infer

an R package for tidy statistical inference

Andrew Bray

Chester Ismay

infer makes p-values easier to compute.

statistical inference

infer makes ~~p-values easier to compute.~~ tidy and transparent.

```
chisq.test(gss$party, gss$space)
```

```
gss %>%
  specify(space ~ party) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "Chisq")
```

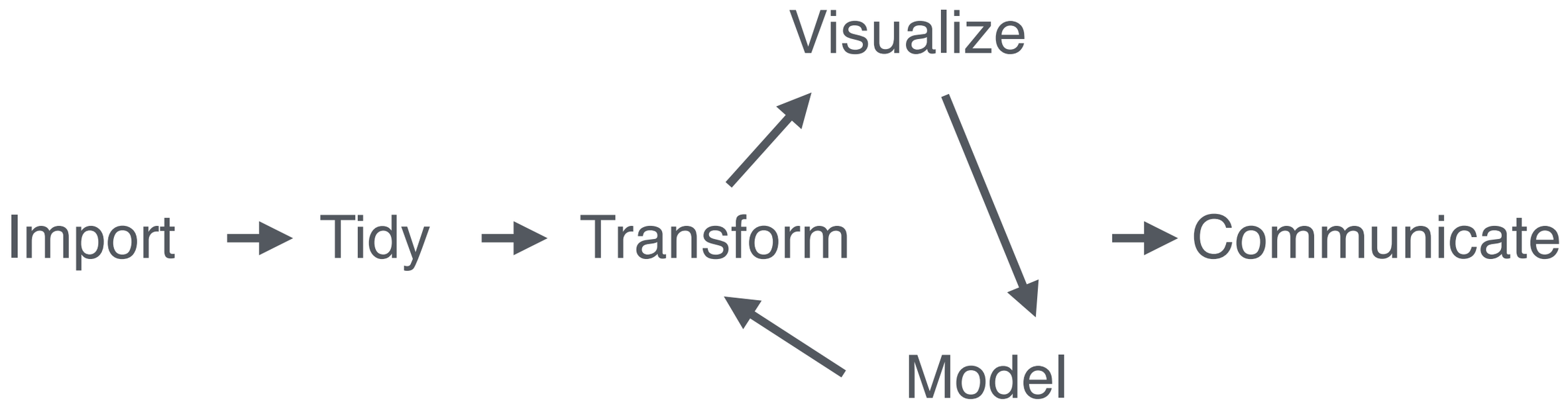
statistical inference

infer makes ~~p-values easier to compute.~~

tidy and transparent.

The Data Science Cycle

Wickham and Grolemund, R for Data Science



The Data Science Cycle

Wickham and Grolemund, R for Data Science

Case Study: Is funding for space exploration a partisan issue?

Import → Tidy → Transform → Visualize

```
library(tidyverse)
gss <- read_csv("gss.csv")
gss
```

```
# A tibble: 23,224 x 27
      id  year age   partyid natspac class degree sex
   <dbl> <int> <chr> <chr>   <chr>   <chr> <chr>  <chr>
 1     1  1982 41    STRONG… ABOUT … WORK… LT HI… MALE
 2     2  1982 49    STRONG… TOO MU… WORK… HIGH … FEMA…
 3     3  1982 27    IND,NE… TOO LI… MIDD… HIGH … FEMA…
 4     4  1982 24    IND,NE… TOO LI… MIDD… HIGH … FEMA…
 5     5  1982 57    STRONG… ABOUT … LOWE… LT HI… MALE
 6     6  1982 29    INDEPE… TOO MU… MIDD… BACHE… MALE
 7     7  1982 21    IND,NE… TOO MU… MIDD… JUNIO… FEMA…
 8     8  1982 68    IND,NE… ABOUT … MIDD… BACHE… MALE
 9     9  1982 54    IND,NE… TOO MU… MIDD… GRADU… FEMA…
10    10  1982 80    STRONG… ABOUT … MIDD… JUNIO… MALE
# ... with 23,214 more rows, and 19 more variables:
```

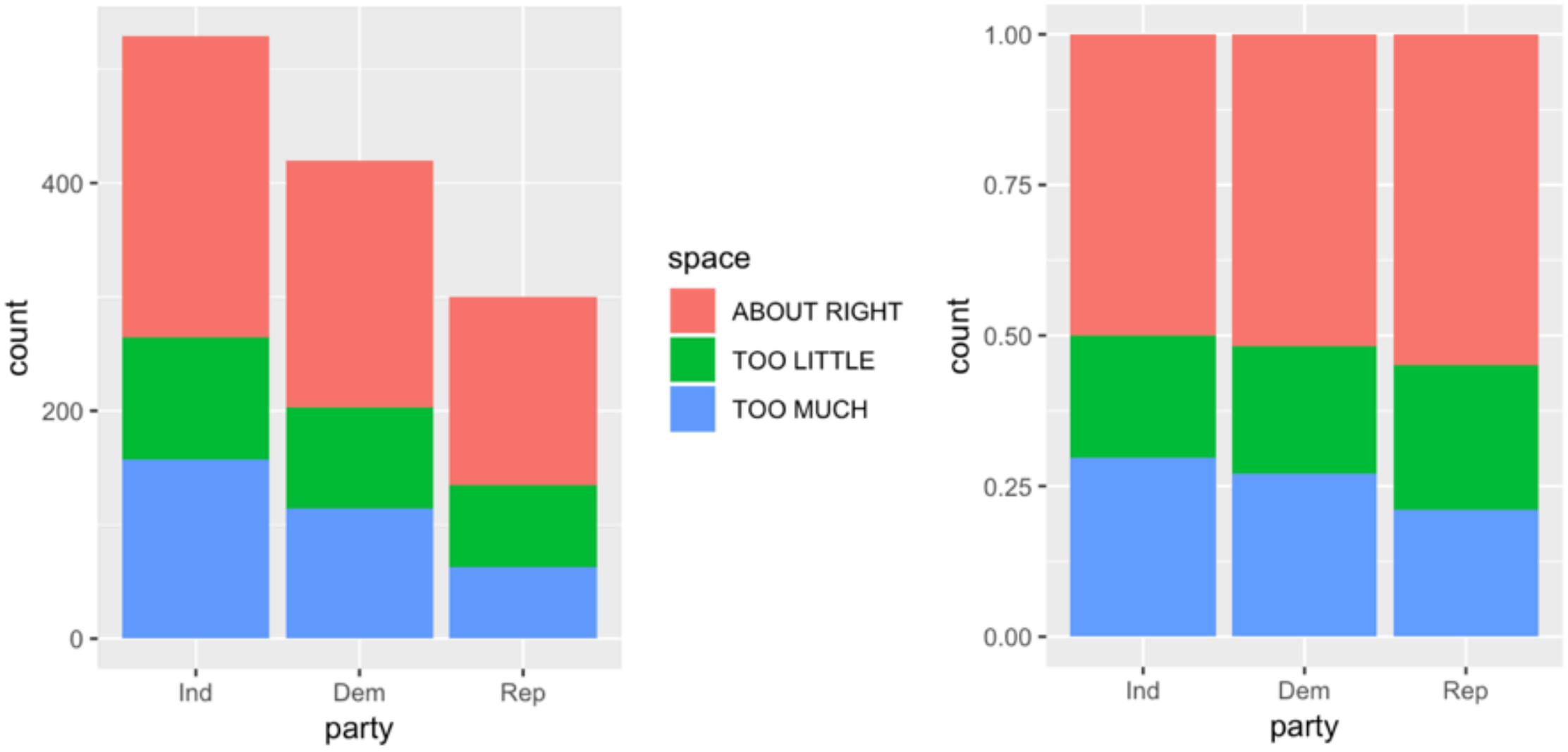
Import → Tidy → Transform → Visualize

```
gss %>%
  filter(year == 2016) %>%
  select(partyid, natspac) %>%
  mutate(space = as.factor(natspac),
         party = as.factor(partyid),
         party = fct_collapse(party,
                              Ind = c("IND,NEAR DEM", "IND,NEAR REP", "INDEPENDENT"),
                              Dem = c("NOT STR DEMOCRAT", "STRONG DEMOCRAT"),
                              Rep = c("NOT STR REPUBLICAN", "STRONG REPUBLICAN")))
```

```
 1 INDEPENDENT         TOO LITTLE  TOO LITTLE  INDEPENDENT
# A tibble: 1,249 x 4                                 REPUBLICAN
   partyid             natspac     space       party  REPUBLICAN
   <chr>               <chr>       <fct>       <fct>  DEMOCRAT
 1 INDEPENDENT         TOO LITTLE  TOO LITTLE  Ind    REPUBLICAN
 2 NOT STR REPUBLICAN  TOO LITTLE  TOO LITTLE  Rep     DEM
 3 NOT STR REPUBLICAN  ABOUT RIGHT ABOUT RIGHT Rep    EMOCRAT
 4 NOT STR DEMOCRAT    ABOUT RIGHT ABOUT RIGHT Dem    EPUBLICAN
 5 NOT STR REPUBLICAN  ABOUT RIGHT ABOUT RIGHT Rep    ENT
 6 IND,NEAR DEM        ABOUT RIGHT ABOUT RIGHT Ind    ENT
 7 STRONG DEMOCRAT     ABOUT RIGHT ABOUT RIGHT Dem
 8 STRONG REPUBLICAN   TOO LITTLE  TOO LITTLE  Rep
 9 INDEPENDENT         TOO MUCH    TOO MUCH    Ind
10 INDEPENDENT         ABOUT RIGHT ABOUT RIGHT Ind
# ... with 1,239 more rows
```

Import → Tidy → Transform → Visualize

```
%>%
  ggplot(aes(x = party, fill = space)) +
  geom_bar(position = "fill")
```



Tidyverse Trademarks

- Compose with pipes
- Write for humans
- Use dataframes
- Express general operations
- Reusable parts

```
read_csv("gss.csv") %>%
  filter(year == 2016) %>%
  select(partyid, natspac) %>%
mutate(space = as.factor(natspac),
       party = as.factor(partyid),
       party = fct_collapse(party,
                            Ind = c(...),
                            Dem = c(...),
                            Rep = c(...))) %>%
ggplot(aes(x = party, fill = space)) +
geom_bar(position = "fill")
```

Optimistic effort I

```
chisq.test(data = gss, x = party, y = space)
```

```
Error in chisq.test(data = gss, x = party, y = space) :
  unused argument (data = gss)
```

. . . optimistic effort II

```
chisq.test(space ~ party, data = gss)
```

```
Error in chisq.test(space ~ party, data = gss) :
  unused argument (data = gss)
```

…after looking at the help file

```
chisq.test(gss$party, gss$space)
```

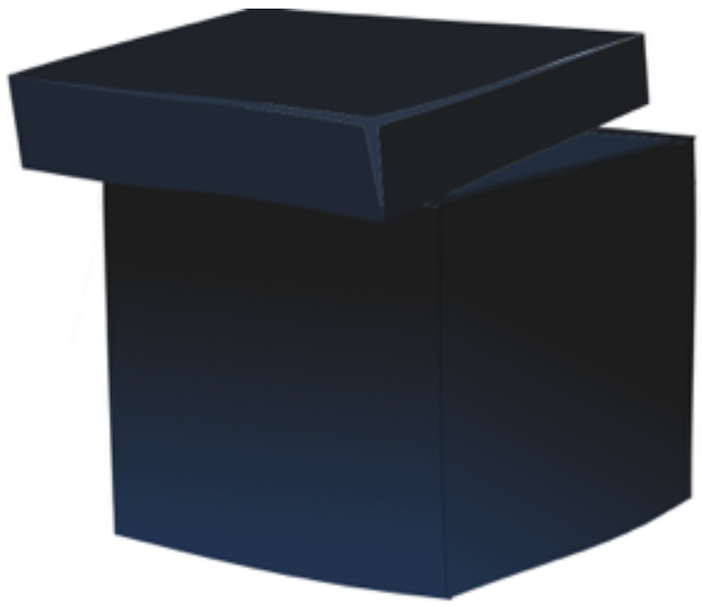
```
        Pearson's Chi-squared test

data:  gss$party and gss$space
X-squared = 7.5682, df = 4, p-value = 0.1087
```

statistical inference

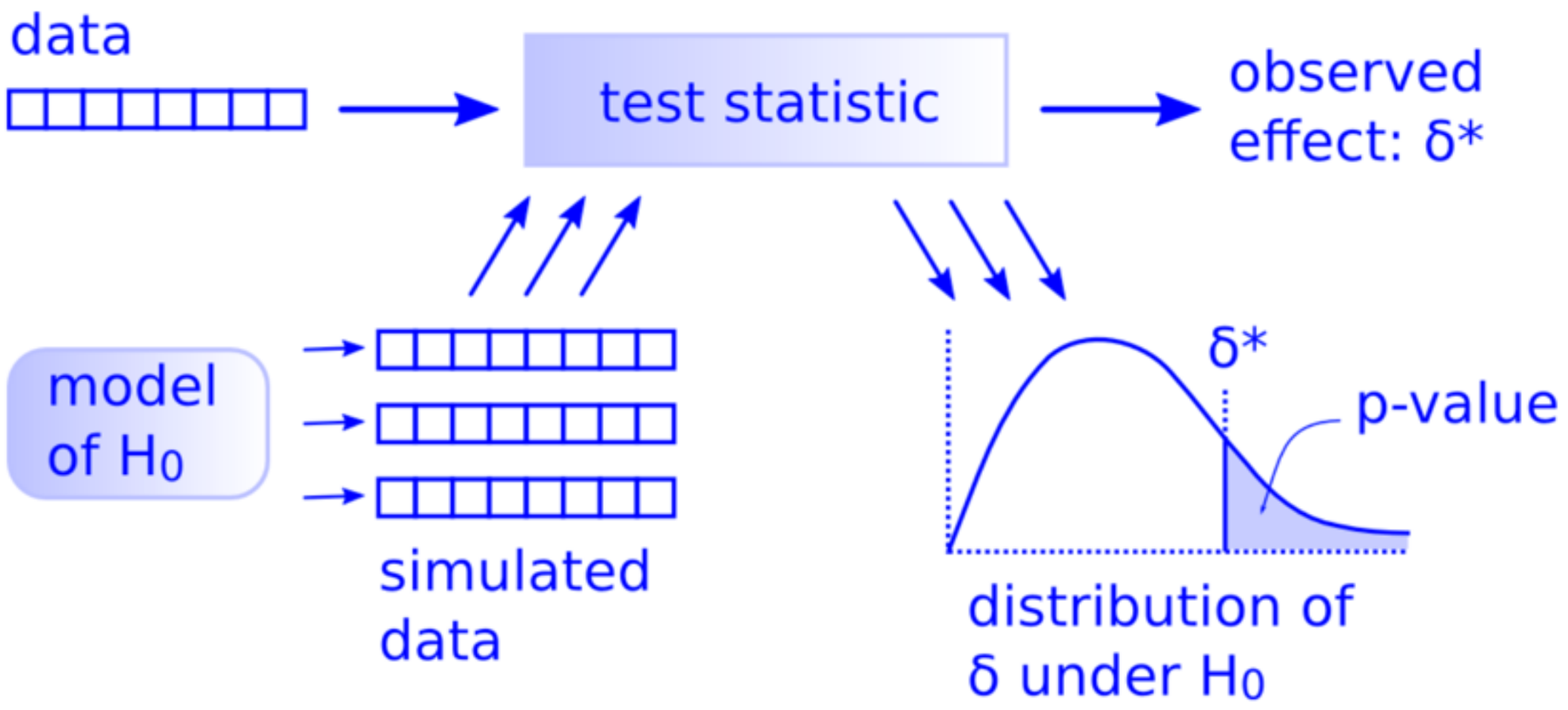
infer makes ~~p-values easier to compute~~.

tidy and transparent.

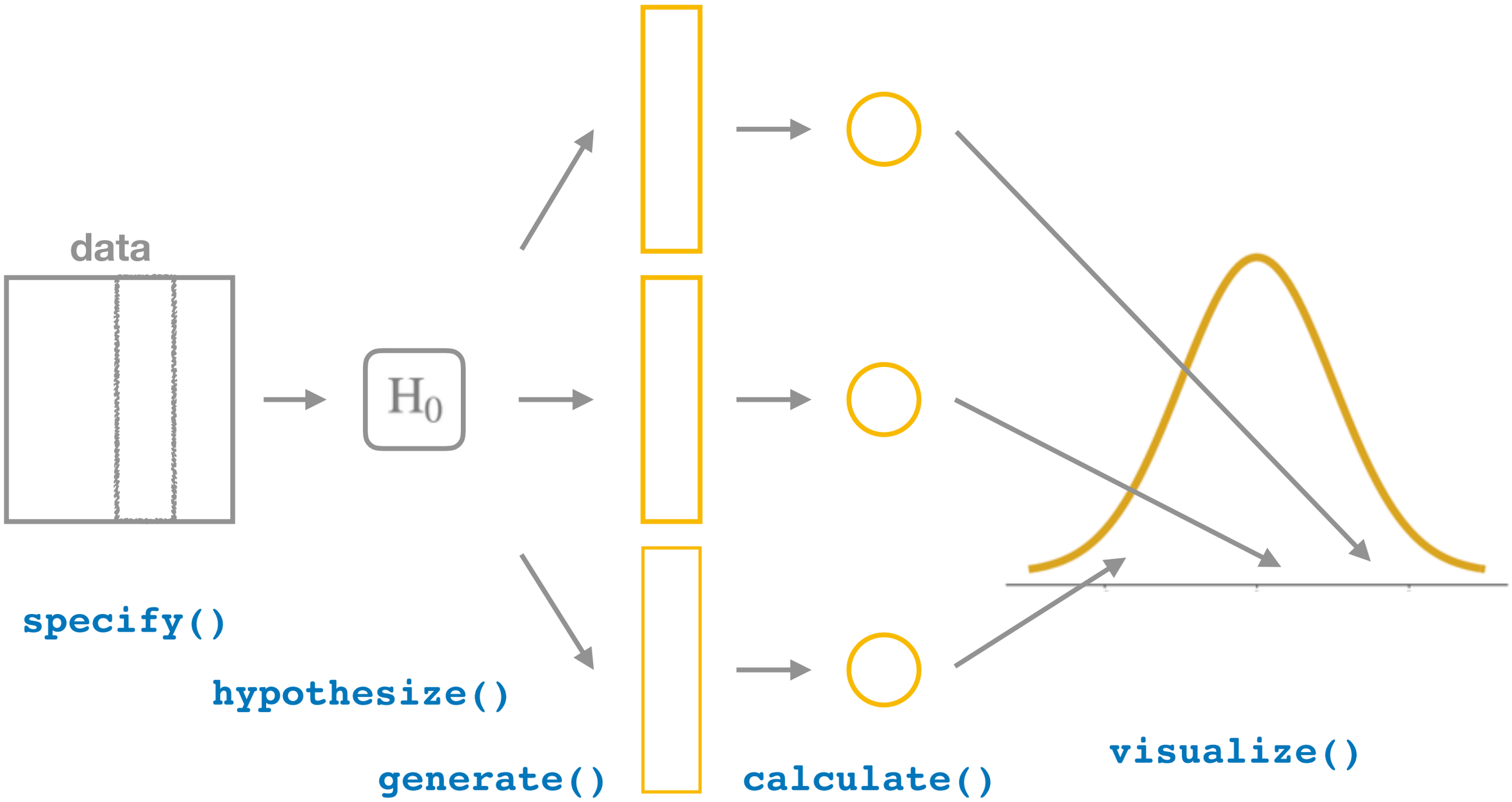


There is only one test

- Allen Downey



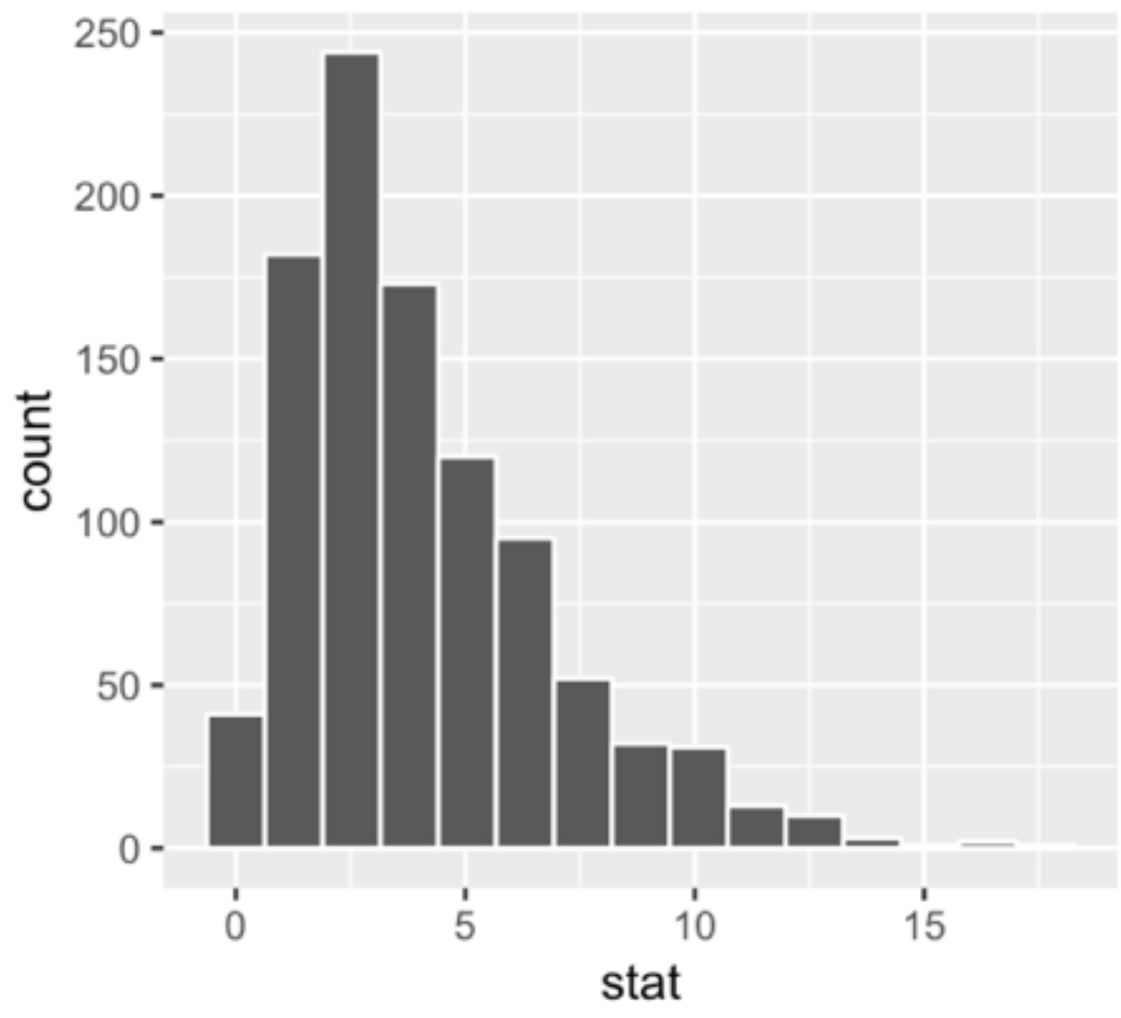
The `infer` verbs



Modeling with `infer`

```
library(infer)
gss %>%
  specify(space ~ party) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "Chisq") %>%
  visualize()
```

```
# A tibble: 1,000 x 2
   replicate  stat
       <int> <dbl>
 1         1 0.563
 2         2 6.31
 3         3 9.81
 4         4 4.29
 5         5 1.46
 6         6 5.77
 7         7 1.53
 8         8 1.08
 9         9 2.36
10        10 5.34
# ... with 990 more rows
```

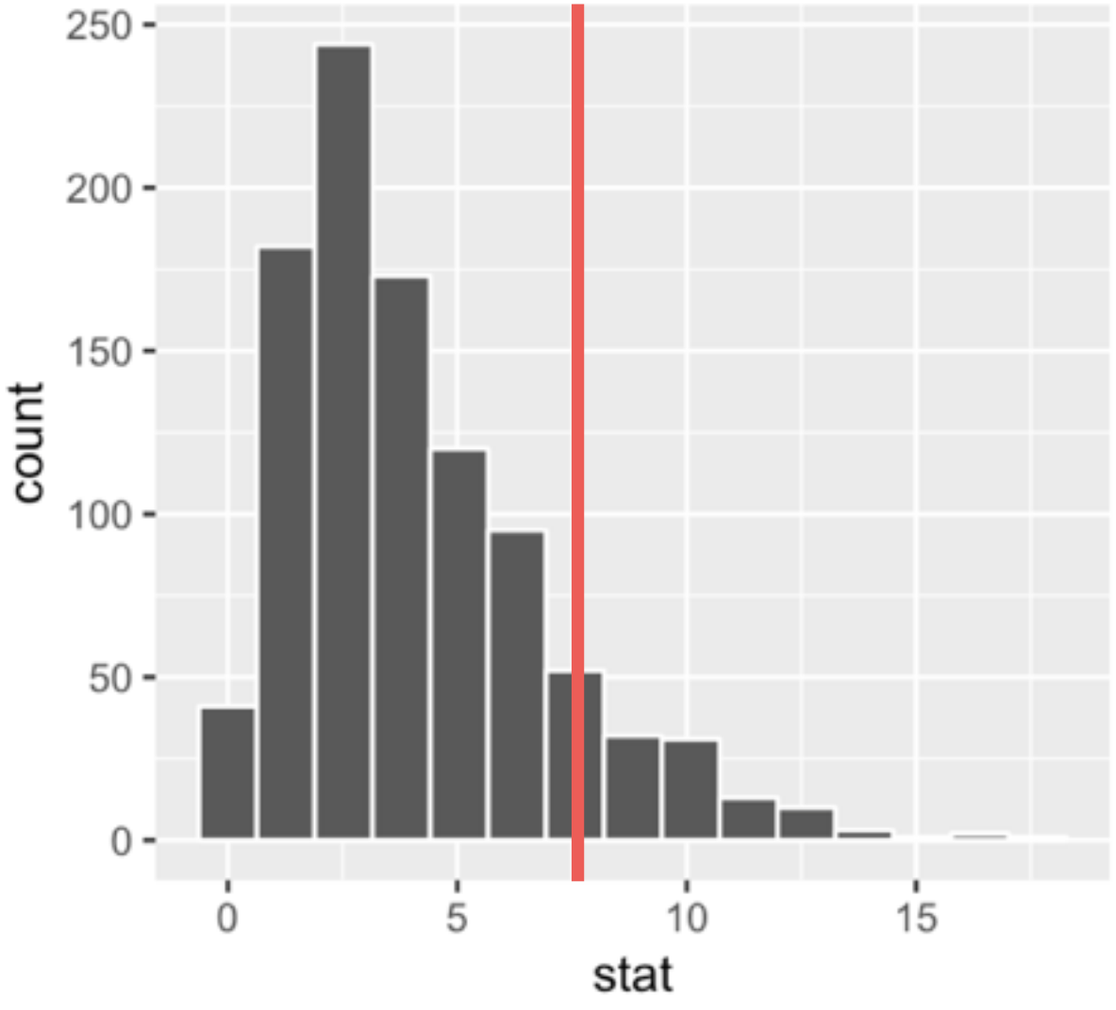


Modeling with `infer`

```
library(infer)
gss %>%
  specify(space ~ party) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "Chisq") %>%
  summarize(p_val = mean(stat > chisq_obs))
```

```
# A tibble: 1 x 1
  p_val
  <dbl>
1 0.117
```

```
chisq_obs <- gss %>%
  specify(space ~ party) %>%
  calculate(stat = "Chisq")
```



Reusable Parts

```
gss %>%
  specify(space ~ party) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "Chisq")
```

Permutation Chi-squared

```
gss %>%
  specify(space ~ party) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "Chisq")
```

Approximation Chi-squared

```
gss %>%
  specify(space ~ party) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "Chisq")
```

*fiddle

Permutation p1 - p2

```
gss %>%
  specify(happy ~ party, success = "HAPPY") %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in props")
```

"diff in props"

Interval for p1 - p2

"bootstrap"

statistical inference

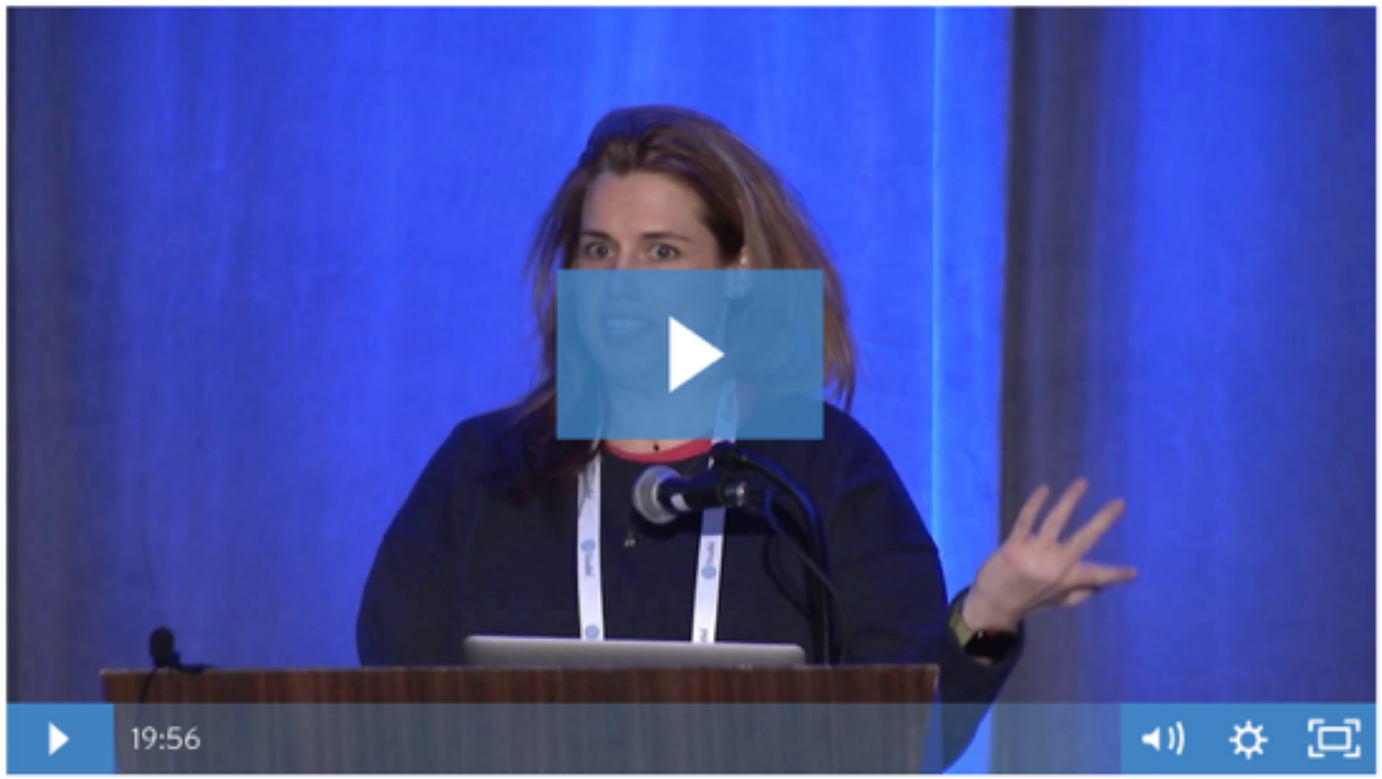
infer makes ~~p-values easier to compute.~~ tidy and transparent.

```
chisq.test(gss$party, gss$space)
```


```
gss %>%
  specify(sp
  hypothesiz
  generate(r
  calculate(
```

- Thanks to Chester Ismay, Ben Baumer, Mine Cetinkaya-Rundel, Jo Hardin, and the other contributors.
- infer.netlify.com

Watch this



https://bit.ly/2NXN9f7