STA130H1F

Class #5

Prof. Nathalie Moon

2018-10-15

Announcement

Midterm text next Friday (October 26th) during the tutorial time

- More details on the course website
- We'll talk about it more in next week's class
- To prepare:
  - review course notes
  - review practice problems
  - review last year's midterms (posted on the website)

Review: hypothesis testing

Criminal trial analogy

- The person on trial must be judged: guilty or not guilty
- Initially: presumption of innocence
- Only **strong evidence** to the contrary leads to rejecting the *not guilty* claim in favour of a *guilty* verdict
- The phrase *beyond a reasonable doubt* is used to set the cutoff value for when enough data has been given to convict

→ how strong is our evidence

"moderate evidence"

strong

very strong

Review: hypothesis testing

Criminal trial analogy

- The decision about the population parameter must be judged to follow one of two hypotheses: H_0 or H_A
- Initially: assume H_0 is true
- We only reject H_0 if we have strong enough evidence that H_0 is false - otherwise, we can't reject H_0.
- The threshold for *beyond a reasonable doubt* is what is known as the *significance level*. Often this is 0.05 or 0.01. If we want to require more evidence before rejecting H_0, we pick a stricter (smaller) value for the significance level

Last class

Testing:

- $H_0: p = p_0$ vs
- $H_A: p \neq p_0$

p_0 is some # between 0 and 1 (a proportion)

Steps:

1. Formulate hypotheses
2. Calculate test statistic
3. Simulate data under the null hypothesis
4. Calculate the p-value
5. Make a conclusion

Today's Class

Question: *If we see a difference between two groups, is it meaningful? Or could it just be due to chance?*

- Comparing two proportions
- Comparing two means
- Type I and Type II errors
- Interpretation of p-values

Comparing two proportions

Example 1: Is Tylenol or Aspirin Better for Headache Relief?

- Consider a sample of four people with a headache.
- Two people are randomly assigned to Aspirin and two assigned to Tylenol. This is called **randomization**.
- This randomization could be carried out by shuffling a deck of four cards: 2 cards marked with T and 2 cards with A, and assigning each person a card.

Is Tylenol or Aspirin Better for Headache Relief?

- Red card -> Tylenol
- Black card -> Aspirin
- After an hour a researcher asked if they still had pain.

Subject	Drug	Pain
1	T	No
2	T	No
3	A	Yes
4	A	No

Tylenol

Aspirin

both ok

← pain

Is Tylenol or Aspirin Better for Headache Relief?

- Red card -> Tylenol
- Black card -> Aspirin
- After an hour a researcher asked if they still had pain.

Subject	Drug	Pain
1	T	No
2	T	No
3	A	Yes
4	A	No

$\hat{p}_T$ = prop. who took Tylenol and are pain free = 1

$\hat{p}_A = 0.5$

Question: Do Tylenol and Aspirin have a different effect on headache relief?

$\Rightarrow$ not just due to random chance

Is Tylenol or Aspirin Better for Headache Relief?

- The null hypothesis is that changing the treatment for a subject has no effect on curing headaches (in other words, both treatments have the same effect)

Is Tylenol or Aspirin Better for Headache Relief?

- The null hypothesis is that changing the treatment for a subject has no effect on curing headaches (in other words, both treatments have the same effect)
- Assuming the null hypothesis is true, Tylenol (T) and Aspirin (A) are mere labels and don't affect the outcome.

Is Tylenol or Aspirin Better for Headache Relief?

- The null hypothesis is that changing the treatment for a subject has no effect on curing headaches (in other words, both treatments have the same effect)
- Assuming the null hypothesis is true, Tylenol (T) and Aspirin (A) are mere labels and don't affect the outcome.
- For example, subject 1 would have no pain whether they had Tylenol or Aspirin.

Is Tylenol or Aspirin Better for Headache Relief?

- The null hypothesis is that changing the treatment for a subject has no effect on curing headaches (in other words, both treatments have the same effect)
- Assuming the null hypothesis is true, Tylenol (T) and Aspirin (A) are mere labels and don't affect the outcome.
- For example, subject 1 would have no pain whether they had Tylenol or Aspirin.
- The alternative hypothesis is that the proportion of subjects without pain is different for Tylenol and Aspirin.

Is Tylenol or Aspirin Better for Headache Relief?

H_0: no difference between T & A

- The null hypothesis is that changing the treatment for a subject has no effect on curing headaches (in other words, both treatments have the same effect)
- Assuming the null hypothesis is true, Tylenol (T) and Aspirin (A) are mere labels and don't affect the outcome.
- For example, subject 1 would have no pain whether they had Tylenol or Aspirin.
- The alternative hypothesis is that the proportion of subjects without pain is different for Tylenol and Aspirin.
- Test statistic: $\hat{p}_T - \hat{p}_A$

Is Tylenol or Aspirin Better for Headache Relief?

All the possible ways to assign two subject to Aspirin and two subjects to Tylenol.

Under H_0, we can shuffle the 'T' and 'A' labels

Subject	Drug	Pain	R1	R2	R3	R4	R5
1	T	No	T	T	A	A	A
2	T	No	A	A	T	T	A
3	A	Yes	T	A	T	A	T
4	A	No	A	T	A	T	T
$\hat{p}_T$		1	0.5	1	0.5	1	0.5
$\hat{p}_A$		0.5	1	0.5	1	0.5	1
$\hat{p}_T - \hat{p}_A$		0.5	-0.5	0.5	-0.5	0.5	-0.5

Example 2: Gender Bias in Promotion

- 1972 study on "sex role stereotypes on personnel decisions".
- 48 male managers were asked to rate whether several candidates were suitable for promotion.
- Managers were randomly assigned to review the file of either a male or female candidate. The files were otherwise identical.

B. Rosen and T.H. Jerdee (1974). Influence of sex role stereotypes on personnel decisions. *Journal of Applied Psychology* **59**(1), 9-14.

What they found

Observed results	Male	Female	Total
Promoted	21	14	35
Not promoted	3	10	13
Total	24	24	48

What they found

Observed results	Male	Female	Total
Promoted	21	14	35
Not promoted	3	10	13
Total	24	24	48

- 21/24 = 87.5% of males were recommended for promotion
- 14/24 = 58.3% of females were recommended for promotion
- This suggested that men were more likely to be recommended for promotion... but the sample size is small! Is the difference 87.5% - 58.3% = 29.2% due to gender or due to chance?
- If many similar studies were conducted, assuming there is no difference between male and female promotion rates, how many of these studies would produce a difference as extreme as this study?

Review: The Logic of Hypothesis Testing

1. The hypotheses

Two claims:

1.There is nothing going on. This is the **null hypothesis**, written .

For the gender bias in promotion study: males and females

$H_0: p_m = p_f$ or $p_m - p_f = 0$ or $p_f - p_m = 0$

where p_m = prop. of males who get promoted, p_f is prop. of females who get promoted

2.There is something going on. This is the **alternative hypothesis**, written H_A (or H_a or H_1).
The alternative is almost always what the research wants to find evidence for.

For the gender bias in promotion study:

$H_A: p_m \neq p_f$

2. The test statistic

The **test statistic** is a number, calculated from the data, that captures what we're interested in.

For the gender bias promotion example, what would be a useful test statistic?

$\hat{p}_m - \hat{p}_f = 0.292$

2. The test statistic

The **test statistic** is a number, calculated from the data, that captures what we're interested in.

For the gender bias promotion example, what would be a useful test statistic?

Is it possible that the value of the test statistic occured just by chance and there was really no difference between genders in being recommended for promotion?

To answer this, simulate possible values of the test statistic assuming there's no difference (i.e., the null hypothesis is true).

3. Simulate what H_0 predicts will happen

- If H_0 is true, then females and males are equally likely to be promoted

3. Simulate what H_0 predicts will happen

- If H_0 is true, then females and males are equally likely to be promoted
- Imagine we had 24 cards labelled with "F" and 24 labelled with "M"

3. Simulate what predicts will happen

- If is true, then females and males are equally likely to be promoted
- Imagine we had 24 cards labelled with "F" and 24 labelled with "M"
- Shuffle the cards...

3. Simulate what predicts will happen

- If is true, then females and males are equally likely to be promoted
- Imagine we had 24 cards labelled with "F" and 24 labelled with "M"
- Shuffle the cards...
- Assign the cards to the 48 people, then calculate the difference in the proportion of males vs females that were promoted
  - this is one *simulated* value of the test statistic $\Rightarrow \hat{p}_m - \hat{p}_f$ (simulated)

3. Simulate what predicts will happen

- Shuffle the cards again...

3. Simulate what predicts will happen

- Shuffle the cards again...
- Assign the cards to the 48 people, then calculate the difference in the proportion of males vs females that were promoted
  - this is another *simulated* value of the test statistic

recalculate $\hat{p}_m - \hat{p}_f$

3. Simulate what predicts will happen

- Shuffle the cards again...
- Assign the cards to the 48 people, then calculate the difference in the proportion of males vs females that were promoted
    - this is another *simulated* value of the test statistic

Repeat many times:

1. Shuffle
2. Assign cards
3. Calculate difference

Gender bias data

Data are in the dataframe `bias` (which I created)

```
glimpse(bias)
```

```
## Observations: 48
## Variables: 2
## $ gender   <chr> "male", "male", "male", "male", "male", "male", "male...
## $ promoted <chr> "yes", "yes", "yes", "yes", "yes", "yes", "yes", "yes...
```

- How many variables are in the data frame?

2 : gender and promoted

- Are the variables numerical or categorical?

both are categorical.

Code to calculate the proportion of males and females promoted

```
n_female <- bias %>% filter(gender=="female") %>% count()
n_male <- bias %>% filter(gender=="male") %>% count()

yes_female <- bias %>%
  filter(promoted=="yes" & gender=="female") %>% count()
as.numeric(yes_female) # treat as a number (not a dataframe)
```

```
## [1] 14
```

```
yes_male <- bias %>%
  filter(promoted=="yes" & gender=="male") %>%
  count()
as.numeric(yes_male)
```

```
## [1] 21
```

```
p_diff <- yes_female/n_female - yes_male/n_male
as.numeric(p_diff)
```

```
## [1] -0.2916667
```

Is the difference between the proportion of males and females promoted meaningful?

- The difference in the proportions of people who were deemed suitable for promotion between the females and males is

$$-0.292$$

- This suggests that the males were more likely to be recommended for promotion.
- But the sample size is small. Could this difference just be due to chance?
- Repeat the experiment assuming it's just due to chance (using simulation), and see what happens

How to shuffle "Gender" in R

other arguments
- prob
- replace
- size.

The `sample()` command by default produces a random sample of the same length of the data without replacement

```
# illustration of sample
a_vector <- c(1,1,1,2,2)
a_vector
```

vector of length 5

```
## [1] 1 1 1 2 2
```

```
sample(a_vector)
```

```
## [1] 1 2 2 1 1
```

```
sample(a_vector)
```

```
## [1] 1 1 1 2 2
```

```
sample(a_vector)
```

```
## [1] 2 2 1 1 1
```

Before the shuffle

```
bias$gender # the values of gender in the data
```

```
##  [1] "male"   "male"   "male"   "male"   "male"   "male"   "male"
##  [8] "male"   "male"   "male"   "male"   "male"   "male"   "male"
## [15] "male"   "male"   "male"   "male"   "male"   "male"   "male"
## [22] "male"   "male"   "male"   "female" "female" "female" "female"
## [29] "female" "female" "female" "female" "female" "female" "female"
## [36] "female" "female" "female" "female" "female" "female" "female"
## [43] "female" "female" "female" "female" "female" "female"
```

```
bias$promoted
```

```
##  [1] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
## [12] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "no"
## [23] "no"  "no"  "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
## [34] "yes" "yes" "yes" "yes" "yes" "no"  "no"  "no"  "no"  "no"  "no"
## [45] "no"  "no"  "no"  "no"
```

After the shuffle

shuffle the group labels
↳ because under H_0, there's no difference between groups.

new simulated data

```
sim <- bias %>% mutate(gender = sample(gender)) #shuffle gender labels
sim$gender
```

↳ no new variable, just shuffling values

```
##  [1] "female" "male"   "male"   "male"   "female" "female" "male"
##  [8] "female" "male"   "female" "male"   "male"   "female" "male"
## [15] "female" "female" "male"   "female" "male"   "male"   "female"
## [22] "male"   "male"   "female" "male"   "female" "female" "female"
## [29] "male"   "male"   "female" "female" "female" "male"   "male"
## [36] "female" "male"   "male"   "female" "male"   "male"   "female"
## [43] "female" "male"   "male"   "female" "female" "female"
```

```
sim$promoted
```

```
##  [1] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
## [12] "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "no"
## [23] "no"  "no"  "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes" "yes"
## [34] "yes" "yes" "yes" "yes" "yes" "no"  "no"  "no"  "no"  "no"  "no"
## [45] "no"  "no"  "no"  "no"
```

After the shuffle: calculate

```
yes_female <- sim %>%
  filter(promoted == "yes" &
  gender == "female") %>%  # only promoted females
  summarize(n())  # count
  as.numeric(yes_female)
```

```
## [1] 17
```

```
  yes_male <- sim %>%
  filter(promoted == "yes" &
  gender == "male") %>%  # only promoted males
  summarize(n())  # count
  as.numeric(yes_male)
```

```
## [1] 18
```

```
  # calculate the difference in the proportion of
  # people promoted by gender
  p_diff <- yes_female / n_female - yes_male / n_male
  as.numeric(p_diff)
```

```
## [1] -0.04166667
```

Set up simulation in R

```
set.seed(130)

repetitions <- 1000  # "many times" will be 1000
# create a vector of missing values to store results
# rep() is the replicate function
# NA means a missing value
simulated_stats <- rep(NA, repetitions) # 1000 missing values

# initialize some values
n_female <- bias %>% filter(gender == "female") %>% summarise(n())
n_male <- bias %>% filter(gender == "male") %>% summarise(n())
```

Calculate the observed value of test statistic

Test statistic

```
# calculate the test statistic
yes_female <- bias %>%
  filter(promoted == "yes" &
           gender == "female") %>%  # only promoted females
  summarize(n())  # count

yes_male <- bias %>%
  filter(promoted == "yes" &
           gender == "male") %>%  # only promoted males
  summarize(n())  # count
test_stat <- as.numeric(yes_female / n_female -
                          yes_male / n_male)
```

Shuffle, Assign, Calculate difference, Repeat...

```
for (i in 1:repetitions)
{
  sim <- bias %>%
    mutate(gender = sample(gender)) # shuffle gender labels

  yes_female <- sim %>%
    filter(promoted == "yes" & gender == "female") %>%
    count()

  yes_male <- sim %>%
    filter(promoted == "yes" & gender == "male") %>%
    count()

  # calculate the difference in the proportion of people
  # promoted by gender in the simulation

  p_diff <- yes_female / n_female - yes_male / n_male

  # add the new simulated value to the ith entry in the
  # vector of results

  simulated_stats[i] <- as.numeric(p_diff) #treat result as a number
}

# turn results into a data frame for plotting
```

where we're shuffling group labels

$\hat{p}_f$ (simulated)

$\hat{p}_m$ (simulated)

$\hat{p}_f - \hat{p}_m$ (simulated)

vector of sim.values

Distribution of simulated values of $\hat{p}_f - \hat{p}_m$ assuming H_0 is true

The for loop on the previous slide produced a simulated distribution of values for $\hat{p}_f - \hat{p}_m$. This can be visualized using a histogram.

vector of $\hat{p}_f - \hat{p}_m$ simulated values

```
ggplot(sim, aes(x = p_diff)) + geom_histogram(binwidth = 0.1,
                                              colour = "black",
                                              fill = "grey")
```

Around what value is this distribution centred? Does this make sense?

0 - makes sense bcs we're assuming H_0.

4. The P-value

- Assuming that the null hypothesis is true, the **P-value** gives a measure of the probability of getting data that are at least as unusual as the sample data.

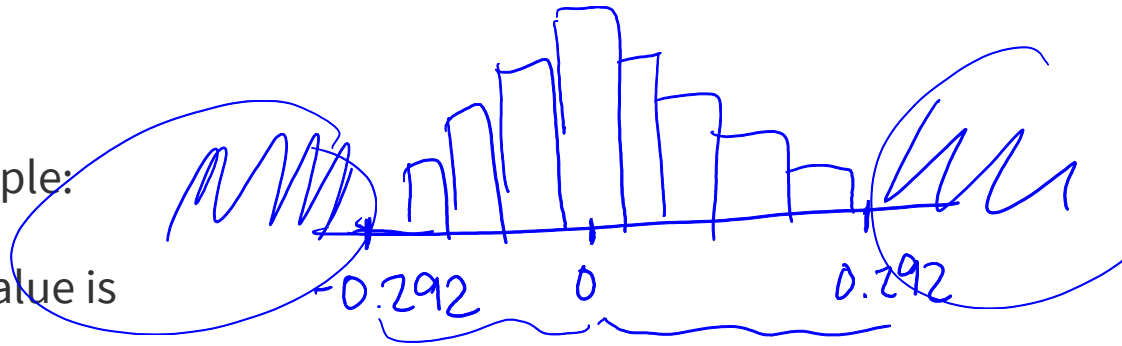
4. The P-value

- Assuming that the null hypothesis is true, the **P-value** gives a measure of the probability of getting data that are at least as unusual as the sample data.
- What does "at least as unusual" mean?

4. The P-value

- Assuming that the null hypothesis is true, the **P-value** gives a measure of the probability of getting data that are at least as unusual as the sample data.
- What does "at least as unusual" mean?
    - Values that are as far away or even farther from the null hypothesis value than the test statistic.

4. The P-value - Gender Bias Example



For the gender bias example:

- the null hypothesis value is
- the observed estimate from the data (the test statistic) is

- values at least as unusual as the data values includes all values *greater than or equal to 0.292* and all values *less than or equal to -0.292*
- This is a **two-sided test** because it considers differences from the null hypothesis that are both larger and smaller than what you observed.

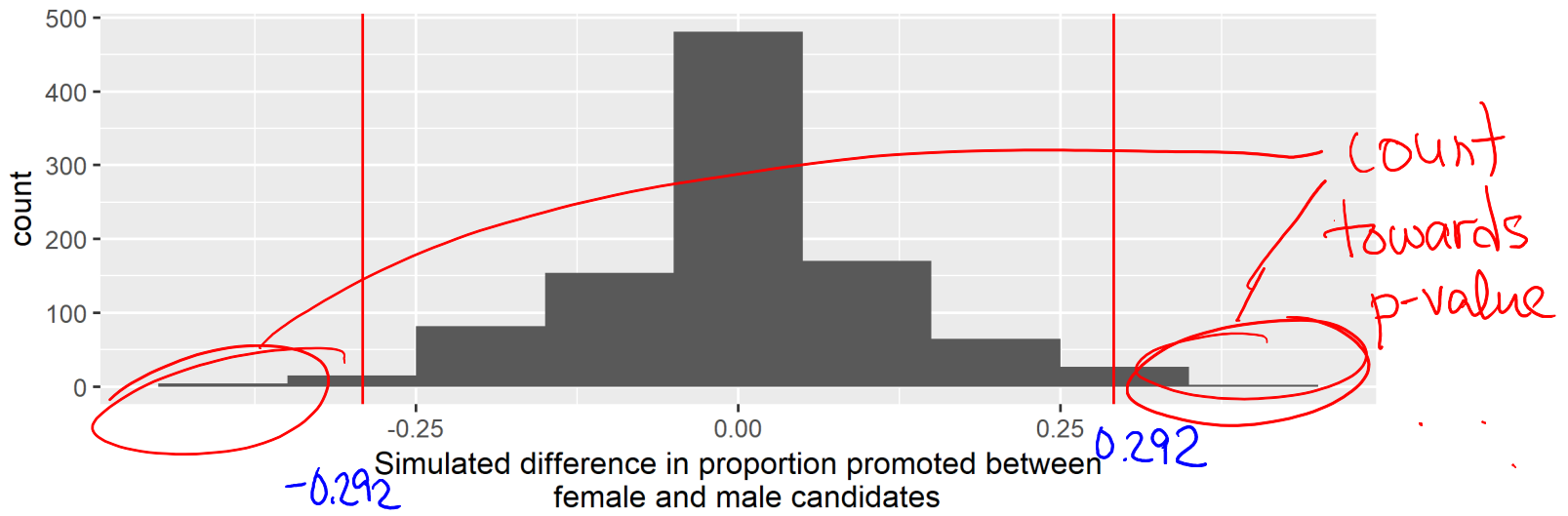
Values more extreme than the test statistic

```
test_stat
```

```
## [1] -0.2916667
```

```
ggplot(sim, aes(p_diff)) +
  geom_histogram(binwidth = 0.1) +
  geom_vline(xintercept = test_stat, color = "red") +
  geom_vline(xintercept = -1 * test_stat, color = "red") +
  labs(x = "Simulated difference in proportion promoted between
  female and male candidates")
```

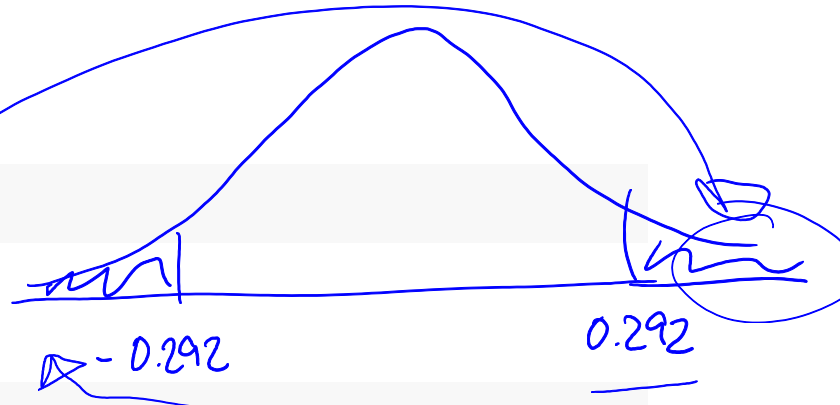
} vertical lines



Calculate P-value

```
test_stat
```

```
## [1] -0.2916667
```




```
extreme_count <- sim %>%
  filter(p_diff >= abs(test_stat) | p_diff <= -1*abs(test_stat)) %>%
  count()

as.numeric(extreme_count)
```

```
## [1] 48
```

```
p_value <- as.numeric(extreme_count)/repetitions
as.numeric(p_value)
```

```
## [1] 0.048
```

5. Make a conclusion

- A small P-value means the data are inconsistent with the null hypothesis.
  - *beyond a reasonable doubt*
- A large P-value means the data are consistent with the null hypothesis.
  - Not enough evidence to be *beyond a reasonable doubt*
  - It's possible that H_A is true, but we don't have *enough* evidence to conclude this

5. Make a conclusion

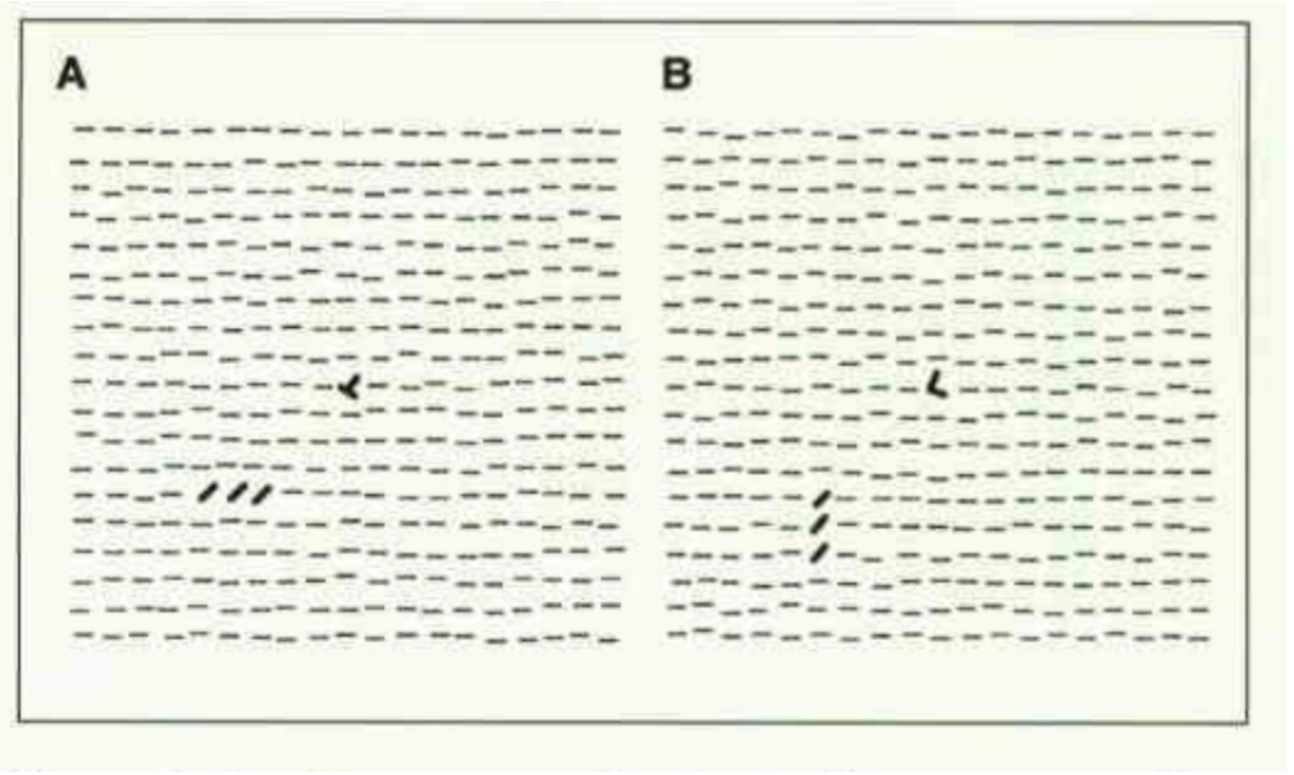
- A small P-value means the data are inconsistent with the null hypothesis.
  - *beyond a reasonable doubt*
- A large P-value means the data are consistent with the null hypothesis.
  - Not enough evidence to be *beyond a reasonable doubt*
  - It's possible that is true, but we don't have *enough* evidence to conclude this
- The P-value is 0.048 for our test that the proportion of people promoted is the same for females and males.
- We conclude that there is moderate evidence of a difference between genders in being chosen for promotion.

↳ reject H_0

Hypothesis testing for comparing a characteristic of a numerical variable between two groups

Example: Sleep and performance on a visual discrimination task

Stickgold, James and Hobson (2000). Visual discrimination learning requires sleep after training. *Nature Neuroscience* **3**(12), 1237-8



Can you recover from an all-nighter after a couple of days of good sleep?

- Subjects: 21 student volunteers (ages 18 to 25)
- Subjects were trained on a visual discrimination task
- Subjects were then randomly assigned into two groups:
  - *sleep deprived*: kept up all night after the training and then not allowed to sleep until 9pm the next day (11 people)
  - *unrestricted sleep*: no restrictions on their sleep (10 people) --
- Subjects then were allowed unrestricted sleep for the next two nights
- Subjects were then retested on the visual discrimination task

The visual discrimination task

- Subjects shown "target screen"" A or B for 17 milliseonds
- Then shown blank screen for a variable length of time, the "interstimulus interval" (ISI)
- Then shown "mask screen" with random pattern for 17 milliseconds
- Asked if target screen included an L or a T and whether the slashes were vertical or horizontal
- Score on the task for a subject was the minimum interstimulus interval (ISI) required for the subject to achieve accurate results

The data

vectors

```
sleep <- c(rep("unrestricted",10),rep("deprived",11))
isi_change <- c(25.2,14.5,-7.0,12.6,34.5,45.6,11.6,18.6,12.1,30.5,
                -10.7,4.5,2.2,21.3,-14.7,-10.7,9.6,2.4,21.8,7.2,10.0)
sleep_data <- data_frame(sleep, isi_change)

sleep_data %>% head()
```

```
## # A tibble: 6 x 2
##   sleep        isi_change
##   <chr>             <dbl>
## 1 unrestricted       25.2
## 2 unrestricted       14.5
## 3 unrestricted       -7
## 4 unrestricted       12.6
## 5 unrestricted       34.5
## 6 unrestricted       45.6
```

The data

```
ggplot(sleep_data, aes(x = isi_change, fill = sleep)) +
  geom_dotplot() +
  xlim(-20, 50) + ylim(0, 5) + facet_wrap( ~ sleep, ncol = 1)  +
  theme_bw()
```

high value of isi_change mean there was a large improvement between the two tests

scores

How is the sleep deprivation study similar to the gender discrimination in promotion study?

both examples involve comparing 2 groups

Hypotheses

- What is an appropriate statistic to capture the difference in `isi_change` between the sleep deprived and unrestricted sleep group?

$\hat{\mu}_1 - \hat{\mu}_2$

where $\hat{\mu}_1$ = mean isi_change in the unrestricted group

$\hat{\mu}_2$ = " " restricted group

Hypotheses

- What is an appropriate statistic to capture the difference in `isi_change` between the sleep deprived and unrestricted sleep group?
- Test whether the mean of the change in ISI is the same for students who are sleep deprived and students who had unrestricted sleep

Hypotheses

- What is an appropriate statistic to capture the difference in `isi_change` between the sleep deprived and unrestricted sleep group?

- Test whether the mean of the change in ISI is the same for students who are sleep deprived and students who had unrestricted sleep

$\hat{\mu}_1 - \hat{\mu}_2$ test statistic

- $\hat{\mu}_1$ is the parameter representing what the mean of the change in ISI would be for all students if they were given this task and had unrestricted sleep.

- $\hat{\mu}_2$ is the parameter representing what the mean of the change in ISI would be for all students if they were given this task and underwent sleep deprivation.

Test statistic

Difference in the means of change in ISI between the sleep deprived and unrestricted sleep groups for the 21 students in our sample of students

```
mean_data <- sleep_data %>% group_by(sleep) %>%
  summarise(means = mean(isi_change))
  mean_data
```

```
## # A tibble: 2 x 2
##   sleep        means
##   <chr>        <dbl>
## 1 deprived       3.9
## 2 unrestricted  19.8
```

```
test_stat <- as.numeric(mean_data %>%
                          summarise(test_stat = diff(means)))
test_stat
```

```
## [1] 15.92
```

← test statistic

Simulate what H_0 predicts will happen

Assume H_0 is true: The value of a subject's ISI is the same if they are in the sleep deprived or unrestricted sleep groups

Simulate what predicts will happen

Assume is true: The value of a subject's ISI is the same if they are in the sleep deprived or unrestricted sleep groups Shuffle, Assign, Calculate test statistic, Repeat:

- **Shuffle:** suffle the categorical variable that says to which sleep group each observation belongs
- **Assign:** the shuffled labels to the subjects
- **Calculate the test statistic:** the difference between the means of change in ISI for the observations in each of the new groups
- **Repeat:** lots of times to obtain an empirical distribution for the test statistic if the null hypothesis were true

Simulate what predicts will happen

Assume is true: The value of a subject's ISI is the same if they are in the sleep deprived or unrestricted sleep groups Shuffle, Assign, Calculate test statistic, Repeat:

- **Shuffle:** suffle the categorical variable that says to which sleep group each observation belongs
- **Assign:** the shuffled labels to the subjects
- **Calculate the test statistic:** the difference between the means of change in ISI for the observations in each of the new groups
- **Repeat:** lots of times to obtain an empirical distribution for the test statistic if the null hypothesis were true

After the distribution of simulated values is obtained, compare the test statistic observed from the data to the empirical distribution

One value of what the test statistic could be if the null hypothesis were true

```
sim <- sleep_data %>%
  mutate(sleep = sample(sleep)) # shuffle sleep group labels

sim %>%
  group_by(sleep) %>%
  summarise(means = mean(isi_change)) %>%
  summarise(sim_test_stat = diff(means))
```

```
## # A tibble: 1 x 1
##   sim_test_stat
##           <dbl>
## 1         -3.09
```

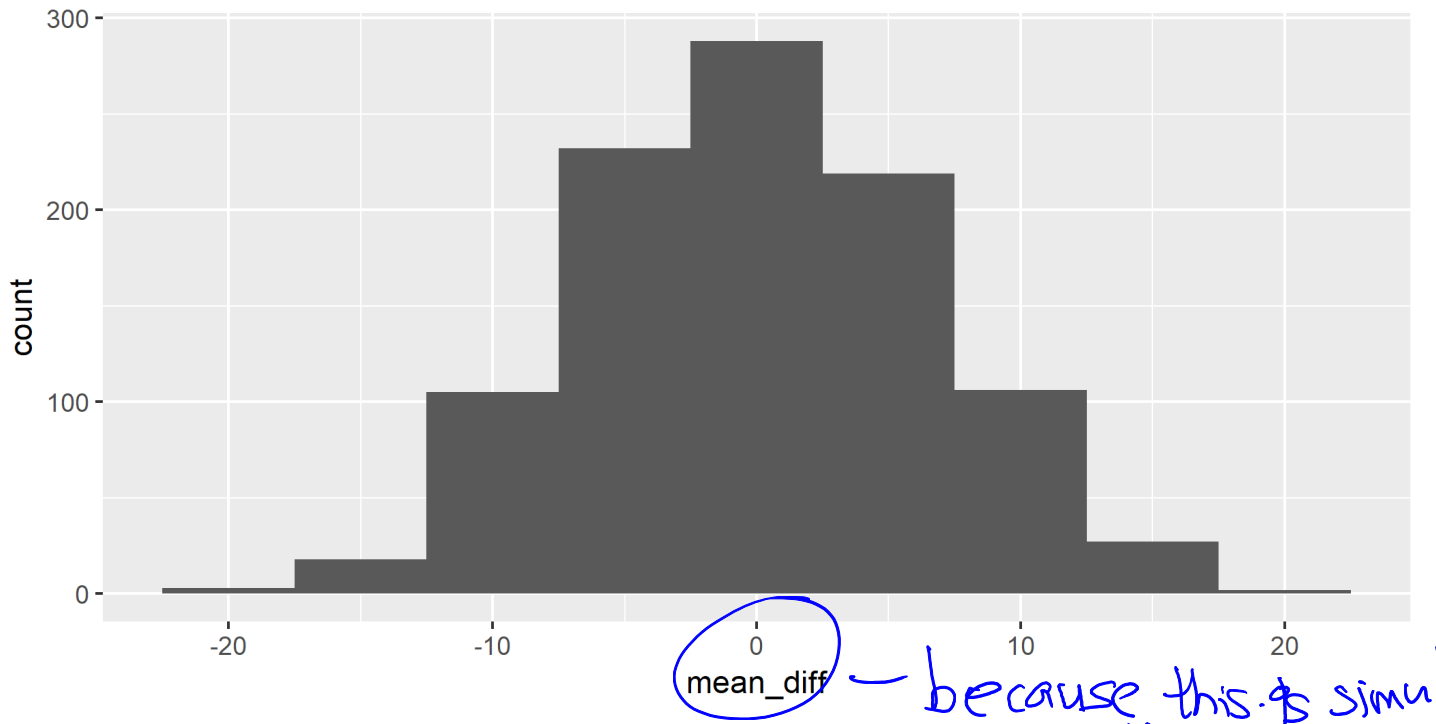
← one simulated value

Many values of what the test statistic could be if the null hypothesis were true

```
set.seed(130) # remove in practice

repetitions <- 1000  # "many times" will be 1000
# create a vector of missing values to store results
simulated_stats <- rep(NA, repetitions) # 1000 missing values

for (i in 1:repetitions)
{
  sim <- sleep_data %>%
    mutate(sleep = sample(sleep))  # shuffle sleep group labels

  # calculate test statistic for new data
  sim_test_stat <- sim %>% group_by(sleep) %>%
    summarise(means = mean(isi_change)) %>%
    summarise(sim_test_stat = diff(means))

  # add result to vector of values of test statistics
  # assuming null hypothesis
  simulated_stats[i] <- as.numeric(sim_test_stat)
}
```

Distribution of simulated values of assuming is true

```
sim <- data_frame(mean_diff=simulated_stats) # turn results
# into a data frame for plotting

ggplot(sim, aes(x=mean_diff)) + geom_histogram(binwidth=5)
```



because this is simulated under H_0: no difference

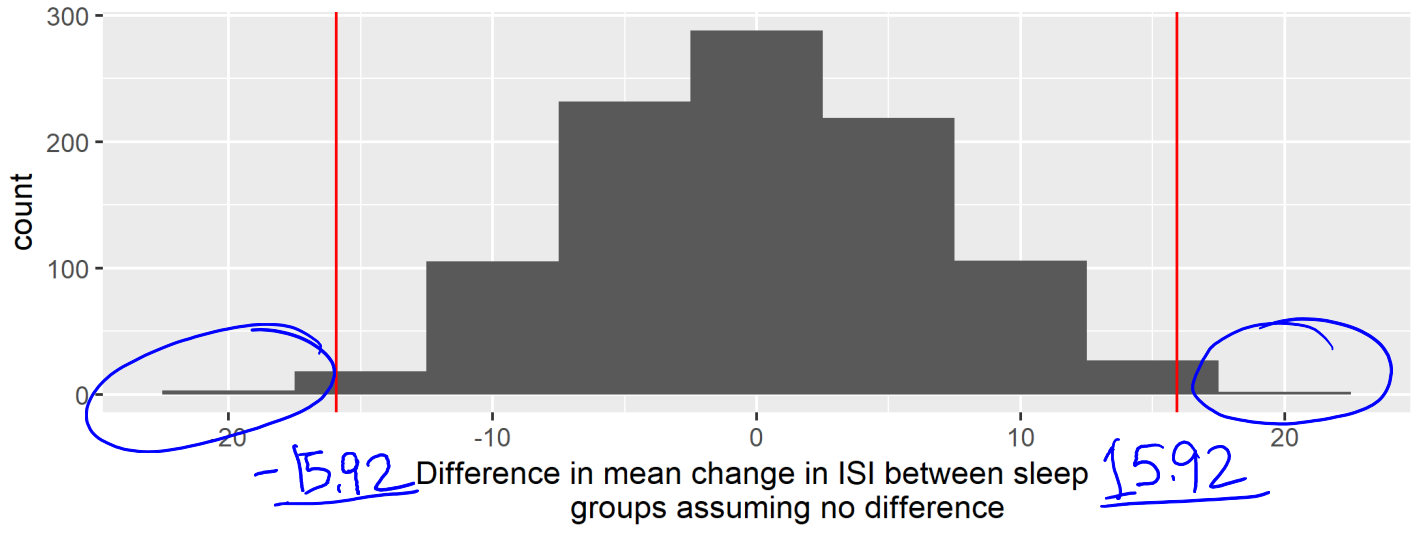
The P-value

P-value is the proportion of observations in the empirical distribution that are greater than or equal to

```
test_stat
```

```
## [1] 15.92
```

```
ggplot(sim, aes(mean_diff)) +
  geom_histogram(binwidth=5) +
  geom_vline(xintercept = test_stat, color="red") +
  geom_vline(xintercept = -1*test_stat, color="red") +
  labs(x = "Difference in mean change in ISI between sleep
       groups assuming no difference")
```



Calculate P-value

```
sim %>%
  filter(mean_diff >= abs(test_stat) |
           mean_diff <= -1*abs(test_stat)) %>%
  summarise(p_value = n() / repetitions)
```

10 values more extreme (in the tails)

1000 repetitions

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1    0.01
```

- Assuming that there is no difference in change in ISI between the sleep deprived and unrestricted sleep groups, the chance of seeing as large a difference in the means of change in ISI or even larger than what we observed is 0.01.
- We have strong evidence that the mean of change in ISI is different between the two sleep groups.

Applet

How many simulations is enough?

- In our examples, we've looked at 1000 simulated values assuming the null hypothesis is true, to compare to the value of our test statistic.
- In practice, the number of simulations is more typically on the order of 10,000.
- But that takes a long time to run.
- (Last set of practice problems asked for 100,000. That would take a very long time with all the shuffles, so it's not recommended!)

Some notes on hypothesis testing:

Type 1 and Type 2 errors

- The P-value gives us the probability of getting the data we got (as summarized by the test statistic) or data that are even less likely if the null hypothesis is true.

Some notes on hypothesis testing:

Type 1 and Type 2 errors

- The P-value gives us the probability of getting the data we got (as summarized by the test statistic) or data that are even less likely if the null hypothesis is true.
- But data values occur randomly (because they are measured on a random sample, or because the measuring process isn't perfect).

Some notes on hypothesis testing:

Type 1 and Type 2 errors

- The P-value gives us the probability of getting the data we got (as summarized by the test statistic) or data that are even less likely if the null hypothesis is true.
- But data values occur randomly (because they are measured on a random sample, or because the measuring process isn't perfect).
- So it's possible to get data that are not consistent with the null hypothesis just by chance and we conclude that the data give evidence against the null hypothesis, but the null hypothesis is actually true. This is called a **Type 1 error**.

Some notes on hypothesis testing:

Type 1 and Type 2 errors

- The P-value gives us the probability of getting the data we got (as summarized by the test statistic) or data that are even less likely if the null hypothesis is true.
- But data values occur randomly (because they are measured on a random sample, or because the measuring process isn't perfect).
- So it's possible to get data that are not consistent with the null hypothesis just by chance and we conclude that the data give evidence against the null hypothesis, but the null hypothesis is actually true. This is called a **Type 1 error**.
- It's also possible that, by chance, the data appear to be consistent with the null hypothesis, but the null hypothesis is actually not true. This is called a **Type 2 error**.

What we observed / What is the truth	H_0 is true	H_0 is false
Test shows data are consistent with H_0		Type 2 error
Test shows evidence against H_0	Type 1 error	

TRUTH

what our data shows (our conclusion)

↳ someone is actually innocent, but they go to jail

Someone is guilty but they don't go to jail

What we observed / What is the truth	is true	is false
Test shows data are consistent with		Type 2 error
Test shows evidence against	Type 1 error	

- Unfortunately, in practice we don't know if we've committed one of these types of errors.
- The more tests you do, the more likely you'll find a Type 1 error. But you won't know which test(s) resulted in Type 1 errors.
- In future statistics courses, you'll learn about ways to control the chance of making of making one of these types of errors.