Section 001
CDS 101 *Introduction to Computational and Data Sciences*

Class 20: Foundations for inference I

November 6, 2017

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.

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?

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 relatonship 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 relatonship between promotion and gender?

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

% of males promoted: $21/24 = 0.875$

% of females promoted: $14/24 = 0.583$

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).

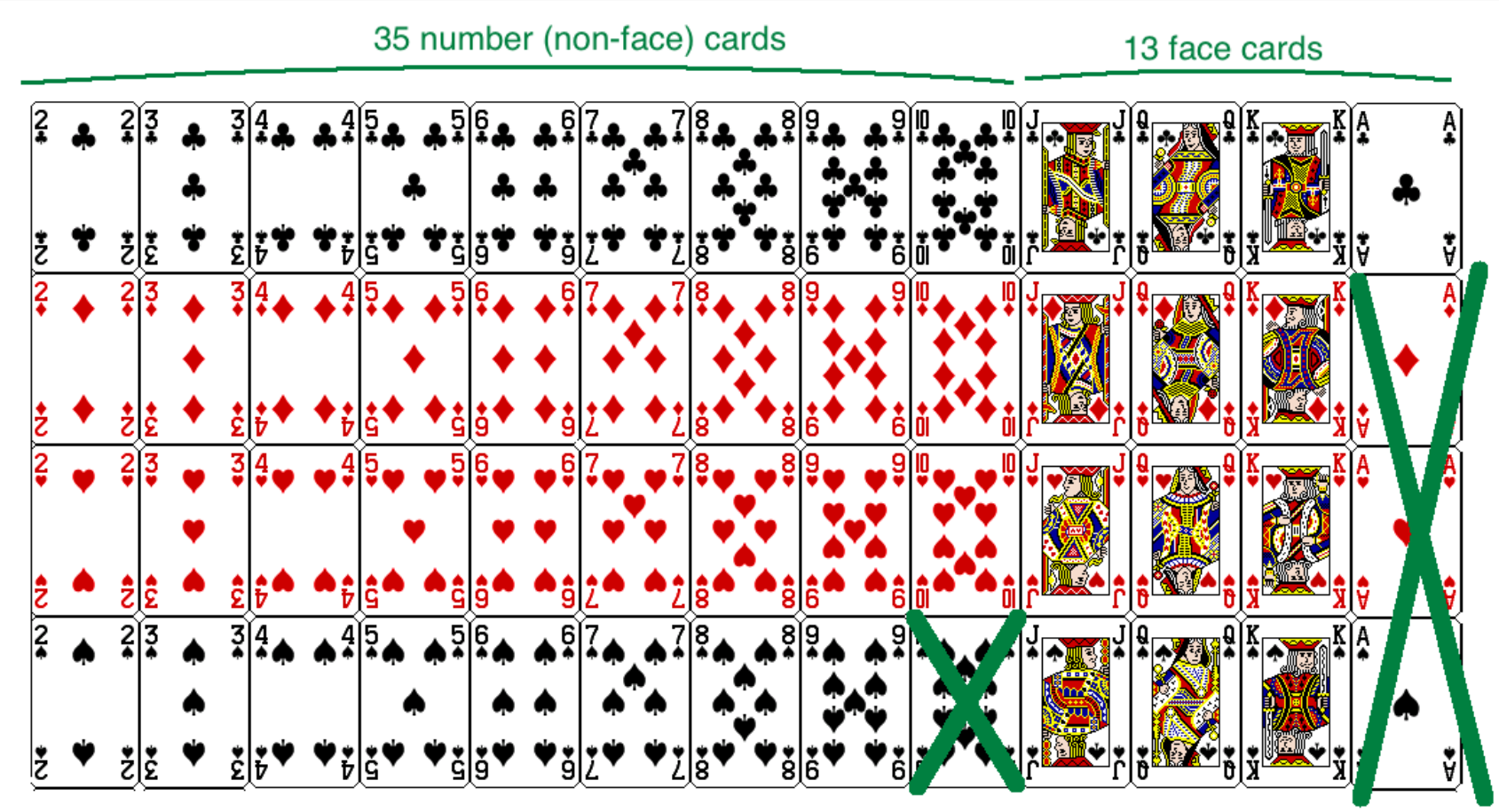
If the results from the simulations based on the chance model do not look like the data, then we can determine that the difference between the proportions of promoted files between males and females was not due to chance, but *due to an actual effect of gender* (promotion and gender are dependent).

Simulating the experiment with a deck of cards

Pretend for a moment that we didn't have our computers available, how could we simulate this experiment using playing cards?

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 $\rightarrow$ there are exactly 13 face cards left in the deck (face cards: A, K, Q, J).
    - Take out a number card $\rightarrow$ 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.

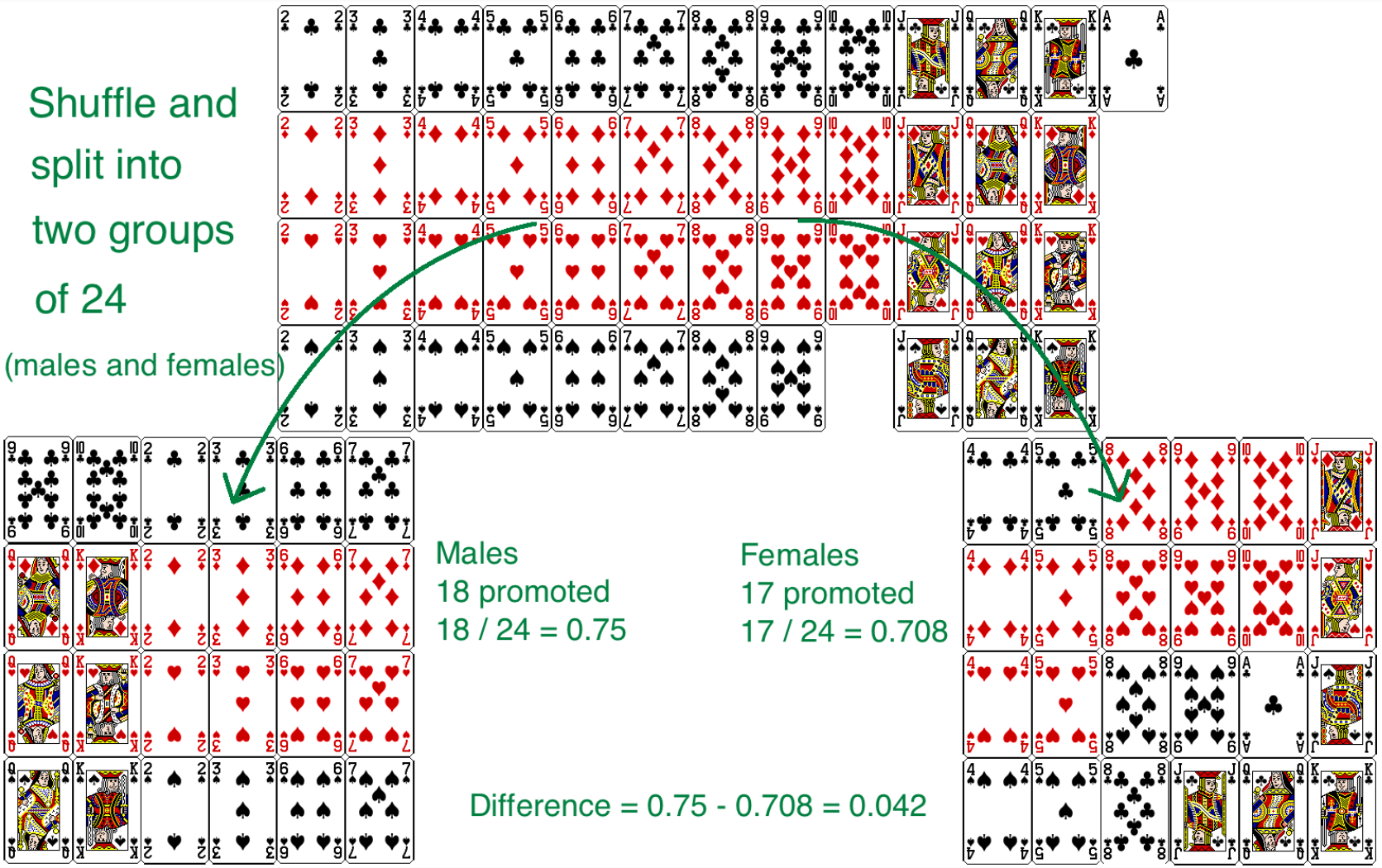
Step 1



Step 2 - 4

Shuffle and split into two groups of 24

(males and females)



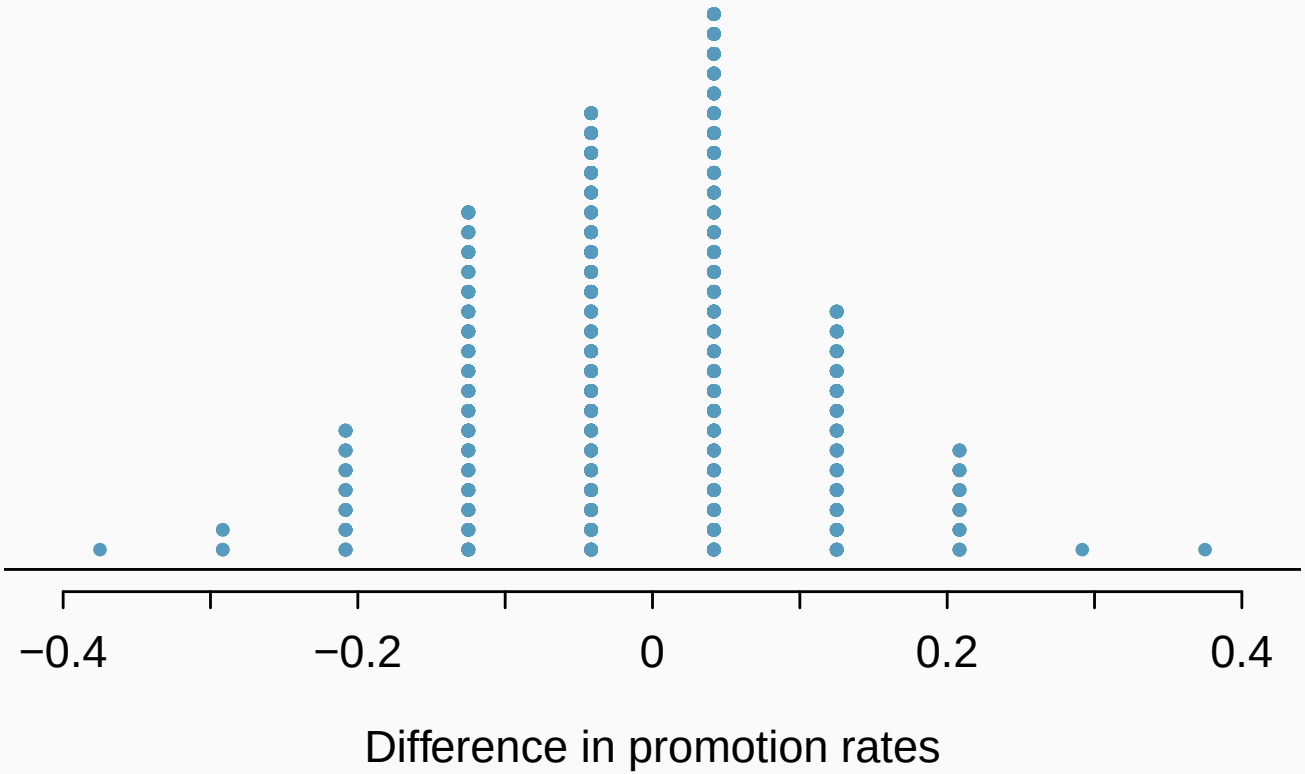
Males
18 promoted
18 / 24 = 0.75

Females
17 promoted
17 / 24 = 0.708

Difference = 0.75 - 0.708 = 0.042

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.



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.

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.*

Hypothesis testing

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 GMU 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 GMU students.

Setting the hypotheses

- The *parameter of interest* is the average number of schools applied to by all GMU 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 GMU 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 GMU students apply to is 8 (as recommended)

$$H_0 : \mu = 8$$

Setting the hypotheses

- The *parameter of interest* is the average number of schools applied to by all GMU 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 GMU students apply to is 8 (as recommended)

$$H_0 : \mu = 8$$

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

$$H_A : \mu > 8$$

Statistical significance

Say that we conducted this study by polling an independent and representative sample of GMU students about how many colleges they applied to, and obtained a sample mean of 9.7.

Statistical significance

Say that we conducted this study by polling an independent and representative sample of GMU students about how many colleges they applied to, and obtained a sample mean of 9.7.

The national average is 8.

Statistical significance

Say that we conducted this study by polling an independent and representative sample of GMU students about how many colleges they applied to, and obtained a sample mean of 9.7.

The national average is 8.

Is this result statistically significant?

Statistical significance

Say that we conducted this study by polling an independent and representative sample of GMU students about how many colleges they applied to, and obtained a sample mean of 9.7.

The national average is 8.

Is this result statistically significant?

In order to evaluate if the observed sample mean is unusual for the hypothesized sampling distribution, we do the following:

- Choose a value for the significance level α (a common choice is 5%)

Statistical significance

Say that we conducted this study by polling an independent and representative sample of GMU students about how many colleges they applied to, and obtained a sample mean of 9.7.

The national average is 8.

Is this result statistically significant?

In order to evaluate if the observed sample mean is unusual for the hypothesized sampling distribution, we do the following:

- Choose a value for the significance level α (a common choice is 5%)
- Determine the percentile rank of the observed sample mean relative to the null distribution

p-values

- We then use the percentile 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.

p-values

- We then use the percentile 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 *lower* than the significance level α, we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence *reject* H_0.

p-values

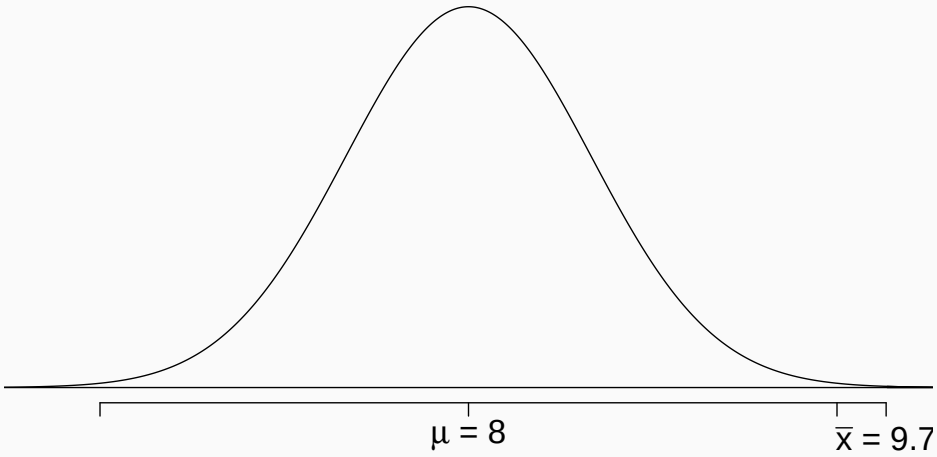
- We then use the percentile to calculate the *p-value*, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.
- If the p-value is *lower* than the significance level α, we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence *reject* H_0.
- If the p-value is *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).

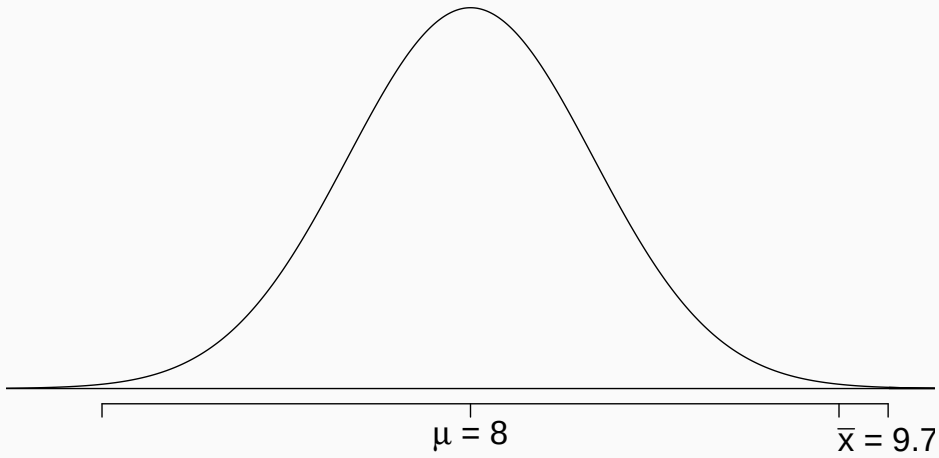
Number of college applications - p-value

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).



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).




```
1 - pnorm(9.7, mean = 8, sd = 7 / sqrt(206))
```

```
[1] 0.0002
```

Number of college applications - Making a decision

- p-value = 0.0002

Number of college applications - Making a decision

- p-value = 0.0002
  - If the true average of the number of colleges GMU students applied to is 8, there is only 0.02% chance of observing a random sample of 206 GMU students who on average apply to 9.7 or more schools.

Number of college applications - Making a decision

- p-value = 0.0002
  - If the true average of the number of colleges GMU students applied to is 8, there is only 0.02% chance of observing a random sample of 206 GMU 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 = 0.0002
  - If the true average of the number of colleges GMU students applied to is 8, there is only 0.02% chance of observing a random sample of 206 GMU 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 = 0.0002
  - If the true average of the number of colleges GMU students applied to is 8, there is only 0.02% chance of observing a random sample of 206 GMU 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 GMU students apply to more than 8 schools on average.

Number of college applications - Making a decision

- p-value = 0.0002
  - If the true average of the number of colleges GMU students applied to is 8, there is only 0.02% chance of observing a random sample of 206 GMU 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 GMU 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.

A poll by the National Sleep Foundation found that college students average about 7 hours of sleep per night. A sample of 169 college students taking an introductory statistics class yielded an average of 6.88 hours, with a standard deviation of 0.94 hours. Assuming that this is a random sample representative of all college students *(bit of a leap of faith?)*, a hypothesis test was conducted to evaluate if college students on average sleep less than 7 hours per night. The p-value for this hypothesis test is 0.0485. Which of the following is correct?

(a) Fail to reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.

(b) Reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.

(c) Reject H_0, the data prove that college students sleep more than 7 hours on average.

(d) Fail to reject H_0, the data do not provide convincing evidence that college students sleep less than 7 hours on average.

(e) Reject H_0, the data provide convincing evidence that college students in this sample sleep less than 7 hours on average.

A poll by the National Sleep Foundation found that college students average about 7 hours of sleep per night. A sample of 169 college students taking an introductory statistics class yielded an average of 6.88 hours, with a standard deviation of 0.94 hours. Assuming that this is a random sample representative of all college students *(bit of a leap of faith?)*, a hypothesis test was conducted to evaluate if college students on average sleep less than 7 hours per night. The p-value for this hypothesis test is 0.0485. Which of the following is correct?

(a) Fail to reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.

(b) *Reject H_0, the data provide convincing evidence that college students sleep less than 7 hours on average.*

(c) Reject H_0, the data prove that college students sleep more than 7 hours on average.

(d) Fail to reject H_0, the data do not provide convincing evidence that college students sleep less than 7 hours on average.

(e) Reject H_0, the data provide convincing evidence that college students in this sample sleep less than 7 hours on average.

Two-sided hypothesis testing with p-values

- If the research question was “Do the data provide convincing evidence that the average amount of sleep college students get per night is *different* than the national average?”, the alternative hypothesis would be different.

$$H_0 : \mu = 7$$

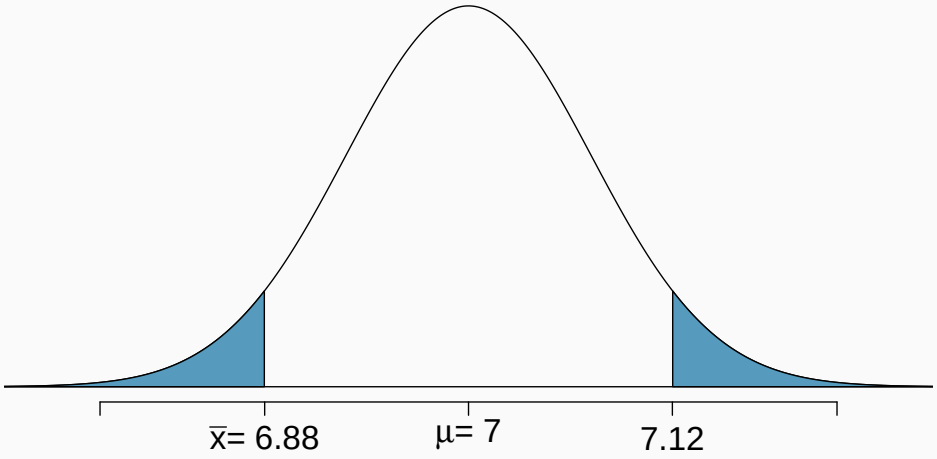
$$H_A : \mu \neq 7$$

Two-sided hypothesis testing with p-values

- If the research question was “Do the data provide convincing evidence that the average amount of sleep college students get per night is *different* than the national average?”, the alternative hypothesis would be different.

$$H_0 : \mu = 7$$
$$H_A : \mu \neq 7$$

- Hence the p-value would change as well:



p-value
$= 0.0485 \times 2$
$= 0.097$