STA130H1S - Class #5

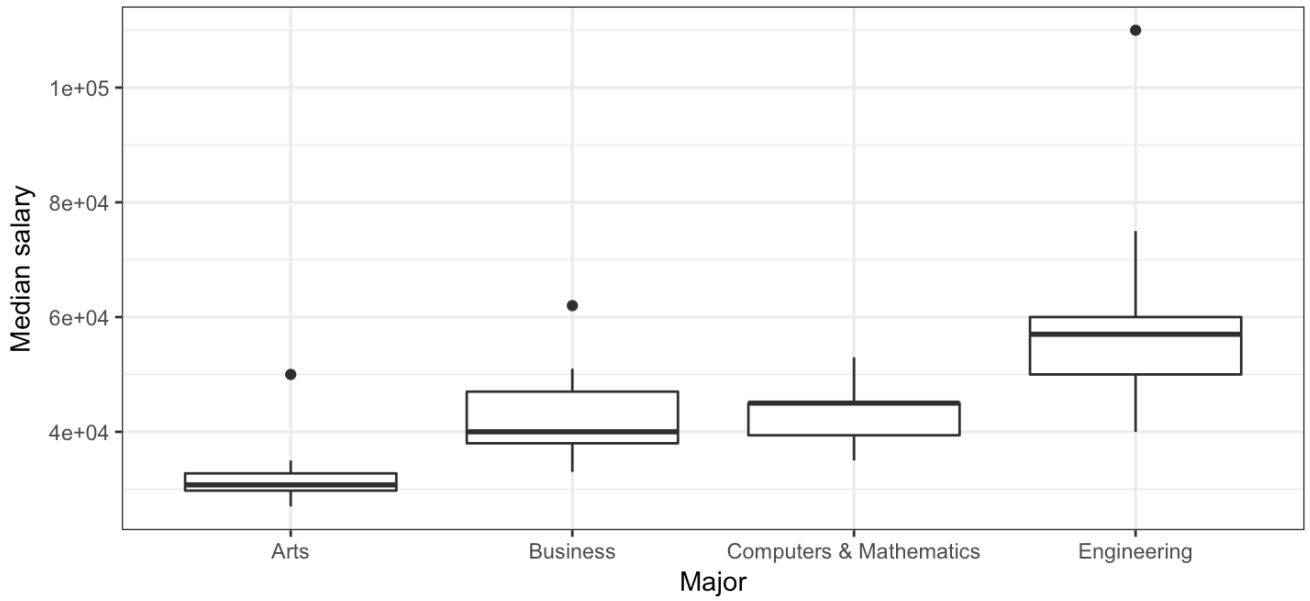
Inferential Thinking Part 2: Testing hypotheses on two groups

Prof. A. Gibbs

February 5, 2018

Another look at the "Economic Guide to Picking a Major" from second week:

- Is there evidence that salaries are, on average, higher for Engineering graduates than Arts graduates?
- Is there evidence that salaries are, on average, higher for Computers & Mathematics graduates than Business graduates?



Today

Answering the question:
if we see a difference between two groups, is it meaningful?
or could it just be due to chance?

Examples for:

- proportions for two groups
- means for two groups

Can extend to comparing any statistic between two groups

Recommended reading:
Sections 2.1, 2.2, 2.3 (excluding 2.3.4) of *Introductory Statistics with Randomization and Simulation* from OpenIntro
(a free open-source textbook)

Statistical Inference

- Imagine we have a "real world" where we observe data, and a "theoretical world" (a population or scientific model) that we want to make conclusions about.
- Inference connects what we have observed in the real world to what we can say about the theoretical world.
- *Last class:* made inferences about one proportion
- *Today:* our theoretical world models will be that two groups are the same in some way, and we'll test to see if our data are consistent with that

Hypothesis Testing for Two Proportions: Comparing a proportion between two groups

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 were recommended for promotion
- 14/24 = 58.3% of females were recommended for promotion

The 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?

- Are the variables numerical or categorical?

Code to calculate the proportion of males and females promoted

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

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

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

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

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

Code to calculate the proportion of males and females 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] -0.2916667
```

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

Review: The Logic of Hypothesis Testing

1. The hypotheses

Two claims:

1. There is nothing going on. This is the **null hypothesis**, written H_0.

For the gender bias in promotion study:

1. 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:

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

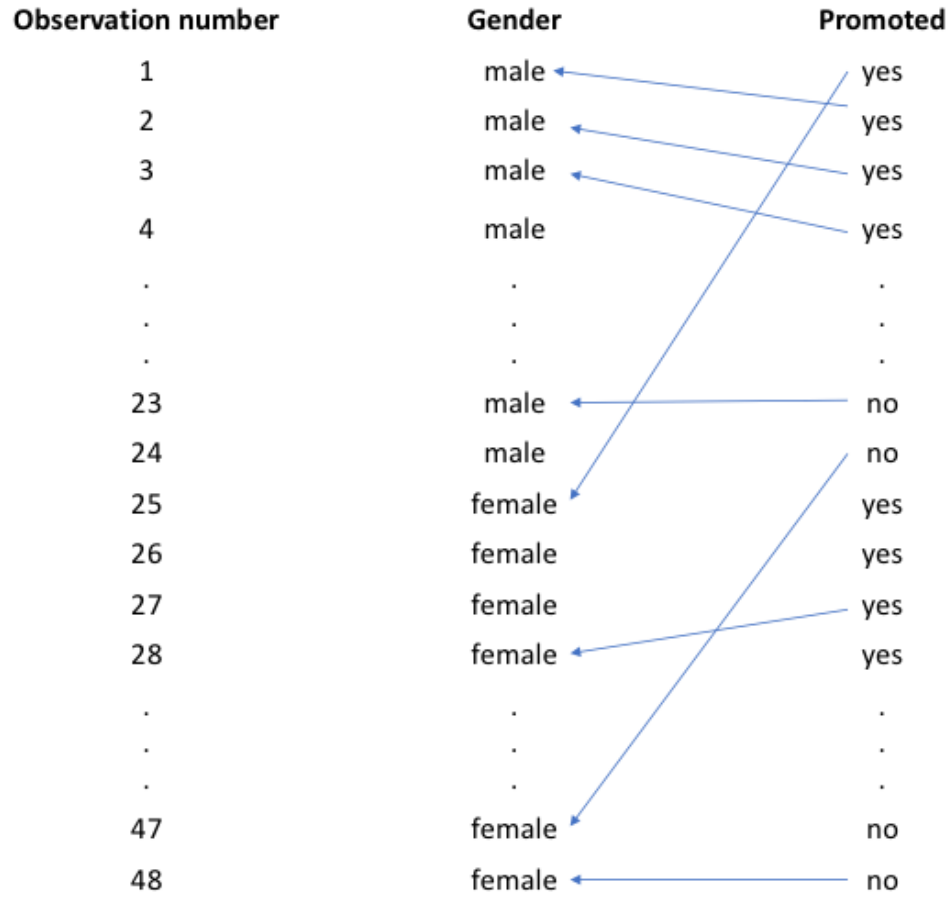
We want to simulate many many possible values of what the test statistic might have looked like if the null hypothesis were true to know the distribution of its possible values.

How can we do this?
If the null hypothesis is true, any observation (promoted or not promoted) is just as likely to be for one gender as the other gender. And all ways the observations could be arranged among the two genders are equally likely.

Our data

Observation number	Gender	Promoted
1	male	yes
2	male	yes
3	male	yes
4	male	yes
.	.	.
.	.	.
.	.	.
23	male	no
24	male	no
25	female	yes
26	female	yes
27	female	yes
28	female	yes
.	.	.
.	.	.
.	.	.
47	female	no
48	female	no

What if any promotion recommendation was equally likely to be for a male or female candidate?



Randomly shuffling gender gives us possible data that could have occurred if any promotion recommendation was equally likely to be for a female or male candidate

Observation number	Gender	Promoted
1	female	yes
2	male	yes
3	male	yes
4	male	yes
.	.	.
.	.	.
.	.	.
23	male	no
24	female	no
25	male	yes
26	female	yes
27	female	yes
28	female	yes
.	.	.
.	.	.
.	.	.
47	female	no
48	female	no

3. Simulate what H_0 predicts will happen

- shuffle the categorical variable that says to which gender each observation belongs
- calculate the difference in the proportions of people who were promoted in the new groups
- repeat lots of times

How to shuffle

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] 1 1 1 2 2
```

```
sample(a_vector)
```

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

```
sample(a_vector)
```

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

```
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"
```

After the shuffle

```
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
```

Simulate the possible values of the difference in the proportions many times, assuming the null hypothesis is true

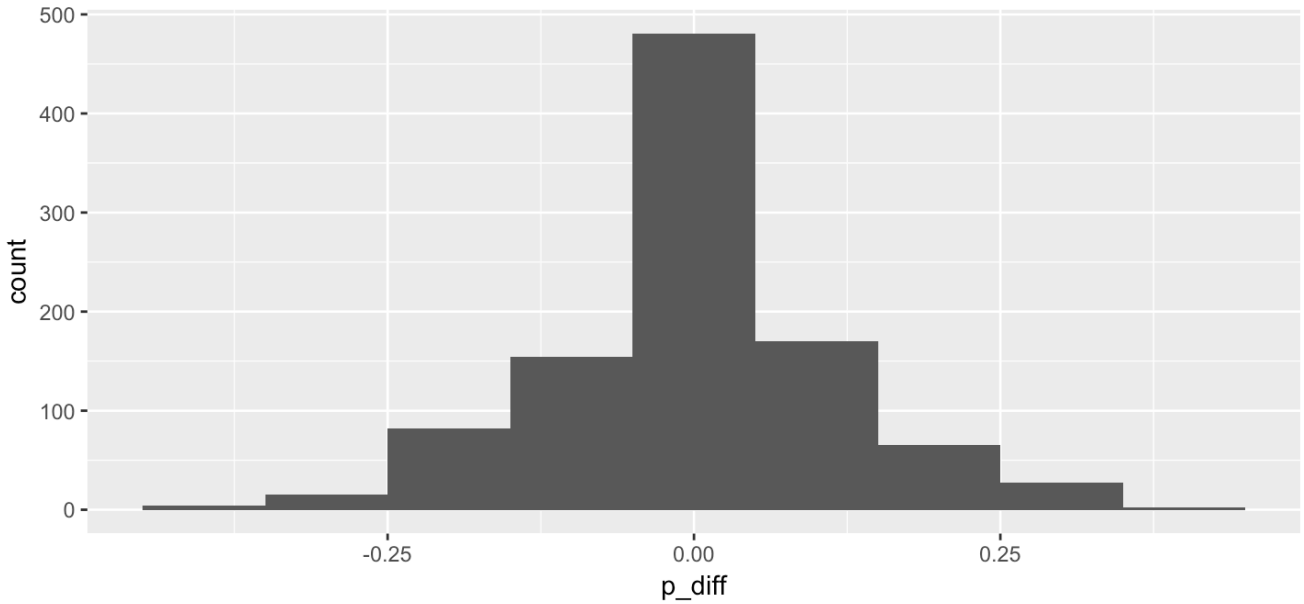
```
set.seed(130) # remove in practice

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") %>% summarise(n())
n_male <- bias %>% filter(gender=="male") %>% summarise(n())
```

```r
# 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) # treat result as a number
```

```
for (i in 1:repetitions)
{
  sim <- bias %>% mutate(gender = sample(gender)) # shuffle gender labels
  yes_female <- sim %>%
    filter(promoted=="yes" & gender=="female") %>%  # only promoted females
    summarize(n())  # count
  yes_male <- sim %>%
    filter(promoted=="yes" & gender=="male") %>%  # only promoted males
    summarize(n())  # 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)
```

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

```
ggplot(sim, aes(x=p_diff)) + geom_histogram(binwidth=0.1)
```



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

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.

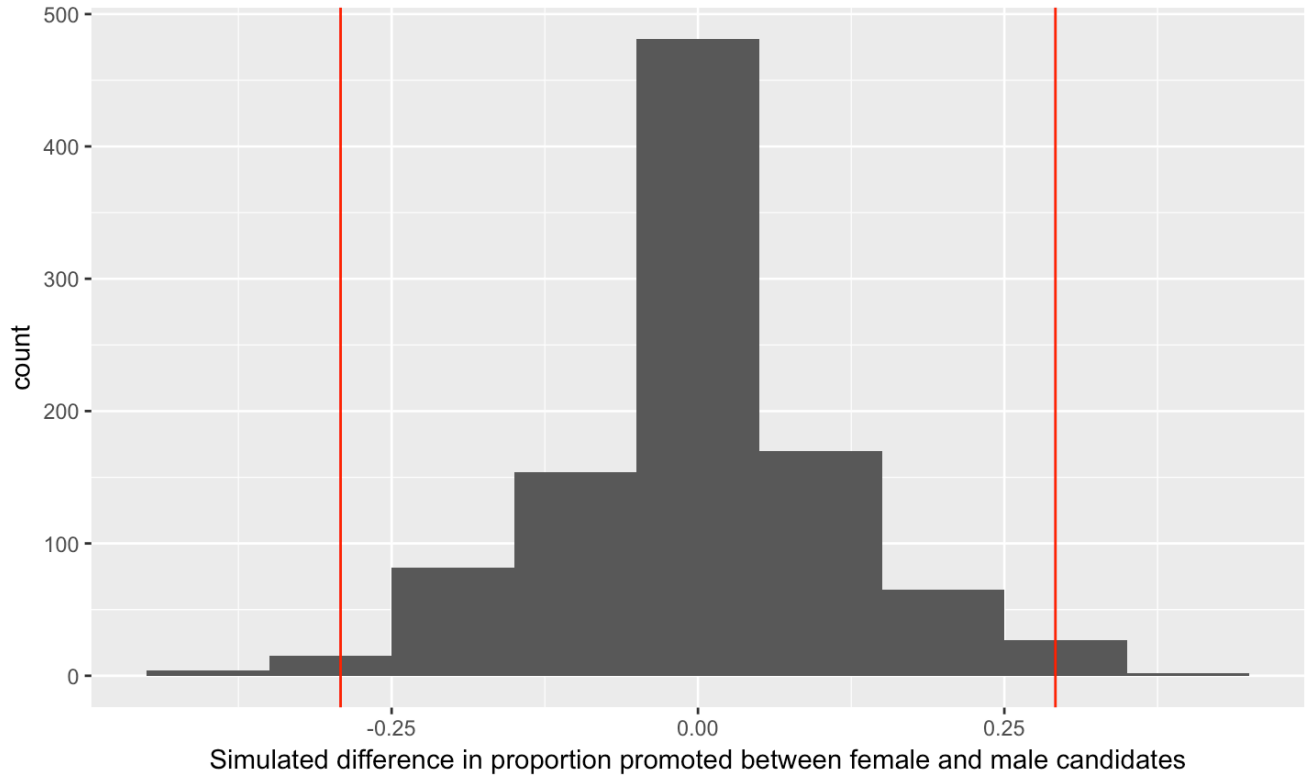
For the gender bias example:

- 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 inclues 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

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")
```

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


500
400
300
200
100
0
count
-0.25
0.00
0.25
Simulated difference in proportion promoted between female and male candidates

Calculate P-value

```
test_stat
```

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

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

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
  - *unrestricted sleep*: no restrictions on their sleep
  - 11 subjects were in the sleep deprived group and 10 subjects were in the unrestricted sleep group
- Subjects then were allowed unrestricted sleep for the next two nights
- Subjects were then retested on the visual discrimination task

The visual discrimination task

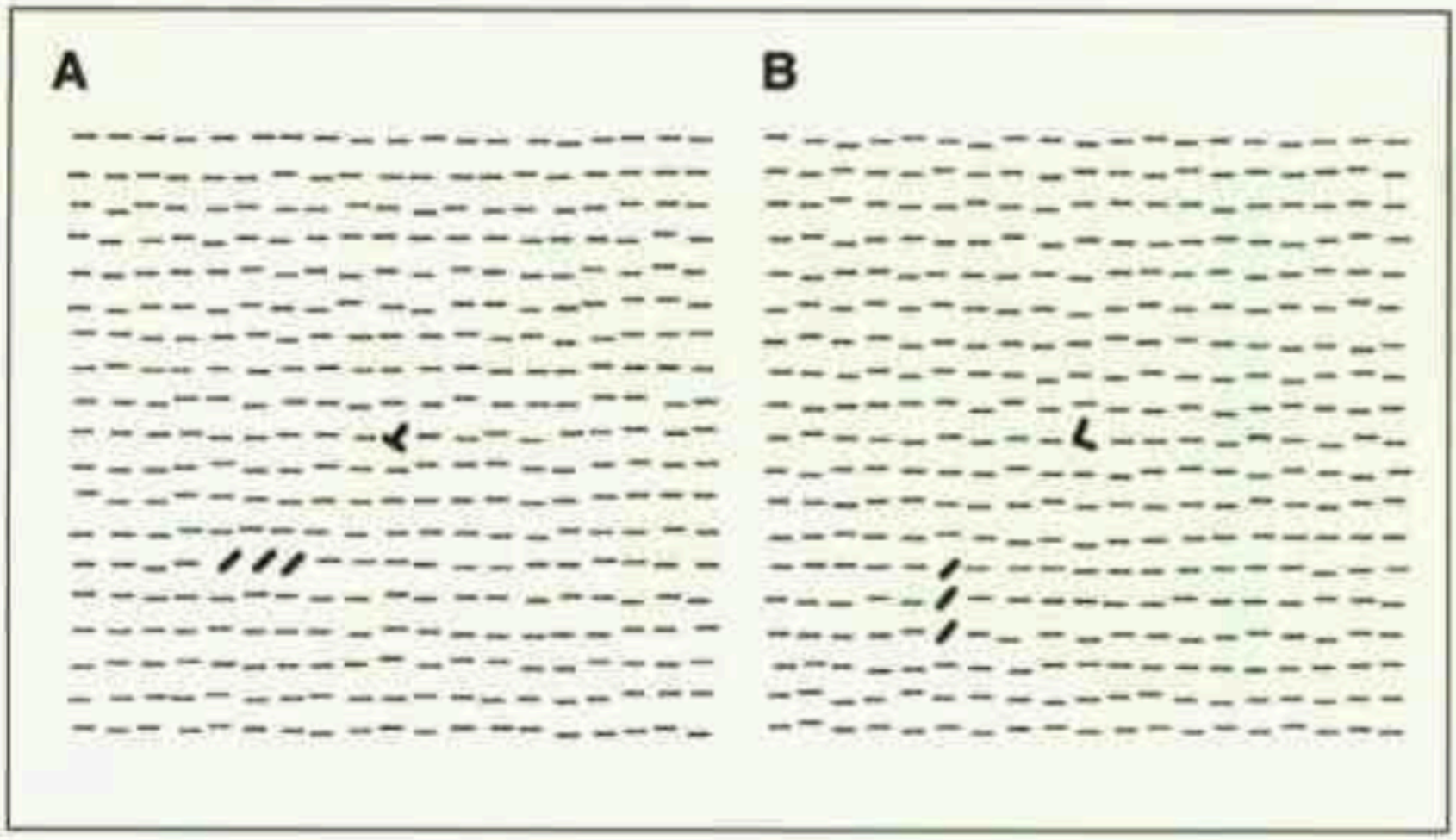


Figure 1. Sample target screens for the visual discrimination task. Each target screen contained a rotated "T" (as in A) or "L" (as in B) at

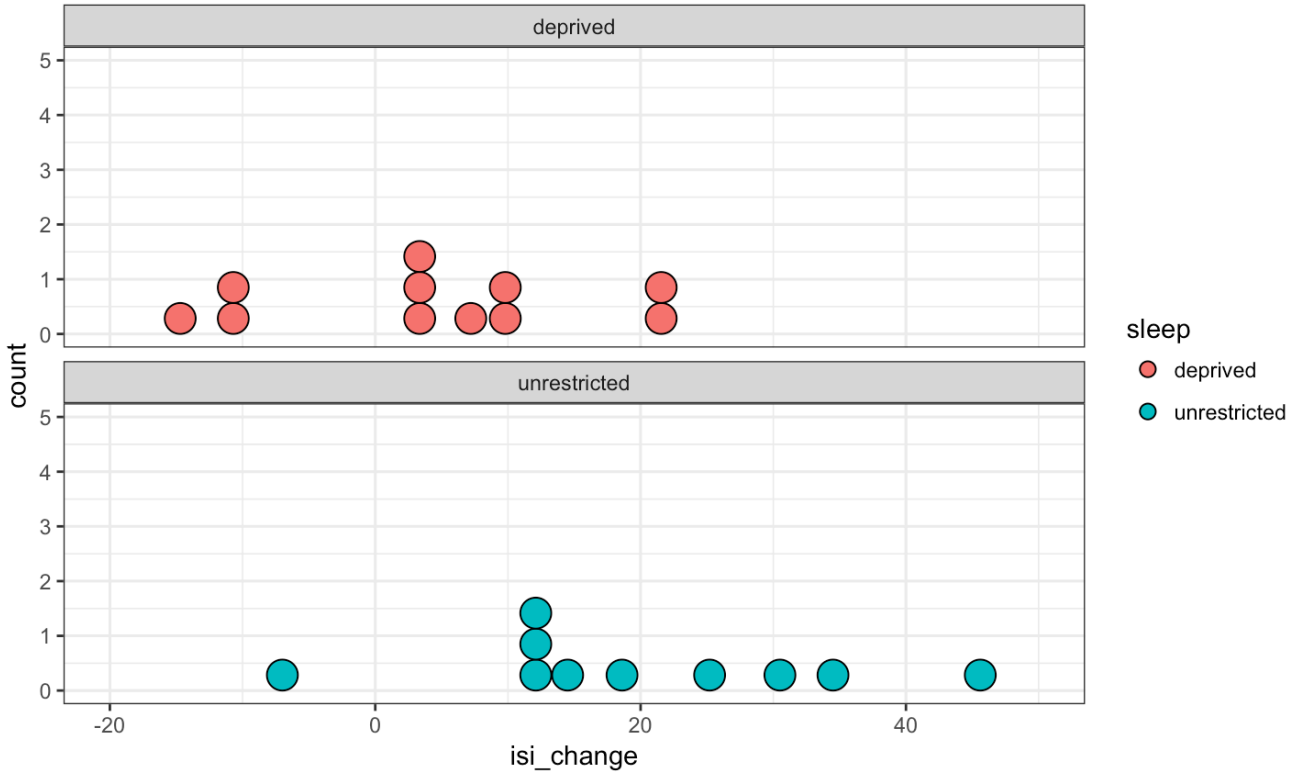
The data

```
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)

glimpse(sleep_data)
```

```
## Observations: 21
## Variables: 2
## $ sleep      <chr> "unrestricted", "unrestricted", "unrestricted", "un...
## $ isi_change <dbl> 25.2, 14.5, -7.0, 12.6, 34.5, 45.6, 11.6, 18.6, 12....
```

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?

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

Hypotheses to test whether the mean of the change in ISI is the same for students who are sleep deprived and students who had unrestricted sleep

$H_0 : \mu_1 = \mu_2$

$H_A : \mu_1 \neq \mu_2$

where 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 and 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

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

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  3.90
## 2 unrestricted 19.82
```

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

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

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

Simulate what H_0 predicts will happen

If there is no difference in change in ISI between the sleep deprived and unrestricted sleep groups, every value of change in ISI that we observed should be equally likely to be from a student who was sleep deprived or from a student who had unrestricted sleep.

So:

- shuffle the categorical variable that says to which sleep group each observation belongs
- calculate the difference in the means of change in ISI for the observations in each of these new groups; this gives a value of what the test statistic would be if the null hypotheses were true
- repeat lots of times giving an empirical distribution for the test statistic if the null hypothesis were true
- compare the test statistic observed from the data to the empirical disribution

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))
```

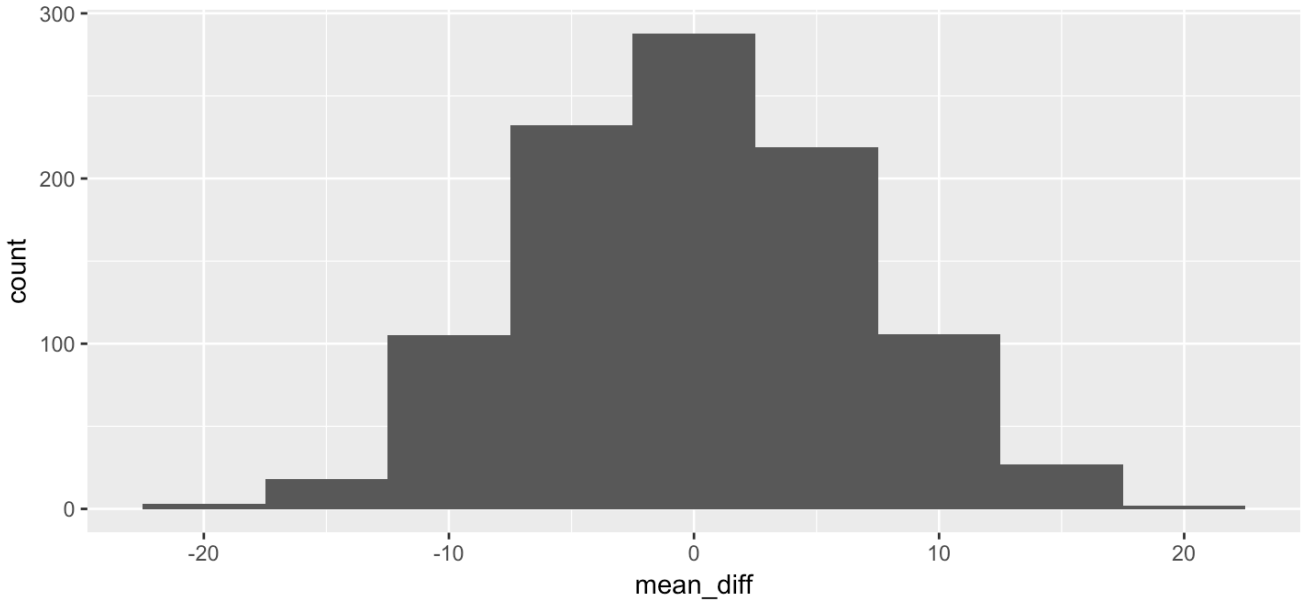
```
## # A tibble: 1 x 1
##   sim_test_stat
##           <dbl>
## 1     -3.094545
```

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)) %>% summaris

  # 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 $\hat{\mu}_1 - \hat{\mu}_2$ 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)
```



The P-value

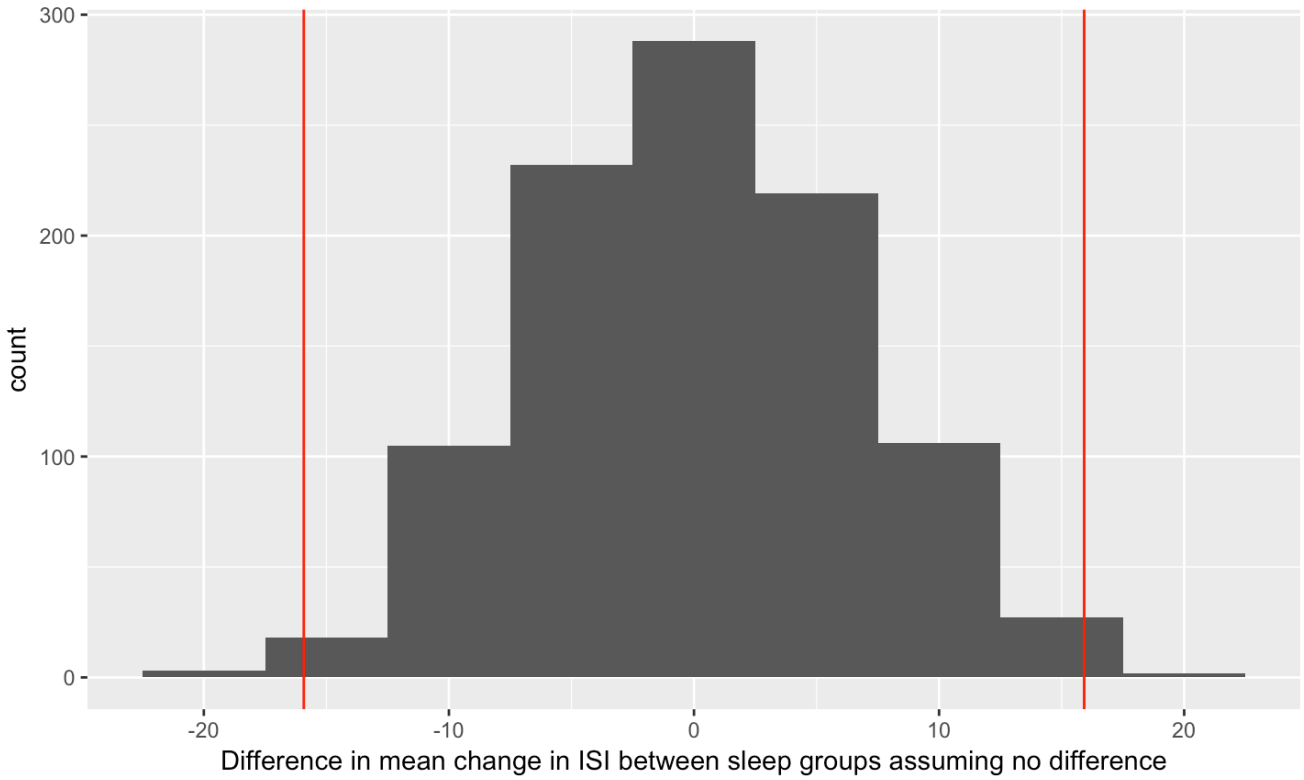
P-value is the proportion of observations in the empirical distribution that are greater than or equal to

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

```
test_stat
```

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

```
ggplot(sim, aes(mean_diff)) +
  geom_histogram(binwidth=5) +
  geom_vline(xintercept = test_stat, color="red") + geom_vline(xintercept = -1*test_stat, colo
  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)
```

```
## # 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.

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.
- 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	

- 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.

Some notes on hypothesis testing: Some cautions

- In the last few years, P-values have received some criticism.
- Often this criticism arises because of over-dependence on them, rather than using and interpreting them appropriately as part of the scientific process.

Some principles related to interpreting P-values from the American Statistical Association:

1. *P-values can indicate how incompatible the data are with a specified statistical model.* (what P-values are)
2. *P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.* (what P-values aren't)
3. *Scientific conclusions and business or policy decisions should not be based only on whether a P-value passes a specific threshold.* (it's not that simple; the scientific context and the quality of the study matter)

1. *P-values can indicate how incompatible the data are with a specified statistical model.*
2. *P-values do not measure the probability that the studied hypothesis is true, or the probability that the data were produced by random chance alone.*
3. *Scientific conclusions and business or policy decisions should not be based only on whether a P-value passes a specific threshold.*
4. *Proper inference requires full reporting and transparency.* (sometimes researchers publish only results where they found evidence against the null hypothesis and don't report everything else they looked at)
5. *A P-value does not measure the size of an effect or the importance of a result.* (just because you got a small P-value, doesn't mean the result is meaningful in a practical way)
6. *By itself, a P-value does not provide a good measure of evidence regarding a model or hypothesis.* (we need the whole scientific story)

A clinical oncologist is investigating the efficacy of a new treatment on reduction in tumour size. She randomly assigns patients to the new treatment or old treatment and compares the mean of the reduction in tumour size between the two groups. She carries out a statistical test and the P-value is 0.001. How many of the following are valid interpretations of the P-value?

1. The probability of observing a difference between the treatment groups as large or larger than she observed if the new treatment has the same efficacy as the old treatment.
2. The probabiliy that the new treatment works the same as the old treatment.
3. The probability that the new treatment, on average, reduces tumour size more than the old treatment.

- None
- One
- Two
- Three

Permutation test

Rather than looking at random shuffles of whether the promotion candidates were male or female, couldn't we have looked at what the proportions would be for every possible arrangement of the 24 females and 24 males?
Yes! This is called a **permutation test**. (*Randomization test* and *permutation test* are sometimes used interchangeably. There is a subtle difference that we won't worry about.)

Why might we not want to do this?
Need to know how many ways to choose the 24 females from the 48 candidates or promotion.
Answer:

```
choose(48, 24)
```

```
## [1] 3.22476e+13
```

Over 30 trillion!

What is `choose(48, 24)`?

This is called a combination, one of the fundamental ideas in counting problems. Knowing how to solve counting problems can be useful in probability problems.

Fundamental Principles of Counting:

- *The Multiplication Principle*: If there are m possible outcomes or choices for a first experiment or decision and n possible outcomes for a second experiment or decision and the two choices / decisions are independent, then the number of ways the series of decisions can be made (or the number of outocmes for the series of experiments) is $m \times n$.
- *Permutations*: The number of ways n things can be ordered is $n \times (n-1) \times (n-2) \times \ldots \times 3 \times 2 \times 1$. This is $n!$; in words: "n factorial".
- The number of ways we can choose k things from n without replacement when we care about their order is $n \times (n-1) \times (n-2) \times \ldots \times (n-k+1)$.

- *Combinations*: If we don't care about the order in which we choose them, the number of ways we can choose k things from n without replacement is

$$\frac{n \times (n-1) \times (n-2) \times \ldots \times (n-k+1)}{k!}$$

since we don't care about the $k!$ ways they can be ordered. In words, this is "n choose k" and can by calculated by the `choose` function in `R`.