Lecture 22

Non-Parametric Testing

Randomization and Permutation Testing

Today we are going to talk about an alternative to the methods we've discussed so far in MATH 1051H. All of the confidence interval and hypothesis tests we've been computing rely on assumptions about the **distributions** of our statistics: these are the **assumptions** we've been requiring that you write for the Written Assignment.

Case Study 1

Case Study: Gender Discrimination (Remember?)

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

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

Data

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

	Promoted		
	Yes	No	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

Simulating the experiment …

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

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

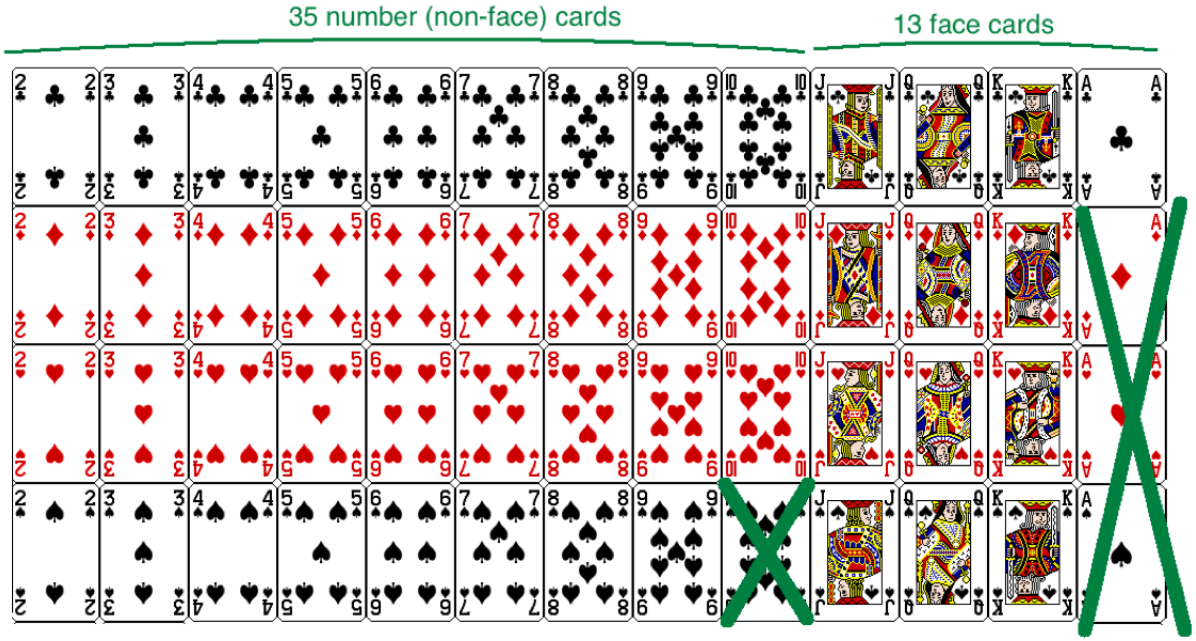
If the results from the simulations based on the chance model do not look like the data, then we can determine that the difference between the proportions of promoted files between males and females was not due to chance, but **due to an actual effect of gender** (promotion and gender are dependent).

Application Activity: Simulating the Experiment

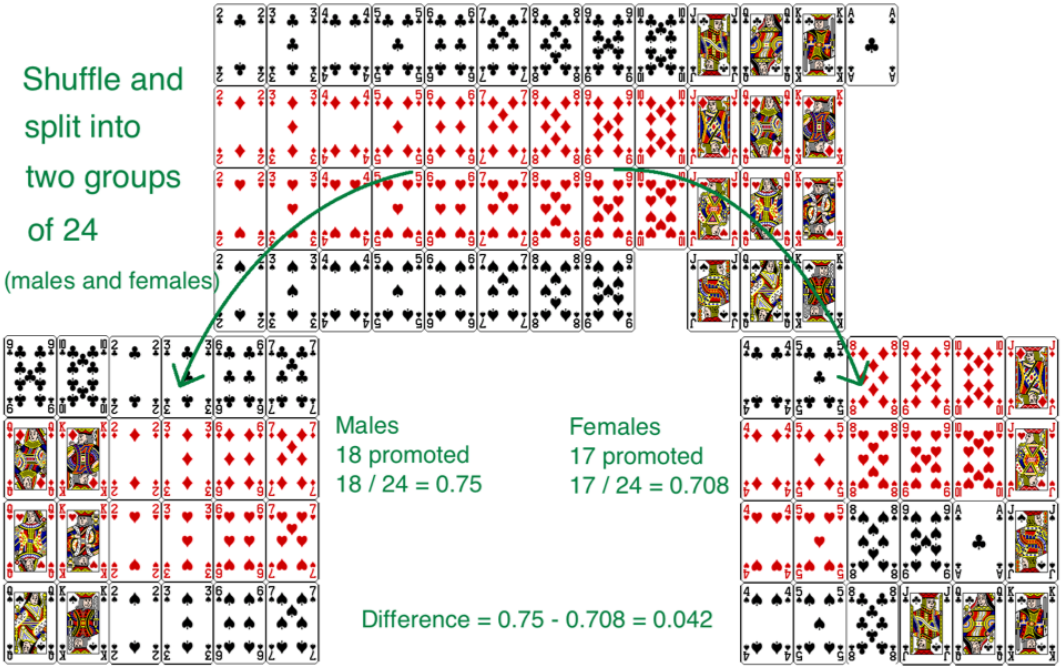
Use a deck of playing cards to simulate this experiment.

- 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).
- Shuffle the cards and deal them intro two groups of size 24, representing males and females.
- Count and record how many files in each group are promoted (number cards).
- Calculate the proportion of promoted files in each group and take the difference (male - female), and record this value.
- Repeat steps 2 - 4 many times.

Step 1



Steps 2-4


Shuffle and
split into
two groups
of 24
(males and females)
Males
18 promoted
18 / 24 = 0.75
Females
17 promoted
17 / 24 = 0.708
Difference = 0.75 - 0.708 = 0.042

Practice

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

1. 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.
2. 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.

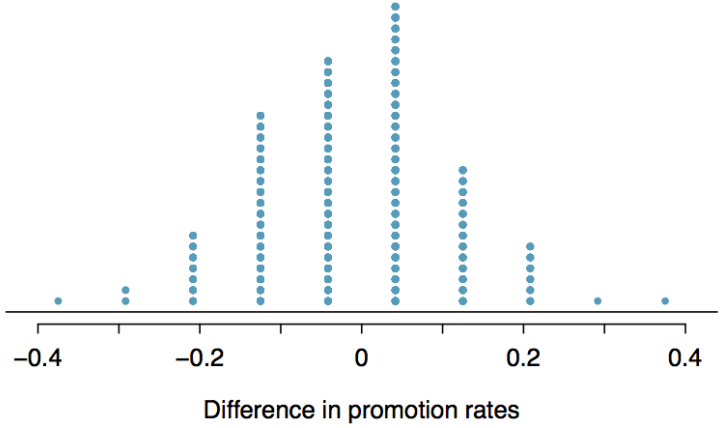
Practice

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?

1. 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.
2. *Yes, the data provide convincing evidence for the alternative hypothesis of gender discrimination against women in promotion decisions. The observed difference between the two proportions was due to a real effect of gender.*

Simulations Using Software

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



Where is our original result (from the J.Applied Psychology paper) on this graph?

Final statement

Given the simulations, we would expect to see a result as extreme as, or more extreme than, the data of 0.292 approximately 2-in-100 times. Reasonably rare!

Thus, we reject the null hypothesis, and conclude that there is evidence supporting gender discrimination against women in promotion decisions.

Case Study 2

Case Study: Tappers and Listeners

A Stanford University graduate student named Elizabeth Newton conducted an experiment using the "tapper-listener" game, where one person (the "tapper") taps a tune on a desk using their fingers, and a second person (the "listener") tries to guess what song the person is tapping out.

In her study, she recruited 120 pairs of tappers-and-listeners. About 50% of her tappers indicated that they expected that the listeners should be able to guess their song. Newton wondered whether 50% was a reasonable expectation.

Described in the book Made to Stick by Chip and Dan Heath.

Tappers & Listeners: Hypotheses

Establish two hypotheses:

H_0: the tappers are correct, and approximately 50% of the listeners will be able to determine the tune. $p_0 = 0.50$.

H_A: the tappers are incorrect, and either more or less than 50% of the listeners will be able to guess the tune. $p_0 \neq 0.50$.

Results

In Newton's study, only 3 of the 120 listeners ($\hat{p} = 0.025$) were able to guess the song!

From the point of view of the null (skeptical) hypothesis: how likely is this to happen by chance?

Simulation

How might we simulate this problem?

To simulate each individual game (tapper-listener pair), we need something that is 50-50 to represent the 50% chance of being one of the listeners who correctly guesses the tune. What do you know that's 50-50?

Simulation (continued)

Use coin flips! Let H = Heads be a successful guess, and T = Tails be a failed (incorrect) guess.

```
rbinom(n = 1, size = 1, prob = 0.5)
```

```
## [1] 0
```

So in our first simulated game, the listener successfully guessed the song (represented by flipping a coin and achieving Heads). Now, we actually need to expand this single game into a full-fledged experiment. Newton had 120 pairs, so we will simulate 120 coin flips, and that will represent a single experiment.

```
experiment <- rbinom(n = 1, size = 120, prob = 0.5) / 120
experiment
```

```
## [1] 0.55
```

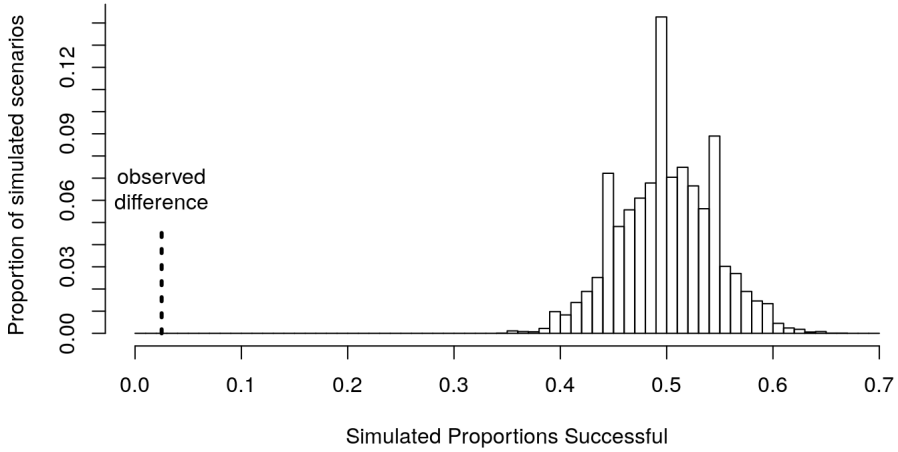
Simulation (continued)

Now, we repeat the experiment 10,000 times to see how likely it might be, by chance, to obtain a result of $\hat{p} = 0.025$.

```
results <- rbinom(n = 10000, size = 120, prob = 0.5) / 120
print(results[1:30])
```

```
##  [1] 0.5333333 0.4500000 0.5166667 0.6083333 0.4500000 0.5250000 0.4916667
##  [8] 0.4833333 0.5250000 0.5000000 0.5083333 0.5583333 0.5166667 0.4666667
## [15] 0.4916667 0.5166667 0.4750000 0.5500000 0.4416667 0.4916667 0.4583333
## [22] 0.4916667 0.5416667 0.4916667 0.5666667 0.5333333 0.4500000 0.4750000
## [29] 0.4416667 0.5416667
```

Plot the Results



Then, of the 10,000 simulated experiments, 0 of them (0) gave differences less-than-or-equal-to the data difference of 0.025.

p-value

So what is our *p*-value in this case? Less than 1-in-10000! And twice that (doubling, to get the two-tailed *p*-value) is still tiny.

Since we have a *p*-value less than 0.0001, we have extremely strong evidence for rejecting the null: the proportion of listeners able to determine which song is being tapped is most definitely not $p_0 = 0.50$.

Note: in a case like this, even though we have a two-tailed hypothesis test, the *p*-value is so small, and the data result so far out into one (lower) tail, that we can actually acknowledge this, and say that this data provides strong evidence that the chance that a listener will guess the correct tune is **less** than 50%.

The infer package

Packages in R

Some of you may have noticed that when you start a new project in RStudio, you often have to install packages to get things like R Markdown to work. This is how R is designed: the basic functionality is available and distributed as one software file, but then people can write lots of extra functions and functionality into **packages**. These packages include:

- **knitr**: the package that takes R Markdown and makes PDFs
- **rmarkdown**: the package that understands what R Markdown is
- **readr**: a package for opening Excel files in R
- **ggplot2**: a package for making really amazing figures and plots

Today we're going to conclude the lecture by exploring the use of another package designed for **statistical inference**, the thing we're doing in this lecture!

The infer Package

For any of you working along, you'll need to install the package to be able to do the following.

```
install.packages("infer")
```

Case Study: Gender Bias (redux)

Example 1: Gender Bias

In our gender bias example, we had:

- Male candidates: 21 yes, 3 no
- Female candidates: 14 yes, 10 no

We wanted to test the hypothesis of difference in proportions, that is:

$H_0 : p_m - p_f = 0$ versus $H_A : p_m - p_f \neq 0$

Our statistic for this is 21/24 - 14/24 = 0.2916667.

Let's use infer to do this!

Example 1: Infer

Setup some data:

```
gender_study <- data.frame(Subject = 1:48,
                           Promotion = c(rep("Promote", 21), rep("Not", 3),
                                         rep("Promote", 14), rep("Not", 10)),
                           Gender = c(rep("M", 24), rep("F", 24)))
gender_study[c(1:4, 22:26), ]
```

```
##    Subject Promotion Gender
## 1        1   Promote      M
## 2        2   Promote      M
## 3        3   Promote      M
## 4        4   Promote      M
## 22      22       Not      M
## 23      23       Not      M
## 24      24       Not      M
## 25      25   Promote      F
## 26      26   Promote      F
```

Example 1: Infer

Now compute the statistic that we already know:

```
diff_hat <- gender_study %>%
   specify(Promotion ~ Gender, success = "Promote") %>%
   calculate(stat = "diff in props", order = c("M", "F"))
diff_hat
```

```
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1 0.292
```

And that's the test statistic we had before: the **difference in proportions** between male promotion rates and female promotion rates.

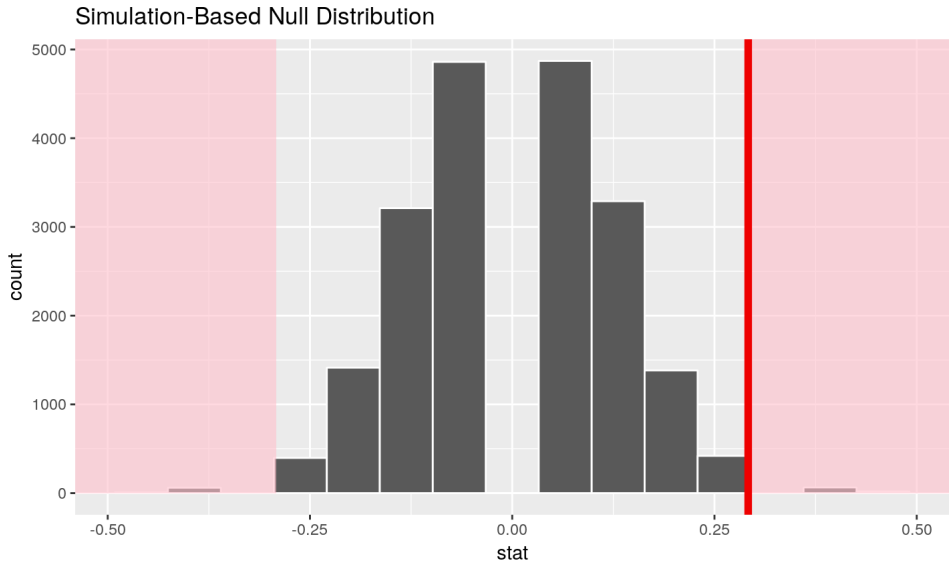
Example 1: Infer (Calculate)

Now, let's compute

```
null_distn <- gender_study %>%
  specify(Promotion ~ Gender, success = "Promote") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 20000) %>%
  calculate(stat = "diff in props", order = c("M", "F"))
```

Example 1: Infer (Plot)

```
visualize(null_distn) +
  shade_p_value(obs_stat = diff_hat, direction = "two_sided")
```

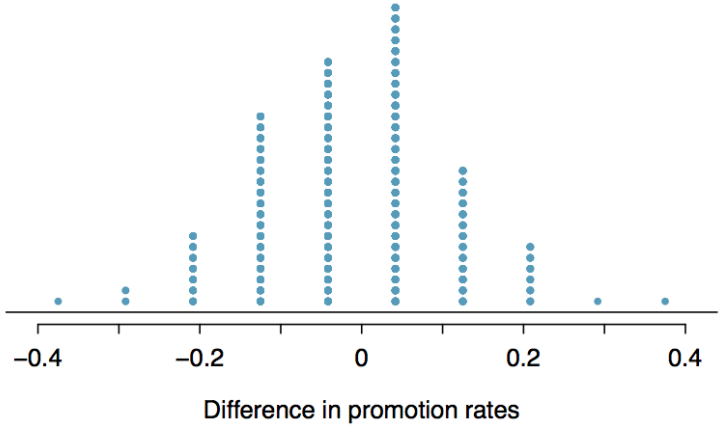


Example 1: Infer (p-value)

```
null_distn %>%
  get_p_value(obs_stat = diff_hat, direction = "two_sided")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1  0.0503
```

Example 1: The Original Plot



Case Study: Tappers and Listeners (redux)

Example 2: Tappers and Listeners

We have data again: 3 successful identifications of the song, and 117 failures. For this approach, we don't permute the data - our null hypothesis is that 50% of the people should get the song identification correct. So we simulate from that null distribution, and see how probable our result was.

Example 2: Infer (setup)

```
tappers <- data.frame(Subject = 1:120,
                      Success = c(rep("Y", 3),
                                  rep("N", 117)))
p_hat <- tappers %>%
  summarize(mean(Success == "Y")) %>%
  pull()
```

Here we again made a data frame with our experimental results in it, and calculated our $\hat{p}$ value: 0.025, as before.

Example 2: Infer (compute)

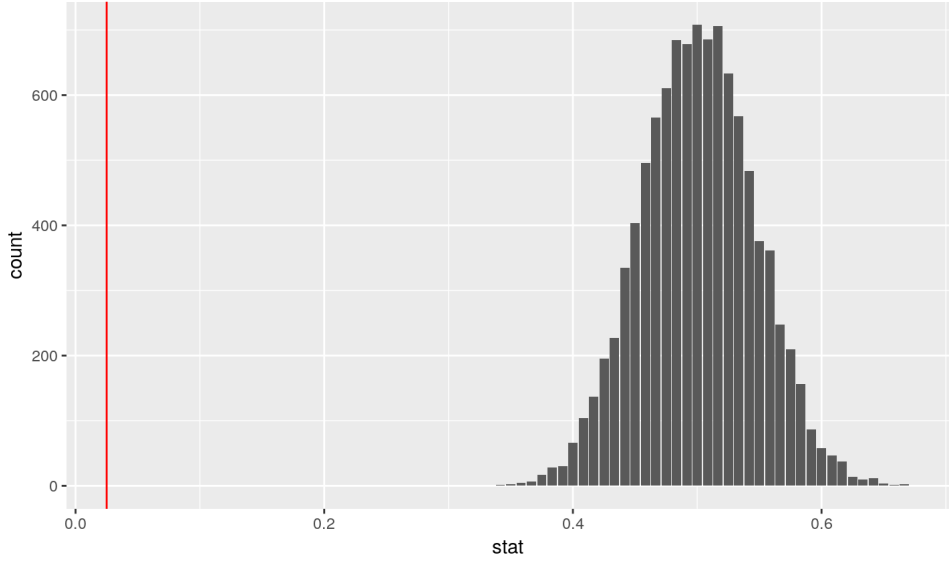
```
null_distn <- tappers %>%
  specify(response = Success, success = "Y") %>%
  hypothesize(null = "point", p = 0.5) %>%
  generate(reps = 10000, type = "simulate") %>%
  calculate(stat = "prop")

null_distn %>%
  get_p_value(obs_stat = p_hat, direction = "two_sided")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1       0
```

Example 2: Infer (graph)

```
ggplot(null_distn, aes(x = stat)) +
  geom_bar() +
  geom_vline(xintercept = p_hat, color = "red")
```



Example 2: Old Graphic