Online CDS 101 *Introduction to Computational and Data Sciences*

Inference and simulation

Simulating the gender discrimation experiment in R

Download and load the dataset

You can follow along by downloading and loading the dataset by placing the following *setup* code block at the top of a R Markdown file.

````
```{r setup, include = FALSE}
# Load required packages
library(tidyverse)
library(infer)
# Load dataset
applicants_data <- read_rds(
  url("http://data.cds101.com/gender_discrimation.rds")
)
# Observed result
experiment_result <- (21/24) - (14/24)
```
````

Recap of gender discrimination dataset

- Experiment involving 48 male bank supervisors that 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
- **Result: 29.2% more men than women were recommended for promotion**

Is this result statistically significant?

Building the null distribution

- To test the alternative hypothesis that "women were less likely than men to be hired", we need to generate a null distribution using the `infer` package:

```
simulation_results <- applicants_data %>%
  specify(outcome ~ sex, success = "Promoted") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 10000, type = "permute") %>%
  calculate(stat = "diff in props", order = combine("Male", "Female"))
```

Building the null distribution

```
simulation_results <- applicants_data %>%
  specify(outcome ~ sex, success = "Promoted") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 10000, type = "permute") %>%
  calculate(stat = "diff in props", order = combine("Male", "Female"))
```

- In `specify(outcome ~ sex, success = "Promoted")`, the first part `outcome ~ sex` is a formula where the lefthand variable `outcome` is the response and the righthand variable `sex` is explanatory.

Building the null distribution

```
simulation_results <- applicants_data %>%
  specify(outcome ~ sex, success = "Promoted") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 10000, type = "permute") %>%
  calculate(stat = "diff in props", order = combine("Male", "Female"))
```

- In `specify(outcome ~ sex, success = "Promoted")`, the first part `outcome ~ sex` is a formula where the lefthand variable `outcome` is the response and the righthand variable `sex` is explanatory.
- In `hypothesize(null = "independence")`, we specify that we will simulate what will happen if `outcome` and `sex` were independent.

Building the null distribution

```
simulation_results <- applicants_data %>%
  specify(outcome ~ sex, success = "Promoted") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 10000, type = "permute") %>%
  calculate(stat = "diff in props", order = combine("Male", "Female"))
```

- In `specify(outcome ~ sex, success = "Promoted")`, the first part `outcome ~ sex` is a formula where the lefthand variable `outcome` is the response and the righthand variable `sex` is explanatory.
- In `hypothesize(null = "independence")`, we specify that we will simulate what will happen if `outcome` and `sex` were independent.
- In `generate(reps = 10000, type = "permute")`, we specify that we will run 10,000 simulations by permuting the `outcome` and `sex` columns
- To permute, we randomly shuffle the data in the `outcome` column, and then randomly shuffle the data in the `sex` column

Building the null distribution

```
simulation_results <- applicants_data %>%
  specify(outcome ~ sex, success = "Promoted") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 10000, type = "permute") %>%
  calculate(stat = "diff in props", order = combine("Male", "Female"))
```

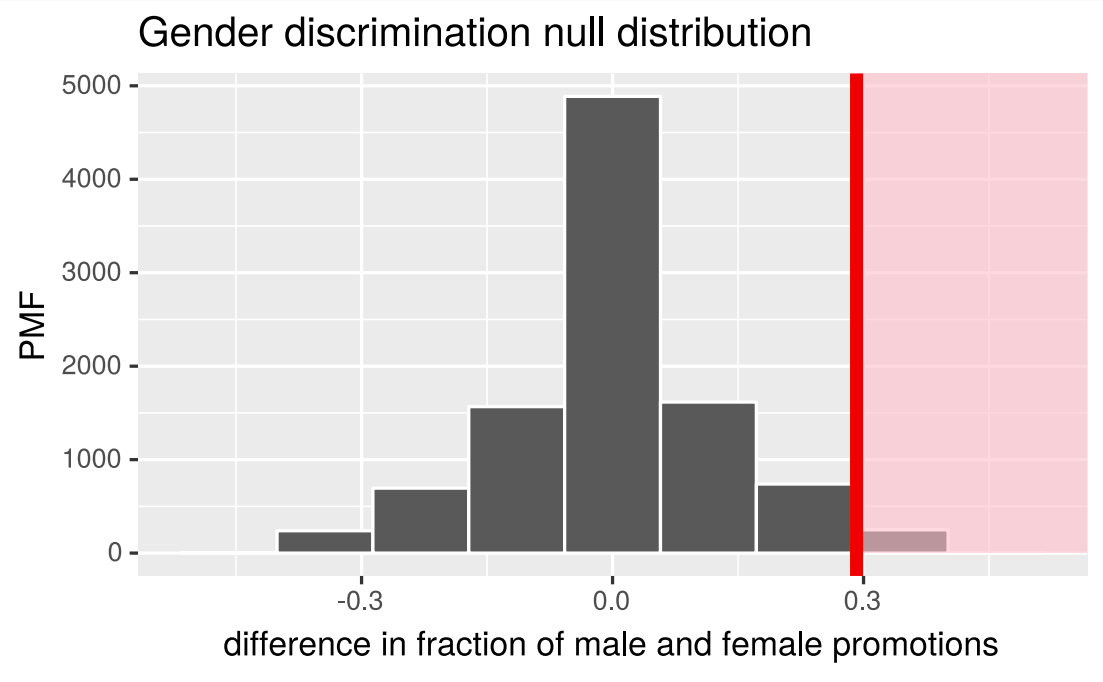
- Using `calculate(stat = "diff in props", order = combine("Male", "Female"))` means, after each simulation, compute:

$$\frac{\text{Promoted Men}}{\text{Total Men}} - \frac{\text{Promoted Men}}{\text{Total Men}}$$

Note that this expression is exactly how `experiment_result` was calculated.

Visualizing the null distribution

```
simulation_results %>%
  visualize(bins = 9) +
  shade_p_value(obs_stat = experiment_result, direction = "right") +
  labs(
    title = "Gender discrimination null distribution",
    x = "difference in fraction of male and female promotions",
    y = "PMF"
  )
```



Probability of randomly getting result

One way we can quantify the probability that the experiment's result is due to statistical variance is to compute how much of the null distribution represents outcomes that are the same or more extreme than the actual experiment:

```
simulation_results %>%
  filter(stat >= experiment_result) %>%
  summarize(random_result_probability = n() / 10000)
```

random_result_probability
0.0253

This value is called the *p*-value, and can be obtained directly using `infer`:

```
simulation_results %>%
  get_pvalue(obs_stat = experiment_result, direction = "right")
```

p_value
0.0253

Conclusions from our simulation

Do the results of the simulation 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 us to reject the null hypothesis in favor of 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.

Conclusions from our simulation

Do the results of the simulation 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 us to reject the null hypothesis in favor of 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.

Credits

License

Acknowledgments

Content adapted from the Chapter 3 OpenIntro Statistics slides developed by Mine Çetinkaya-Rundel and made available under the CC BY-SA 3.0 license.