Tidy Inference in R

Data Science Programming

Alison Hill

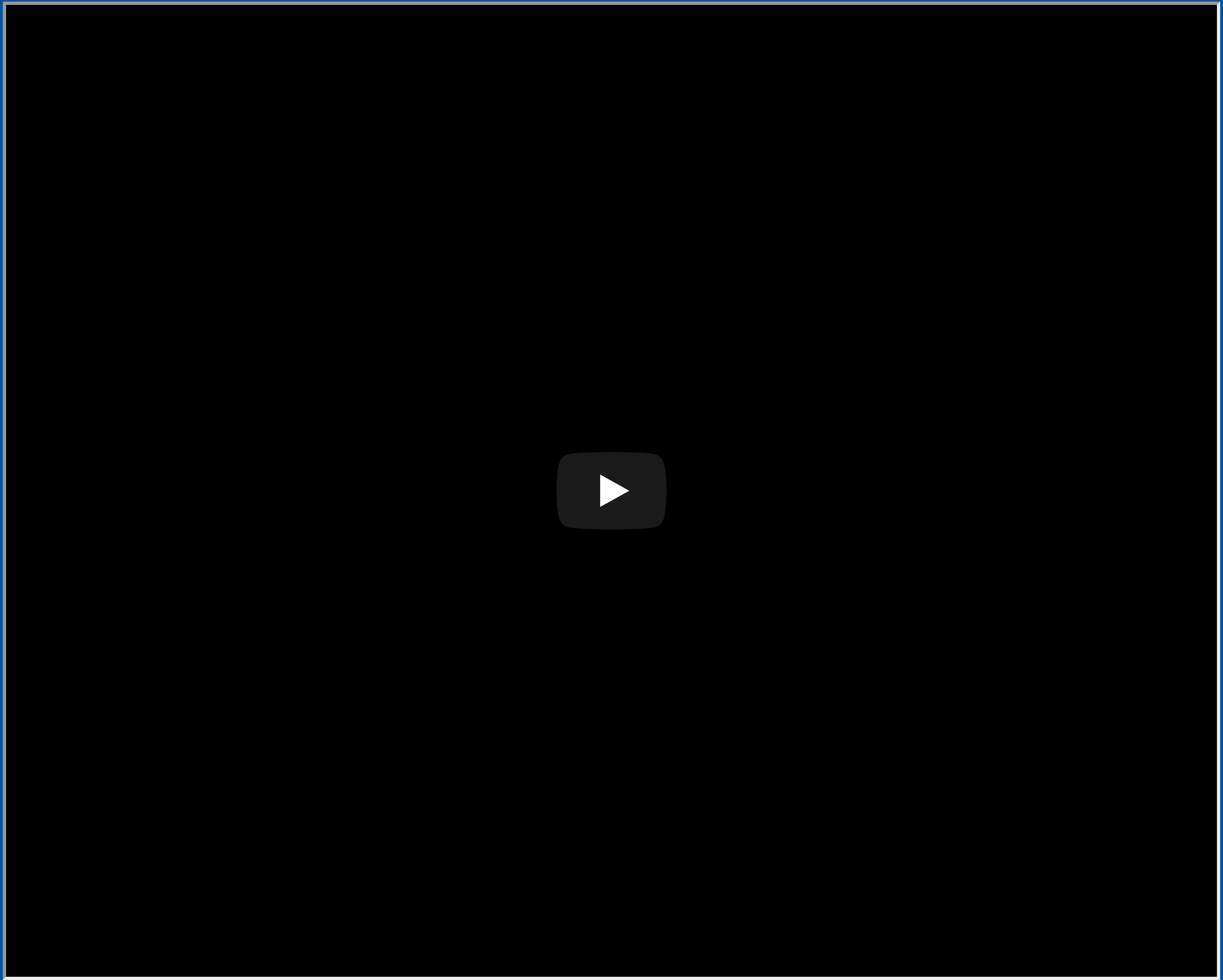
2018-03-09

Install Packages

```r
install.packages("tidyverse")
install.packages("broom")
# today I'll use the development version
install.packages("remotes")
remotes::install_github("andrewpbray/infer", ref = "develop")
install.packages("devtools")
devtools::install_github("sfirke/janitor")
```

Load Packages

```r
library(tidyverse)
library(infer)
library(janitor)
library(broom)
```



Research Question

If you see someone else yawn, are you more likely to yawn?

In an episode of the show **Mythbusters**, they tested the myth that yawning is contagious.

Participants and Procedure

Participants and Procedure

- $N = 50$ adults who thought they were being considered for an appearance on the show.

Participants and Procedure

- $N = 50$ adults who thought they were being considered for an appearance on the show.

- Each participant was interviewed individually by a show recruiter ("confederate") who either

  - yawned, $n = 34$
  - or not, $n = 16$.

Participants and Procedure

- $N = 50$ adults who thought they were being considered for an appearance on the show.
- Each participant was interviewed individually by a show recruiter ("confederate") who either
  - yawned, $n = 34$
  - or not, $n = 16$.
- Participants then sat by themselves in a large van and were asked to wait.

Participants and Procedure

- $N = 50$ adults who thought they were being considered for an appearance on the show.
- Each participant was interviewed individually by a show recruiter ("confederate") who either
  - yawned, $n = 34$
  - or not, $n = 16$.
- Participants then sat by themselves in a large van and were asked to wait.
- While in the van for some set amount of time (unknown), the Mythbusters watched to see if the unaware participants yawned.

Data

Two group design:

- $n = 34$ saw the confederate yawn (*seed*)
- $n = 16$ did not see the confederate yawn (*control*)

Data

Two group design:

- $n = 34$ saw the confederate yawn (*seed*)
- $n = 16$ did not see the confederate yawn (*control*)

```r
group ← c(rep("control", 12), rep("seed", 24),
            rep("control", 4), rep("seed", 10))
yawn ← c(rep(0, 36), rep(1, 14))

yawn_myth ← data_frame(subj = seq(1, 50), group, yawn) %>%
  mutate(yawn = as.factor(yawn))

glimpse(yawn_myth)
```

```
Observations: 50
Variables: 3
$ subj  <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 1...
$ group <chr> "control", "control", "control", "control", "control", "...
$ yawn  <fct> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,...
```

Show 8 entries Search:

	subj	group	yawn
1	1	control	0
2	2	control	0
3	3	control	0
4	4	control	0
5	5	control	0
6	6	control	0
7	7	control	0
8	8	control	0

Showing 1 to 8 of 50 entries

Previous 1 2 3 4 5 6 7 Next

Results

```
yawn_myth %>%
  tabyl(group, yawn) %>%
  adorn_percentages() %>%
  adorn_pct_formatting() %>%
  adorn_ns()
```

```
  group           0          1
control 75.0% (12) 25.0%  (4)
   seed 70.6% (24) 29.4% (10)
```

Conclusion

Conclusion

Finding: CONFIRMED[1]

[1] http://www.discovery.com/tv-shows/mythbusters/mythbusters-database/yawning-contagious/

Really?

> "Though that's not an enormous increase, since they tested 50 people in the field, the gap was still wide enough for the MythBusters to confirm that yawning is indeed contagious."[1]

[1] http://www.discovery.com/tv-shows/mythbusters/mythbusters-database/yawning-contagious/

State the hypotheses

State the hypotheses

H_0:

> There is no difference between the seed and control groups in the proportion of people who yawned.

State the hypotheses

H_0:

> There is no difference between the seed and control groups in the proportion of people who yawned.

H_1 (directional):

> More people (relatively) yawned in the seed group than in the control group.

Test the hypothesis

Which type of hypothesis test would you conduct here?

- Independent samples t-test
- Two proportion test
- Chi-square test of independence
- Analysis of Variance
- I don't know!

Test the hypothesis

Which type of hypothesis test would you conduct here?

- Independent samples t-test
- Two proportion test
- Chi-square test of independence
- Analysis of Variance
- I don't know!

Test the hypothesis

Which type of hypothesis test would you conduct here?

- Independent samples t-test
- Two proportion test
- Chi-square test of independence
- Analysis of Variance
- I don't know!

Answer:

- Two proportion test

Two proportion test

Two proportion test

$$H_0 : p_{seed} = p_{control}$$

Two proportion test

$$H_0 : p_{seed} = p_{control}$$

$$H_1 : p_{seed} > p_{control}$$

The observed difference

```r
yawn_myth %>%
  group_by(group) %>%
  summarize(prop = mean(yawn == 1))
```

```
# A tibble: 2 x 2
  group    prop
  <chr>   <dbl>
1 control 0.250
2 seed    0.294
```

The observed difference

```
yawn_myth %>%
  group_by(group) %>%
  summarize(prop = mean(yawn == 1))
```

```
# A tibble: 2 x 2
  group    prop
  <chr>   <dbl>
1 control 0.250
2 seed    0.294
```

```
(obs_diff <- yawn_myth %>%
  group_by(group) %>%
  summarize(prop = mean(yawn == 1)) %>%
  summarize(diff(prop)) %>%
  pull())
```

```
[1] 0.04411765
```

Is this difference *meaningful*?

Is this difference *meaningful*?

Different question:

Is this difference *meaningful*?

Different question:

Is this difference *significant*?

Modeling the null hypothesis

If...

$$H_0 : p_{seed} = p_{control}$$

is true, then whether or not the participant saw someone else yawn does not matter: there is no association between exposure and yawning.

Original universe

```
# A tibble: 12 x 3
    subj group   yawn
   <int> <chr>   <fct>
 1     1 control 0
 2     2 control 0
 3     3 control 0
 4     4 control 0
 5     5 control 0
 6     6 control 0
 7    15 seed    0
 8    16 seed    0
 9    17 seed    0
10    18 seed    0
11    19 seed    0
12    20 seed    0

   group  0  1 Total
 control 12  4    16
    seed 24 10    34
   Total 36 14    50
```

Original universe

```
# A tibble: 12 x 3
    subj group   yawn
   <int> <chr>   <fct>
 1     1 control 0
 2     2 control 0
 3     3 control 0
 4     4 control 0
 5     5 control 0
 6     6 control 0
 7    15 seed    0
 8    16 seed    0
 9    17 seed    0
10    18 seed    0
11    19 seed    0
12    20 seed    0
```

```
   group  0  1 Total
 control 12  4    16
    seed 24 10    34
   Total 36 14    50
```

Parallel universe

```
# A tibble: 12 x 3
    subj group   alt_yawn
   <int> <fct>   <fct>
 1     1 control 0
 2     2 control 0
 3     3 control 1
 4     4 control 0
 5     5 control 0
 6     6 control 0
 7    15 seed    0
 8    16 seed    1
 9    17 seed    0
10    18 seed    0
11    19 seed    0
12    20 seed    1
```

```
   group  0  1 Total
 control 12  4    16
    seed 24 10    34
   Total 36 14    50
```

1000 parallel universes

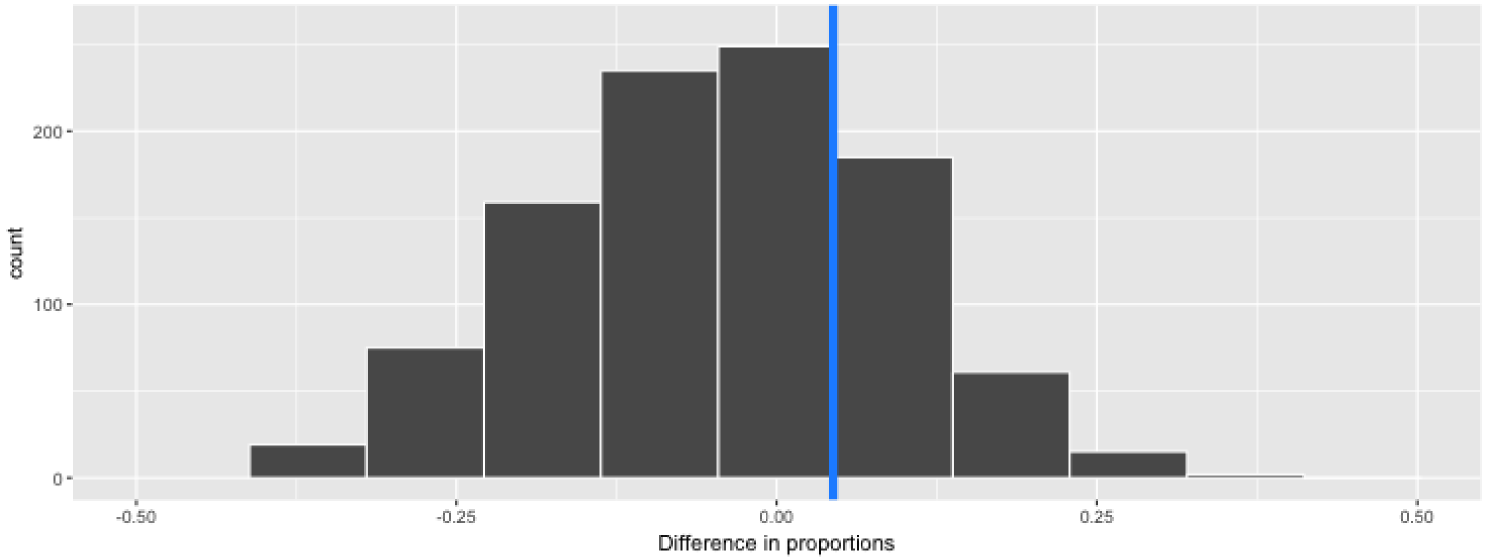
```
# A tibble: 1,000 x 2
   replicate    stat
       <int>   <dbl>
 1         1 -0.140
 2         2 -0.0478
 3         3 -0.0478
 4         4 -0.0478
 5         5 -0.140
 6         6  0.228
 7         7 -0.140
 8         8  0.136
 9         9 -0.232
10        10  0.0441
#  ... with 990 more rows
```

1000 parallel universes

```
# A tibble: 1,000 x 2
   replicate    stat
       <int>   <dbl>
 1         1 -0.140
 2         2 -0.0478
 3         3 -0.0478
 4         4 -0.0478
 5         5 -0.140
 6         6  0.228
 7         7 -0.140
 8         8  0.136
 9         9 -0.232
10        10  0.0441
# ... with 990 more rows
```

```
# A tibble: 11 x 2
   replicate    stat
       <int>   <dbl>
 1       990 -0.232
 2       991 -0.0478
 3       992 -0.232
 4       993  0.136
 5       994 -0.140
 6       995  0.136
 7       996 -0.232
 8       997  0.0441
 9       998  0.0441
10       999 -0.232
11      1000 -0.140
```

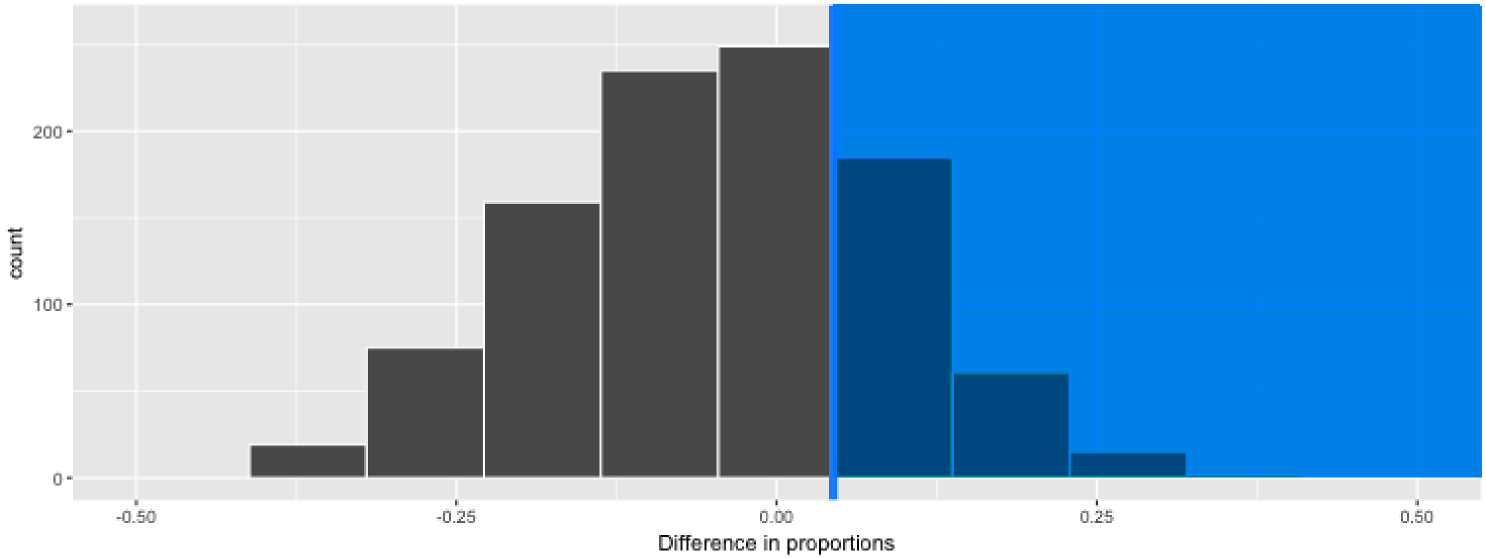
The parallel universe distribution



The distribution of 1000 differences in proportions, if the null hypothesis were *true* and yawning was not contagious.

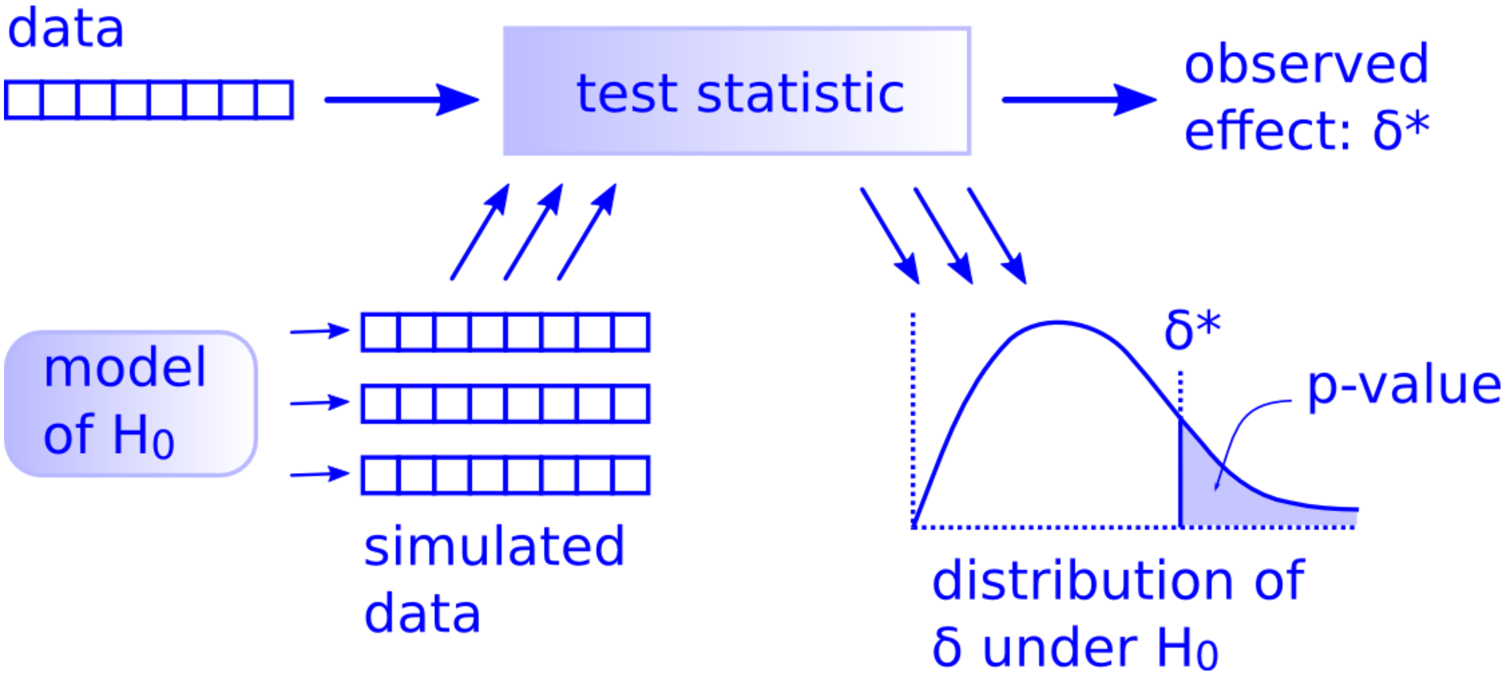
In how many of our "parallel universes" is the difference as big or bigger than the one we observed (0.0441176)?

Calculating the p-value



That proportion is the p-value!

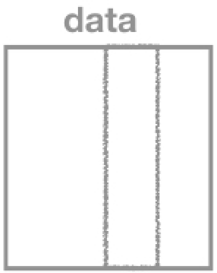
```
# A tibble: 1 x 3
  n_as_big n_total p_value
     <int>   <int>   <dbl>
1      512    1000   0.512
```


data
test statistic
observed effect: δ*
model of H0
simulated data
δ*
p-value
distribution of δ under H0

The tidy way

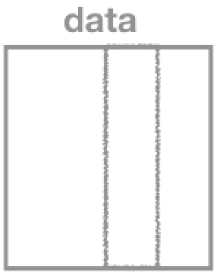
Use the infer package

Hypothesis test



https://github.com/ismayc/talks/tree/master/data-day-texas-infer

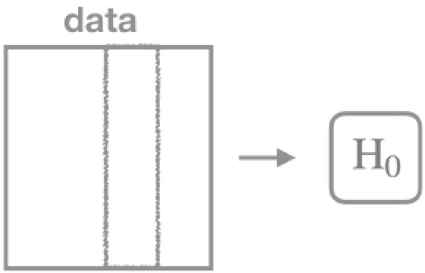
Hypothesis test



`specify()`

https://github.com/ismayc/talks/tree/master/data-day-texas-infer

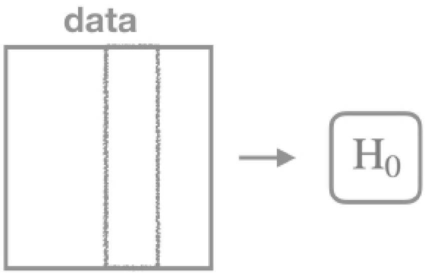
Hypothesis test



`specify()`

https://github.com/ismayc/talks/tree/master/data-day-texas-infer

Hypothesis test

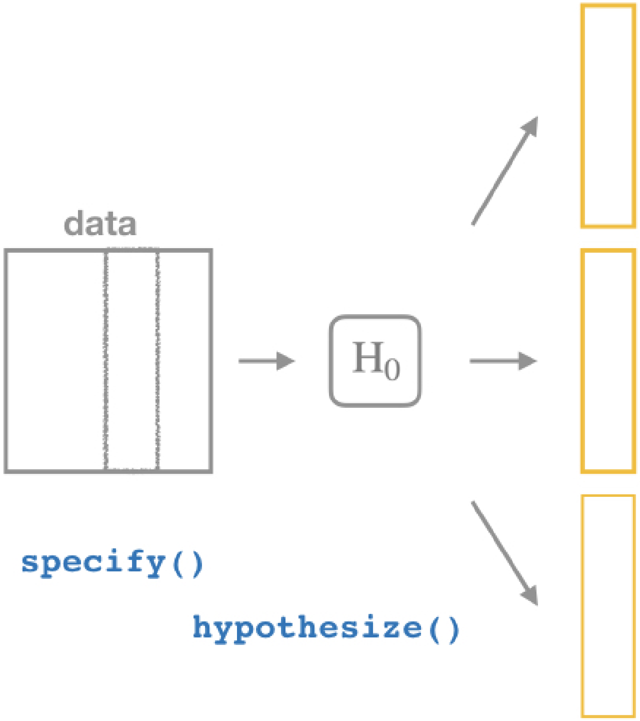


`specify()`

`hypothesize()`

https://github.com/ismayc/talks/tree/master/data-day-texas-infer

Hypothesis test

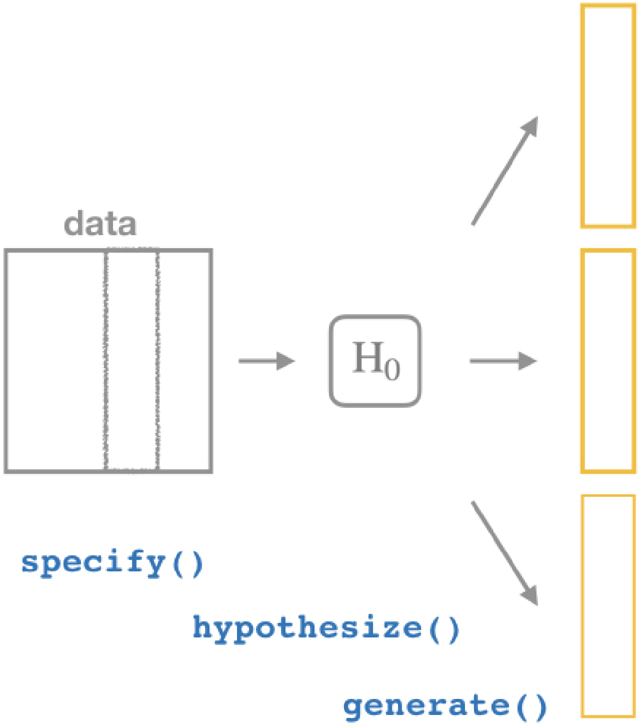


`specify()`

`hypothesize()`

https://github.com/ismayc/talks/tree/master/data-day-texas-infer

Hypothesis test



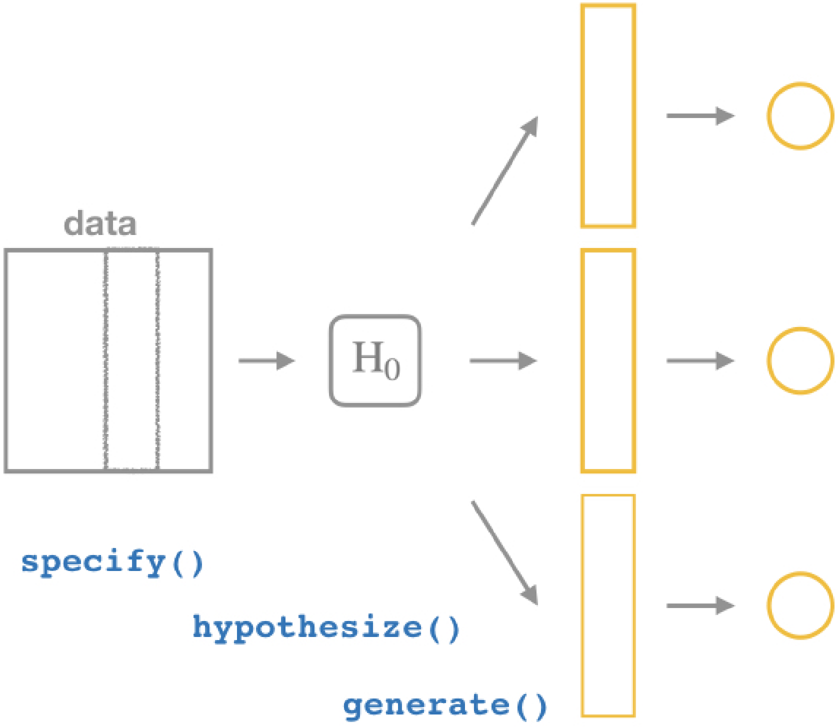
`specify()`

`hypothesize()`

`generate()`

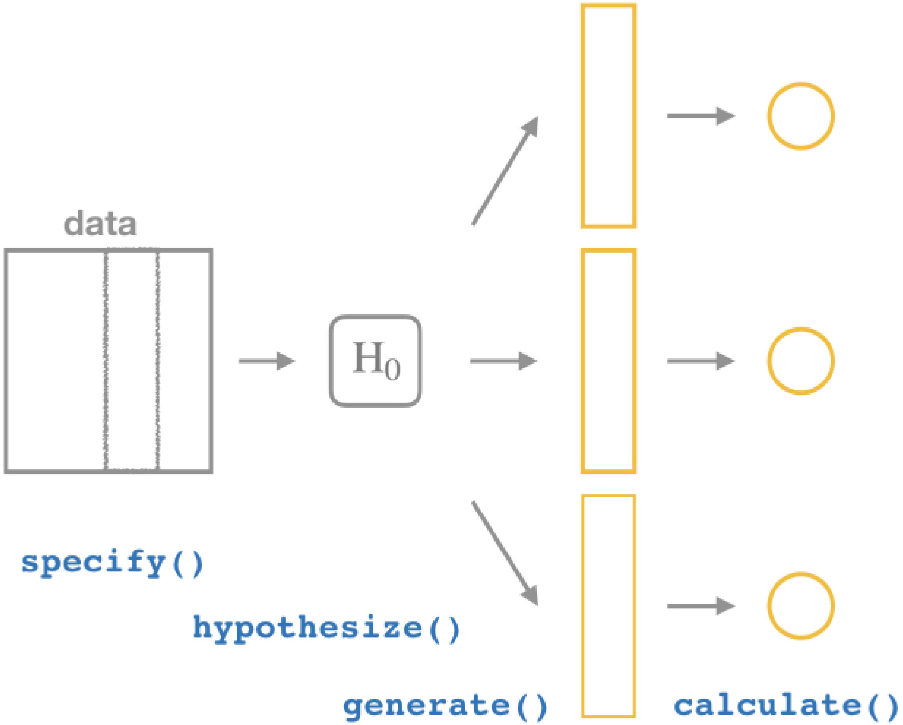
https://github.com/ismayc/talks/tree/master/data-day-texas-infer

Hypothesis test



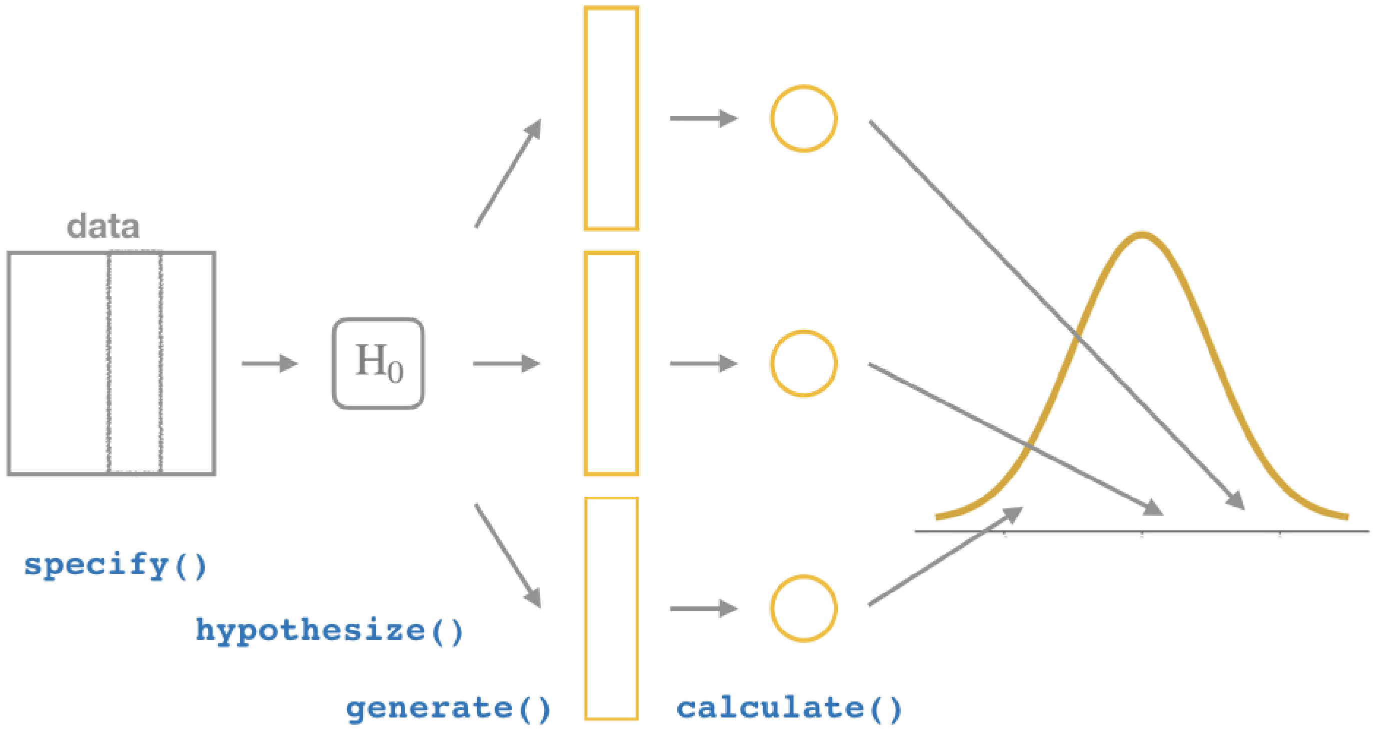
https://github.com/ismayc/talks/tree/master/data-day-texas-infer

Hypothesis test



https://github.com/ismayc/talks/tree/master/data-day-texas-infer

Hypothesis test



data

H_0

`specify()`

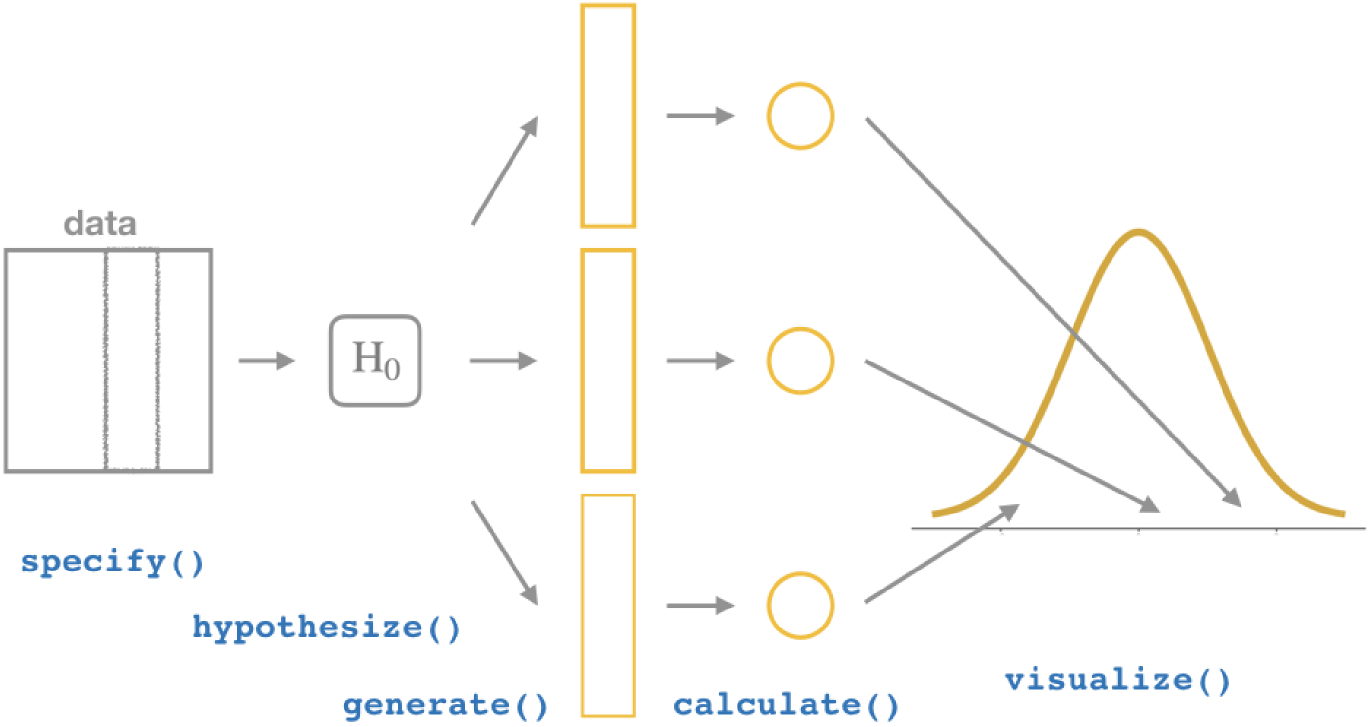
`hypothesize()`

`generate()`

`calculate()`

https://github.com/ismayc/talks/tree/master/data-day-texas-infer

Hypothesis test



`specify()`

`hypothesize()`

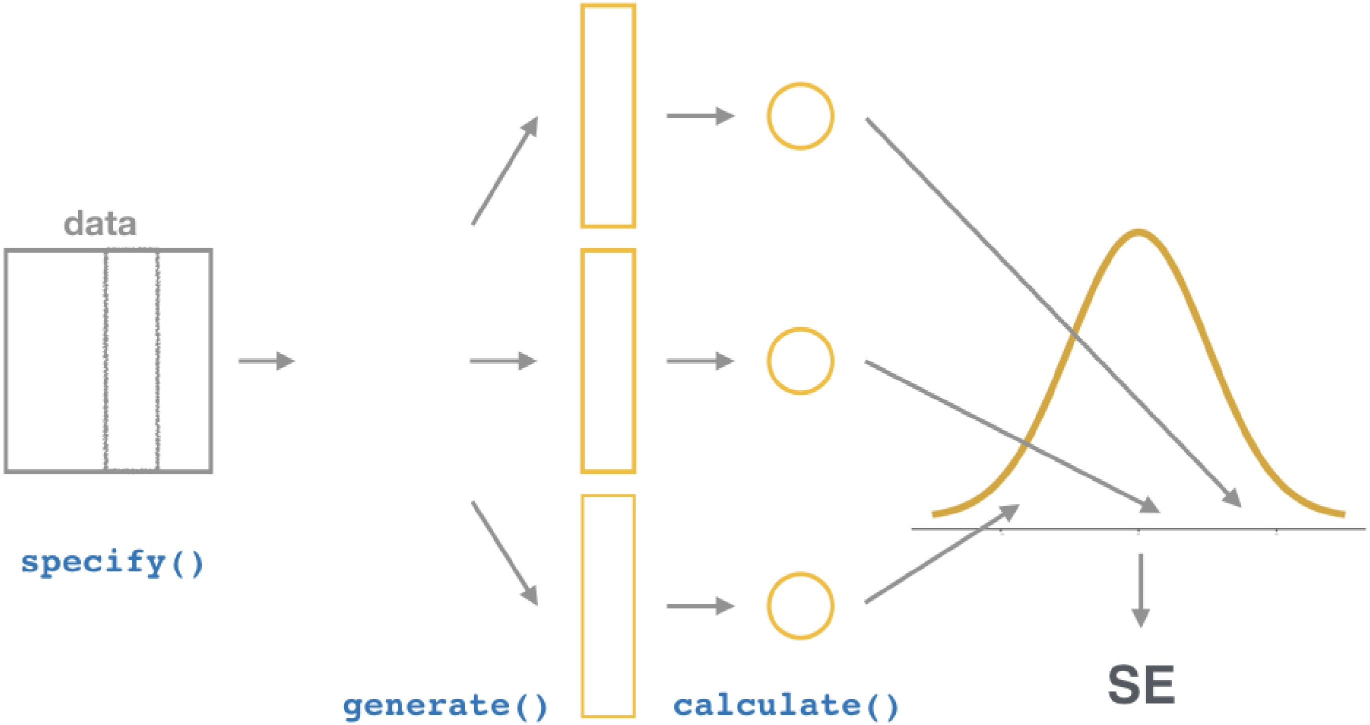
`generate()`

`calculate()`

`visualize()`

https://github.com/ismayc/talks/tree/master/data-day-texas-infer

Confidence Interval



https://github.com/ismayc/talks/tree/master/data-day-texas-infer

infer

5 functions:

- `specify()`
- `hypothesize()`
- `generate()`
- `calculate()`
- `visualize()`

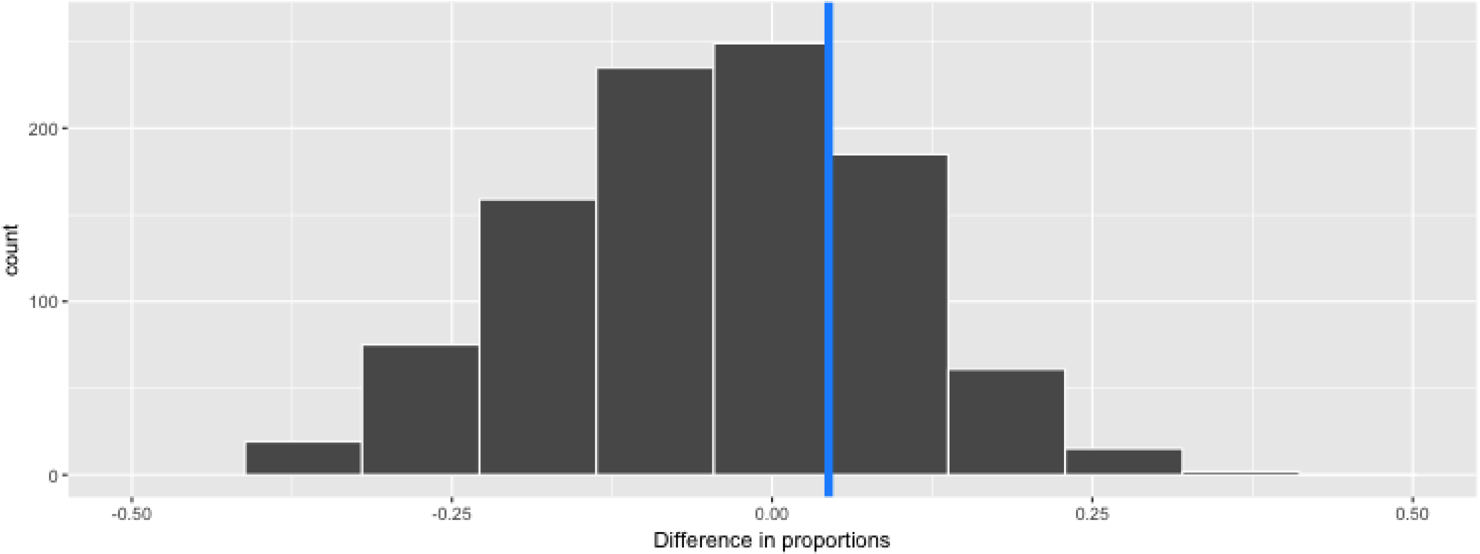
`infer`

- `specify` the response and explanatory variables (`y ~ x`)
- `hypothesize` what the null H_0 is (here, `independence` of `y` and `x`)
- `generate` new samples from parallel universes:
  - Resample from our original data **without replacement**, each time shuffling the `group` (`type = "permute"`)
  - Do this **a ton of times** (`reps = 1000`)
- `calculate` the *new* statistic (`stat = "diff in props"`) for each `rep`

```r
set.seed(8)
null_distn <- yawn_myth %>%
  specify(formula = yawn ~ group, success = "1") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in props", order = c("seed", "control"))
```

Visualize the null distribution

- `visualize` the distribution of the `stat` (here, `diff in props`)



```
visualize(null_distn, bins = 10) +
  geom_vline(xintercept = obs_diff,
             color = "dodgerblue",
             size = 2)
```

Classical inference

Rely on theory to tell us what the null distribution looks like.

```
yawn_myth %>%
  specify(yawn ~ group, success = "1") %>%
  hypothesize(null = "independence") %>%
  # generate() is not needed since we are not doing randomization
  # calculate(stat = "z") is implied based on variable types
  visualize(method = "theoretical") +
  geom_vline(xintercept = obs_stat, color = "orchid", size = 2)
```

Classical vs resampling

Changed the `stat` to calculate to `z` now (before we did `diff in props`).

```r
yawn_myth %>%
  specify(yawn ~ group, success = "1") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "z", order = c("seed", "control")) %>%
  visualize(method = "both", bins = 10) +
  geom_vline(xintercept = obs_stat, color = "orchid", size = 2)
```

Do the test in R

```r
yawn_table ← table(group, yawn)
yz ← prop.test(x = yawn_table, n = nrow(yawn_myth),
          alternative = "greater", correct = FALSE)
yz
```

```
	2-sample test for equality of proportions without continuity
	correction

data:  yawn_table
X-squared = 0.10504, df = 1, p-value = 0.3729
alternative hypothesis: greater
95 percent confidence interval:
 -0.1754872  1.0000000
sample estimates:
   prop 1    prop 2 
0.7500000 0.7058824 
```

Do the test in R

```
yawn_table ← table(group, yawn)
yz ← prop.test(x = yawn_table, n = nrow(yawn_myth),
          alternative = "greater", correct = FALSE)
yz
```

```

	2-sample test for equality of proportions without continuity
	correction

data:  yawn_table
X-squared = 0.10504, df = 1, p-value = 0.3729
alternative hypothesis: greater
95 percent confidence interval:
 -0.1754872  1.0000000
sample estimates:
   prop 1    prop 2 
0.7500000 0.7058824 
```

What does this test assume?

Pull out the z statistic

The observed `z` value is 0.105042.

```r
obs_stat <- yz %>%
  broom::tidy(yz) %>%
  pull(statistic)
```

Now you can use `obs_stat` in `geom_vline(xintercept = obs_stat)` when added to `visualize`!

In class exercise

- Read in the `mazes` data (http://bit.ly/mazes-gist)

```
library(readr)
mazes <- read_csv("http://bit.ly/mazes-gist") %>%
  clean_names() %>%  #janitor package
  filter(dx %in% c("ASD", "TD"))
```

- Use `dplyr::filter` to include only two groups (`DX` if you didn't `clean_names()`; `dx` if you did!).

- Use `infer` to compare a numerical variable between the two groups using:

    - A permutation test and
    - A classical theoretical test.

See: https://infer-dev.netlify.com

About the data: Quantitative analysis of disfluency in children with autism spectrum disorder or language impairment

Classical t-test in R

- Independent samples (`paired = FALSE`)
- Assume equal variances (`var.equal = TRUE`)
- `alternative` is relative to the groups *alphabetically*: so here $H_1 = \mu_{asd} < \mu_{td}$

```
myt ← t.test(viq ~ dx,
              data = mazes,
              var.equal = TRUE,
              paired = FALSE,
              alternative = "less")
myt
```

```

    Two Sample t-test

data:  viq by dx
t = -11.842, df = 308, p-value < 2.2e-16
alternative hypothesis: true difference in means is less than 0
95 percent confidence interval:
      -Inf -18.63839
sample estimates:
mean in group ASD  mean in group TD 
         95.28962         116.94488 
```

Save the t statistic

```
obs_t <- myt %>%
  tidy() %>% # from broom
  pull(statistic) # from dplyr
obs_t
```

```
[1] -11.84247
```

Now you can use `obs_t` in `geom_vline(xintercept = obs_t)` when added to `visualize`!

Thanks!

Slides created via the R package **xaringan**.