Hypothesis testing for comparing two means via simulation

INFERENCE FOR NUMERICAL DATA IN R

Mine Cetinkaya-Rundel
Associate Professor of the Practice,
Duke University

Motivation

- Motivating question: Does a treatment using embryonic stem cells help improve heart function following a heart attack more so than traditional therapy?
- Data: `stem.cell` data from the `openintro` package

```r
library(openintro)
data(stem.cell)
```

	trmt	before	after
1	ctrl	35.25	29.50
2	ctrl	36.50	29.50
3	ctrl	39.75	36.25
	...	...	...
n	esc	53.75	51.00

Analysis outline

Step 1. Calculate `change` for each sheep: difference between before and after heart pumping capacities for each sheep.

```
   trmt   before   after   change
1  ctrl    35.25   29.50   ?
2  ctrl    36.50   29.50   ?
3  ctrl    39.75   36.25   ?
   ...      ...     ...
n  esc     53.75   51.00   ?
```

Analysis outline

Step 2. Set the hypotheses:

$H_0 : \mu_{esc} = \mu_{ctrl}$; There is no difference between average change in treatment and control groups.

$H_A : \mu_{esc} > \mu_{ctrl}$; There is a difference between average change in treatment and control groups.

Analysis outline

Step 3. Conduct the hypothesis test.

- Write the values of `change` on 18 index cards.
- (1) Shuffle the cards and randomly split them into two equal sized decks: treatment and control.
- (2) Calculate and record the test statistic: difference in average `change` between treatment and control.
- Repeat (1) and (2) many times to generate the sampling distribution.
- Calculate p-value as the percentage of simulations where the test statistic is at least as extreme as the observed difference in sample means.

Hypothesis test: generate resamples

Use the `infer` package to conduct the test:

```
library(infer)
```

Hypothesis test: generate resamples

Start with the data frame and **specify** the model:

```
library(infer)

diff_ht_mean <- stem.cell %>%
  specify(__) %>%                    # y ~ x
  ...
```

Hypothesis test: generate resamples

Declare null hypothesis, i.e. no difference between means:

```
library(infer)

diff_ht_mean <- stem.cell %>%
  specify(__) %>%                      # y ~ x
  hypothesize(null = __) %>%           # "independence" or "point"
  ...
```

Hypothesis test: generate resamples

Generate resamples assuming H_0 is true:

```
library(infer)

diff_ht_mean <- stem.cell %>%
  specify(__) %>%                    # y ~ x
  hypothesize(null = __) %>%         # "independence" or "point"
  generate(reps = __, type = __) %>% # "bootstrap", "permute", or "simulate"
  ...
```

Hypothesis test: generate resamples

Calculate test statistic:

```
library(infer)

diff_ht_mean <- stem.cell %>%
  specify(__) %>%                         # y ~ x
  hypothesize(null = __) %>%              # "independence" or "point"
  generate(reps = _N_, type = __) %>%# "bootstrap", "permute", or "simulate"
  calculate(stat = "diff in means")  # type of statistic to calculate
```

Hypothesis test: calculate p-value

Calculate the p-value as the proportion of simulations where the simulated difference between the sample means is at least as extreme as the observed

$$P((\bar{x}_{esc,sim} - \bar{x}_{ctrl,sim}) \geq (\bar{x}_{esc,obs} - \bar{x}_{ctrl,obs}))$$

Let's practice!

INFERENCE FOR NUMERICAL DATA IN R

Bootstrap CI for difference in two means

INFERENCE FOR NUMERICAL DATA IN R

Mine Cetinkaya-Rundel
Associate Professor of the Practice,
Duke University

Bootstrap CI for a difference

1. Take a bootstrap sample *of each sample* - a random sample taken with replacement from each of the original samples, of the same size as each of the original samples.
2. Calculate the bootstrap statistic - a statistic such as *difference* in means, medians, proportion, etc. computed based on the bootstrap samples.
3. Repeat steps (1) and (2) many times to create a bootstrap distribution - a distribution of bootstrap statistics.
4. Calculate the interval using the percentile or the standard error method.

Let's practice!

INFERENCE FOR NUMERICAL DATA IN R

Comparing means with a t-test

INFERENCE FOR NUMERICAL DATA IN R

Mine Cetinkaya-Rundel
Associate Professor of the Practice,
Duke University

A (more) standard measure of pay

Instead of comparing average annual `income`, compare average `hrly_rate`:

- assume 52 weeks in a year
- `hrly_rate = income / (hrs_work * 52)`

Research question and hypotheses

> Do the data provide convincing evidence of a difference between the average hourly rate of citizens and non-citizens in the US?

Let $\mu =$ average hourly pay

$H_0 : \mu_{citizen} = \mu_{non-citizen}$

$H_A : \mu_{citizen} \neq \mu_{non-citizen}$

Summary statistics

```
acs12 %>%
  filter(!is.na(hrly_rate)) %>%
  group_by(citizen) %>%
  summarise(x_bar = round(mean(hrly_rate), 2),
            s = round(sd(hrly_rate), 2),
            n = length(hrly_rate))
```

```
  citizen  x_bar      s   n
1 no       21.19  34.50  58
2 yes      18.52  24.73 901
```

Conducting the test

```
t.test(hrly_rate ~ citizen, data = acs12, null = 0,
       alternative = "two.sided")
```

- Null:
  - $H_0 : \mu_{citizen} = \mu_{non-citizen}$
  - $H_0 : \mu_{citizen} - \mu_{non-citizen} = 0 \rightarrow$ `null = 0`
- $H_A : \mu_{citizen} \neq \mu_{non-citizen} \rightarrow$ `alternative = "two.sided"`

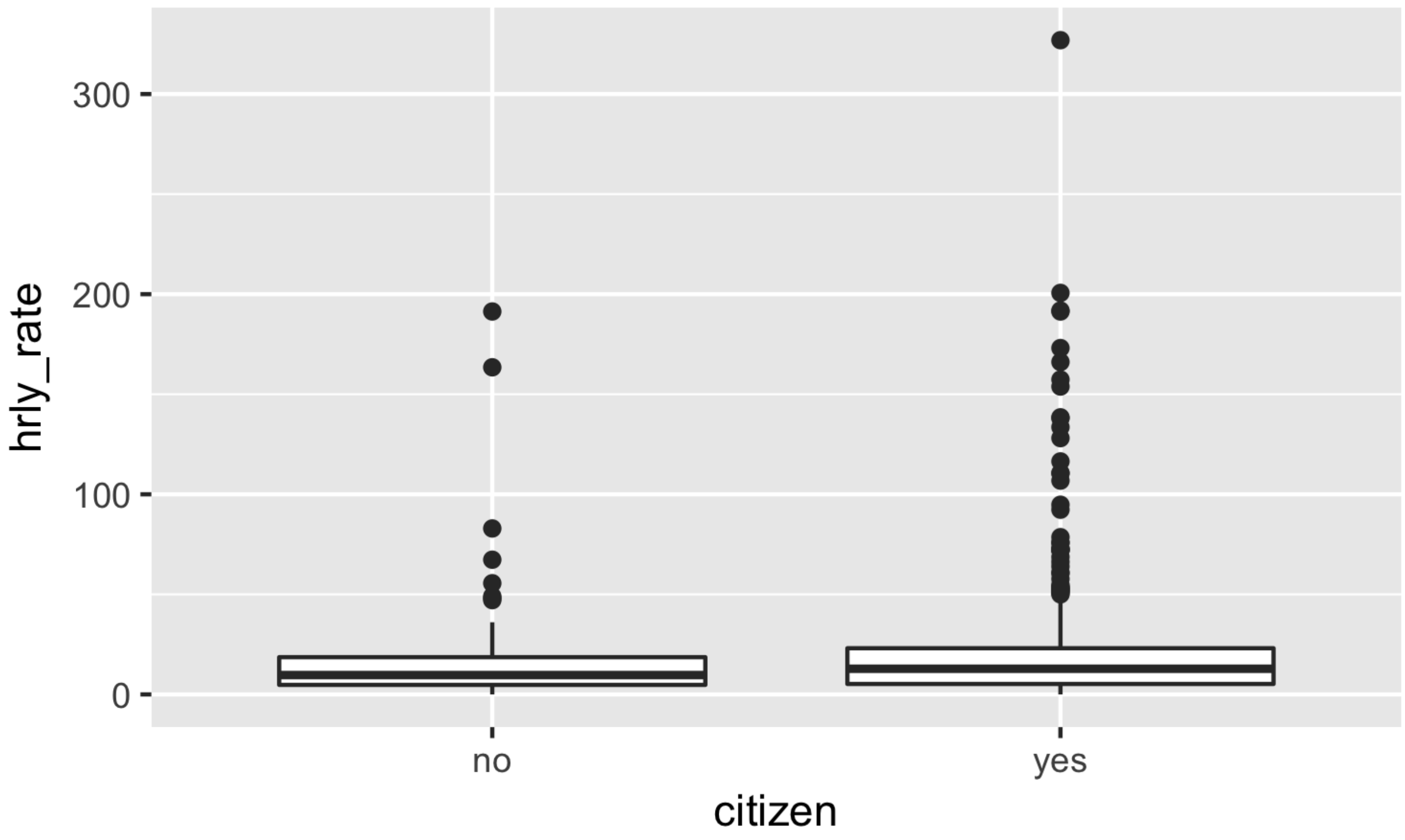
Conducting the test

```
t.test(hrly_rate ~ citizen, data = acs12, null = 0,
       alternative = "two.sided")
```

```
    Welch Two Sample t-test
data:  hrly_rate by citizen
t = 0.58058, df = 60.827, p-value = 0.5637
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -6.53483 11.88170
sample estimates:
 mean in group no mean in group yes
         21.19494          18.52151
```

Conditions

- Independence:
  - Observations in each sample should be independent of each other.
  - The two samples should be independent of each other.
- Sample size / skew: The more skewed the original data, the higher the sample size required to have a symmetric sampling distribution.



Let's practice!

INFERENCE FOR NUMERICAL DATA IN R