Contingency tables

INFERENCE FOR CATEGORICAL DATA IN R

Andrew Bray

Assistant Professor of Statistics at Reed College

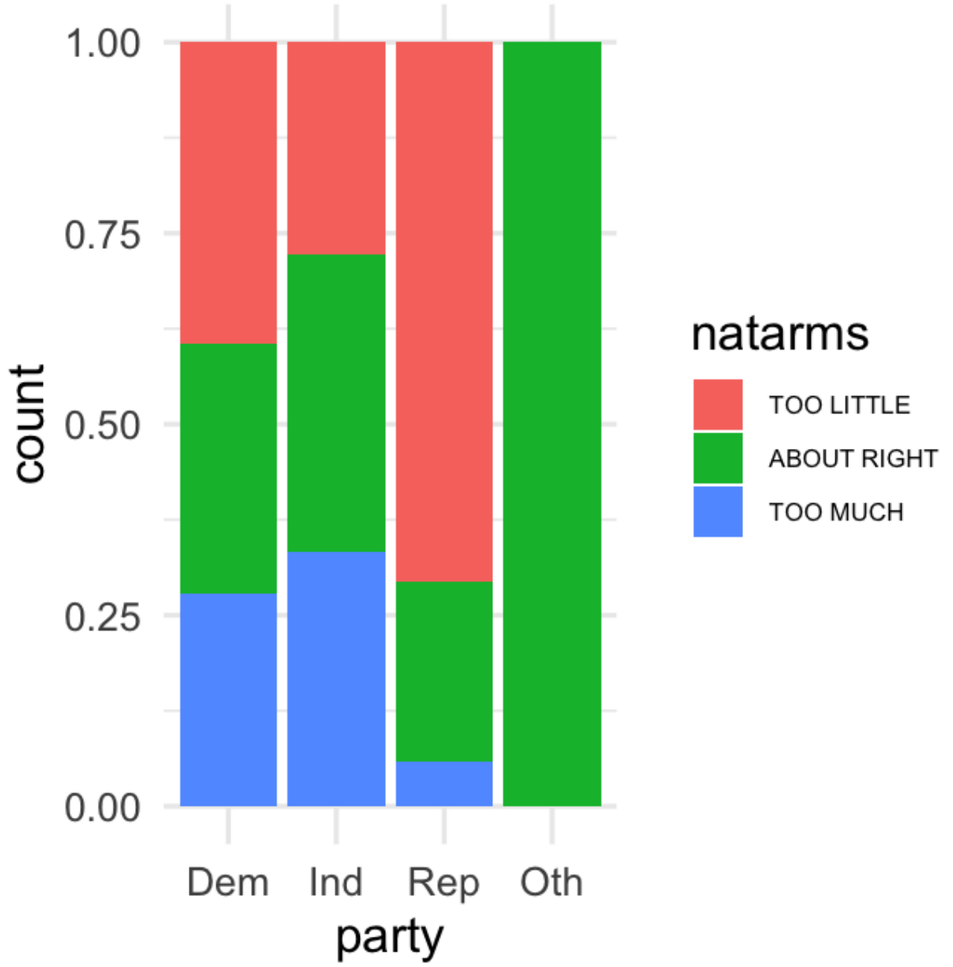
Politics and military spending

```
gss2016 %>%
  select(party, natarms) %>%
  glimpse()
```

```
Observations: 150
Variables: 2
$ party   <fct> Ind, Ind, Dem, Ind, Ind, Ind, Ind, Dem, Dem, Ind,...
$ natarms <fct> TOO LITTLE, TOO MUCH, TOO MUCH,...
```

Politics and military spending

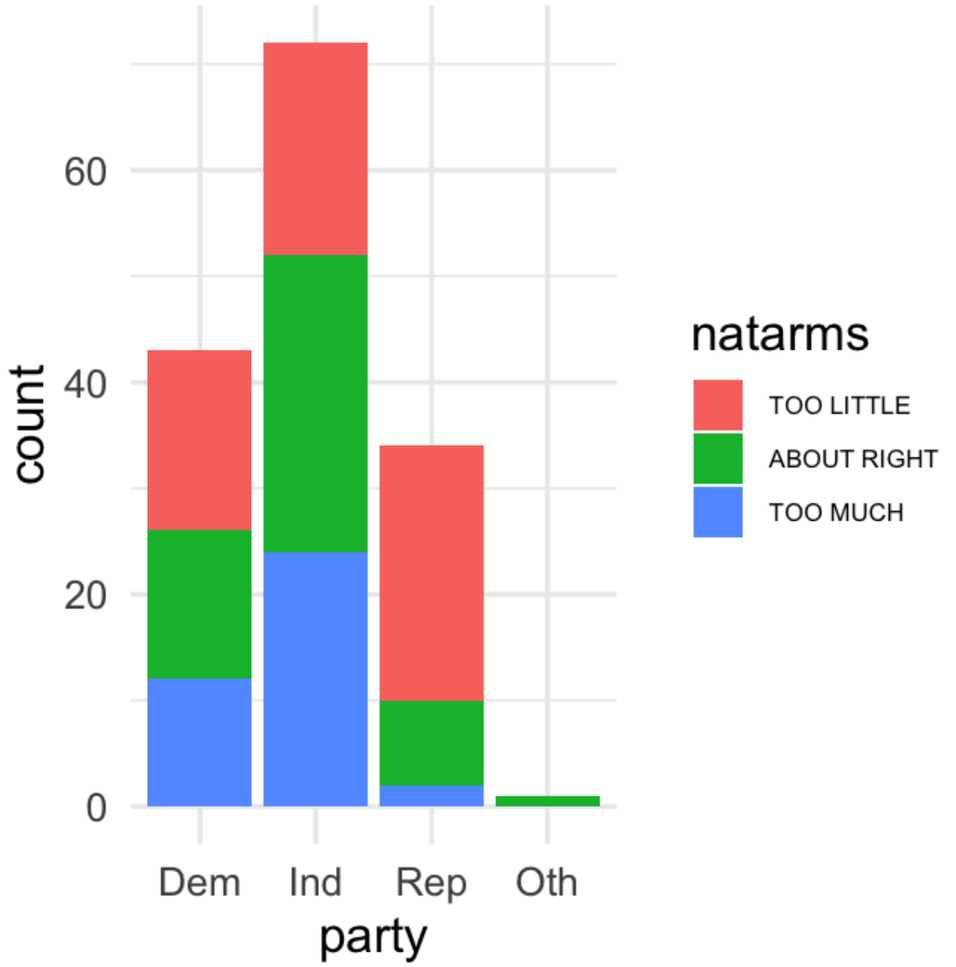
```
ggplot(gss2016, aes(x = party, fill = natarms)) +
  geom_bar(position = "fill")
```



Politics and military spending

```
ggplot(gss2016, aes(x = party, fill = natarms)) +
  geom_bar()
```

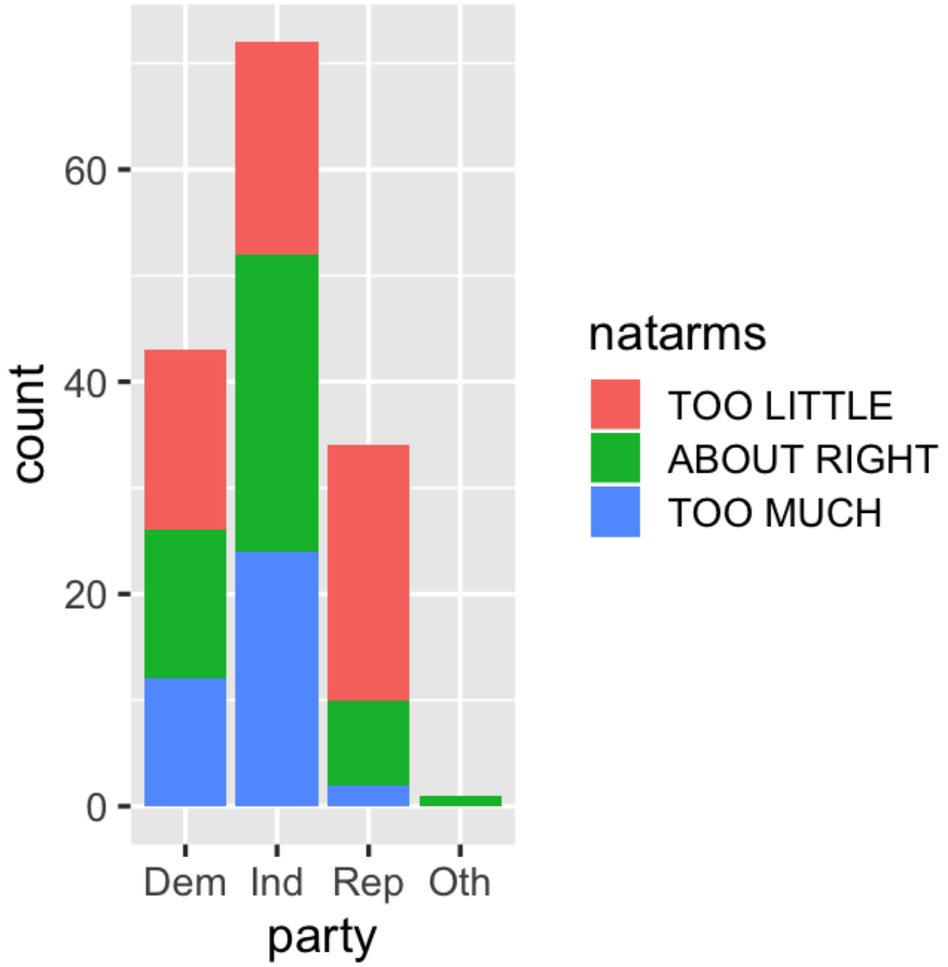
```
library(broom)
```



Tables and tidy data

```
tab <- gss2016 %>%
  select(natarms, party) %>%
  table()
tab
```

```
             party
natarms          Dem  Ind  Rep  Oth
  TOO LITTLE      17   20   24    0
  ABOUT RIGHT     14   28    8    1
  TOO MUCH        12   24    2    0
```



Tables and tidy data

```
tab <- gss2016 %>%
  select(natarms, party) %>%
  table()
tab
```

```
             party
natarms        Dem  Ind  Rep  Oth
  TOO LITTLE    17   20   24    0
  ABOUT RIGHT   14   28    8    1
  TOO MUCH      12   24    2    0
```

```
tab %>%
  tidy()
```

```
# A tibble: 12 x 3
   natarms     party     n
   <chr>       <chr> <int>
 1 TOO LITTLE  Dem      17
 2 ABOUT RIGHT Dem      14
 3 TOO MUCH    Dem      12
 4 TOO LITTLE  Ind      20
 5 ABOUT RIGHT Ind      28
 6 TOO MUCH    Ind      24
 7 TOO LITTLE  Rep      24
 8 ABOUT RIGHT Rep       8
 9 TOO MUCH    Rep       2
10 TOO LITTLE  Oth       0
```

Tables and tidy data

```
tab <- gss2016 %>%
  select(natarms, party) %>%
  table()
tab
```

```
                party
natarms          Dem   Ind   Rep   Oth
  TOO LITTLE      17    20    24     0
  ABOUT RIGHT     14    28     8     1
  TOO MUCH        12    24     2     0
```

```
tab %>%
  tidy() %>%
  uncount(n)
```

```
# A tibble: 150 x 2
   natarms    party
   <chr>      <chr>
 1 TOO LITTLE Dem
 2 TOO LITTLE Dem
 3 TOO LITTLE Dem
 4 TOO LITTLE Dem
 5 TOO LITTLE Dem
 6 TOO LITTLE Dem
```

Let's practice!

INFERENCE FOR CATEGORICAL DATA IN R

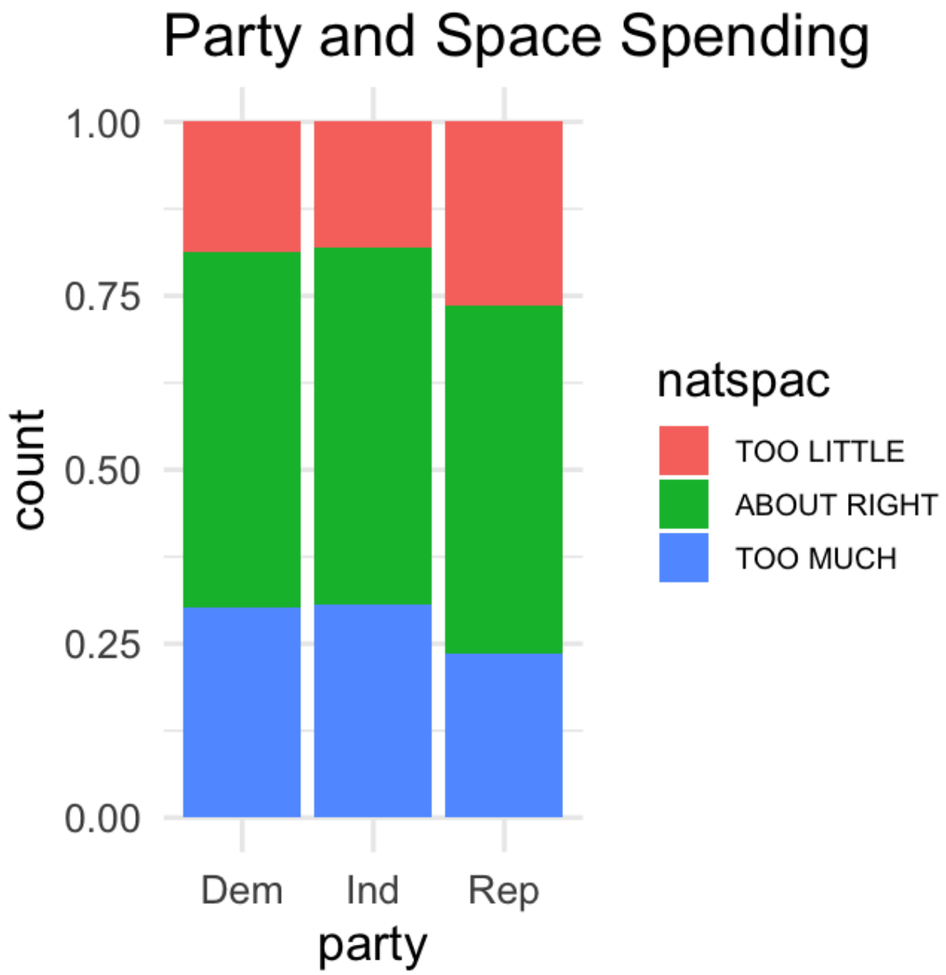
Chi-squared test statistic

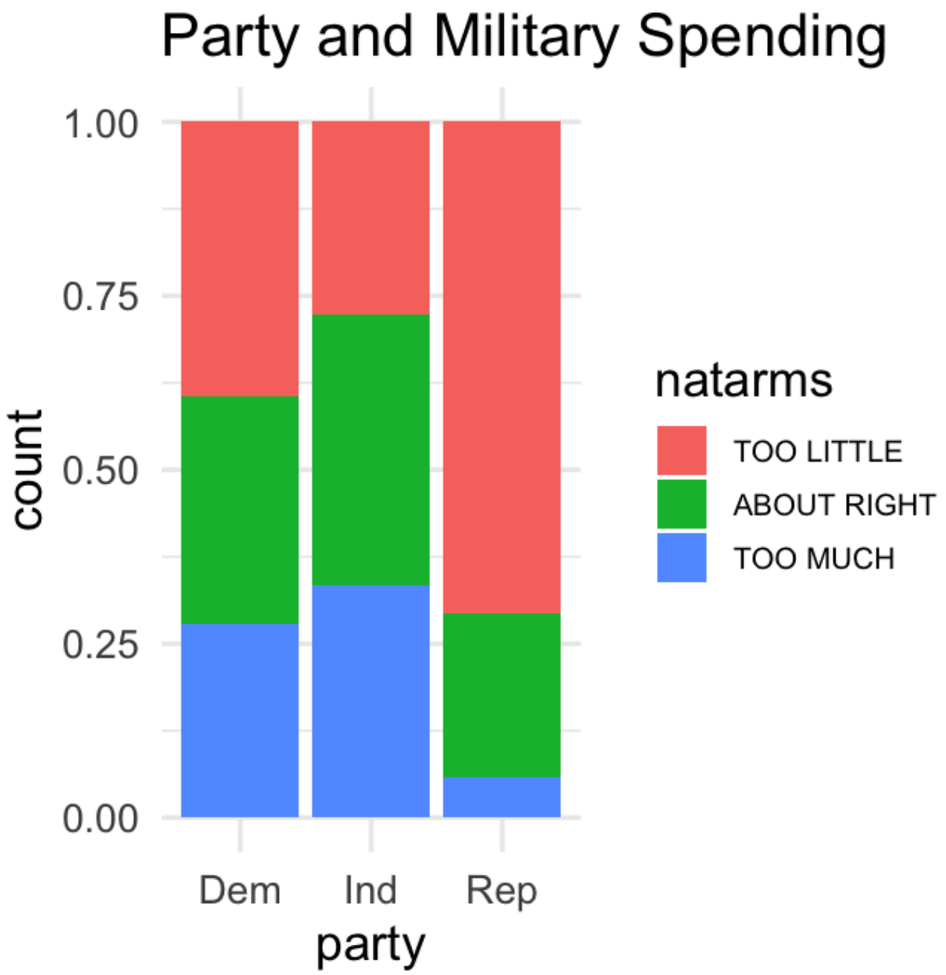
INFERENCE FOR CATEGORICAL DATA IN R

Andrew Bray

Assistant Professor of Statistics at Reed College

Comparing bar plots

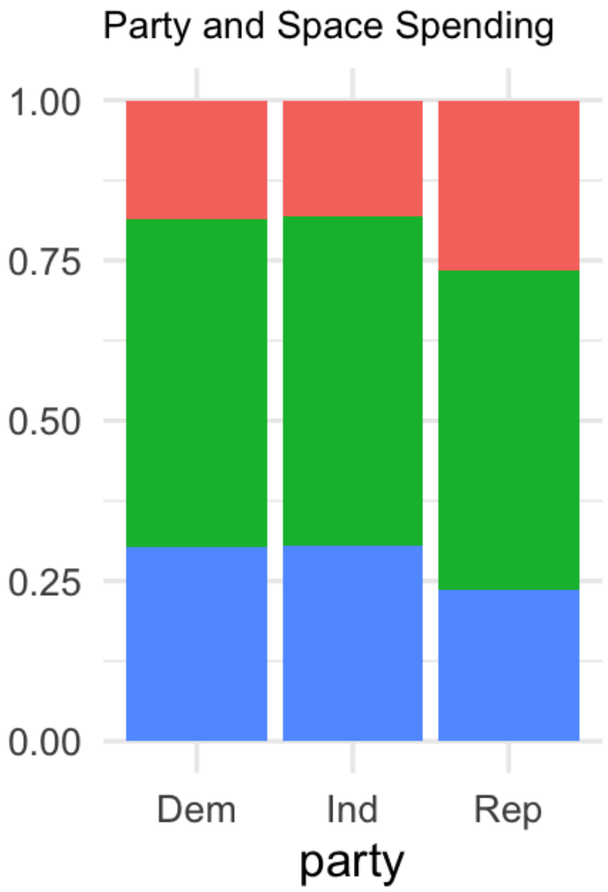


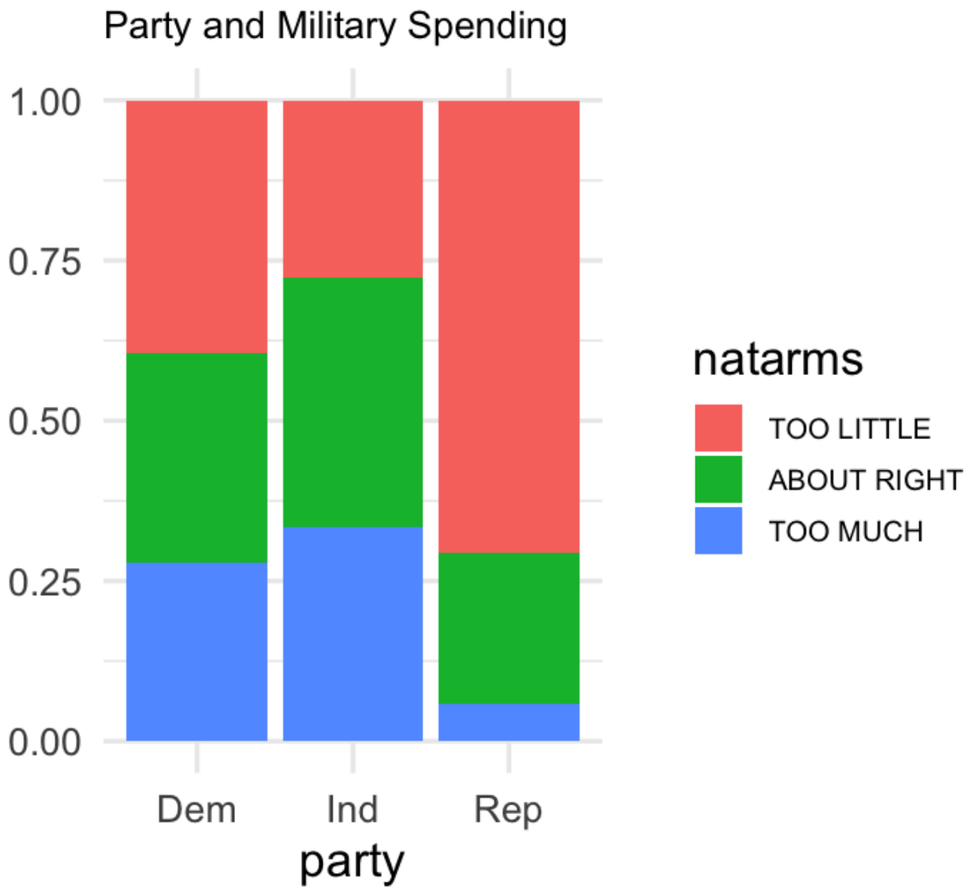


Hypothesis test

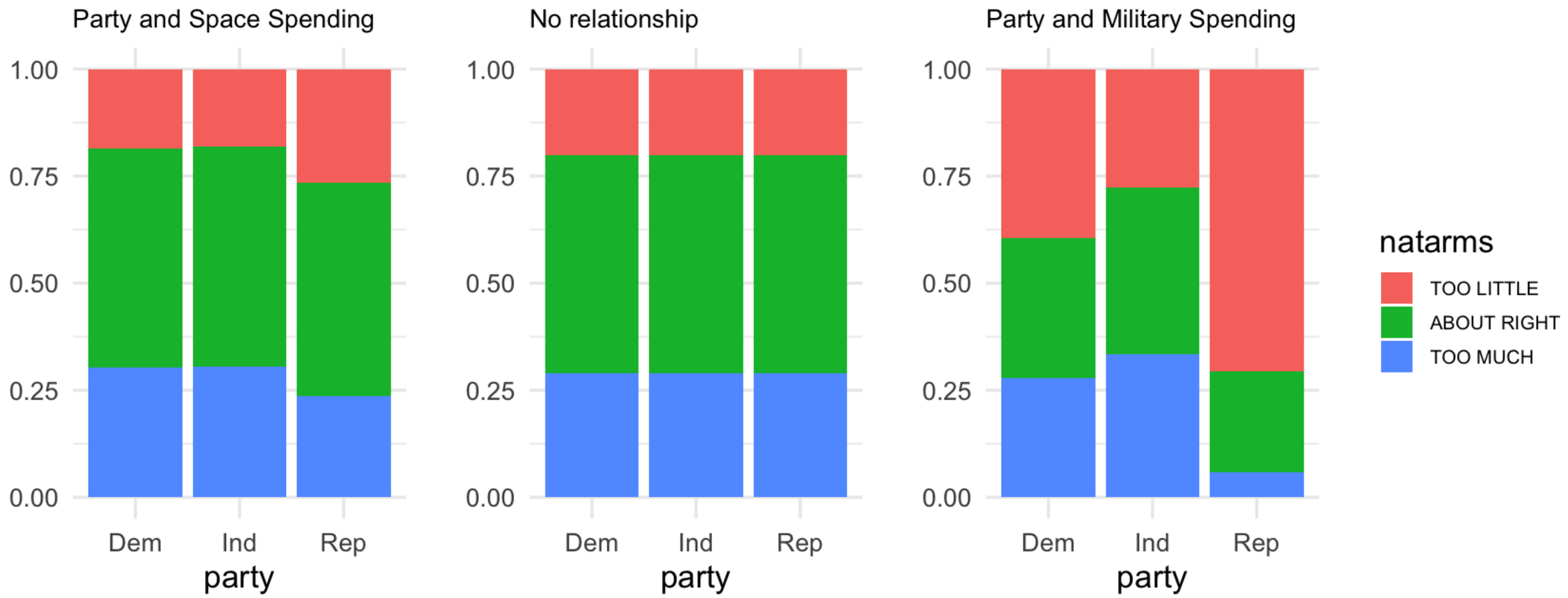
```
null <- data %>%
  specify(var1 ~ var2) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 100, type = "permute") %>%
  calculate(stat = ?)
```

Choosing a statistic





Choosing a statistic



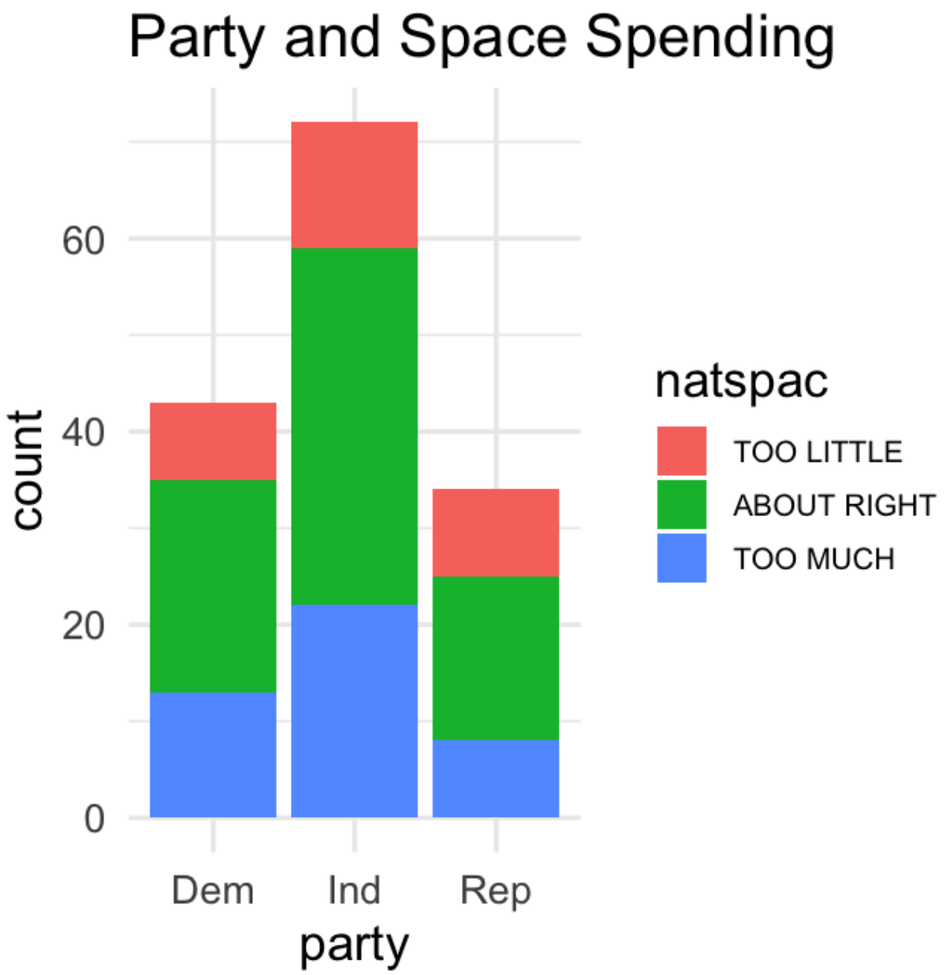
Choosing a statistic

```
observed_counts
```

```
             party
natspac        Dem Ind Rep
  TOO LITTLE     8  13   9
  ABOUT RIGHT   22  37  17
  TOO MUCH      13  22   8
```

```
expected_counts
```

```
             party
natspac         Dem  Ind  Rep
  TOO LITTLE    8.7 14.5  6.8
  ABOUT RIGHT 21.9 36.7 17.3
  TOO MUCH    12.4 20.8  9.8
```



Choosing a statistic

```
observed_counts
```

```
            party
natspac       Dem Ind Rep
  TOO LITTLE    8  13   9
  ABOUT RIGHT  22  37  17
  TOO MUCH     13  22   8
```

```
expected_counts
```

```
            party
natspac        Dem  Ind  Rep
  TOO LITTLE   8.7 14.5  6.8
  ABOUT RIGHT 21.9 36.7 17.3
  TOO MUCH    12.4 20.8  9.8
```

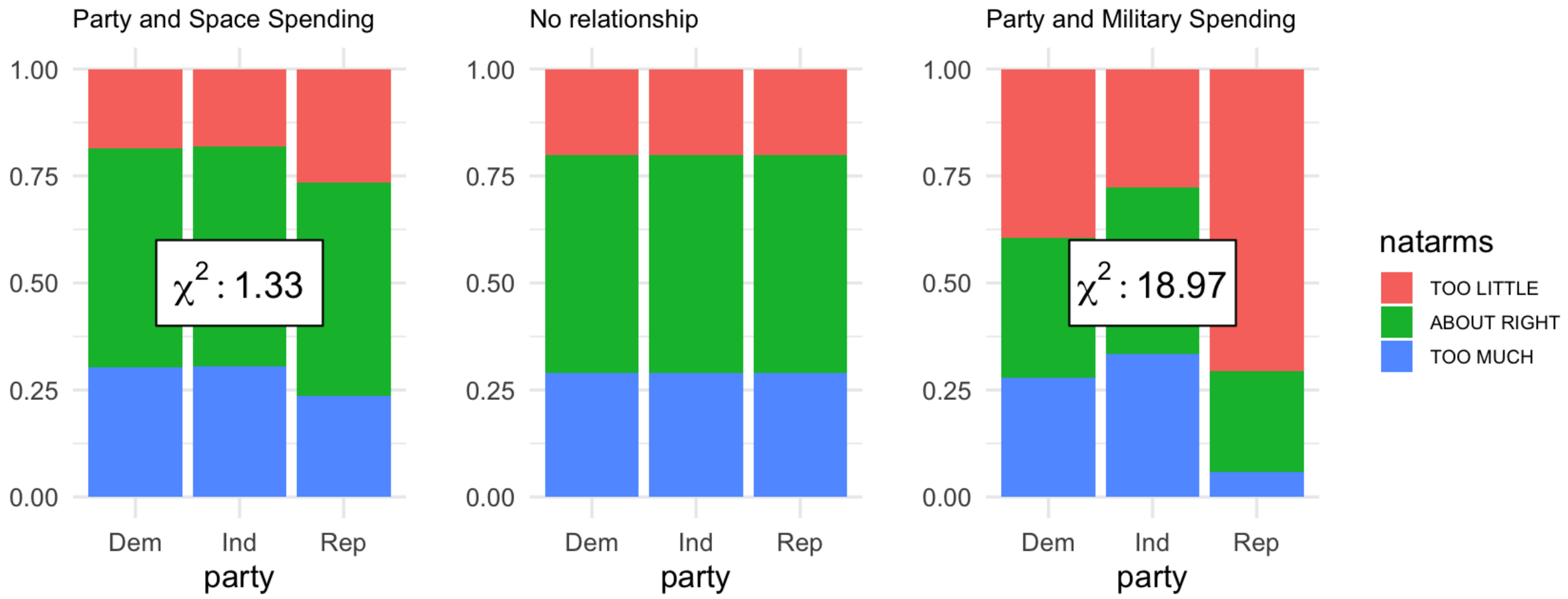
```
(observed_counts - expected_counts) ^ 2
```

```
            party
natspac         Dem   Ind   Rep
  TOO LITTLE  0.433 2.240 4.641
  ABOUT RIGHT 0.005 0.076 0.117
  TOO MUCH    0.349 1.492 3.284
```

```
sum((observed_counts - expected_counts) ^ 2)
```

```
12.63565
```

Chi-squared distance



Let's practice!

INFERENCE FOR CATEGORICAL DATA IN R

Alternate method: the chi-squared distribution

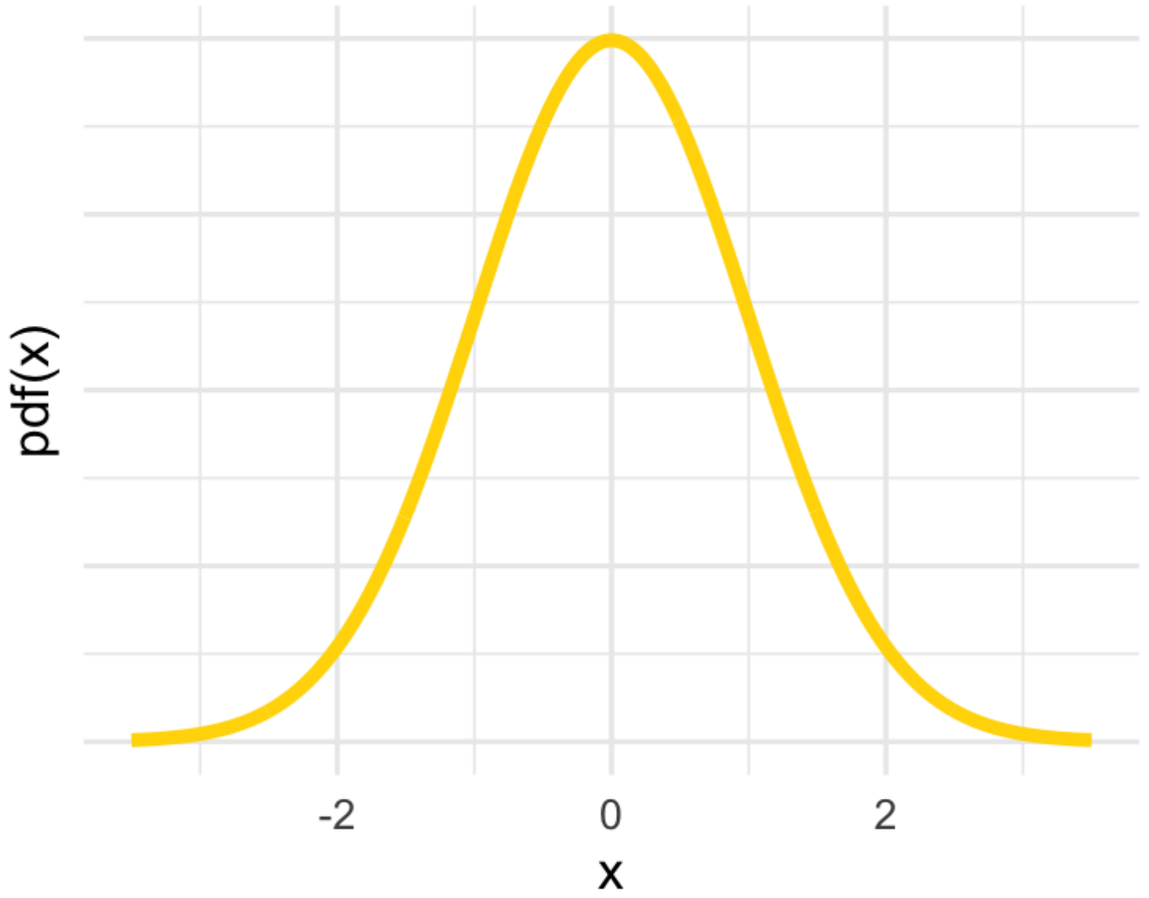
INFERENCE FOR CATEGORICAL DATA IN R

Andrew Bray
Assistant Professor of Statistics at Reed College

datacamp

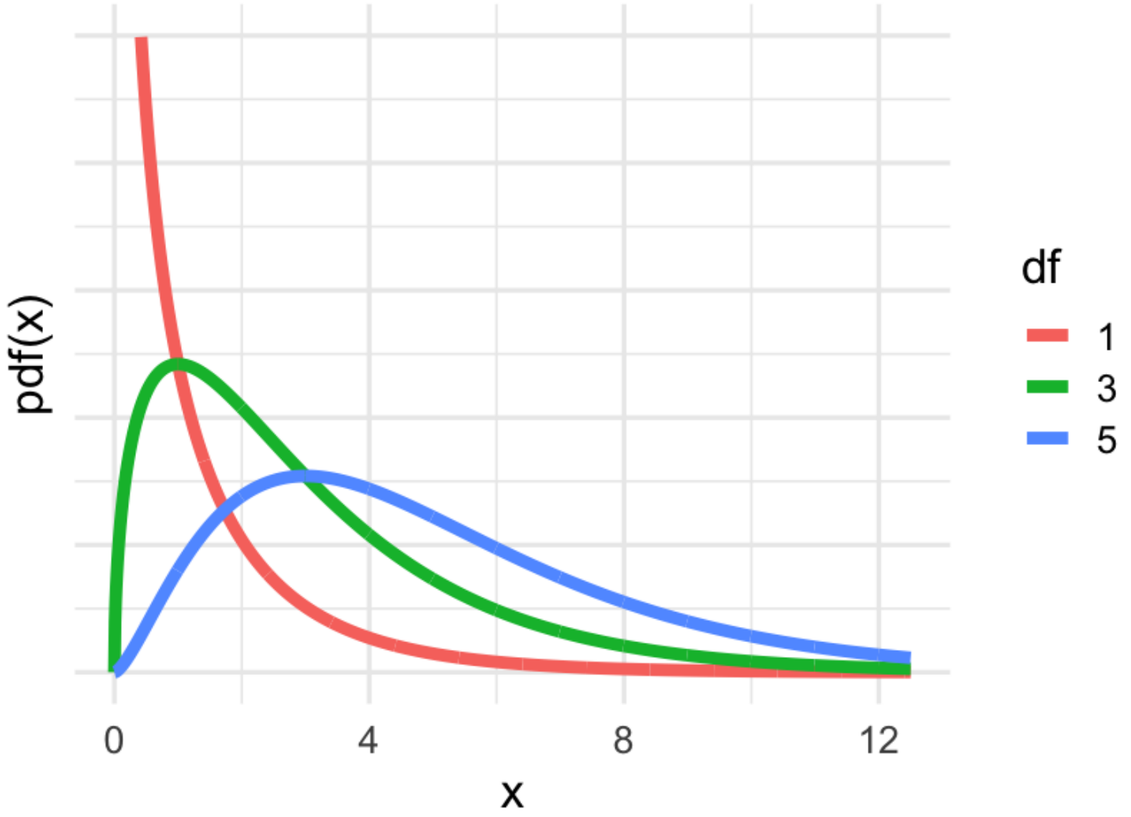
Approximation distributions: normal

- Statistics: $\hat{p}$, $\hat{p}_1 - \hat{p}_2$



Approximation distributions: chi-squared

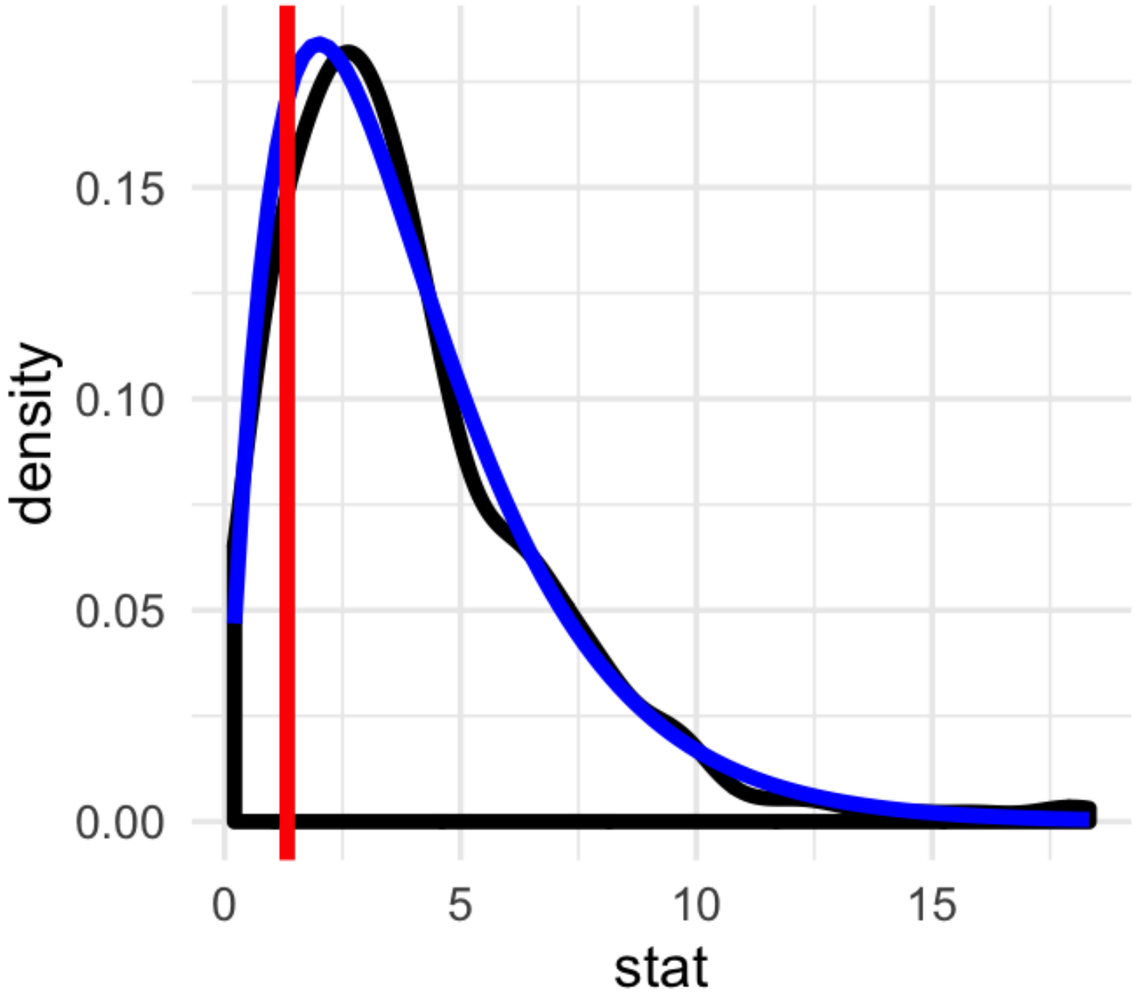
- Statistics: $\hat{x}^2$
- Shape is determined by degrees of freedom
- $df = (nrows - 1) \times (ncols - 1)$



H-test via approximation

```
null_spac <- gss_party %>%
  specify(natspac ~ party) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 100, type = "permute") %>%
  calculate(stat = "Chisq")
```

```
ggplot(null_spac, aes(x = stat)) +
  geom_density() +
  stat_function(
    fun = dchisq,
    args = list(df = 4),
    color = "blue"
  ) +
  geom_vline(xintercept = chi_obs_spac, color = "red")
```



H-test via approximation

```
gss_party %>%
  select(natarms, party) %>%
  table()
```

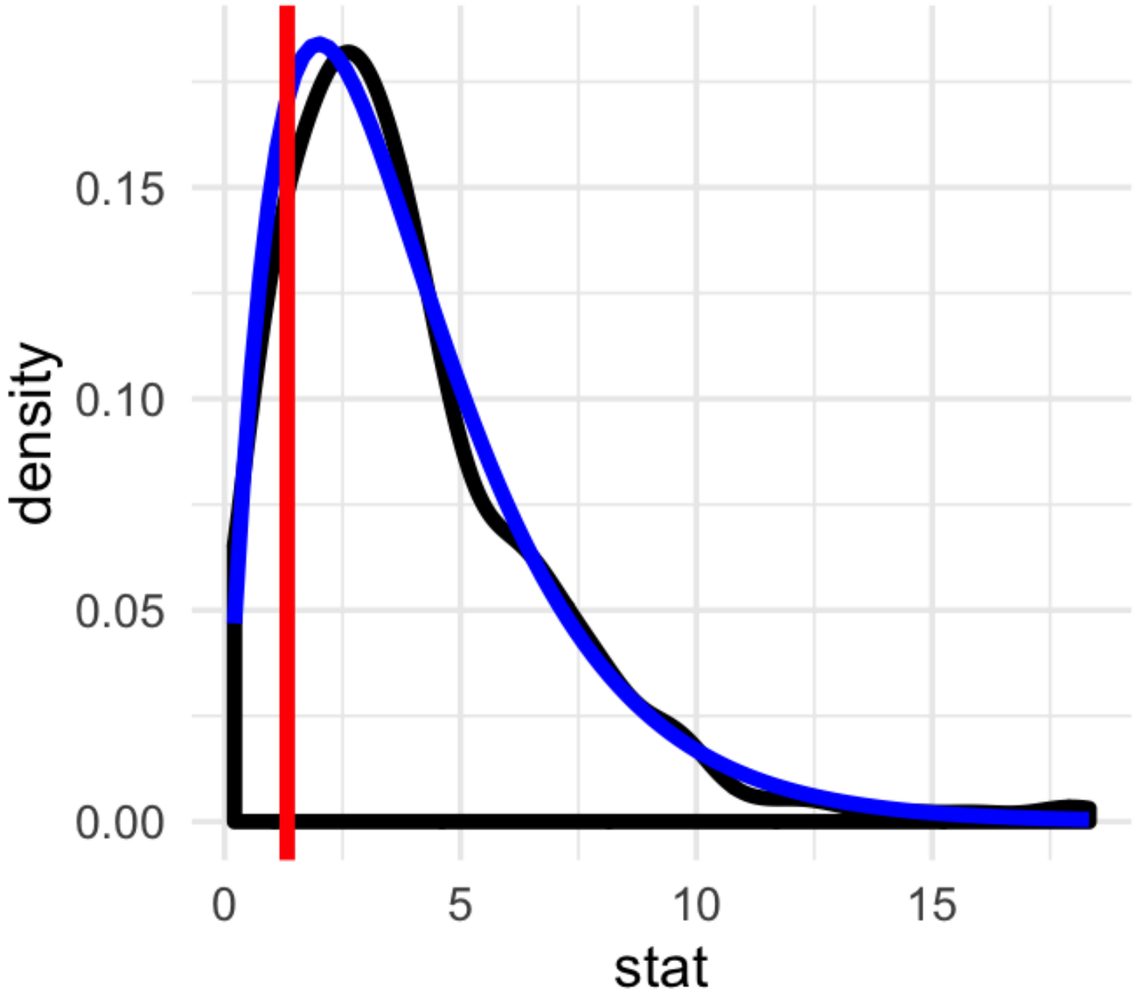
```
            party
natarms        D  I  R
  TOO LITTLE  17 20 24
  ABOUT RIGHT 14 28  8
  TOO MUCH    12 24  2
```

```
pchisq(chi_obs_spac, df = 4)
```

```
X-squared
0.1430612
```

```
1 - pchisq(chi_obs_spac, df = 4)
```

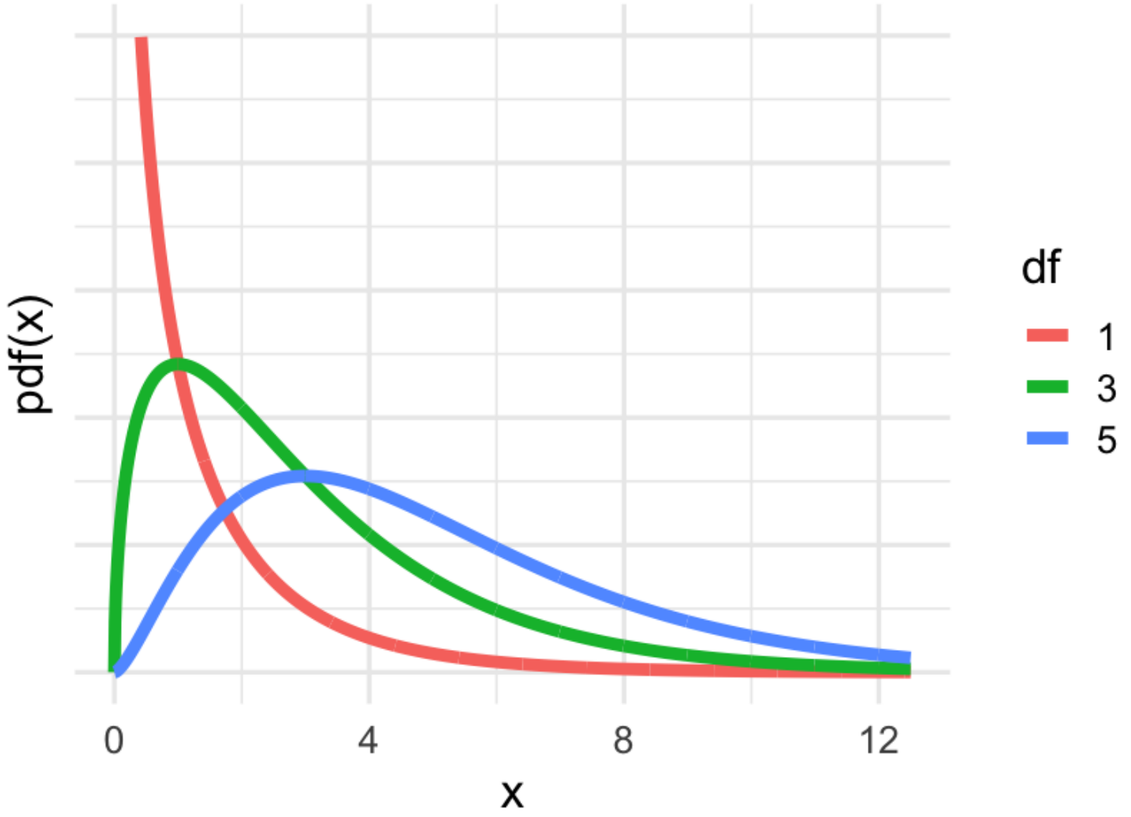
```
X-squared
0.8569388
```



The chi-squared distribution

Becomes a good approximation when:

- $expected\_count >= 5$
- $df >= 2$



Let's practice!

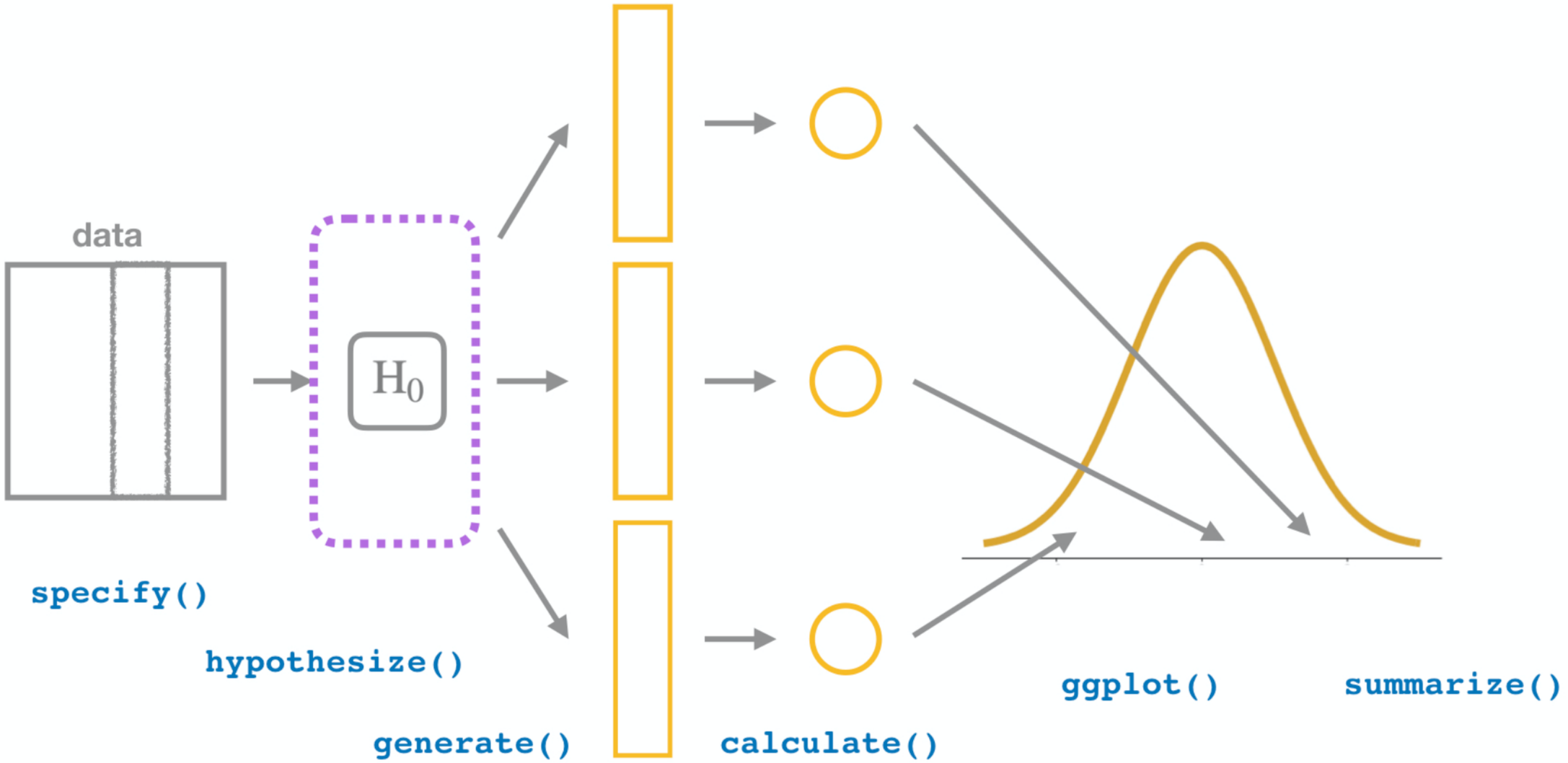
INFERENCE FOR CATEGORICAL DATA IN R

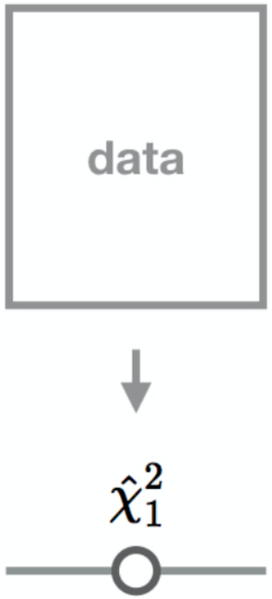
Intervals for the chi-squared distribution

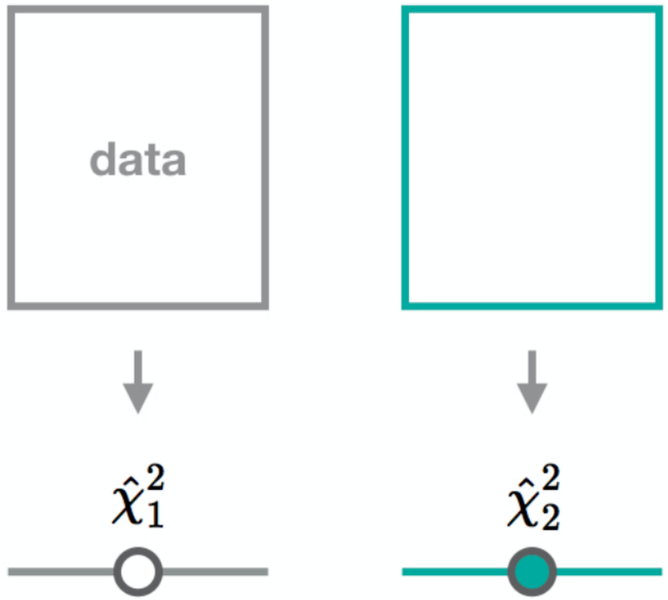
INFERENCE FOR CATEGORICAL DATA IN R

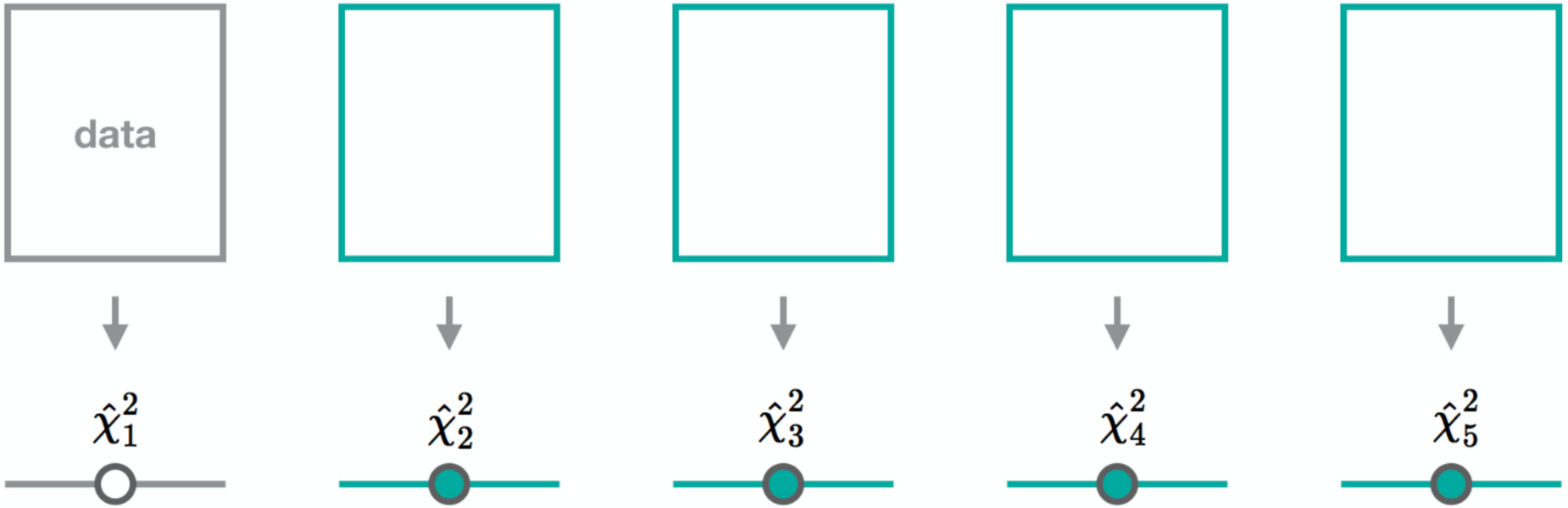
Andrew Bray

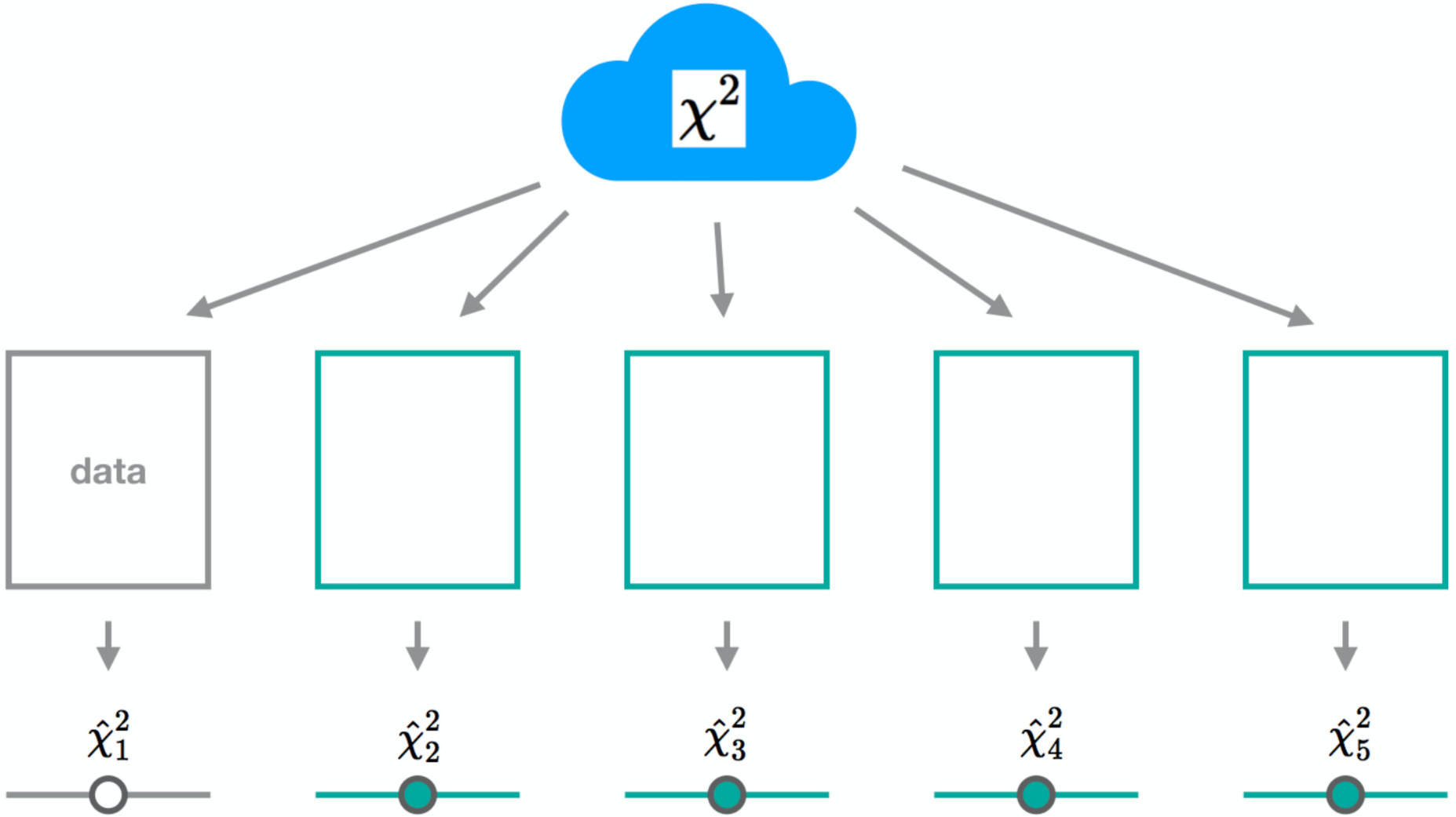
Assistant Professor of Statistics at Reed College

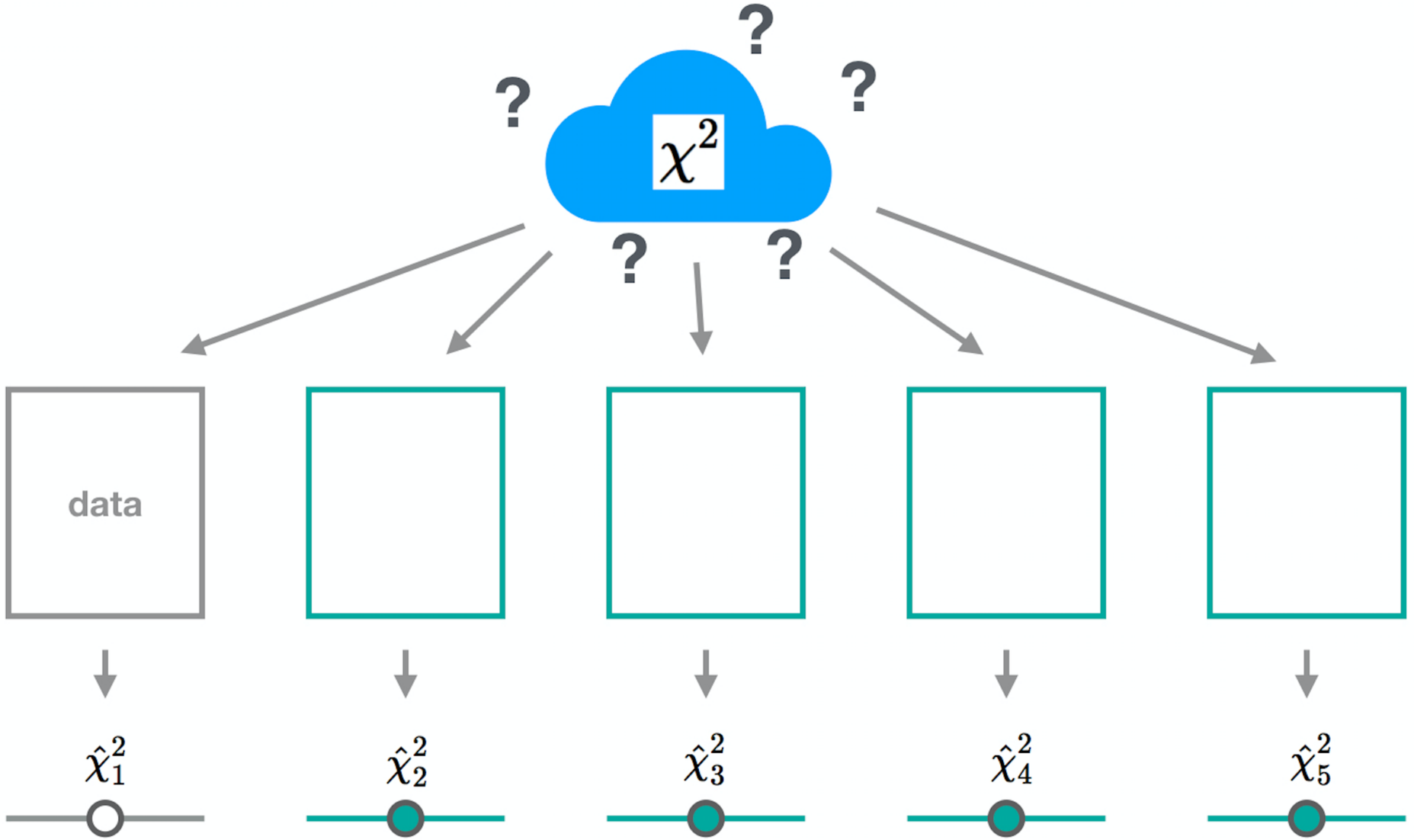

data
H_0
specify()
hypothesize()
generate()
calculate()
ggplot()
summarize()

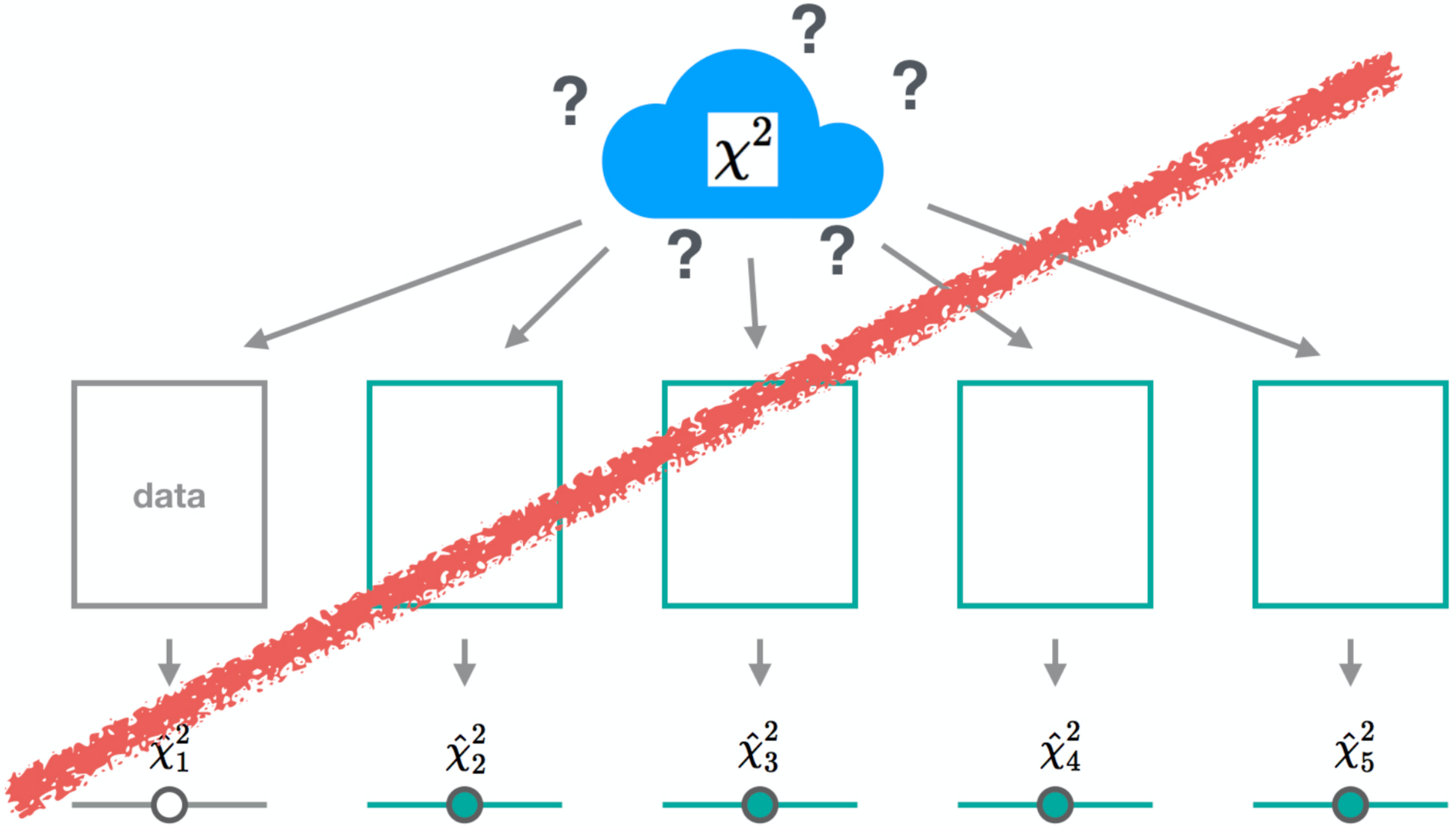

data
$\hat{\chi}_1^2$


data
$\hat{\chi}_1^2$
$\hat{\chi}_2^2$


data
$\hat{\chi}^2_1$
$\hat{\chi}^2_2$
$\hat{\chi}^2_3$
$\hat{\chi}^2_4$
$\hat{\chi}^2_5$


χ^2
data
$\hat{\chi}_1^2$
$\hat{\chi}_2^2$
$\hat{\chi}_3^2$
$\hat{\chi}_4^2$
$\hat{\chi}_5^2$


?
?
?
χ^2
?
?
data
$\hat{\chi}_1^2$
$\hat{\chi}_2^2$
$\hat{\chi}_3^2$
$\hat{\chi}_4^2$
$\hat{\chi}_5^2$


?
?
?
?
?
χ^2
data
$\hat{\chi}_1^2$
$\hat{\chi}_2^2$
$\hat{\chi}_3^2$
$\hat{\chi}_4^2$
$\hat{\chi}_5^2$

Let's practice!

INFERENCE FOR CATEGORICAL DATA IN R