Online CDS 101 *Introduction to Computational and Data Sciences*

Inference and simulation

One-sided hypothesis tests using `infer`

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
college_apps <- read_rds(
  url("http://data.cds101.com/college_applications.rds")
)
```
````

Number of college applications

A survey asked how many colleges students applied to, and 206 students responded to this question. This sample yielded an average of 9.7 college applications with a standard deviation of 7. College Board website states that counselors recommend students apply to roughly 8 colleges. Do these data provide convincing evidence that the average number of colleges all GMU students apply to is *higher* than recommended?

http://www.collegeboard.com/student/apply/the-application/151680.html

Setting the hypotheses

- The **parameter of interest** is the average number of schools applied to by *all* GMU students.
- There may be two explanations why our sample mean is higher than the recommended 8 schools.
  - The true population mean is different
  - The true population mean is 8, and the difference between the true population mean and the sample mean is simply due to natural sampling variability
- We start with the assumption the average number of colleges GMU students apply to is 8 (as recommended)

$$H_0 : \mu = 8$$

- We test the claim that the average number of colleges GMU students apply to is greater than 8

$$H_A : \mu > 8$$

Statistical significance

Say that we conducted this study by polling an independent and representative sample of GMU students about how many colleges they applied to, and obtained a sample mean of 9.7.

The national average is 8.

Is this result statistically significant?

Statistical significance

Say that we conducted this study by polling an independent and representative sample of GMU students about how many colleges they applied to, and obtained a sample mean of 9.7.

The national average is 8.

Is this result statistically significant?

In order to evaluate if the observed sample mean is unusual for the hypothesized sampling distribution, we do the following:

- Choose a value for the significance level α (a common choice is 5%)
- Determine the percentile rank of the observed sample mean relative to the null distribution

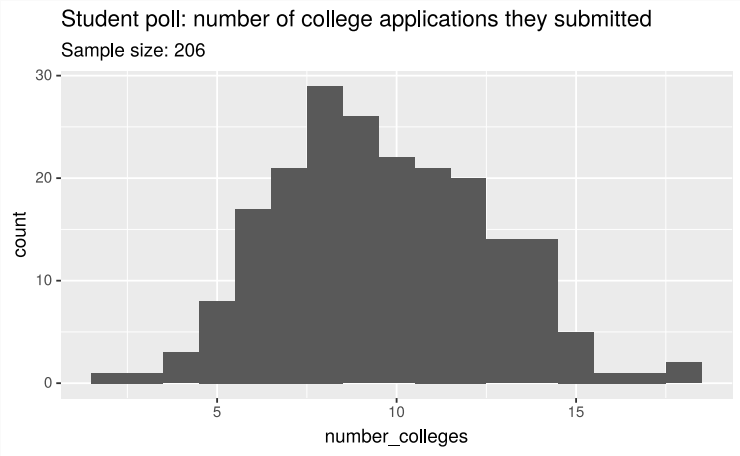
p-values

- We then use the percentile to calculate the **p-value**, the probability of observing data at least as favorable to the alternative hypothesis as our current data set, if the null hypothesis were true.
- If the p-value is **lower** than the significance level α, we say that it would be very unlikely to observe the data if the null hypothesis were true, and hence **reject H_0**.
- If the p-value is **higher** than α, we say that it is likely to observe the data even if the null hypothesis were true, and hence **do not reject H_0**.

Number of college applications p-value

p-value

probability of observing data at least as favorable to H_A as our current data set (a sample mean greater than 9.7), if in fact H_0 were true (the true population mean was 8).



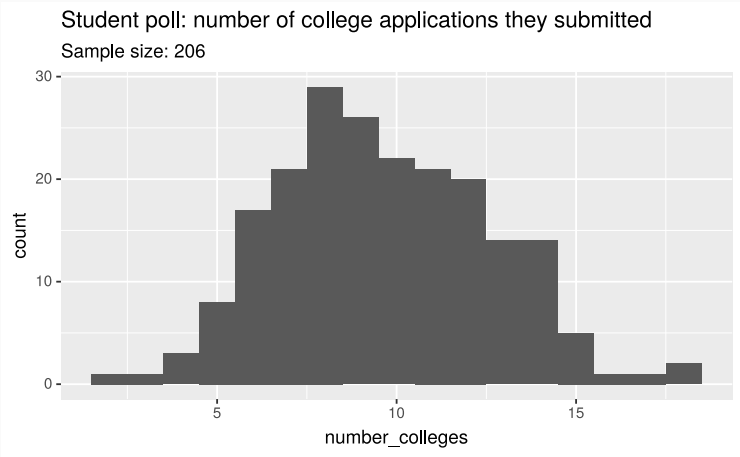

```
college_apps_null <- college_apps %>%
  specify(formula = number_colleges ~ NULL) %>%
  hypothesize(null = "point", mu = 8) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "mean")

college_apps_p_value <- college_apps_null %>%
  get_pvalue(obs_stat = 9.7, direction = "right")
```

Number of college applications p-value

p-value

probability of observing data at least as favorable to H_A as our current data set (a sample mean greater than 9.7), if in fact H_0 were true (the true population mean was 8).



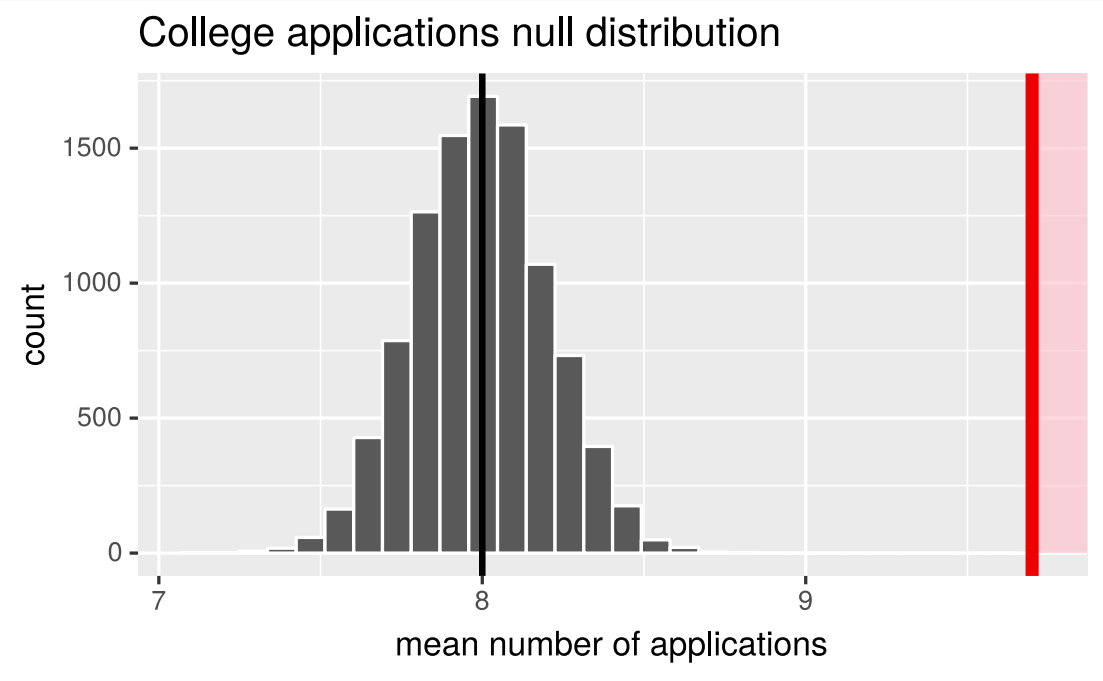

```
college_apps_null <- college_apps %>%
  specify(formula = number_colleges ~ NULL) %>%
  hypothesize(null = "point", mu = 8) %>%
  generate(reps = 10000, type = "bootstrap") %>%
  calculate(stat = "mean")

college_apps_p_value <- college_apps_null %>%
  get_pvalue(obs_stat = 9.7, direction = "right")
```

p-value = 0

Number of applications p-value

```
college_apps_null %>%
  visualize(bins = 30) +
  shade_p_value(obs_stat = 9.7, direction = "right") +
  geom_vline(xintercept = 8, size = 1) +
  labs(
    x = "mean number of applications",
    title = "College applications null distribution"
  )
```



Number of college applications - Making a decision

- p-value = 0
- If the true average of the number of colleges GMU students applied to is 8, there is a 0% chance of observing a random sample of 206 GMU students who on average apply to 9.7 or more schools.
- This is a pretty low probability for us to think that a sample mean of 9.7 or more schools is likely to happen simply by chance.
- Since p-value is **low** (lower than 5%) we **reject H_0**.
- The data provide convincing evidence that GMU students apply to more than 8 schools on average.
- The difference between the null value of 8 schools and observed sample mean of 9.7 schools is **not due to chance** or sampling variability.

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.