

STA130H1F

Class #6

Prof. Nathan Taback

2018-10-11

Today's Class

Answering the question:

$$\left\{ \begin{array}{l} H_0: p = 0.5 \\ H_A: p \neq 0.5 \end{array} \right\}$$

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

— Errors in hypothesis testing.

Midterm test : Oct. 26 during tutorial time.

More details will be posted on Course Website this week.

What will be Covered?

- All material up to and including today's class.

What is the format?

Short answer questions. Similar to last year's mt. format.

Will there be Coding on the test?

No. But, you will be asked to understand code and output from running code.

Next class will be a review class.

Comparing two proportions

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 four cards - 2 red suit cards marked and 2 black suit cards, and assigning each person a card.



treatments.

Is Tylenol or Aspirin Better for Headache Relief?

- If a person receives a card with a red suit then they receive Tylenol, and if they receive a card with a black suit then they receive an 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

if there is no diff. between T and A then outcome doesn't change (pain)

Is Tylenol or Aspirin Better for Headache Relief?

Null hypothesis.
 $H_0: P_T = P_A$

- The null hypothesis is that changing the treatment for a subject has no effect on pain, in particular, no effect on the proportion that have no pain.
- Assuming this null hypothesis is true Tylenol (T) and Aspirin (A) are mere labels and don't affect the outcome.
- For example, assuming H_0 is true, subject 1 would have no pain if they had Tylenol or Aspirin.
- The alternative hypothesis is that the proportion of subjects without pain is different for Tylenol and Aspirin.

$$H_A: P_T \neq P_A$$

alternative hypothesis

Is Tylenol or Aspirin Better for Headache Relief?

$$H_0: P_T = P_A$$

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

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	
$\hat{p}_T - \hat{p}_A$			0.5	-0.5	0.5	-0.5	0.5	-0.5

observed another possibility.

$\binom{4}{2} = \#$ of ways to assign 2 treatments to four people = 6 possible randomizations.

observed diff.

Assume H_0 is true.

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

- $21/24 = 87.5\%$ of males and $14/24 = 58.3\%$ of females were recommended for promotion.
- This suggests that the males 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 chance?
- If many similar studies were conducted, assuming there is no difference between male and female promotion rates, then 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 no difference between the two groups. This is the **null hypothesis**, written H_0 .

For the gender bias in promotion study:

$$H_0: P_M = P_F, \text{ where } P_M = \text{proportion of males promoted}$$
$$P_F = \text{proportion of females promoted.}$$

2. There is a difference between the two groups. This is the **alternative hypothesis**, written H_A (or H_a or H_1). The alternative is almost always corresponds to the research question.

For the gender bias in promotion study:

$$H_A: P_M \neq P_F, \quad P_M > P_F \text{ or } P_M < P_F$$

two-sided alternative hypothesis.

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?

$$\begin{aligned} 29.2 &= 87.5 - 58.3 \\ &= \hat{p}_M - \hat{p}_F \\ &= \text{Observed value} \\ &\quad \text{of the test} \\ &\quad \text{Statistic.} \end{aligned}$$

\hat{p}_M, \hat{p}_F are the proportions of Males and females promoted (resp.) in the Study we have data on.

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 occurred 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 have 24 cards labelled with an "F" and 24 cards labelled with an "M".

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 have 24 cards labelled with an "F" and 24 cards labelled with an "M".
- Shuffle the cards ...

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 have 24 cards labelled with an "F" and 24 cards labelled with an "M".
- Shuffle the cards ...
- Assign the cards to the 48 people then calculate the difference in the proportion of males versus females that were promoted. This is one simulated value of the test statistic.

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 have 24 cards labelled with an "F" and 24 cards labelled with an "M".
- Shuffle the cards ...
- Assign the cards to the 48 people then calculate the difference in the proportion of males versus females that were promoted. This is one simulated value of the test statistic.
- Shuffle the cards again ...

- Assign the cards to the 48 people then calculate the difference in the proportion of males versus females that were promoted. This is another simulated value of the test statistic.

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

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

- Assign the cards to the 48 people then calculate the difference in the proportion of males versus females that were promoted. This is another simulated value of the test statistic.
- Shuffle the cards again ...
- Assign the cards to the 48 people then calculate the difference in the proportion of males versus females that were promoted. This is another simulated value of the test statistic.
- Repeat: shuffle, assign cards, calculate difference

Gender Bias Data

Need the raw data to simulate - Shuffle, assign cards, Calculate diff.

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

```
# create dataframe
bias <- data_frame(gender = c(rep("male", 24), rep("female", 24)),
                  promoted = c(rep("yes", 21), rep("no", 3),
                              rep("yes", 14), rep("no", 10)))
```

repeat Male 24 times.

```
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

- Are the variables numerical or categorical?

Categorical.

Calculate the proportion of males and females promoted

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

total # of males (females)

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

only want promoted females.

Counts the number of observations.

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

Convert to numeric.

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

yes_male is a data frame.

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

Check that data corresponds to table.

```
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

$$\hat{p}_{female} - \hat{p}_{male} = 0.583 - 0.875 = -0.292$$

- 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

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
```

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

```
sample(a_vector)
```

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

```
sample(a_vector)
```

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

```
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" "female"
```

```
bias$promoted
```

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

After the shuffle

gender is shuffled.

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

```
## [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"
```

outcome of promoted (not promoted stays the same.)

After the shuffle

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

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

of females promoted

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

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

of males promoted.

```
# 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
```

if shuffled again then another value of test Stat would occur.

Set up Simulation in R

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

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


Calculate Observed Value of Test Statistic

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

yes_male <- bias %>%
  # only promoted males
  filter(promoted == "yes" & gender == "male") %>%
  count() # 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
sim <- data_frame(p_diff = simulated_stats)
```

Count #
of
females
and
males
promoted

1000

Shuffle Cards, assign Cards to
people in Study

assuming H_0
is true.

$H_0: P_M = P_F$

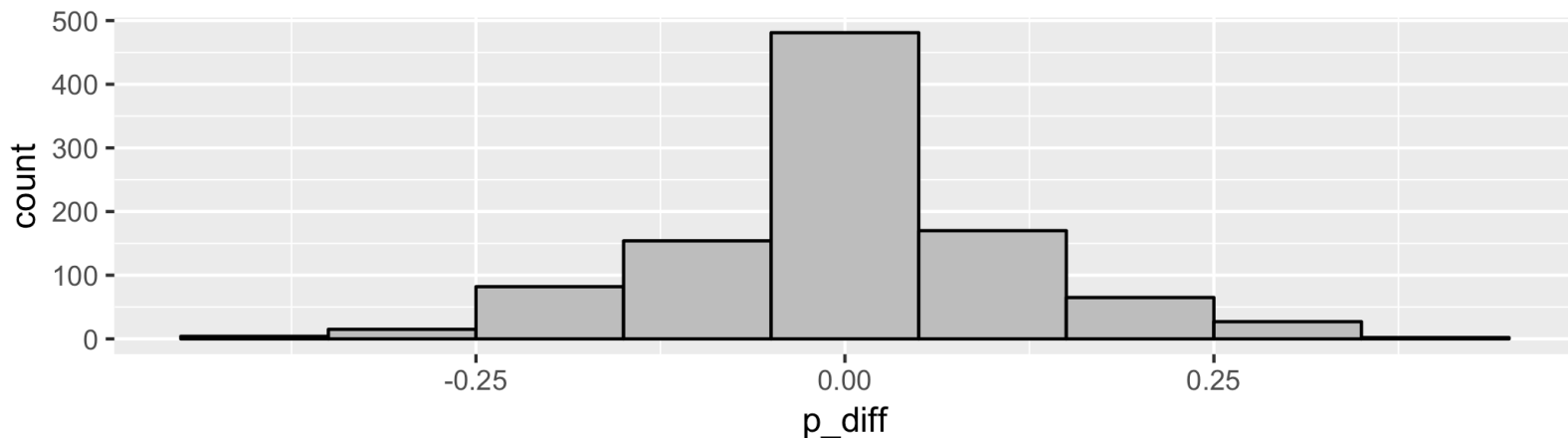
Calculate
test Stat.

i^{th} value of Simulated Stats.

Distribution of simulated values of $\hat{p}_{female} - \hat{p}_{male}$ assuming H_0 is true

The for loop in the previous slide produced a simulated distribution of differences in proportion of males and females promoted. This can be visualized using a histogram.

```
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?

We have assumed $P_m - P_f = 0$

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.

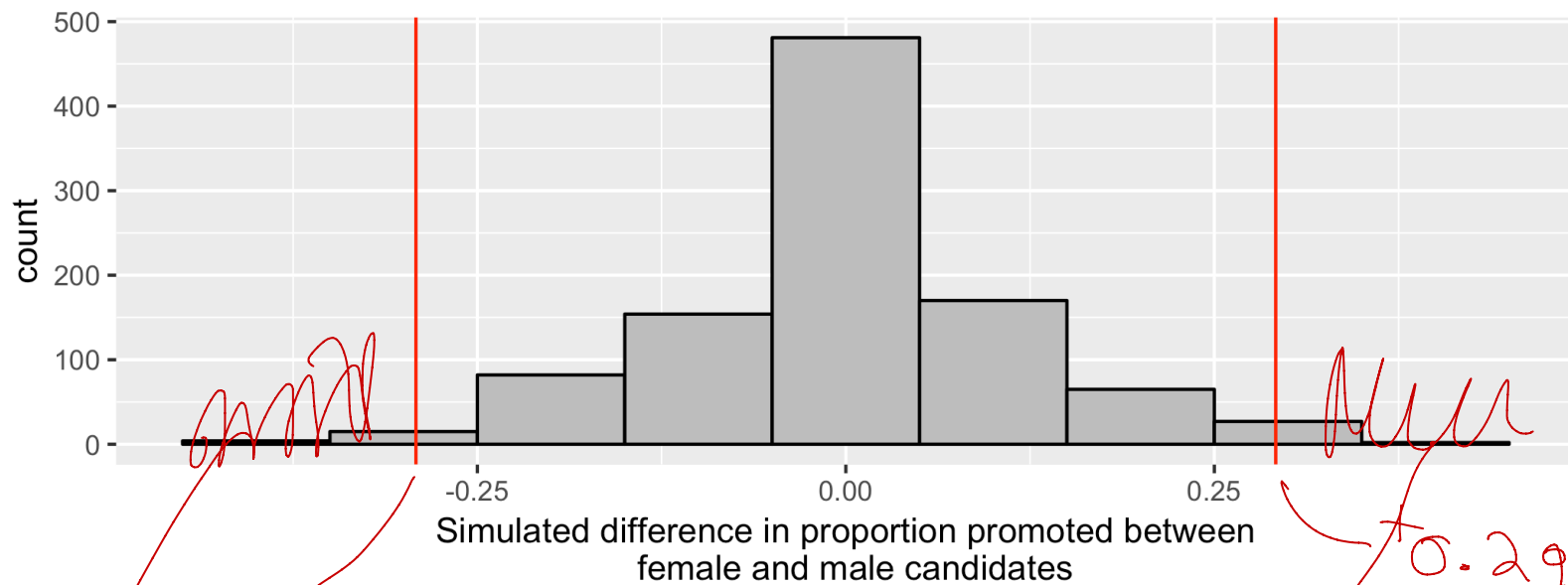
4. The P-value - Gender Bias Example

For the gender bias example:

- the null hypothesis value is $p_1 - p_2 = 0$
- the observed estimate from the data (the test statistic) is $\hat{p}_1 - \hat{p}_2 = -0.292$
- 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

[1] -0.2916667



-0.2916667

0.2916667

more extreme differences than observed in data -

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
```

1000

5. Make a conclusion

- A large P-value means the data are consistent with the null hypothesis.
- A small P-value means the data are inconsistent with the null hypothesis.
- 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.

one of two possibilities:

① H_0 is true $P_M = P_F$

or ② we observed something rare.

what is rare?

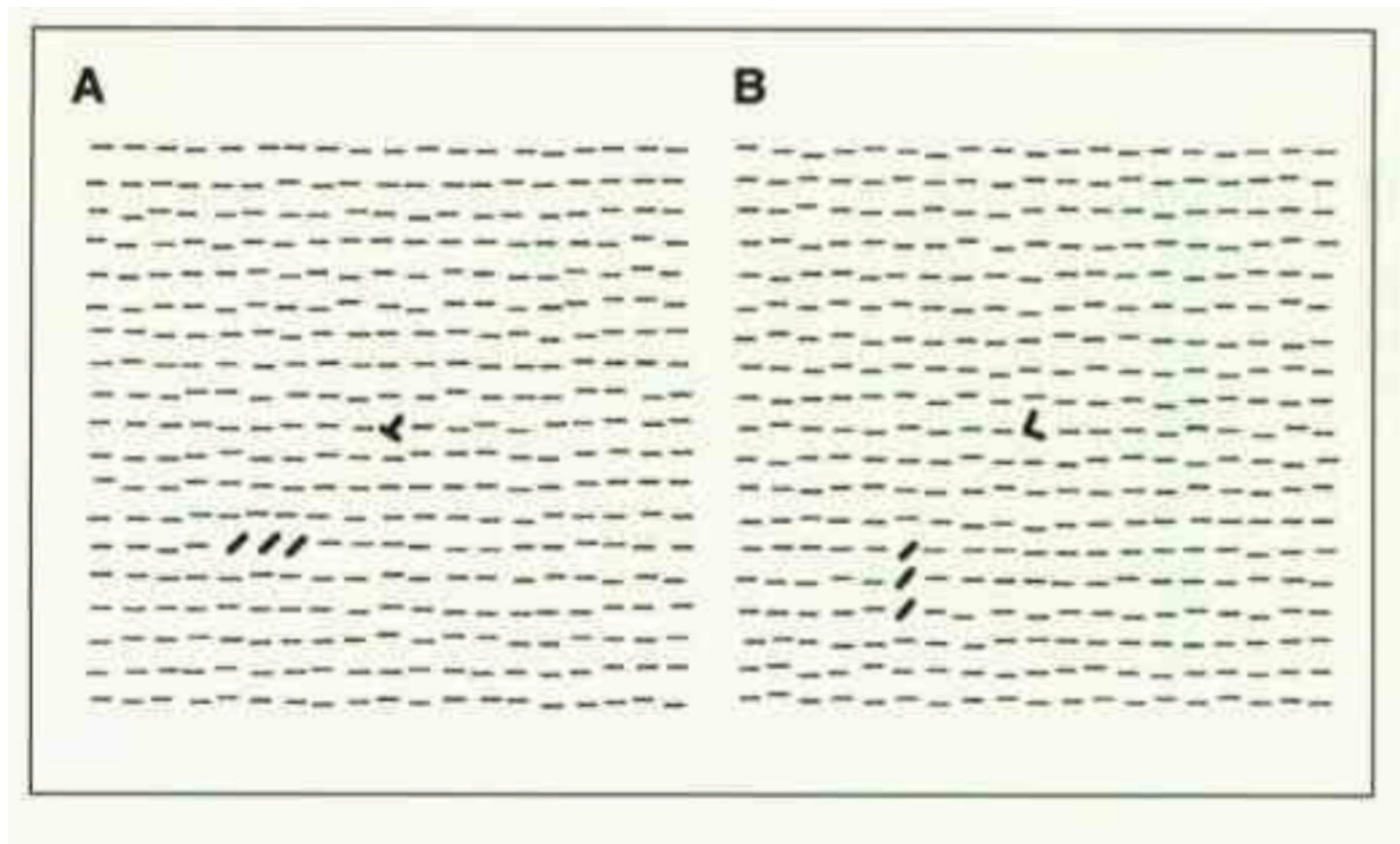
Conclusion: We did 1000 Simulations to test $H_0: P_m = P_f$, assuming H_0 is true. 48 of the 1000 Simulations were more extreme than the observed value of the test statistic.

∴ We have moderate evidence that there is a difference in males being promoted compared to females.

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 milliseconds
- 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

Labels for the group.

```
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()
```

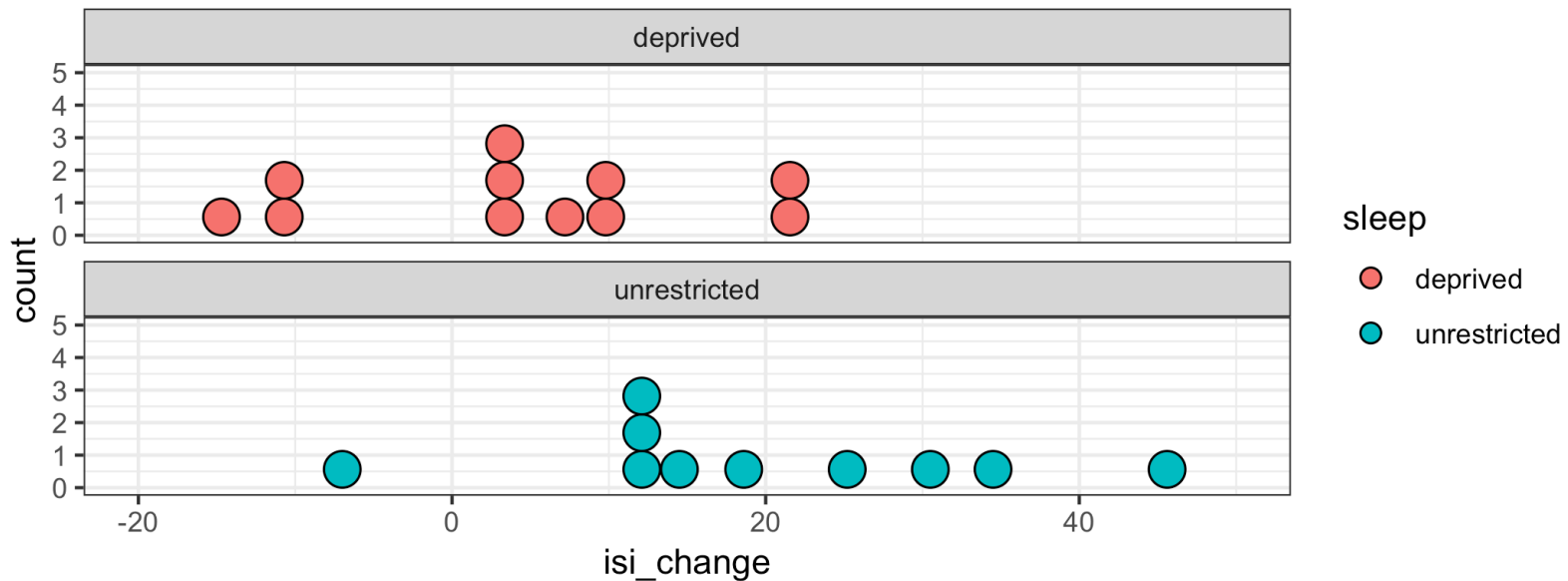
outcome.

data frame.

```
## # 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()
```



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

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

$$H_0 : \mu_1 = \mu_2 \text{ vs. } H_A : \mu_1 \neq \mu_2$$

- μ_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**.
- μ_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**.

H_0 : are the means the same

H_A : are the means different

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

$$\text{Test statistic} = \hat{\mu}_1 - \hat{\mu}_2$$

```
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
```

$$\text{Test statistic} = \hat{\mu}_1 - \hat{\mu}_2 = 19.82 - 3.90 = 15.92$$

observed mean
in unrestricted
group

observed mean
in restricted
group.

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

```
test_stat
```

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

Calculate diff.
ln means.

Simulate what H_0 predicts will happen

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

Simulate what H_0 predicts will happen

Assume H_0 is true: The value of a subject's ISI is same if they are in the sleep deprived or unrestricted sleep groups. Shuffle, Assign, Calculate Test Statistic, Repeat:

- **Shuffle:** shuffle 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 in the means of change in ISI for the observations in each of these new groups
- **Repeat:** lots of times giving an empirical distribution for the test statistic if the null hypothesis were true

Simulate what H_0 predicts will happen

Assume H_0 is true: The value of a subject's ISI is same if they are in the sleep deprived or unrestricted sleep groups. Shuffle, Assign, Calculate Test Statistic, Repeat:

- **Shuffle:** shuffle 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 in the means of change in ISI for the observations in each of these new groups
- **Repeat:** lots of times giving 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.

Compare observed value to lots of other possible values assuming H_0 is true!

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  
  
one_sim <- sim %>%  
  group_by(sleep) %>%  
  summarise(means = mean(isi_change))  
one_sim
```

Shuffle, assign

get means of each group.

```
## # A tibble: 2 x 2  
##   sleep      means  
##   <chr>    <dbl>  
## 1 deprived    13.0  
## 2 unrestricted  9.86
```

```
one_sim$means[2] - one_sim$means[1]
```

Calculate test Stat.
 $\hat{\mu}_1 - \hat{\mu}_2$

```
## [1] -3.094545
```


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)
}
```

Shuffle, assign.

Calc. test Stat.

Store values.

Distribution of simulated values of

$\hat{\mu}_1 - \hat{\mu}_2$ **assuming H_0 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, colour = "black", fill = "grey")
```

missing plot ??

The P-value

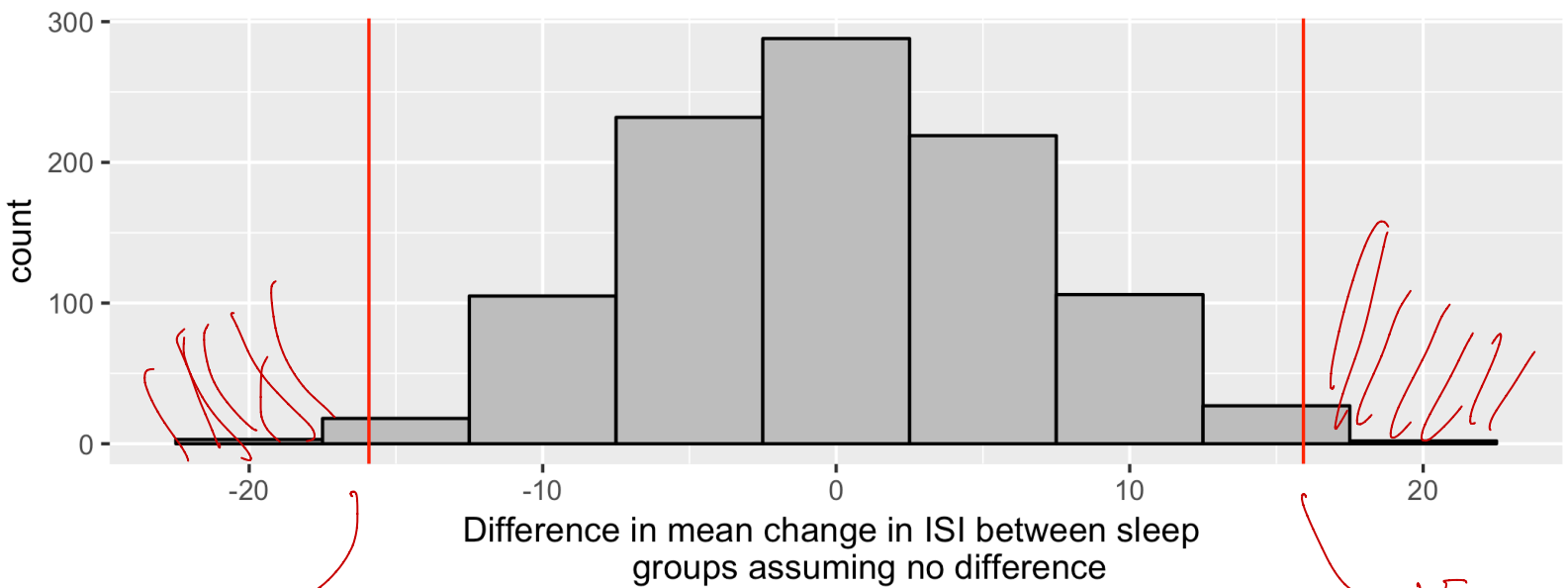
P-value is the proportion of observations in the empirical distribution that are greater than or equal to

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

```
test_stat
```

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

```
ggplot(sim, aes(mean_diff)) +  
  geom_histogram(binwidth=5, colour = "black", fill = "grey") +  
  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")
```



-15.92

15.92

Calculate P-value

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

greater than obs.

less than obs.

```
## # 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.

1. Assume H_0 is true i.e., $\mu_1 = \mu_2$
2. Simulated 1000 values of test Stat. under H_0 ,
3. Count # of Simulations as extreme as observed in data.
4. \therefore only 1% of Sims as extreme $\Rightarrow H_0$ is not true!

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!)

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.

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).

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**.

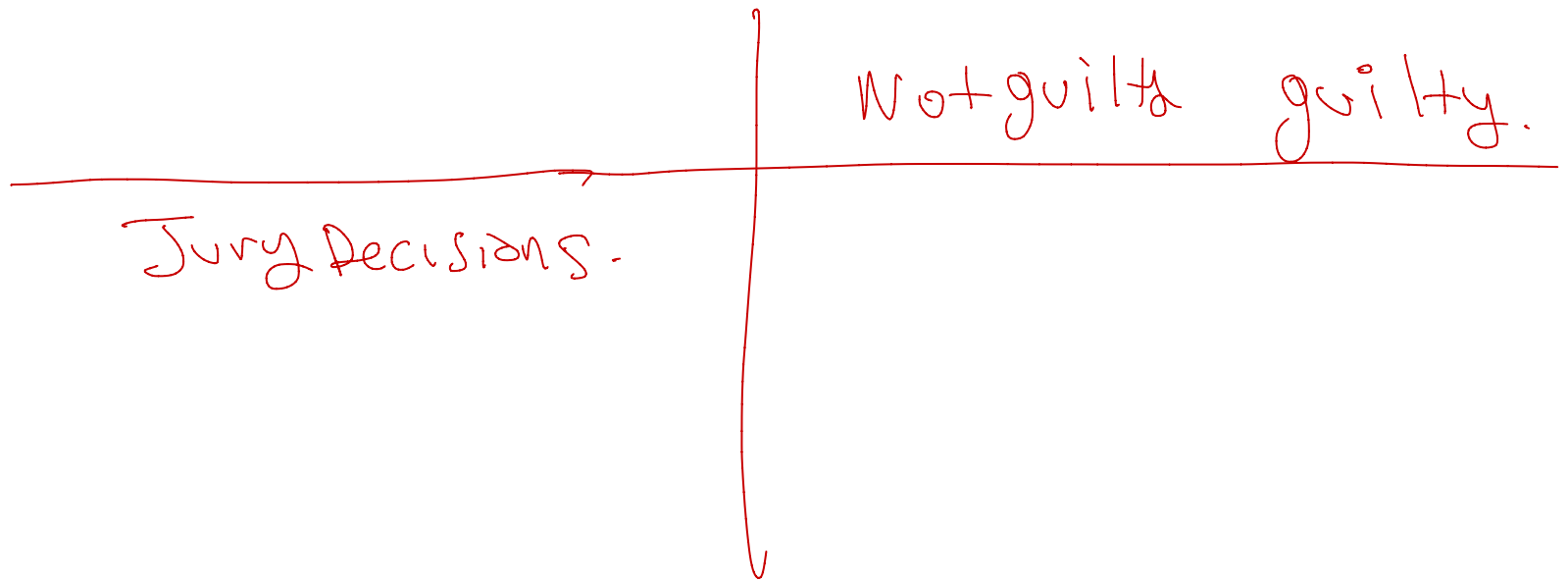
Type 1 and Type 2 Errors

- 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**.

Truth.

Observed

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	✓



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	

- 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.