Online CDS 101 *Introduction to Computational and Data Sciences*

Inference and simulation

Simulating the confidence interval

Example: Constructing a confidence interval

How do we construct a 95% confidence interval for the gender discrimination experiment?

Example: Constructing a confidence interval

How do we construct a 95% confidence interval for the gender discrimination experiment?

We can use the bootstrap simulation from `infer`:

Example: Constructing a confidence interval

How do we construct a 95% confidence interval for the gender discrimination experiment?

We can use the bootstrap simulation from `infer`:

1. Copy over code used to generate 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"))
```

Example: Constructing a confidence interval

How do we construct a 95% confidence interval for the gender discrimination experiment?

We can use the bootstrap simulation from `infer`:

1. Copy over code used to generate the null distribution.
2. Change the assigned variable name.

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

Example: Constructing a confidence interval

How do we construct a 95% confidence interval for the gender discrimination experiment?

We can use the bootstrap simulation from `infer`:

1. Copy over code used to generate the null distribution.
2. Change the assigned variable name.
3. Remove the `hypothesize()` line.

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

Example: Constructing a confidence interval

How do we construct a 95% confidence interval for the gender discrimination experiment?

We can use the bootstrap simulation from `infer`:

1. Copy over code used to generate the null distribution.
2. Change the assigned variable name.
3. Remove the `hypothesize()` line.
4. Change the type keyword to `"bootstrap"` in `generate()`.

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

Example: Constructing a confidence interval

How do we construct a 95% confidence interval for the gender discrimination experiment?

After running the bootstrap simulation, we can obtain the 95% confidence interval using `get_confidence_interval()`,

```
simulation_ci <- simulation_bootstrap %>%
  get_confidence_interval()
```

Example: Constructing a confidence interval

How do we construct a 95% confidence interval for the gender discrimination experiment?

After running the bootstrap simulation, we can obtain the 95% confidence interval using `get_confidence_interval()`,

```
simulation_ci <- simulation_bootstrap %>%
  get_confidence_interval()
```

2.5%	97.5%
0.047619	0.5287136

Example: Constructing a confidence interval

How do we construct a 95% confidence interval for the gender discrimination experiment?

After running the bootstrap simulation, we can obtain the 95% confidence interval using `get_confidence_interval()`,

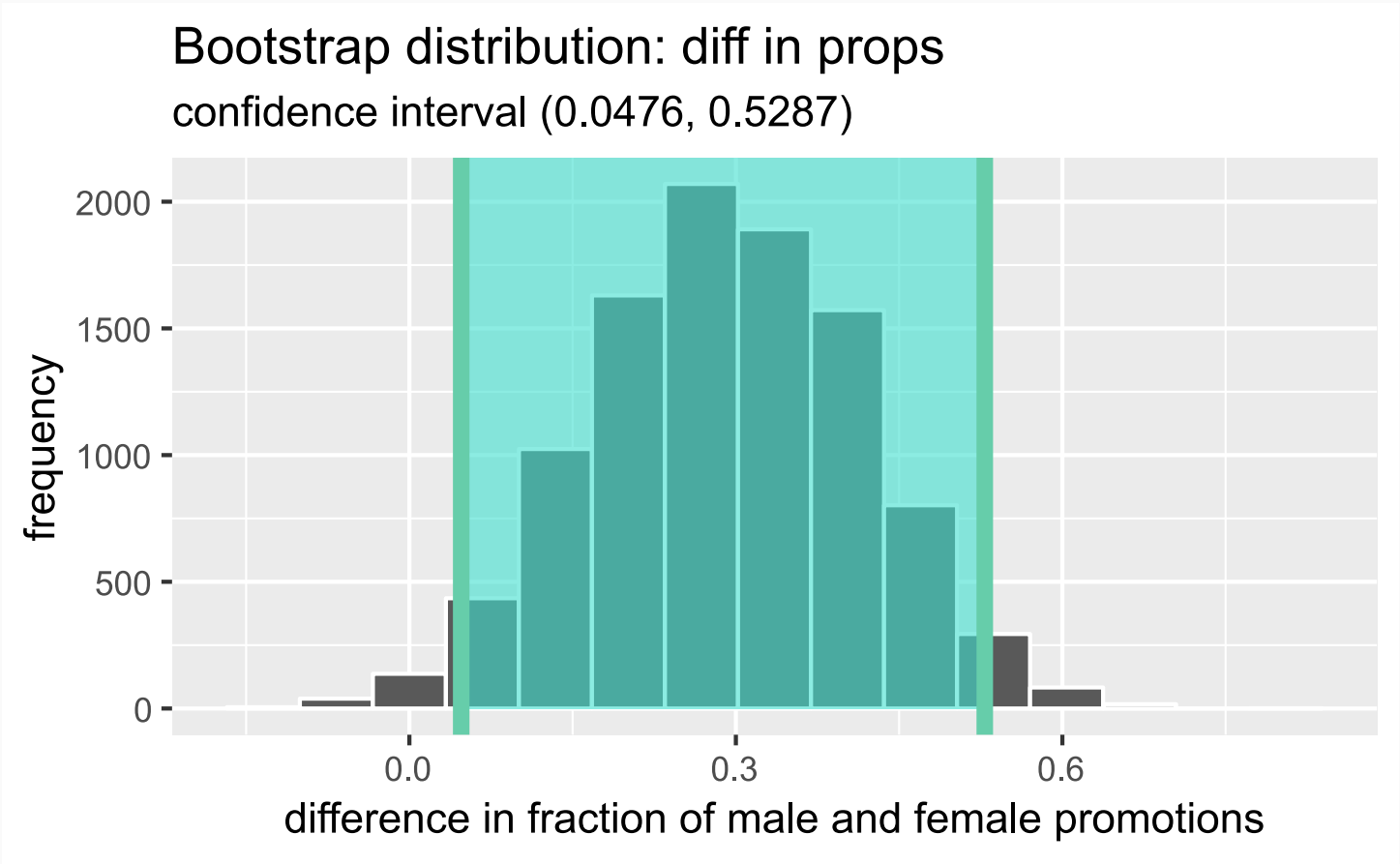
```
simulation_ci <- simulation_bootstrap %>%
  get_confidence_interval()
```

2.5%	97.5%
0.047619	0.5287136

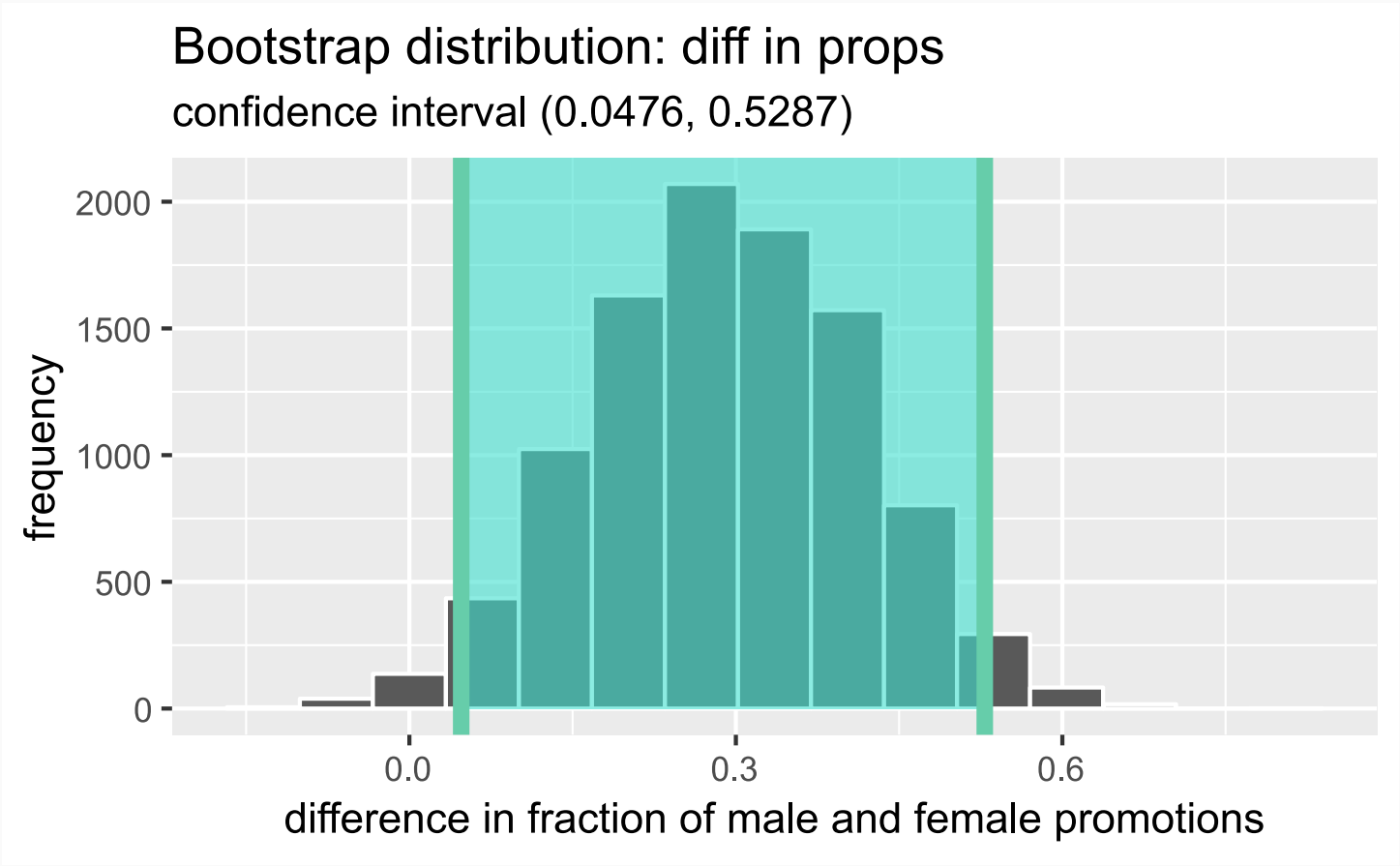
To visualize the confidence interval, we use the `visualize()` and `shade_confidence_interval` functions,

```
simulation_bootstrap %>%
  visualize() +
  shade_confidence_interval(simulation_ci)
```

Example: Constructing a confidence interval



Example: Constructing a confidence interval



Interpretation: All else kept equal, men are, on average, between 4.8% to 52.9% more likely to be promoted than women.

What does 95% confident mean?

- Suppose we generated a series of bootstrap distributions with `infer` and found the 95% confidence interval for each one

What does 95% confident mean?

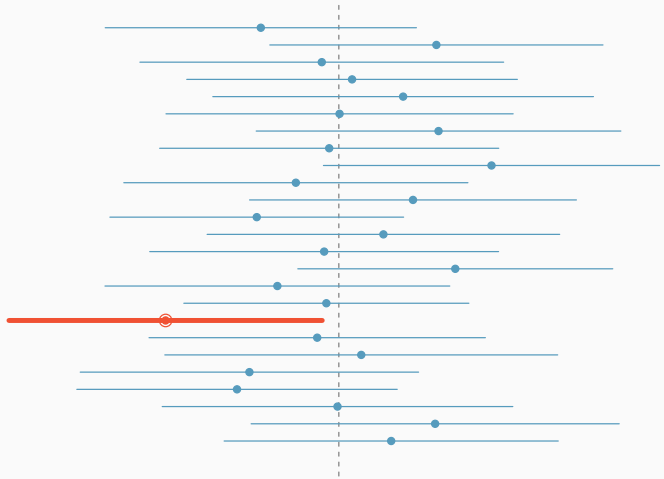
- Suppose we generated a series of bootstrap distributions with `infer` and found the 95% confidence interval for each one
- Let's also assume that, for one reason or another, we somehow knew the value of the true population mean

What does 95% confident mean?

- Suppose we generated a series of bootstrap distributions with `infer` and found the 95% confidence interval for each one
- Let's also assume that, for one reason or another, we somehow knew the value of the true population mean
- Then, the phrase "95% confident" means that about 95% of those intervals would contain the true population mean

What does 95% confident mean?

- Suppose we generated a series of bootstrap distributions with `infer` and found the 95% confidence interval for each one
- Let's also assume that, for one reason or another, we somehow knew the value of the true population mean
- Then, the phrase "95% confident" means that about 95% of those intervals would contain the true population mean
- The figure shows this process with 25 samples, where 24 of the resulting confidence intervals contain the true average number of exclusive relationships, and one does not.



Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

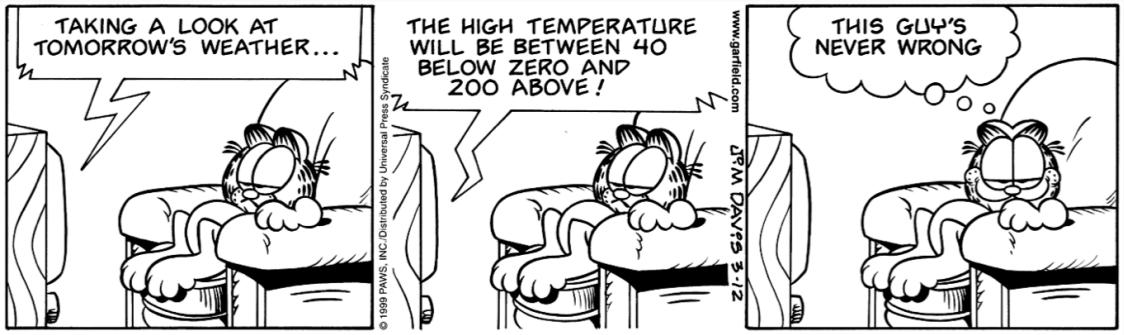
Can you see any drawbacks to using a wider interval?

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Can you see any drawbacks to using a wider interval?



If the interval is too wide it may not be very informative.

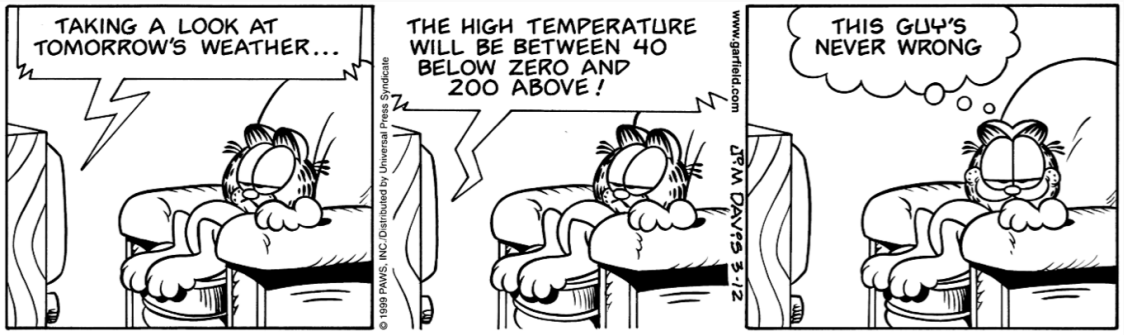
Image source (defunct): http://web.as.uky.edu/statistics/users/earo227/misc/garfield_weather.gif

Width of an interval

If we want to be more certain that we capture the population parameter, i.e. increase our confidence level, should we use a wider interval or a smaller interval?

A wider interval.

Can you see any drawbacks to using a wider interval?



If the interval is too wide it may not be very informative.

Commonly used confidence levels in practice are 90%, 95%, 98%, and 99%.

Image source (defunct): http://web.as.uky.edu/statistics/users/earo227/misc/garfield_weather.gif

Credits

License

Acknowledgments

Content adapted from the Chapter 4 OpenIntro Statistics slides developed by Mine Çetinkaya-Rundel and made available under the CC BY-SA 3.0 license.