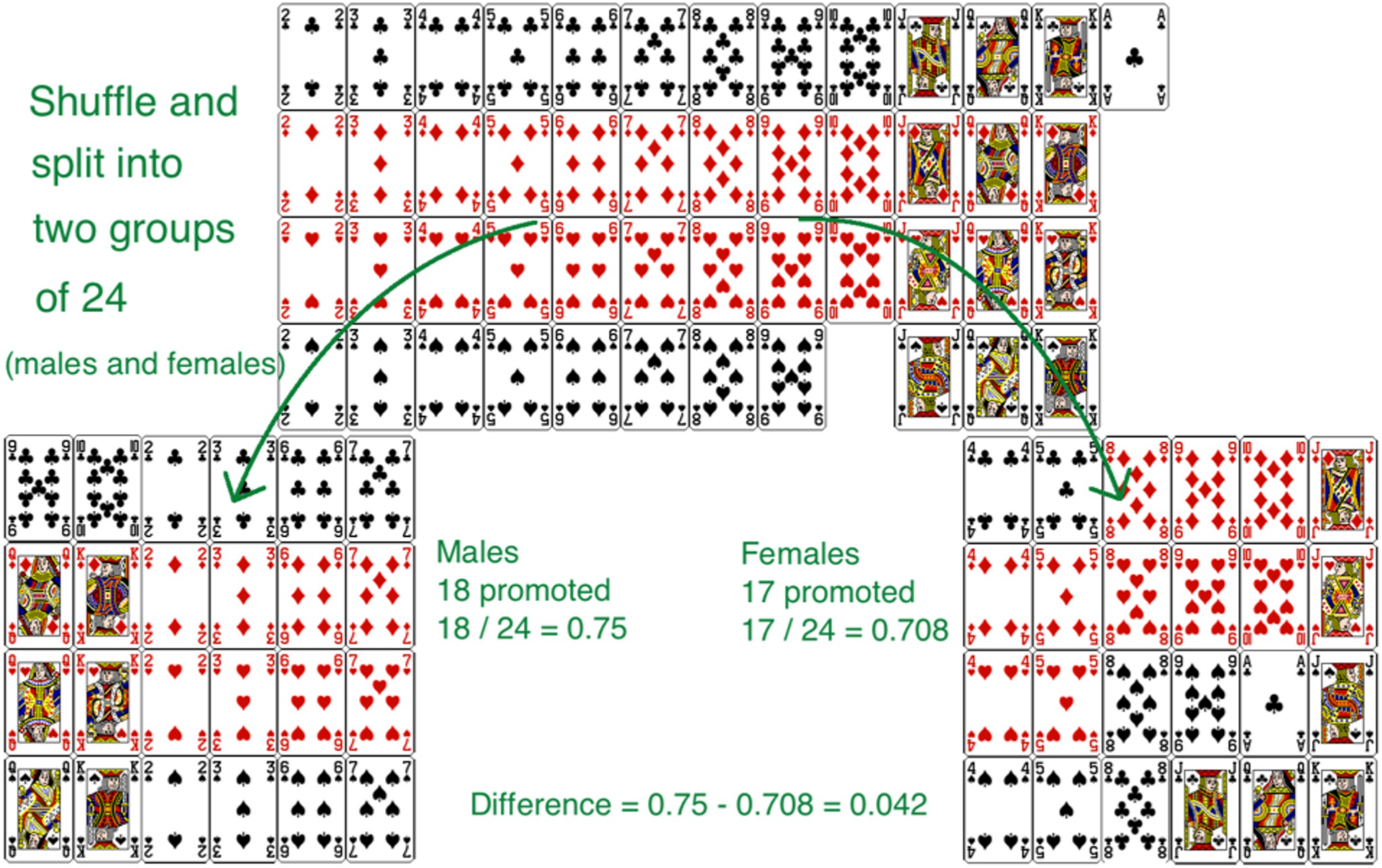
Step 1



Step 2 - 4

Shuffle and split into two groups of 24

(males and females)



Males
18 promoted
18 / 24 = 0.75

Females
17 promoted
17 / 24 = 0.708

Difference = 0.75 - 0.708 = 0.042

Practice

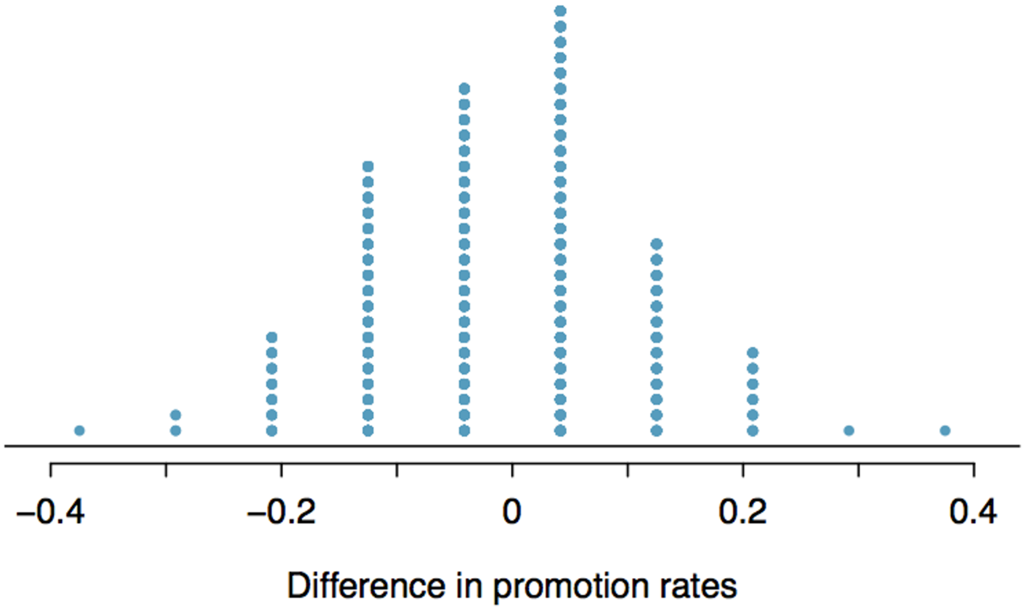
Do the results of the simulation you just ran provide convincing evidence of gender discrimination against women, i.e. dependence between gender and promotion decisions?

A. No, the data do not provide convincing evidence for the alternative hypothesis, therefore we can't reject the null hypothesis of independence between gender and promotion decisions. The observed difference between the two proportions was due to chance.

B. Yes, the data provide convincing evidence for the alternative hypothesis of gender discrimination against women in promotion decisions. The observed difference between the two proportions was due to a real effect of gender.

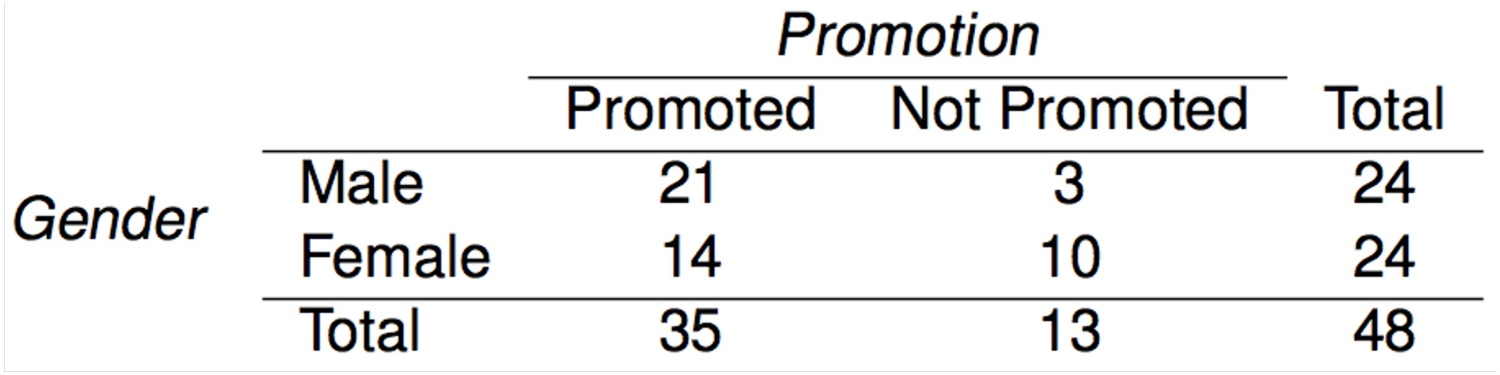
Simulations Using Software

These simulations are tedious and slow to run using the method described earlier. In reality, we use software to generate the simulations. The dot plot below shows the distribution of simulated differences in promotion rates based on 100 simulations.



Remember when...

Gender discrimination experiment:



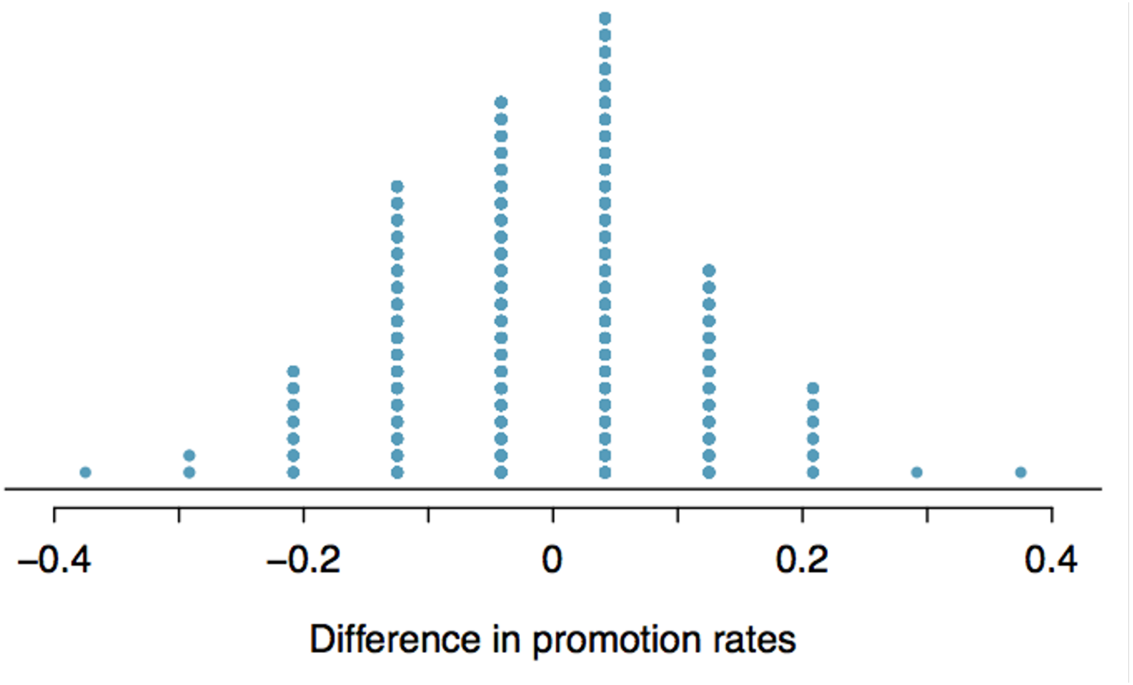
		Promotion		
		Promoted	Not Promoted	Total
Gender	Male	21	3	24
	Female	14	10	24
	Total	35	13	48

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

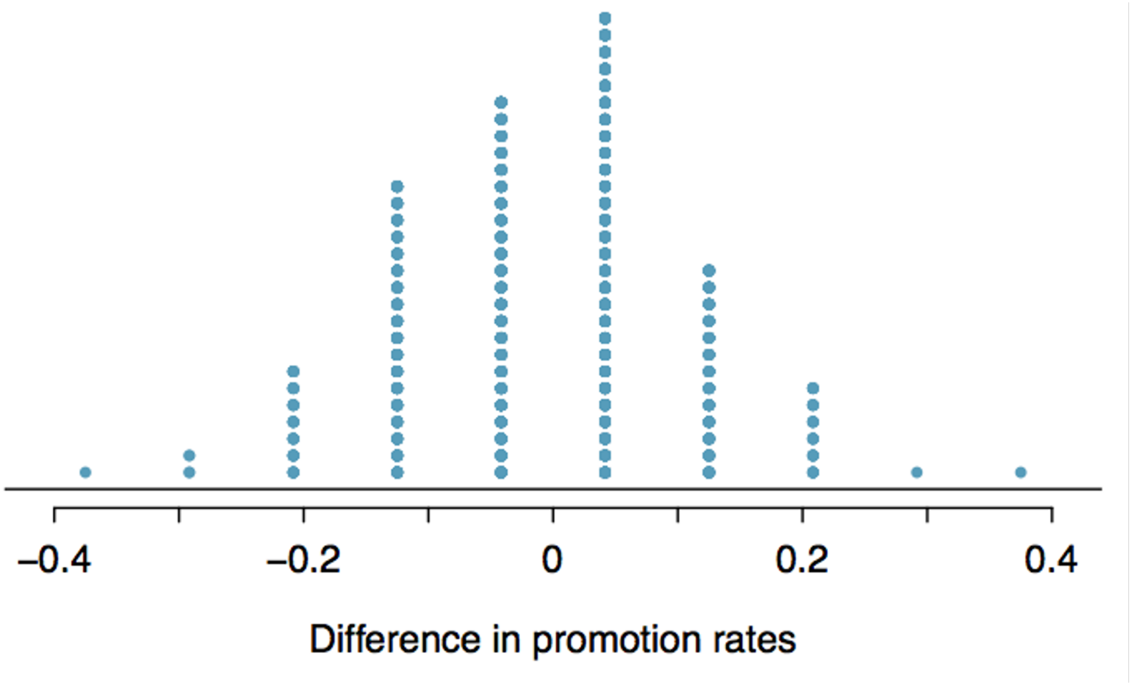
Possible explanations:

- 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.
- 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.

Decision errors

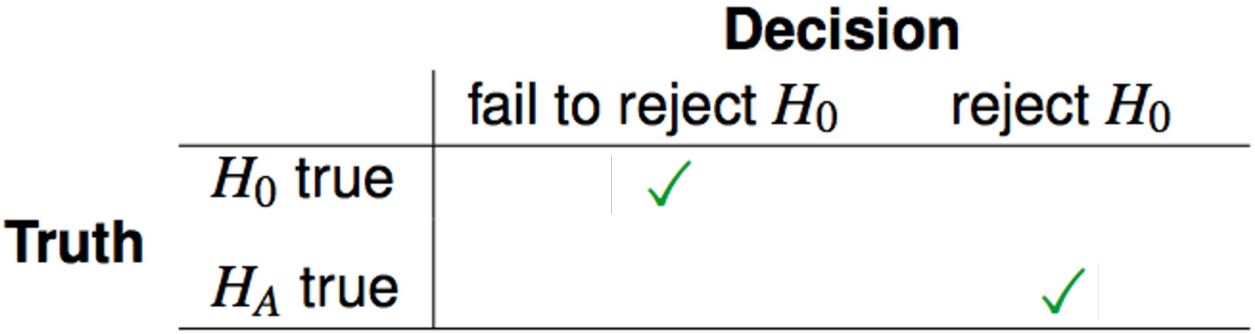
- Hypothesis tests are not flawless.
- In the court system innocent people are sometimes wrongly convicted, and the guilty sometimes walk free.
- Similarly, we can make a wrong decision in statistical hypothesis tests as well.
- The difference is that we have the tools necessary to quantify how often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

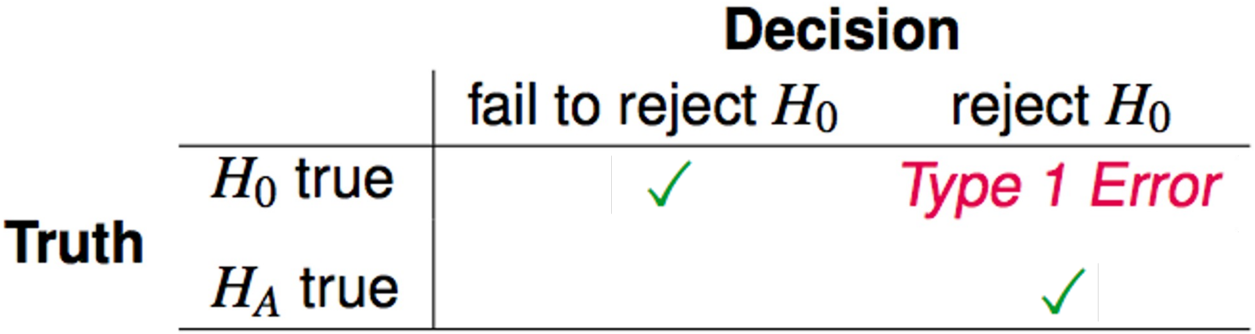
There are two competing hypotheses: the null and the alternative. In a hypothesis test, we make a decision about which might be true, but our choice might be incorrect.



		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.

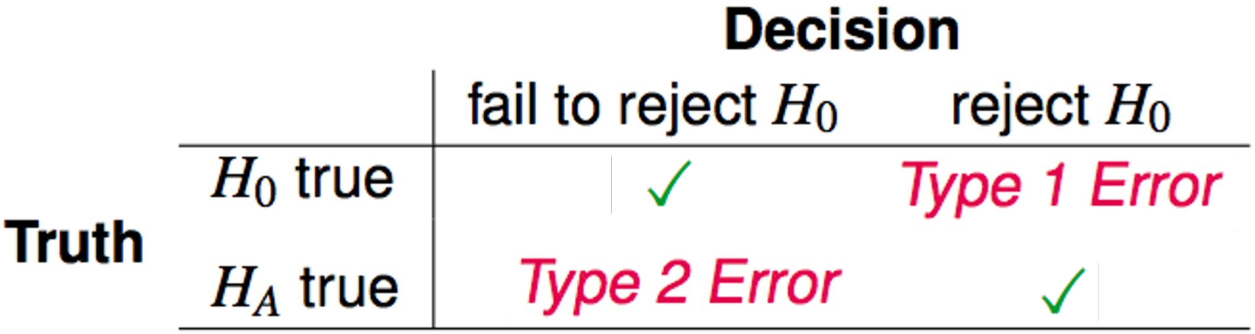


		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true		✓

- 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.

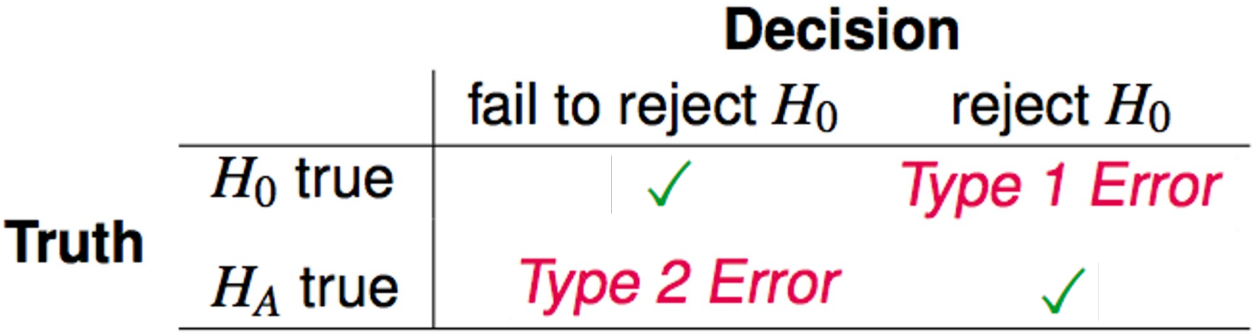


		Decision	
		fail to reject H_0	reject H_0
Truth	H_0 true	✓	*Type 1 Error*
	H_A true	*Type 2 Error*	✓

- 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.

Number of college applications

A survey asked how many colleges students applied to, and 206 students responded to this question. This sample yielded an average of 9.7 college applications with a standard deviation of 7. College Board website states that counselors recommend students apply to roughly 8 colleges. Do these data provide convincing evidence that the average number of colleges all Duke students apply to is higher than recommended?

http://www.collegeboard.com/student/apply/the-application/151680.html

Setting the hypotheses

- The *parameter of interest* is the average number of schools applied to by all Duke students.

Setting the hypotheses

- The *parameter of interest* is the average number of schools applied to by all Duke students.
- There may be two explanations why our sample mean is higher than the recommended 8 schools.
    - The true population mean is different.
    - The true population mean is 8, and the difference between the true population mean and the sample mean is simply due to natural sampling variability

Setting the hypotheses

- The *parameter of interest* is the average number of schools applied to by all Duke students.
- There may be two explanations why our sample mean is higher than the recommended 8 schools.
  - The true population mean is different.
  - The true population mean is 8, and the difference between the true population mean and the sample mean is simply due to natural sampling variability
- We start with the assumption the average number of colleges Duke students apply to is 8 (as recommended)

$$H_0 : \boldsymbol{\mu} = 8$$

Setting the hypotheses

- The *parameter of interest* is the average number of schools applied to by all Duke students.
- There may be two explanations why our sample mean is higher than the recommended 8 schools.
  - The true population mean is different.
  - The true population mean is 8, and the difference between the true population mean and the sample mean is simply due to natural sampling variability
- We start with the assumption the average number of colleges Duke students apply to is 8 (as recommended)

$$H_0 : \boldsymbol{\mu} = 8$$

- We test the claim that the average number of colleges Duke students apply to is greater than 8

$$H_A : \boldsymbol{\mu} > 8$$

Number of college applications - conditions

a) Students in the sample should be independent of each other with respect to how many colleges they applied to.

b) Sampling should have been done randomly.

c) The sample size should be more than 30

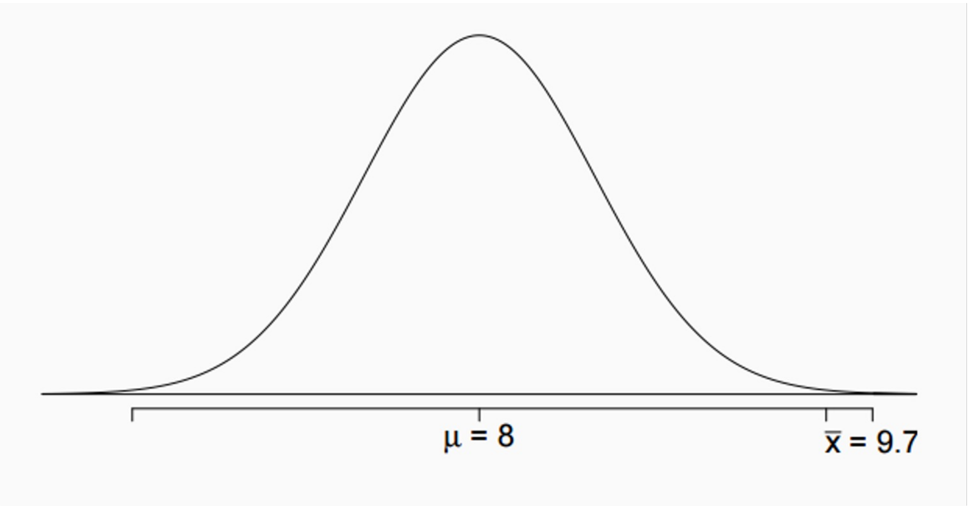
d) The distribution of the number of colleges students apply to should not be extremely skewed.

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*.

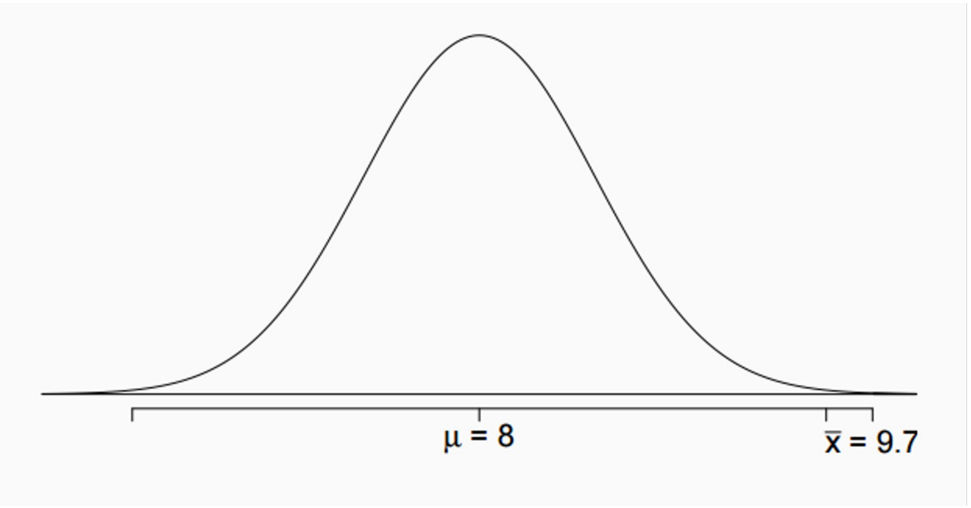
Test Statistic

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*.



Test Statistic

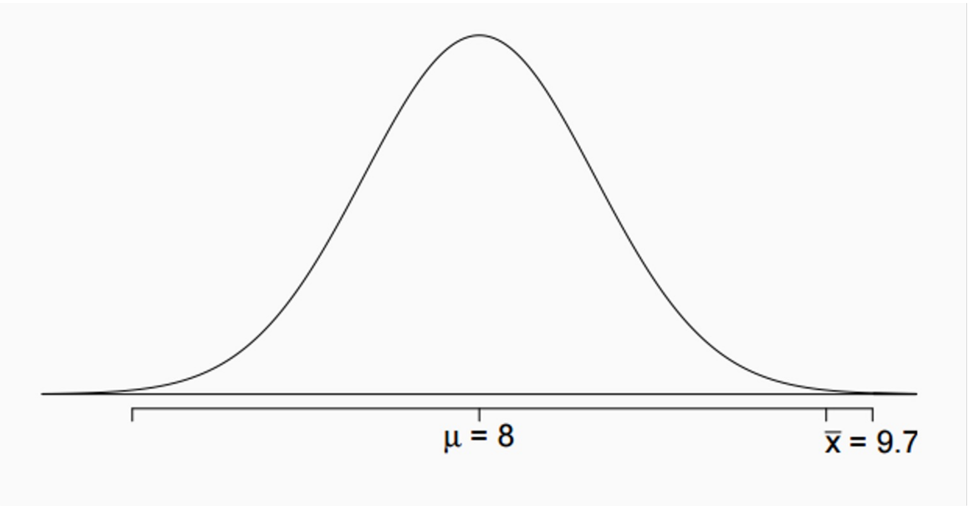
In order to evaluate if the observed sample mean is unusual for the hypothesized sampling distribution, we determine how many standard errors away from the null it is, which is also called the *test statistic*.



$$\bar{x} \sim N\left(\mu = 8, SE = \frac{7}{\sqrt{206}} = 0.5\right)$$

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*.

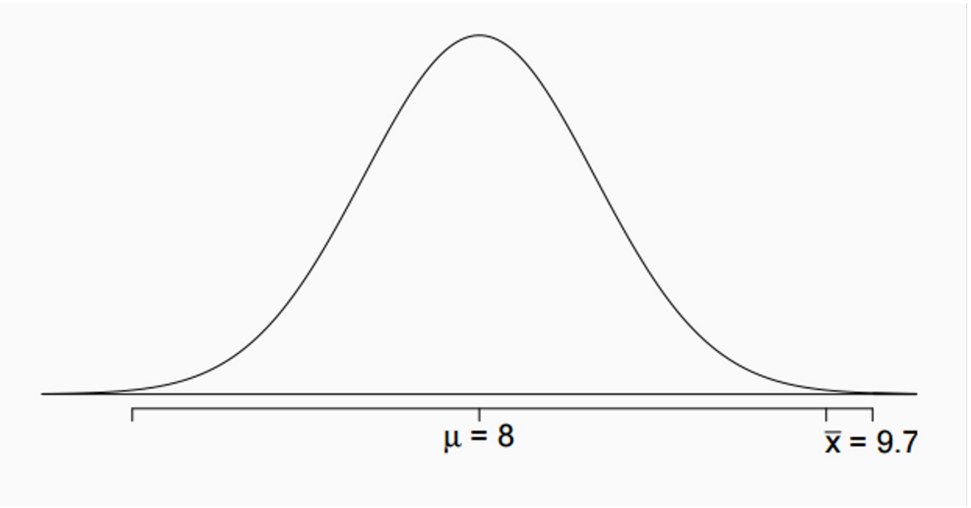


$$\bar{x} \sim N\left(\mu = 8, SE = \frac{7}{\sqrt{206}} = 0.5\right)$$

$$Z = \frac{9.7 - 8}{0.5} = 3.4$$

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*.



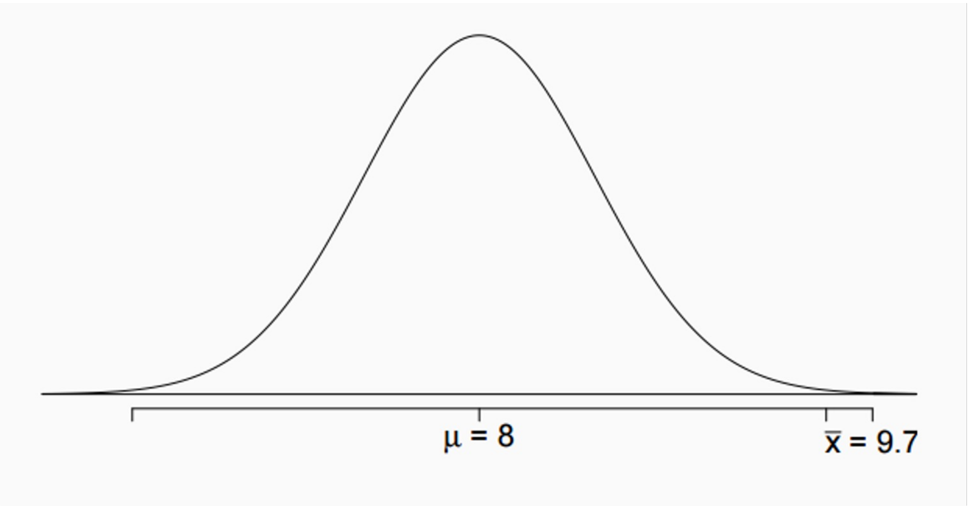
The sample mean is 3.4 standard errors away from the hypothesized value. Is this considered unusually high? That is, is the result *statistically significant*?

$$\bar{x} \sim N\left(\mu = 8, SE = \frac{7}{\sqrt{206}} = 0.5\right)$$

$$Z = \frac{9.7 - 8}{0.5} = 3.4$$

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*.



$$\bar{x} \sim N\left(\mu = 8, SE = \frac{7}{\sqrt{206}} = 0.5\right)$$

$$Z = \frac{9.7 - 8}{0.5} = 3.4$$

The sample mean is 3.4 standard errors away from the hypothesized value. Is this considered unusually high? That is, is the result *statistically significant*?

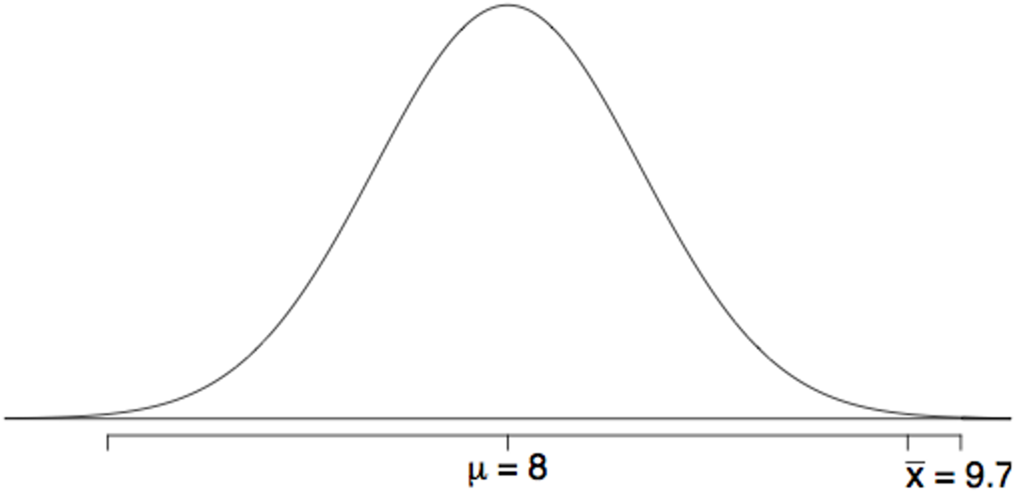
Yes, and we can quantify how unusual it is using a p-value.

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*.

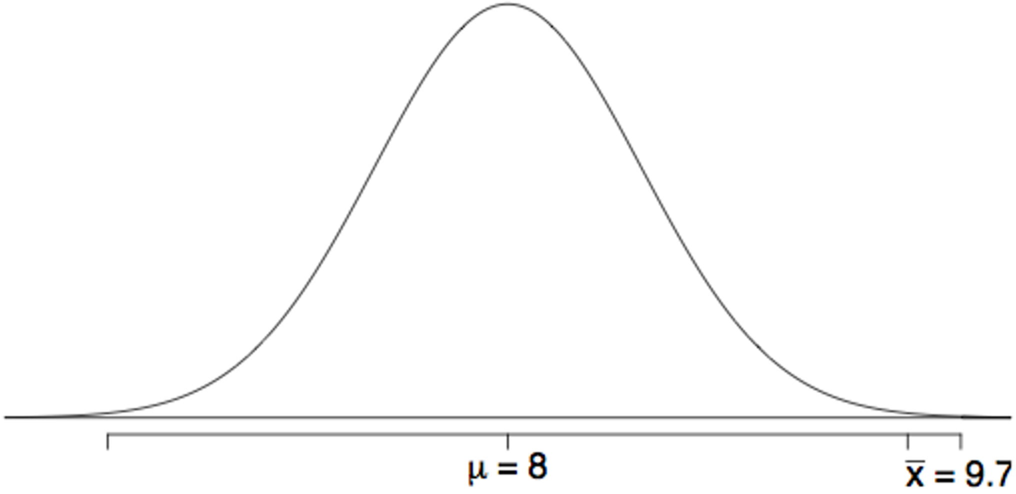
Number of college applications - p-value

p-value: probability of observing data at least as favorable to H_A as our current data set (a sample mean greater than 9.7), if in fact H_0 were true (the true population mean was 8).



Number of college applications - p-value

p-value: probability of observing data at least as favorable to H_A as our current data set (a sample mean greater than 9.7), if in fact H_0 were true (the true population mean was 8).



$$P(\bar{x} > 9.7 \mid \mu = 8) = P(Z > 3.4) = 0.0003$$

Number of college applications - Making a decision

- p-value = 0.0003

Number of college applications - Making a decision

- p-value = 0.0003
    - If the true average of the number of colleges Duke students applied to is 8, there is only 0.03% chance of observing a random sample of 206 Duke students who on average apply to 9.7 or more schools.

Number of college applications - Making a decision

- p-value = 0.0003
  - If the true average of the number of colleges Duke students applied to is 8, there is only 0.03% chance of observing a random sample of 206 Duke students who on average apply to 9.7 or more schools.
  - This is a pretty low probability for us to think that a sample mean of 9.7 or more schools is likely to happen simply by chance.

Number of college applications - Making a decision

- p-value = 0.0003
  - If the true average of the number of colleges Duke students applied to is 8, there is only 0.03% chance of observing a random sample of 206 Duke students who on average apply to 9.7 or more schools.
  - This is a pretty low probability for us to think that a sample mean of 9.7 or more schools is likely to happen simply by chance.
- Since p-value is *low* (lower than 5%) we *reject H_0*.

Number of college applications - Making a decision

- p-value = 0.0003
  - If the true average of the number of colleges Duke students applied to is 8, there is only 0.03% chance of observing a random sample of 206 Duke students who on average apply to 9.7 or more schools.
  - This is a pretty low probability for us to think that a sample mean of 9.7 or more schools 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 Duke students apply to more than 8 schools on average.

Number of college applications - Making a decision

- p-value = 0.0003
  - If the true average of the number of colleges Duke students applied to is 8, there is only 0.03% chance of observing a random sample of 206 Duke students who on average apply to 9.7 or more schools.
  - This is a pretty low probability for us to think that a sample mean of 9.7 or more schools 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 Duke students apply to more than 8 schools on average.
- The difference between the null value of 8 schools and observed sample mean of 9.7 schools is *not due to chance* or sampling variability.

Recap: Hypothesis testing framework

1. Set the hypotheses.
2. Check assumptions and conditions.
3. Calculate a *test statistic* and a p-value.
4. Make a decision, and interpret it in context of the research question.

Recap: Hypothesis testing for a population mean

1. Set the hypotheses

- H_0: μ = null value
- H_A: μ < or > or ≠ null value

2. Calculate the point estimate

3. Check assumptions and conditions

- Independence: random sample/assignment
- Normality: nearly normal population or $n \geq 30$, no extreme skew -- or use the *t* distribution (Ch 5)

4. Calculate a *test statistic* and a p-value (draw a picture!)

$$Z = \frac{\bar{x} - \mu}{SE}, \; where \; SE = \frac{s}{\sqrt{n}}$$

5. Make a decision, and in

- If p-value < α, reject H_0, data provide evidence for H_A
- If p-value > α, do not reject H_0, data do not provide evidence for H_A