Case study: election fraud

INFERENCE FOR CATEGORICAL DATA IN R

Andrew Bray

Assistant Professor of Statistics at Reed College

Election fraud

- Vote buying
- Voting twice
- Altering vote totals

precinct	candidate A	candidate B
1	53	37
2	36	31
3	68	55
4	17	19
5	24	27

[1] The phrase election fraud can mean many things including vote buying, casting two ballots in different locations, and stuffing ballot boxes with fake ballots. We're going to focus on a version of the third, when the vote

Election fraud

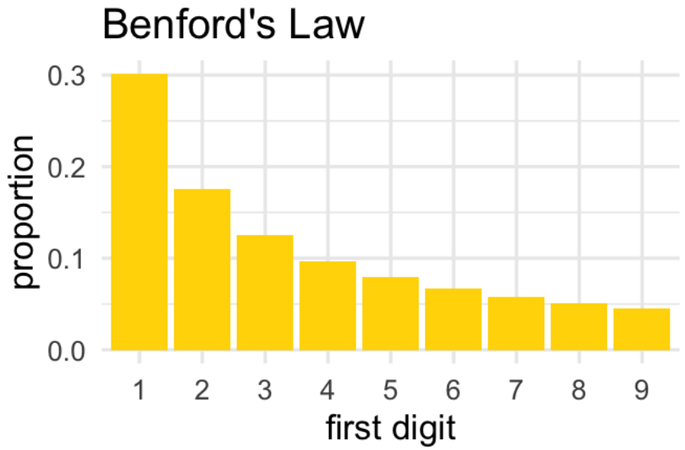
- Vote buying
- Voting twice
- Altering vote totals

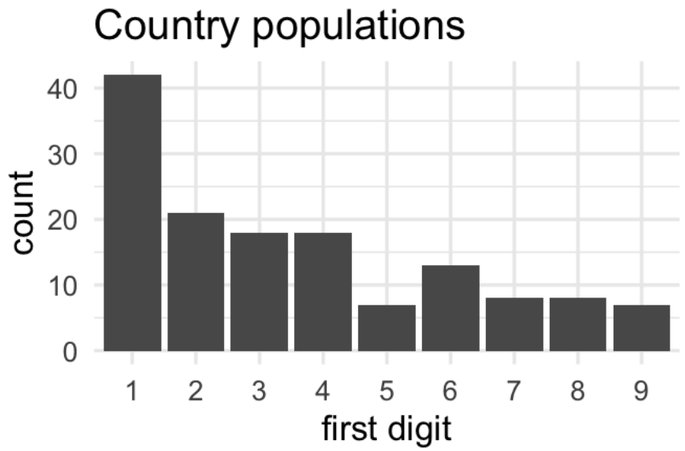
precinct	candidate A	candidate B
1	53	77 ~~37~~
2	36	31
3	68	85 ~~55~~
4	17	19
5	24	27

Benford's Law A.K.A. "the first digit law"

```
library(gapminder)
gapminder %>%
  filter(year == 2007) %>%
  select(country, pop)
```

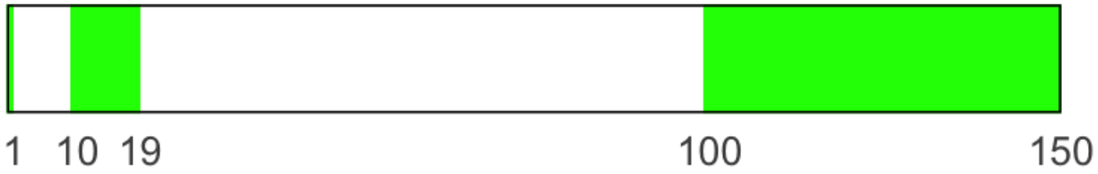
```
# A tibble: 142 x 2
   country           pop
   <fct>           <int>
 1 Afghanistan  31889923
 2 Albania       3600523
 3 Algeria      33333216
 4 Angola       12420476
 5 Argentina    40301927
 6 Australia    20434176
 7 Austria       8199783
 8 Bahrain        708573
 9 Bangladesh  150448339
10 Belgium      10392226
# … with 132 more rows
```

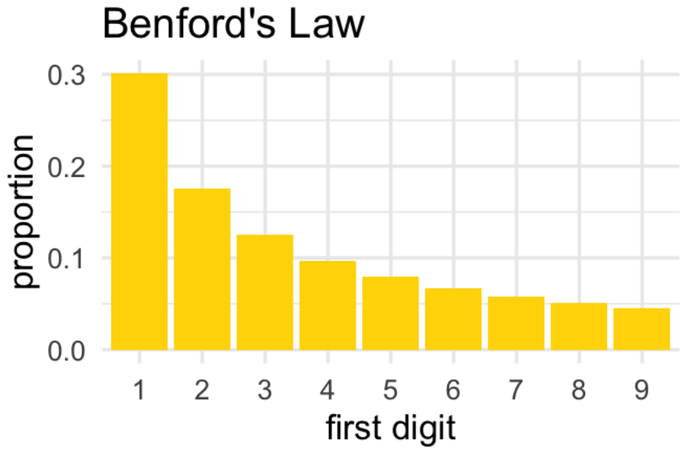




Benford's Law A.K.A. "the first digit law"

- If the election was fair then vote counts should follow Benford's Law.
- If the election was fraudulent then vote counts should not follow Benford's Law.





Iran election 2009

```
iran %>%
  select(city, ahmadinejad, mousavi, total_votes_cast)
```

```
# A tibble: 366 x 4
   city          ahmadinejad mousavi total_votes_cast
   <chr>               <dbl>   <dbl>            <dbl>
 1 Azar Shahr          37203   18312            56712
 2 Asko                32510   18799            52643
 3 Ahar                47938   26220            75500
 4 Bostan Abad         38610   12603            51911
 5 Bonab               36395   33695            71389
 6 Tabriz             435728  419983           876919
 7 Jalfa               20520   14340            35295
 8 Chahar o Imaq       12197    3975            16375
 9 Sarab               53196   17669            72152
10 Shabestar           37099   39182            77459
# … with 356 more rows
```

Ahmadinejad

Mousavi

Let's practice!

INFERENCE FOR CATEGORICAL DATA IN R

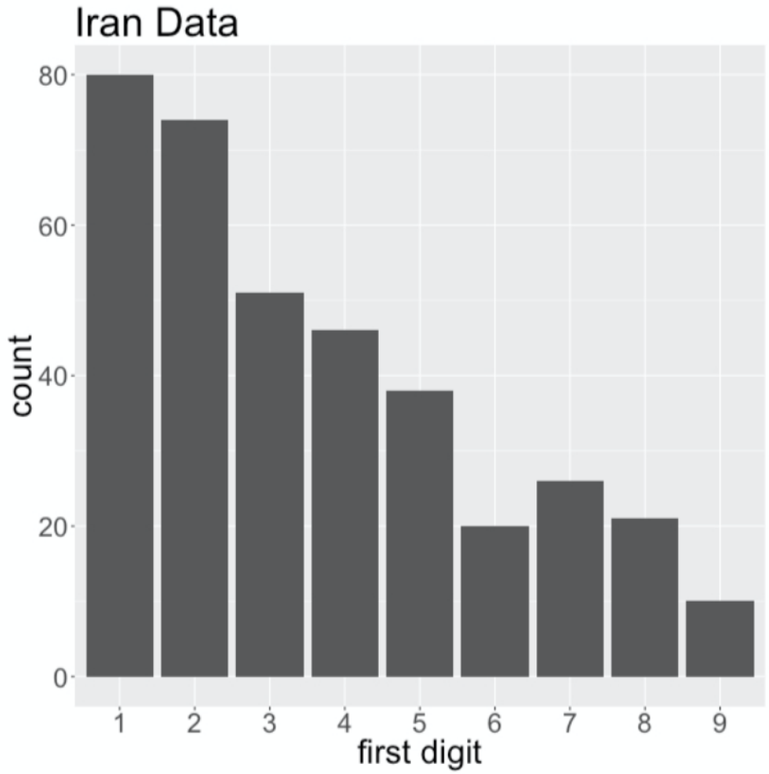
Goodness of fit

INFERENCE FOR CATEGORICAL DATA IN R

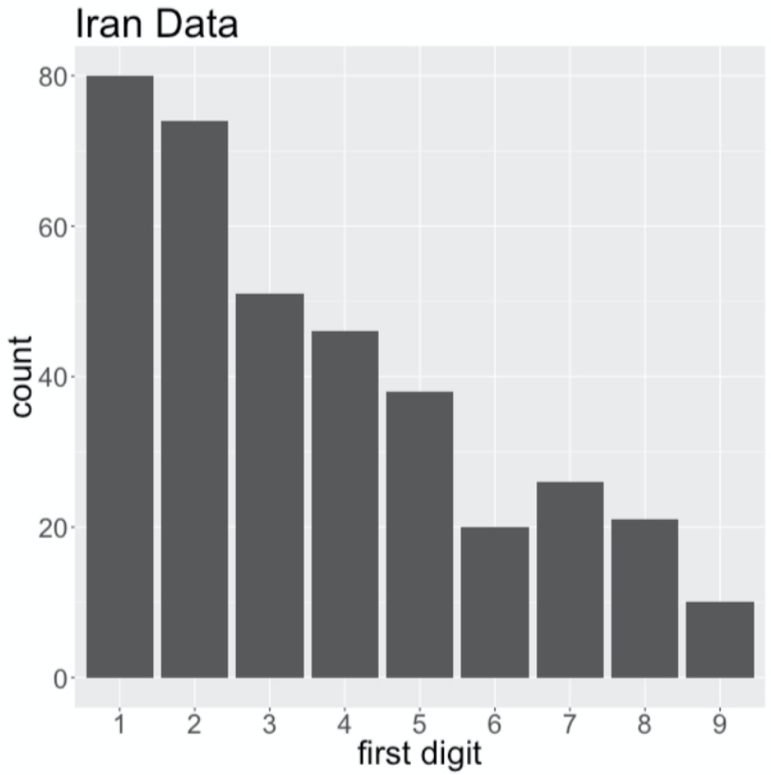
Andrew Bray

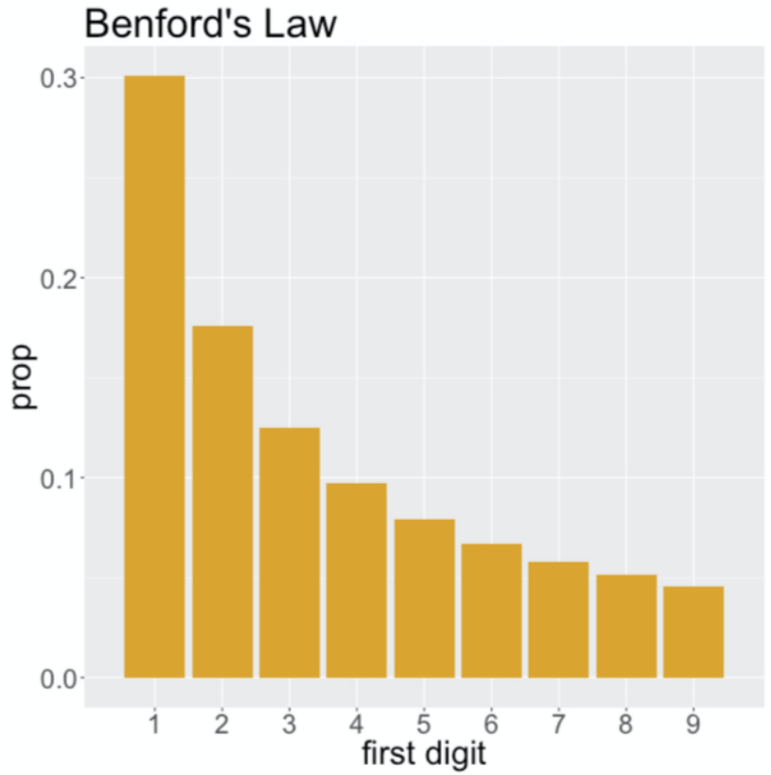
Assistant Professor of Statistics at Reed College

First Digit Distribution

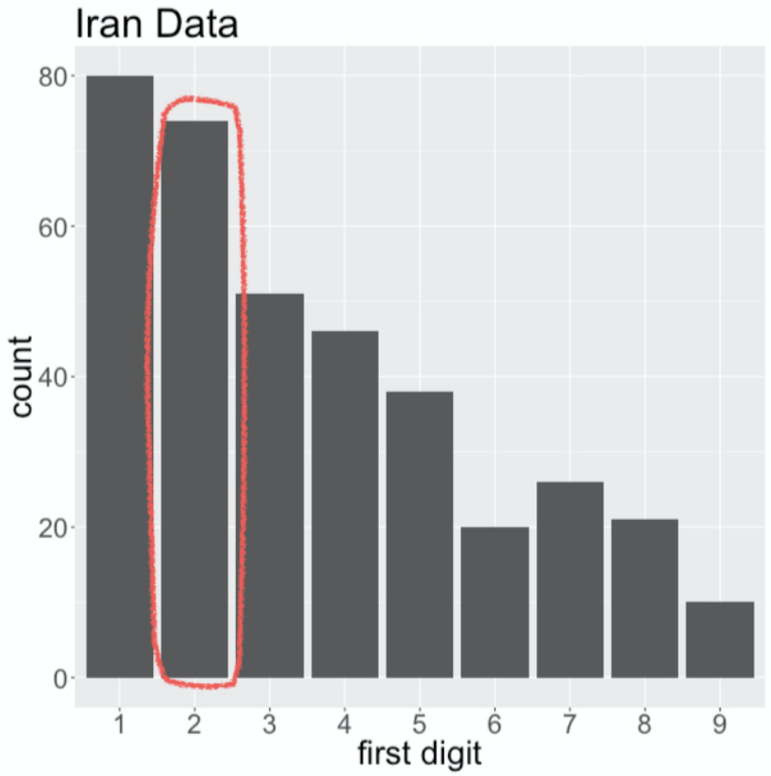


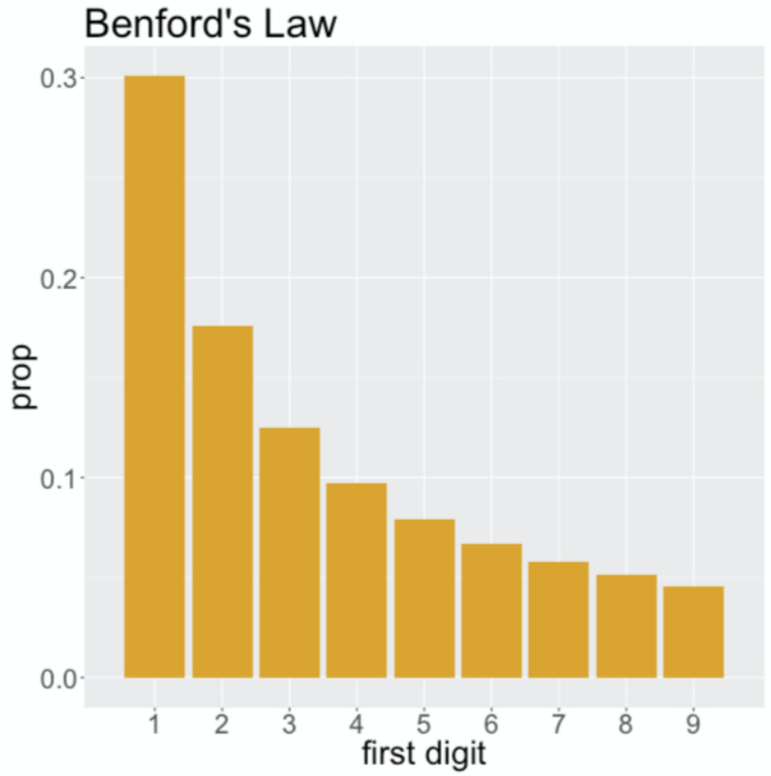
First Digit Distribution



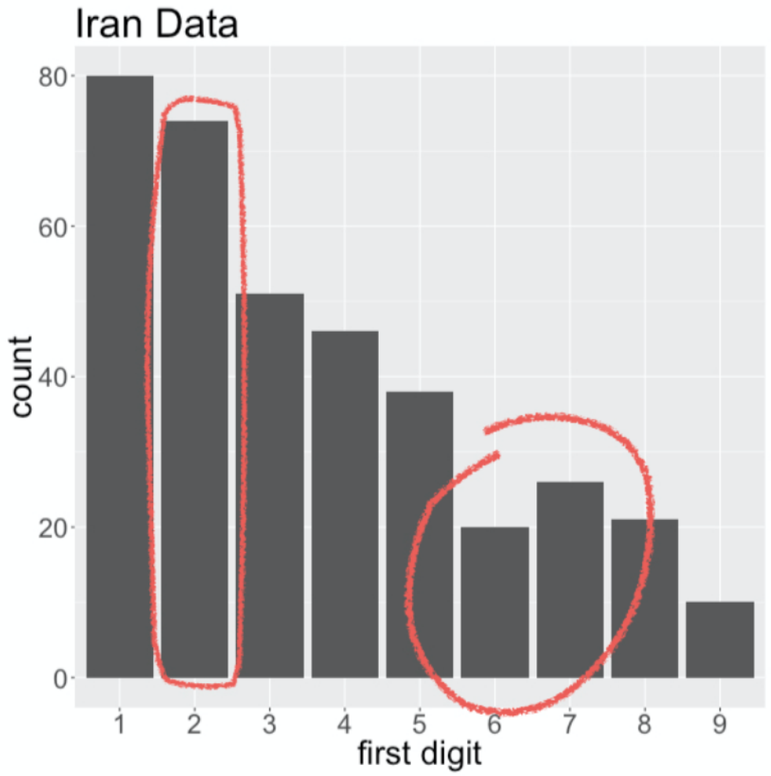


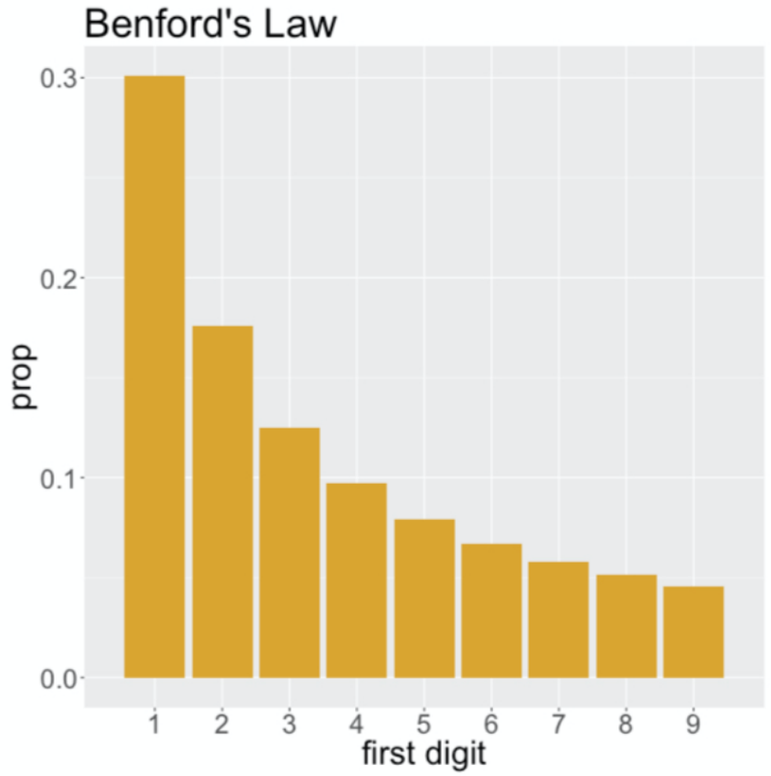
First Digit Distribution



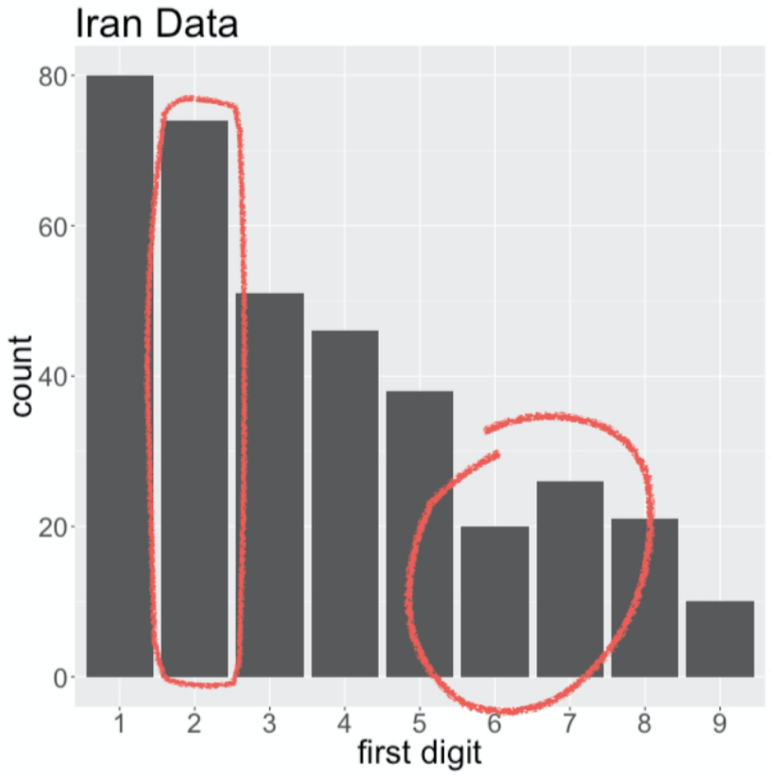


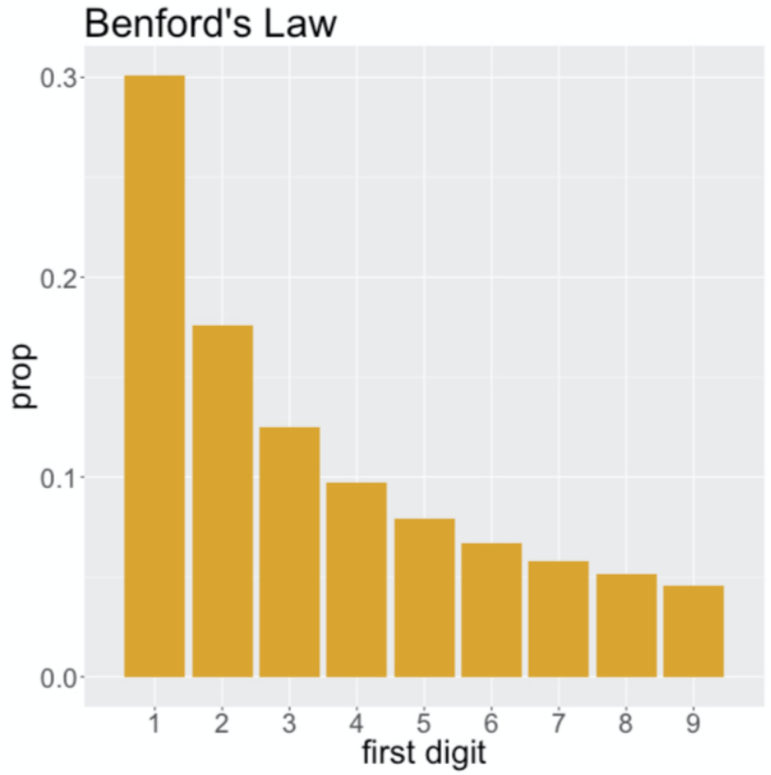
First Digit Distribution





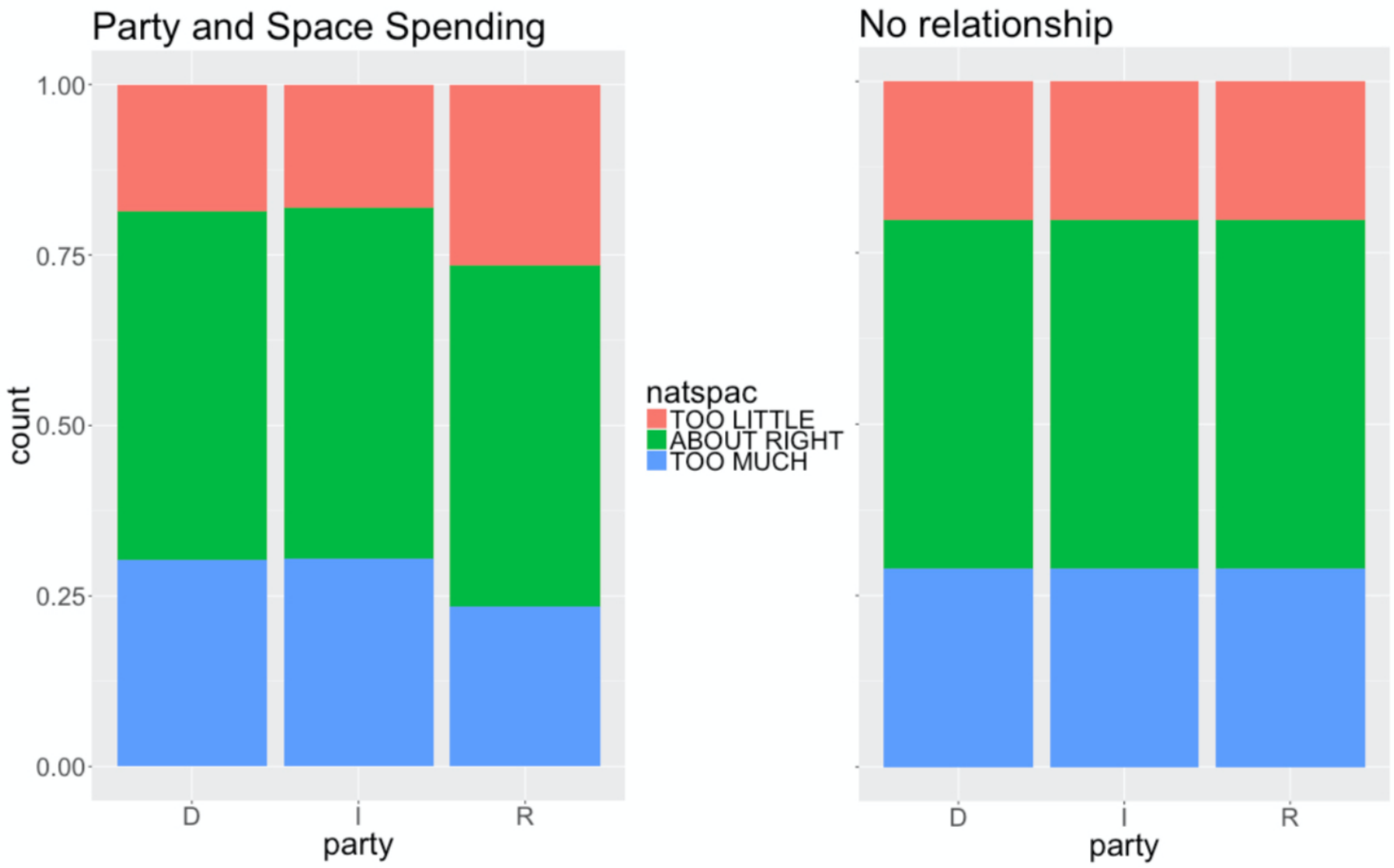
First Digit Distribution



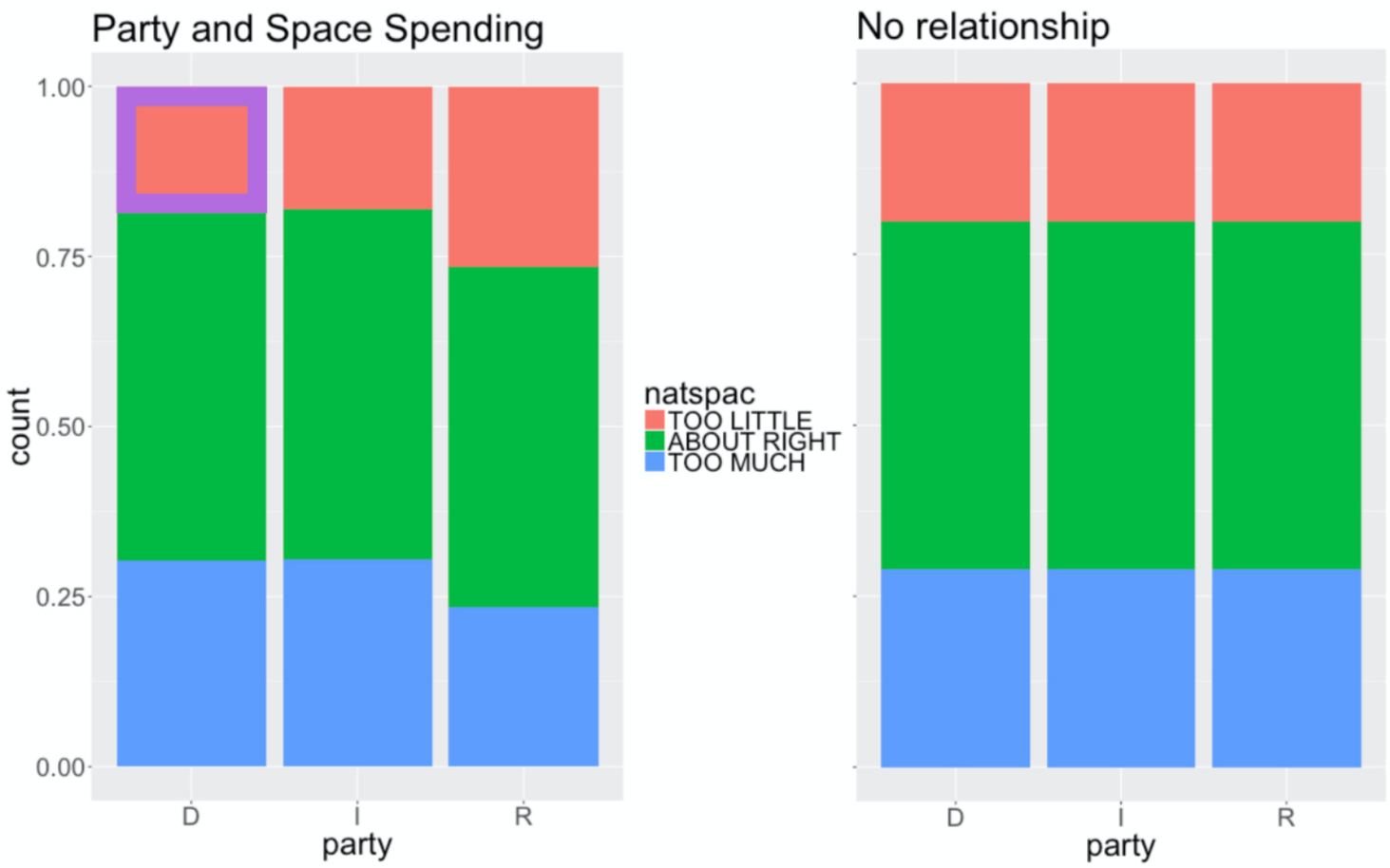


How far apart are these distributions?

Chi-squared distance

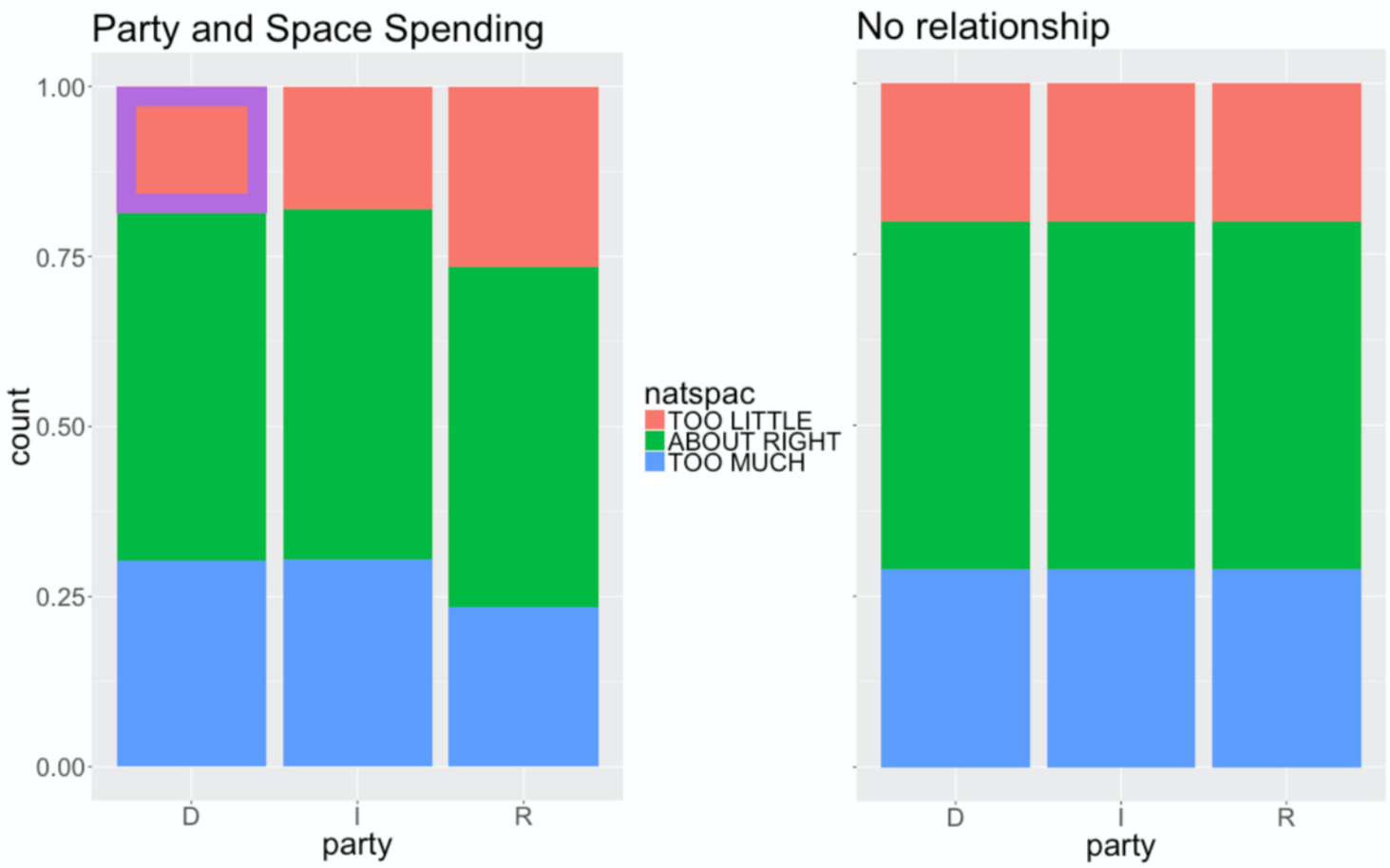


Chi-squared distance



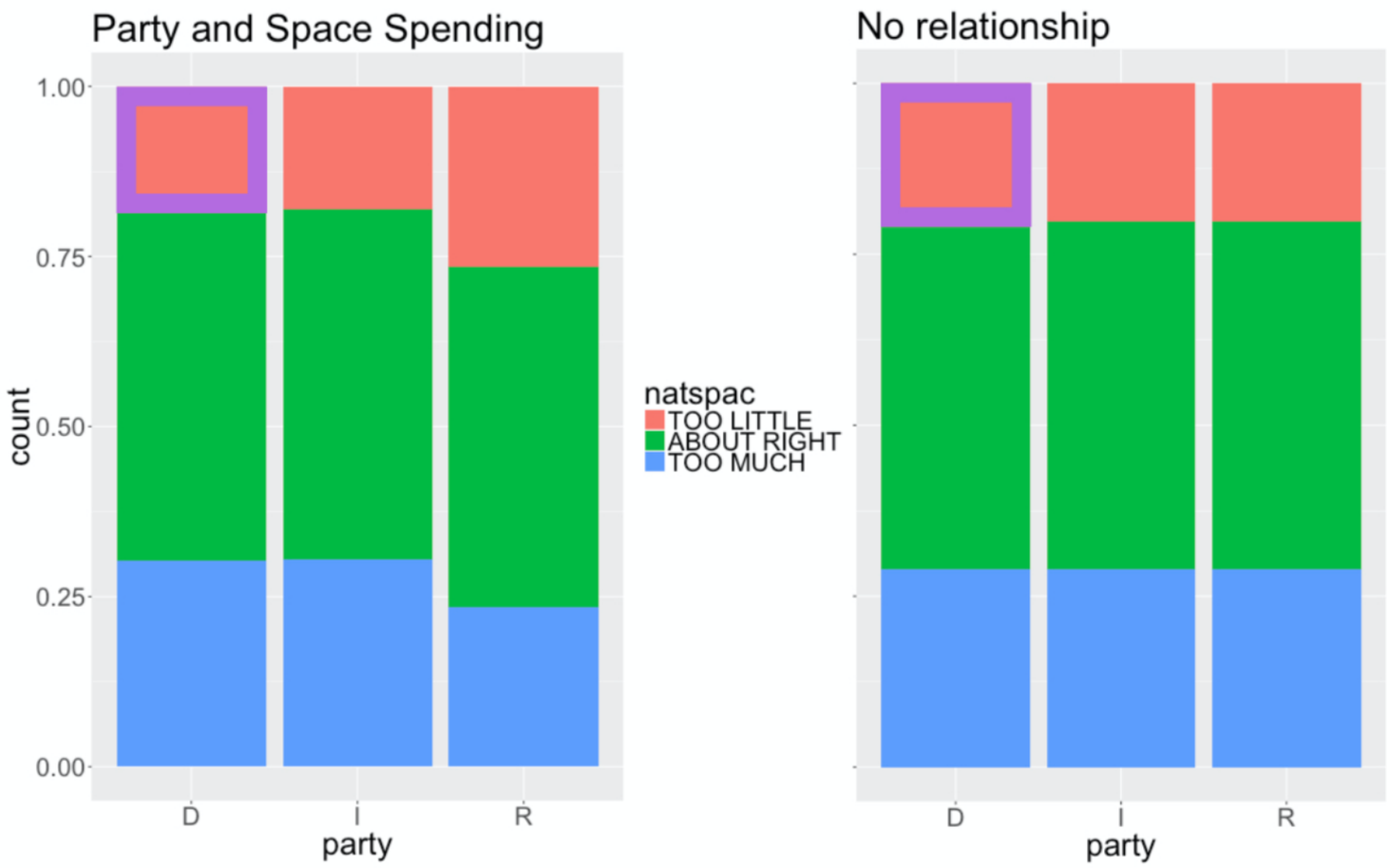
Cell 1:

Chi-squared distance



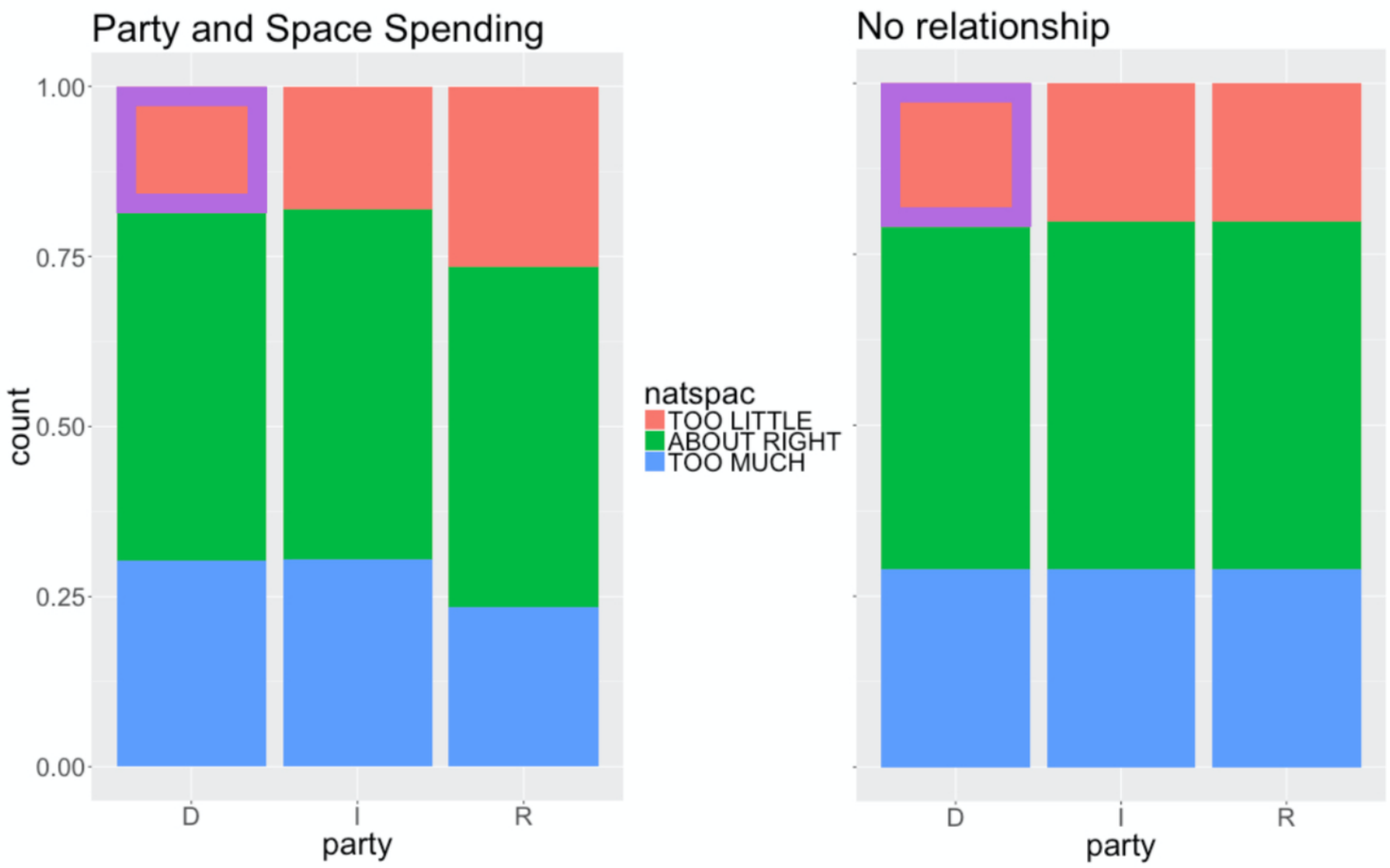
Cell 1: O

Chi-squared distance



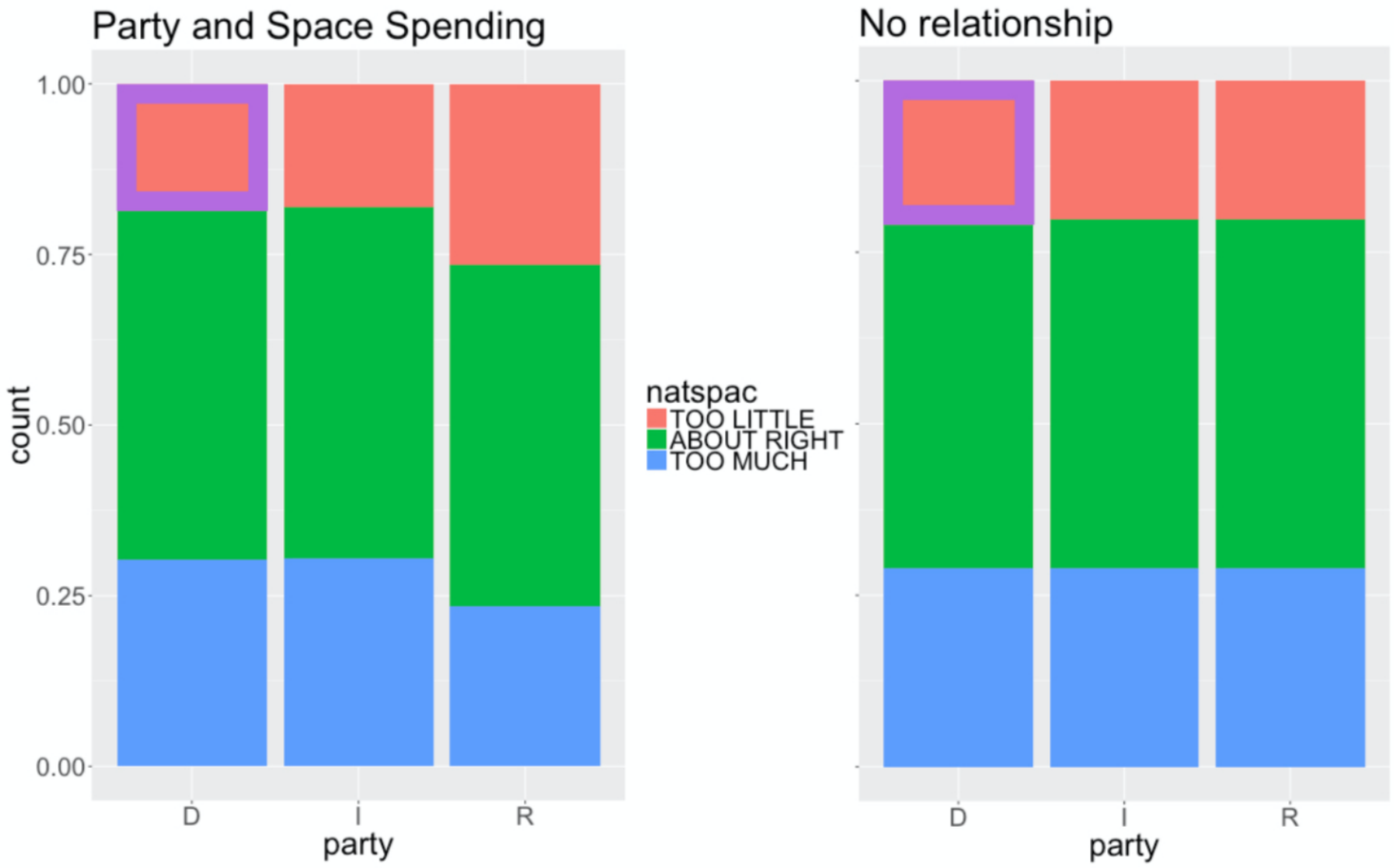
Cell 1: O - E

Chi-squared distance



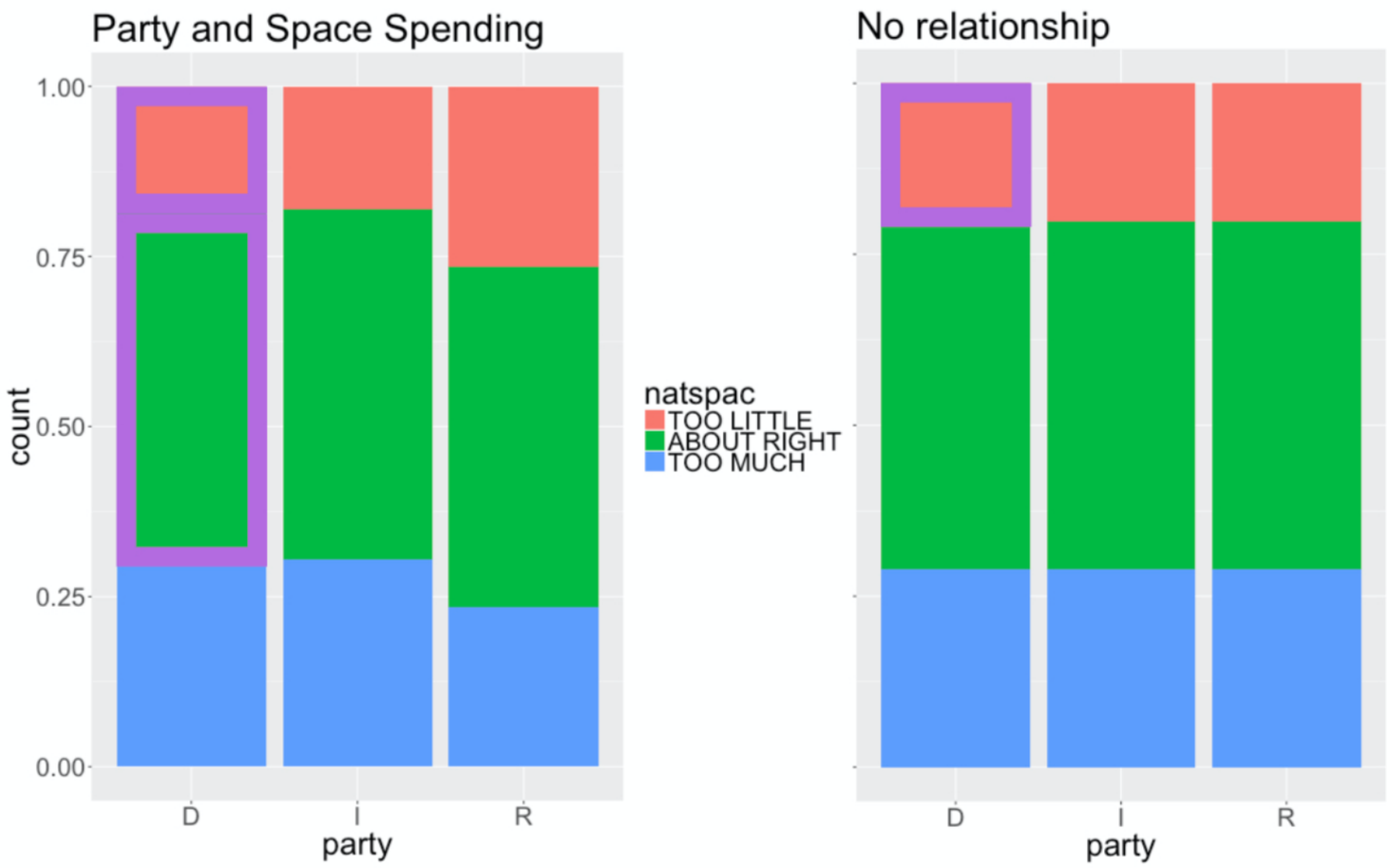
Cell 1: (O - E)^2

Chi-squared distance



Cell 1: $\frac{(O - E)^2}{E}$

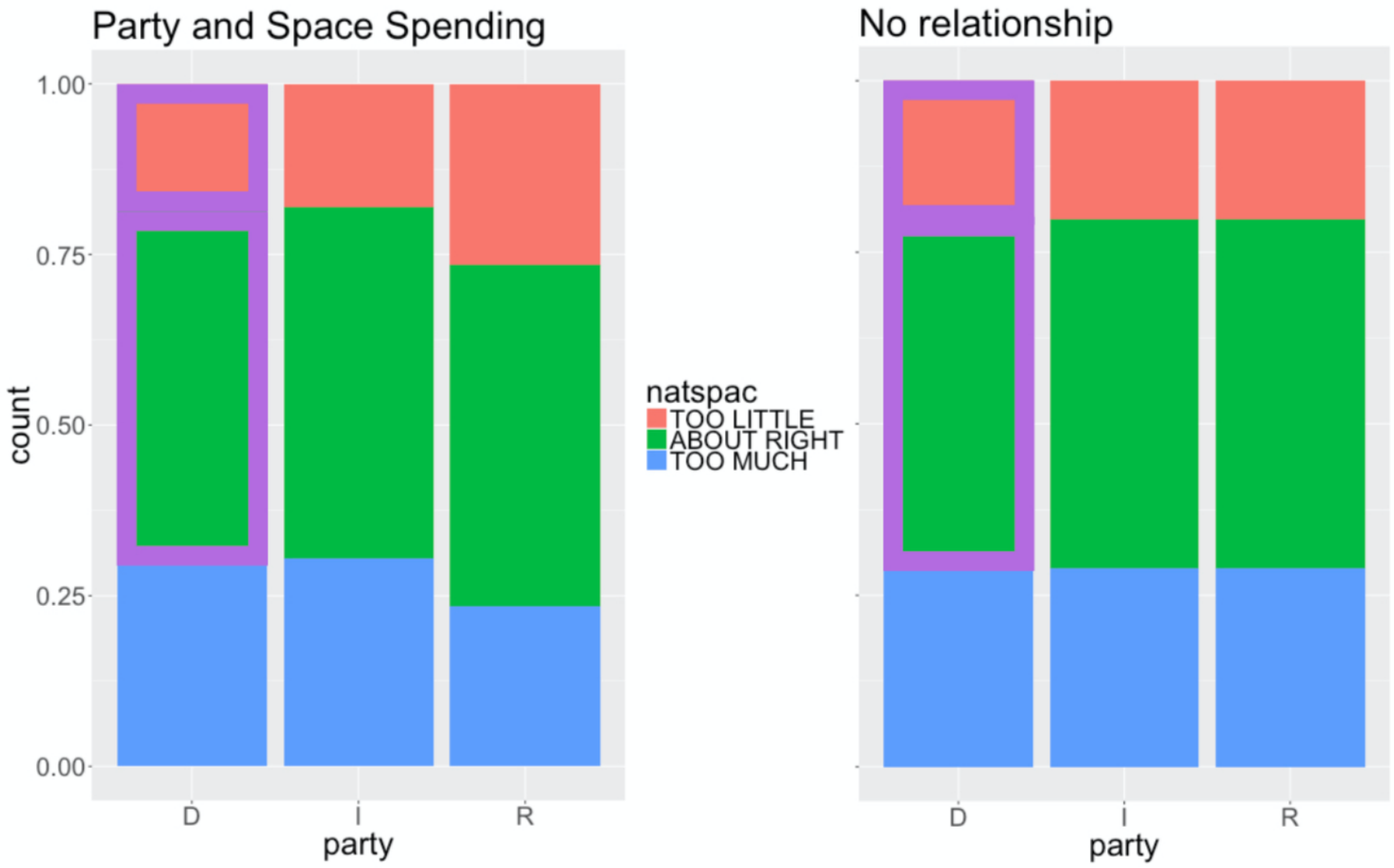
Chi-squared distance



Cell 1: $\frac{(O - E)^2}{E}$

Cell 2: O

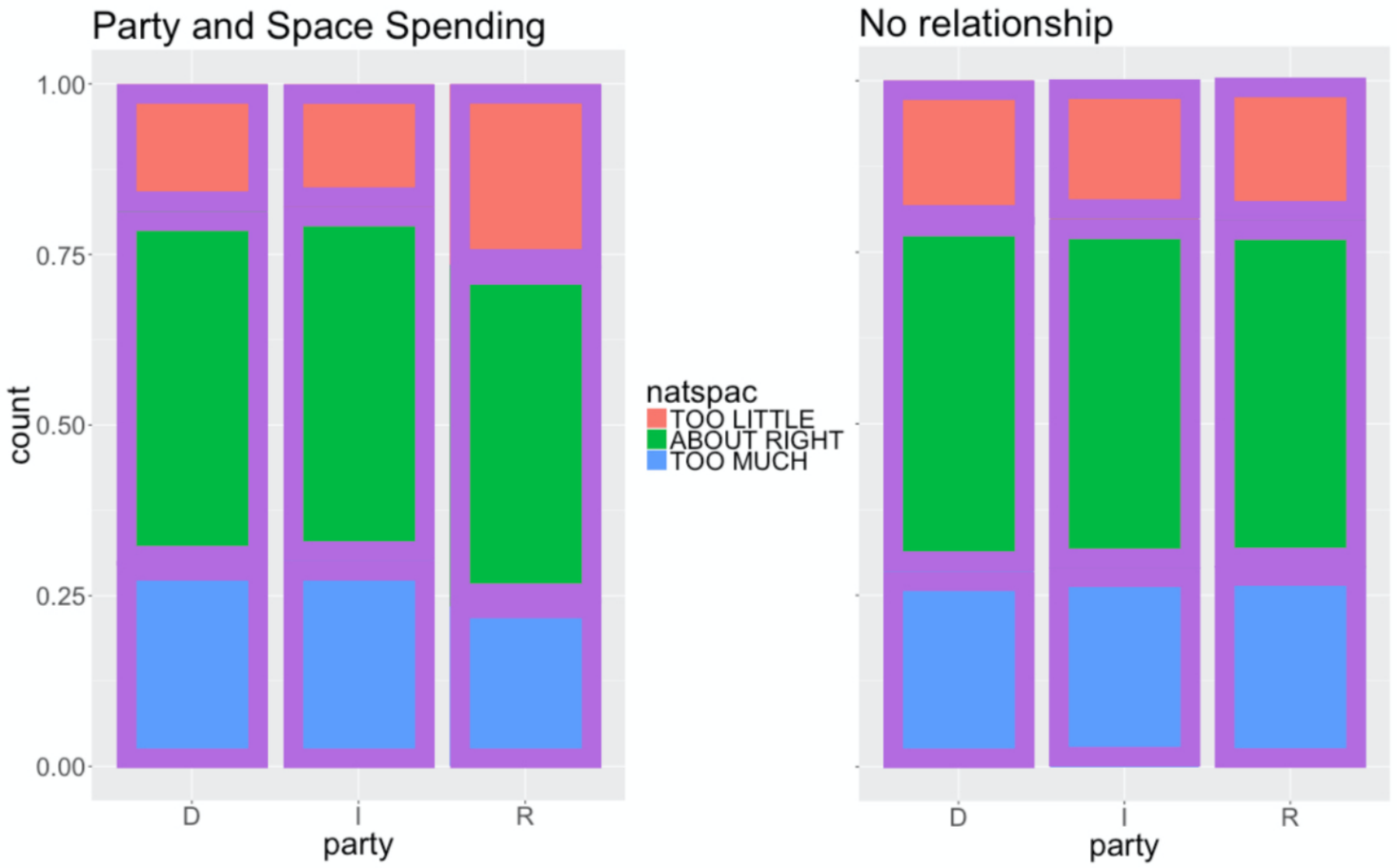
Chi-squared distance



Cell 1: $\frac{(O - E)^2}{E}$

Cell 2: $\frac{(O - E)^2}{E}$

Chi-squared distance



Cell 1: $\frac{(O - E)^2}{E}$

Cell 2: $\frac{(O - E)^2}{E}$

↓

Cell 9: $\frac{(O - E)^2}{E}$

$+$

Chi-squared statistic

Chi-squared distance

Cell 1: $\frac{(O - E)^2}{E}$

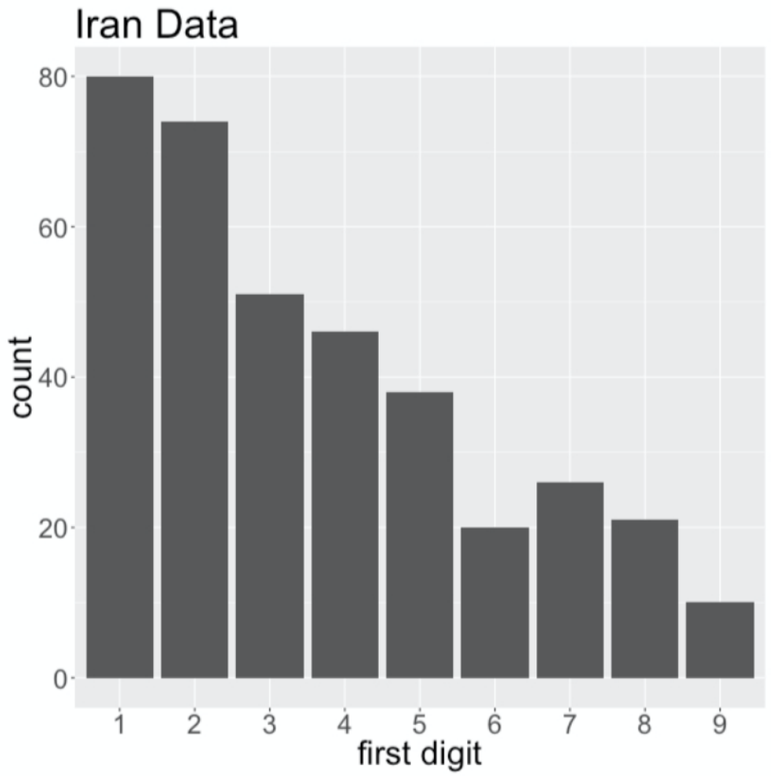
Cell 2: $\frac{(O - E)^2}{E}$

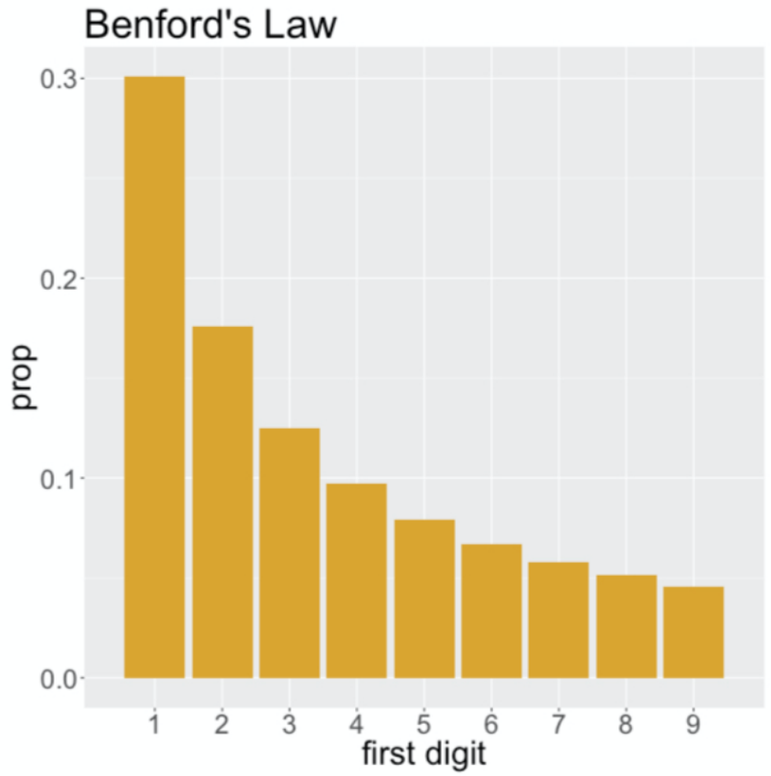
Cell 9: $\frac{(O - E)^2}{E}$

$+$

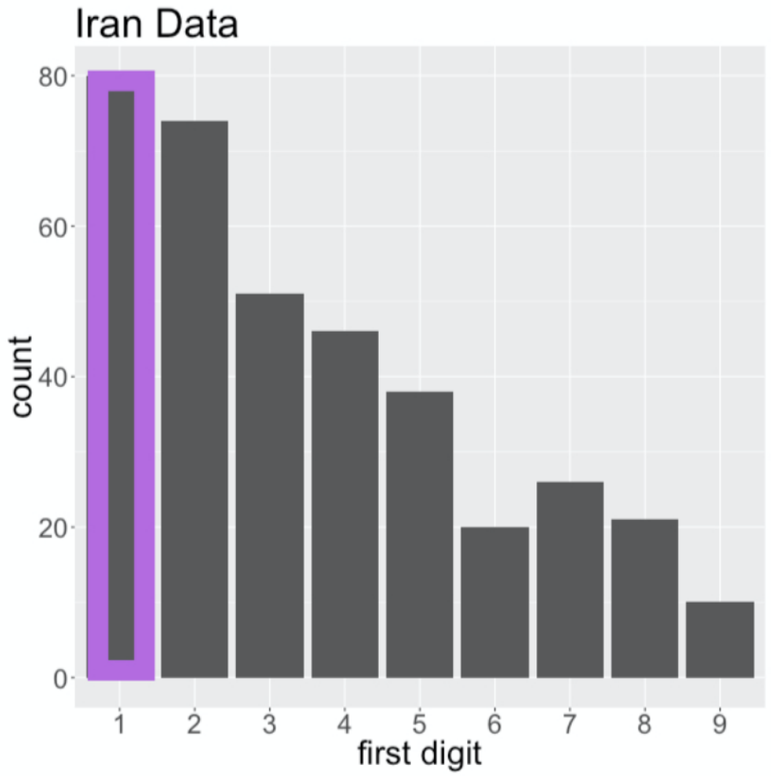
Chi-squared statistic

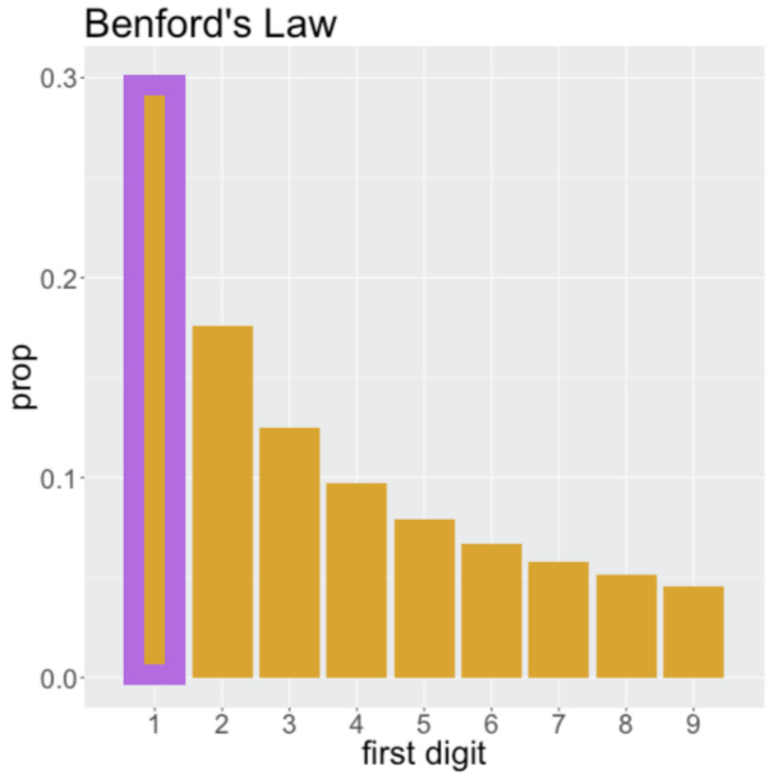
First Digit Distribution





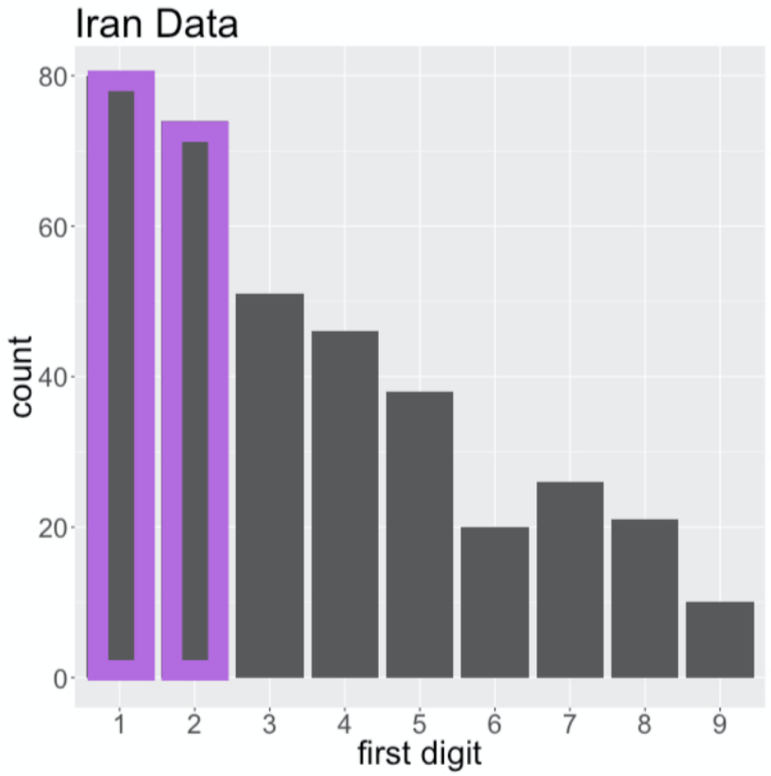
First Digit Distribution

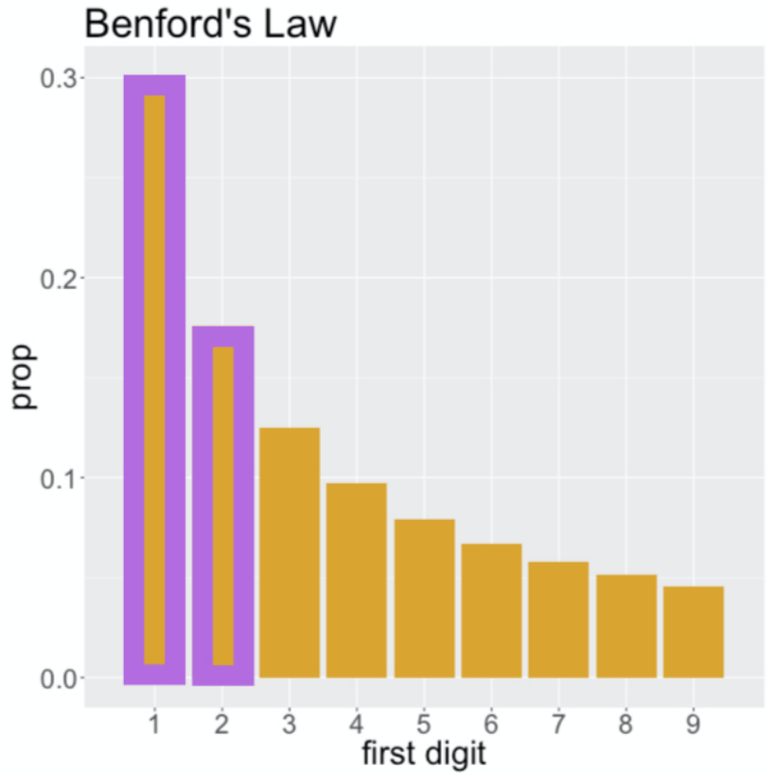




Cell 1: $\frac{(O - E)^2}{E}$

First Digit Distribution

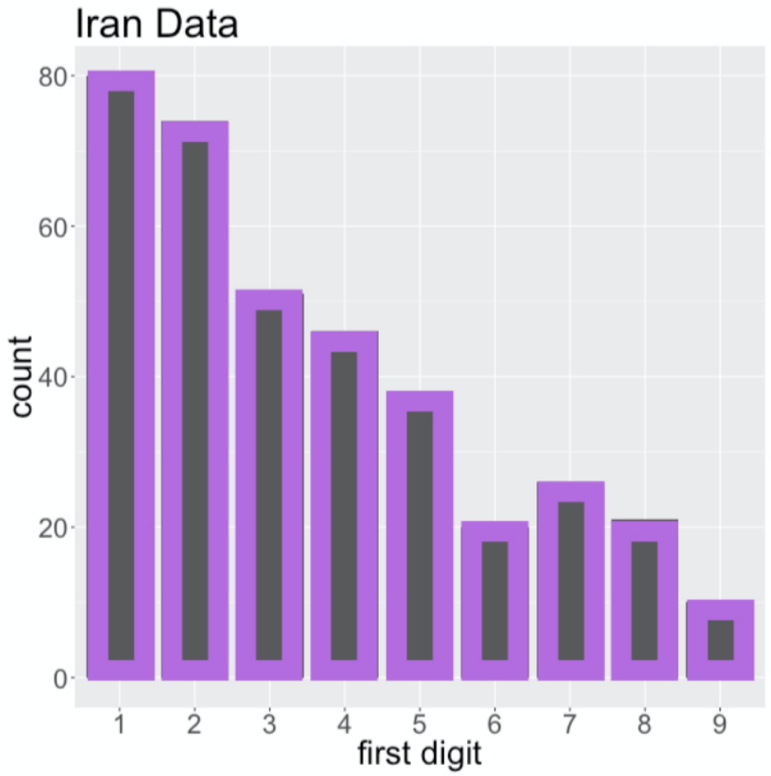


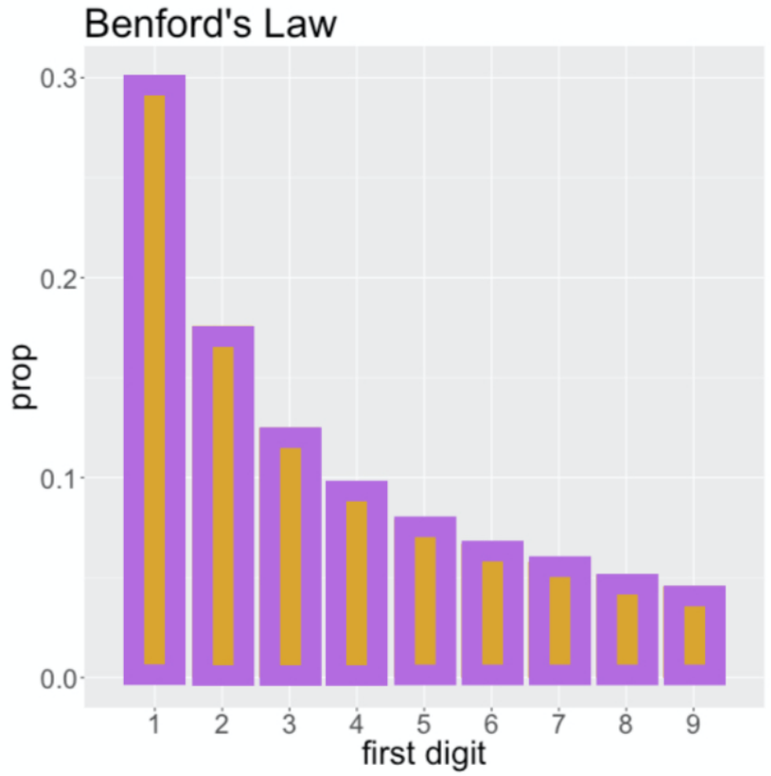


Cell 1: $\frac{(O - E)^2}{E}$

Cell 2: $\frac{(O - E)^2}{E}$

First Digit Distribution





Cell 1: $\frac{(O - E)^2}{E}$

Cell 2: $\frac{(O - E)^2}{E}$

↓

Cell 9: $\frac{(O - E)^2}{E}$

\+

Chi-squared statistic

Example: uniformity of party

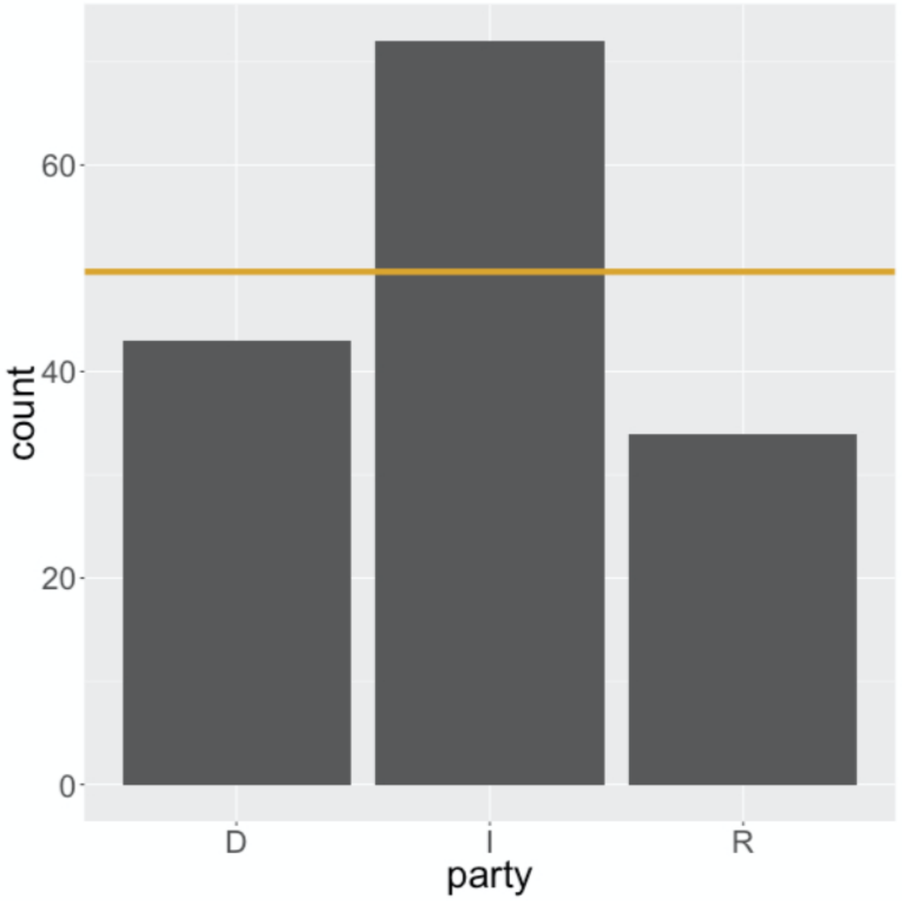
```
ggplot(gss2016, aes(x = party)) +
  geom_bar() +
  geom_hline(yintercept = 149/3, color = "goldenrod", size = 2)
```

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

```
Dem Ind Rep
 43  72  34
```

```
p_uniform <- c(Dem = 1/3, Ind = 1/3, Rep = 1/3)
chisq.test(tab, p = p_uniform)$stat
```

```
X-squared
 15.87919
```



Simulating the null

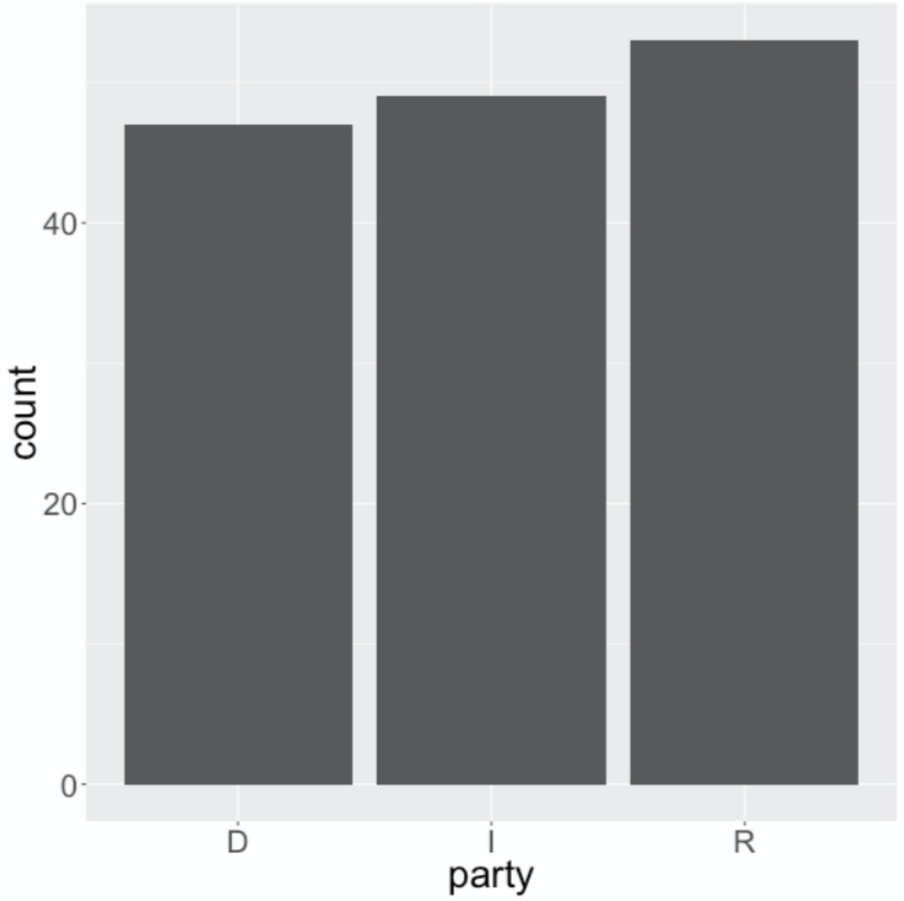
```
gss2016 %>%
  specify(response = party) %>%
  hypothesize(null = "point", p = p_uniform) %>%
  generate(reps = 1, type = "simulate")
```

```
# A tibble: 149 x 2
# Groups:   replicate [1]
   party replicate
   <fct> <fct>
 1 I         1
 2 D         1
 3 I         1
 4 I         1
 5 D         1
 6 R         1
 7 I         1
 8 R         1
 9 D         1
10 I         1
# ... with 139 more rows
```

Simulating the null

```
sim_1 <- gss2016 %>%
   specify(response = party) %>%
   hypothesize(null = “point”, p = p_uniform) %>%
   generate(reps = 1, type = "simulate")
```

```
ggplot(sim_1, aes(x = party)) +
   geom_bar()
```



Let's practice!

INFERENCE FOR CATEGORICAL DATA IN R

And now to US

INFERENCE FOR CATEGORICAL DATA IN R

Andrew Bray

Assistant Professor of Statistics at Reed College

Iran election fraud

H_0: the election was fair (Benford's Law holds)

Iran election fraud

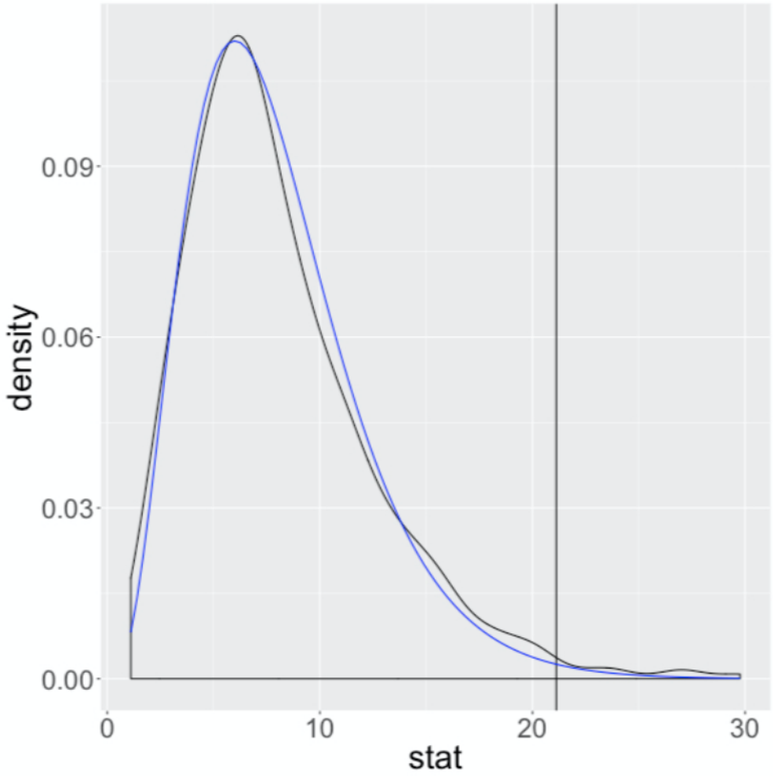
H_0: the election was fair (Benford's Law holds)

H_A: the election was fraudulent (Benford's Law does *not* hold)

Iran election fraud

H_0: the election was fair (Benford's Law holds)

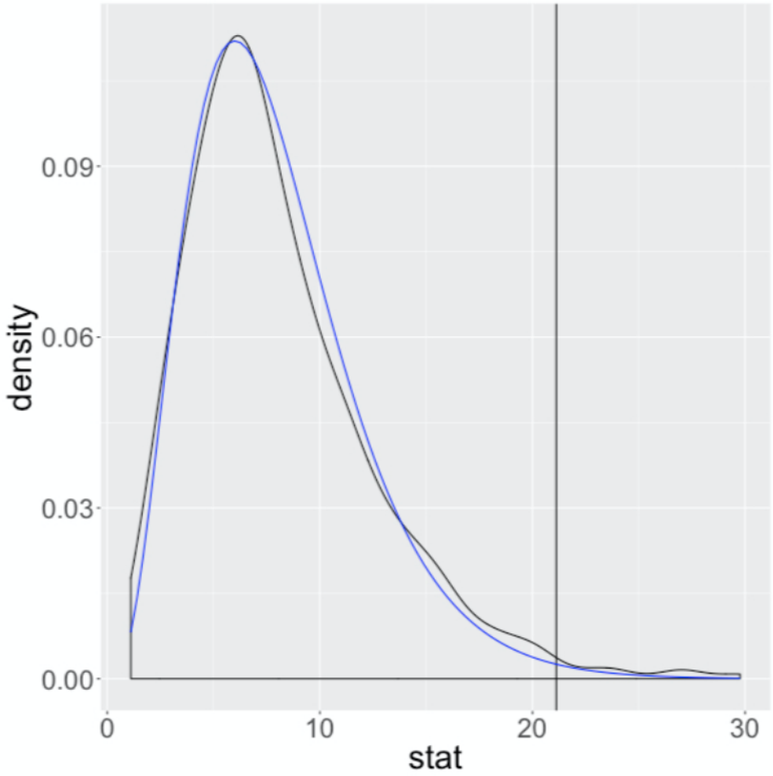
H_A: the election was fraudulent (Benford's Law does *not* hold)



Iran election fraud

H_0: the election was fair (Benford's Law holds)

H_A: the election was fraudulent (Benford's Law does *not* hold)



U.S.A. 2016 election

H_0 : the election was fair (Benford's Law holds)

H_A : the election was fraudulent (Benford's Law does not hold)

Iowa vote totals

[1] By TUBS [CC BY SA 3.0], from Wikimedia Commons

Iowa vote totals

```
iowa
```

```
# A tibble: 1,386 x 5
   office              candidate                          party              county votes
   <chr>               <chr>                              <chr>              <chr>  <dbl>
 1 President/Vice Pre… Evan McMullin / Nathan Johnson     Nominated by Peti… Adair     10
 2 President/Vice Pre… Under Votes                        NA                 Adair     32
 3 President/Vice Pre… Gary Johnson / Bill Weld           Libertarian        Adair    127
 4 President/Vice Pre… Over Votes                         NA                 Adair      5
 5 President/Vice Pre… Gloria La Riva / Dennis J. Banks   Socialism and Lib… Adair      0
 6 President/Vice Pre… Darrell L. Castle / Scott N. Bra… Constitution       Adair     10
 7 President/Vice Pre… Hillary Clinton / Tim Kaine        Democratic         Adair   1133
 8 President/Vice Pre… Jill Stein / Ajamu Baraka          Green              Adair     14
 9 President/Vice Pre… Rocky Roque De La Fuente / Micha… Nominated by Peti… Adair      3
10 President/Vice Pre… Donald Trump / Mike Pence          Republican         Adair   2461
# … with 1,376 more rows
```

Let's practice!

INFERENCE FOR CATEGORICAL DATA IN R

Election fraud in Iran and Iowa: debrief

INFERENCE FOR CATEGORICAL DATA IN R

Andrew Bray

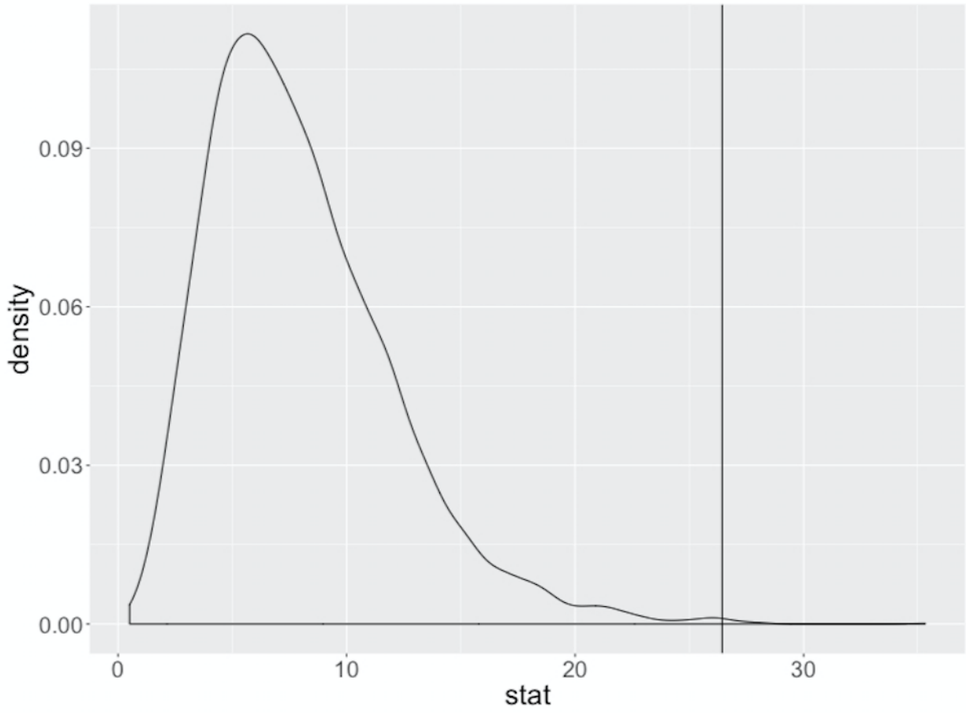
Assistant Professor of Statistics at Reed College

datacamp

Iowa election fraud

H_0: the election was fair (Benford's Law holds)

H_A: the election was fraudulent (Benford's Law does not hold)

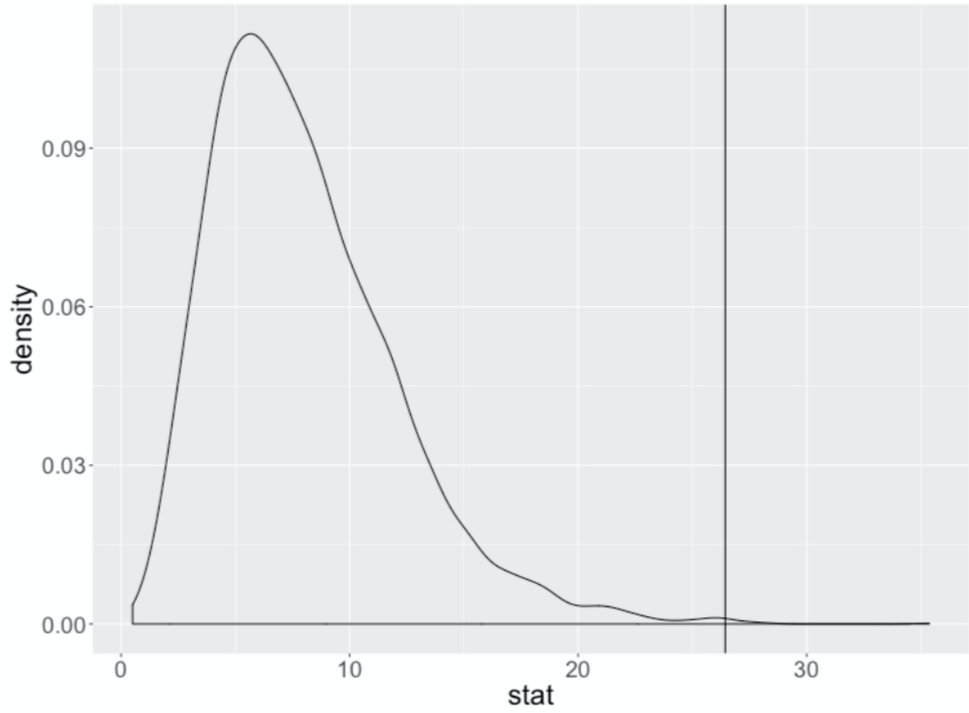


Iowa election fraud

Rejected

H_0: the election was fair (Benford's Law holds)

H_A: the election was fraudulent (Benford's Law does not hold)



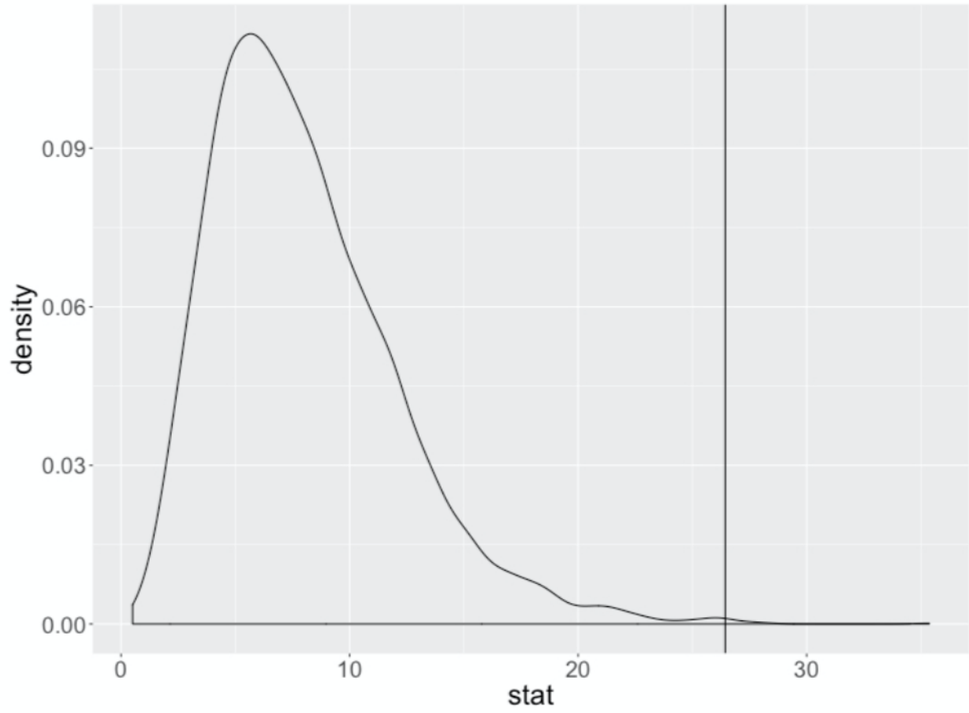
Iowa election fraud

Rejected

H_0: the election was fair (Benford's Law holds)

H_A: the election was fraudulent (Benford's Law does not hold)

What went wrong?



What went wrong?

Possibility 1: type I error

What went wrong?

Possibility 1: type I error

- Rejecting the null hypothesis when it is true, just due to chance

What went wrong?

Possibility 1: type I error

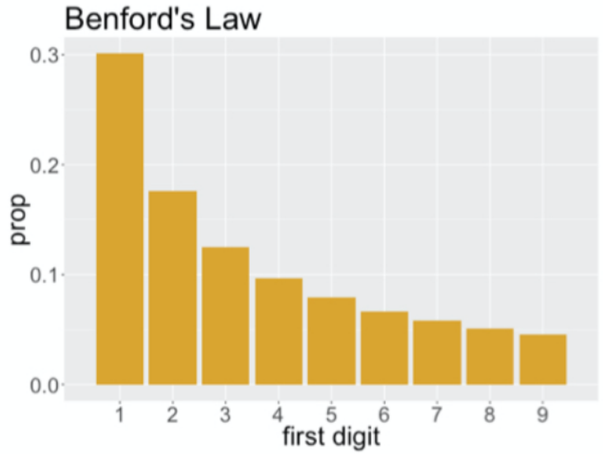
- Rejecting the null hypothesis when it is true, just due to chance

Possibility 2: Bedford's Law doesn't apply to vote totals

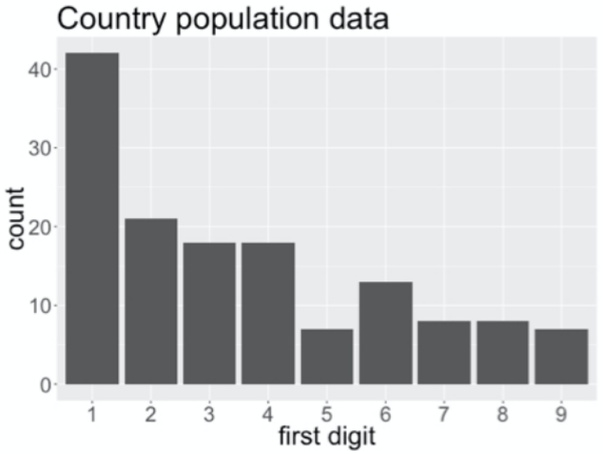
What went wrong?

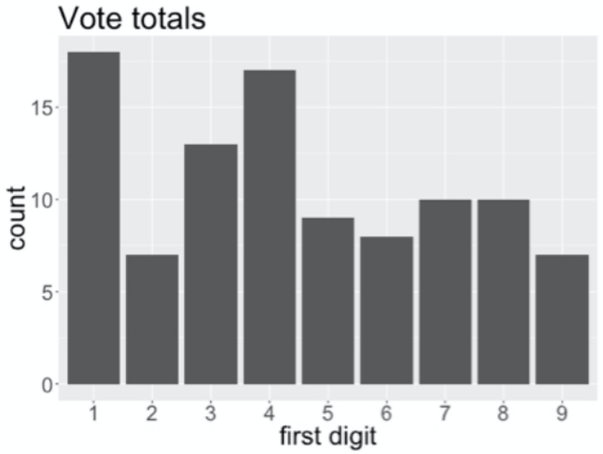
Possibility 1: type I error

- Rejecting the null hypothesis when it is true, just due to chance



Possibility 2: Bedford's Law doesn't apply to vote totals

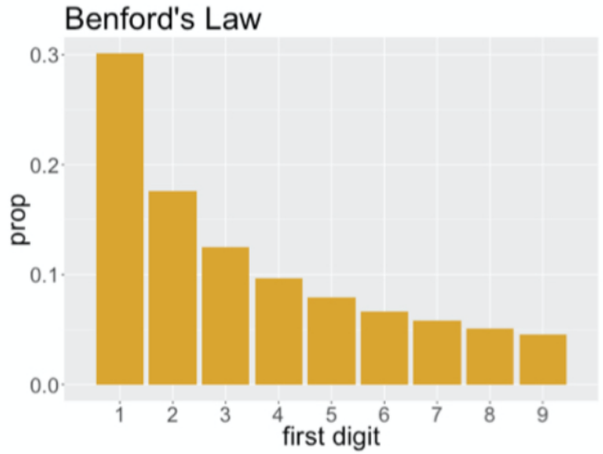




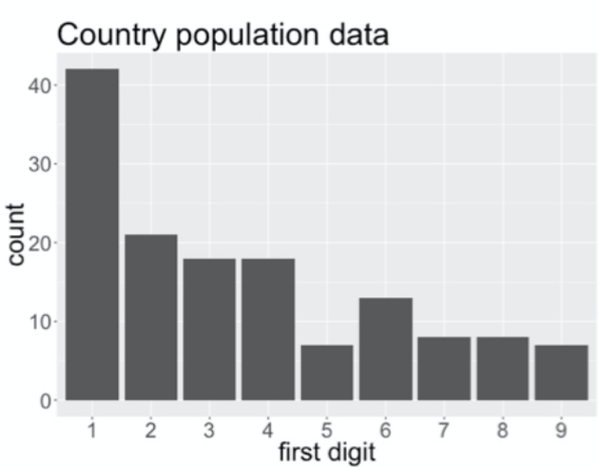
What went wrong?

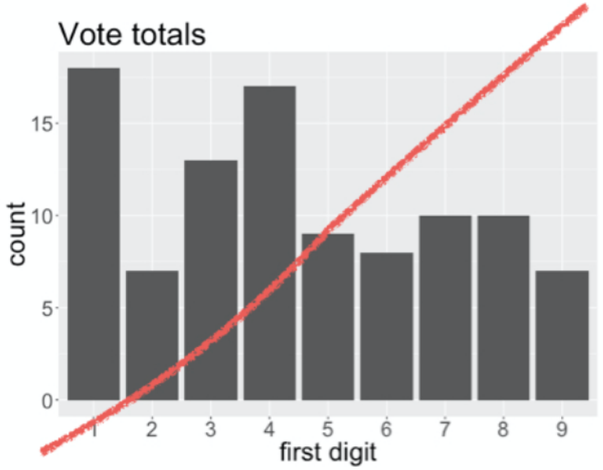
Possibility 1: type I error

- Rejecting the null hypothesis when it is true, just due to chance



Possibility 2: Bedford's Law doesn't apply to vote totals





Take-home lesson

The statistical tool must be appropriate for the task.

[1] By TUBS [CC BY-SA 3.0], from Wikimedia Commons [2] By P30Carl [GFDL] or [CC BY-SA 3.0], from Wikimedia Commons

Methods for categorical data

Confidence Intervals

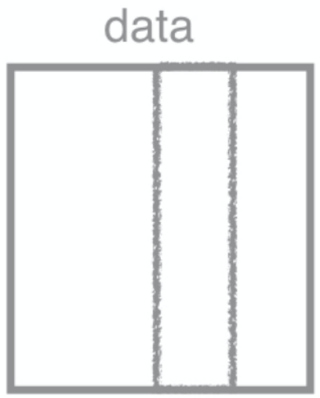
- One proportion
- Difference in proportions

Hypothesis tests

- One proportion
- Difference in proportions
- Test of independence
- Goodness of fit

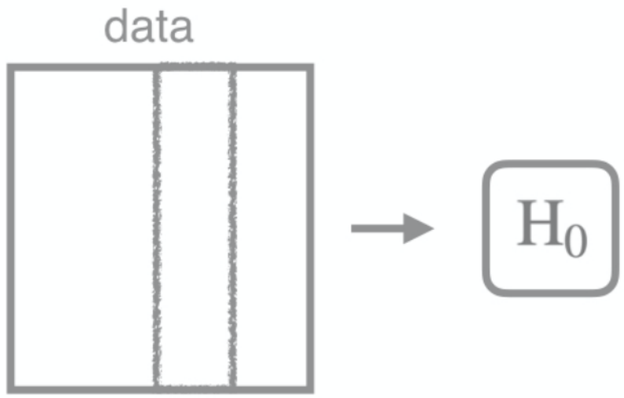
Hypothesis Test

Hypothesis Test



specify()

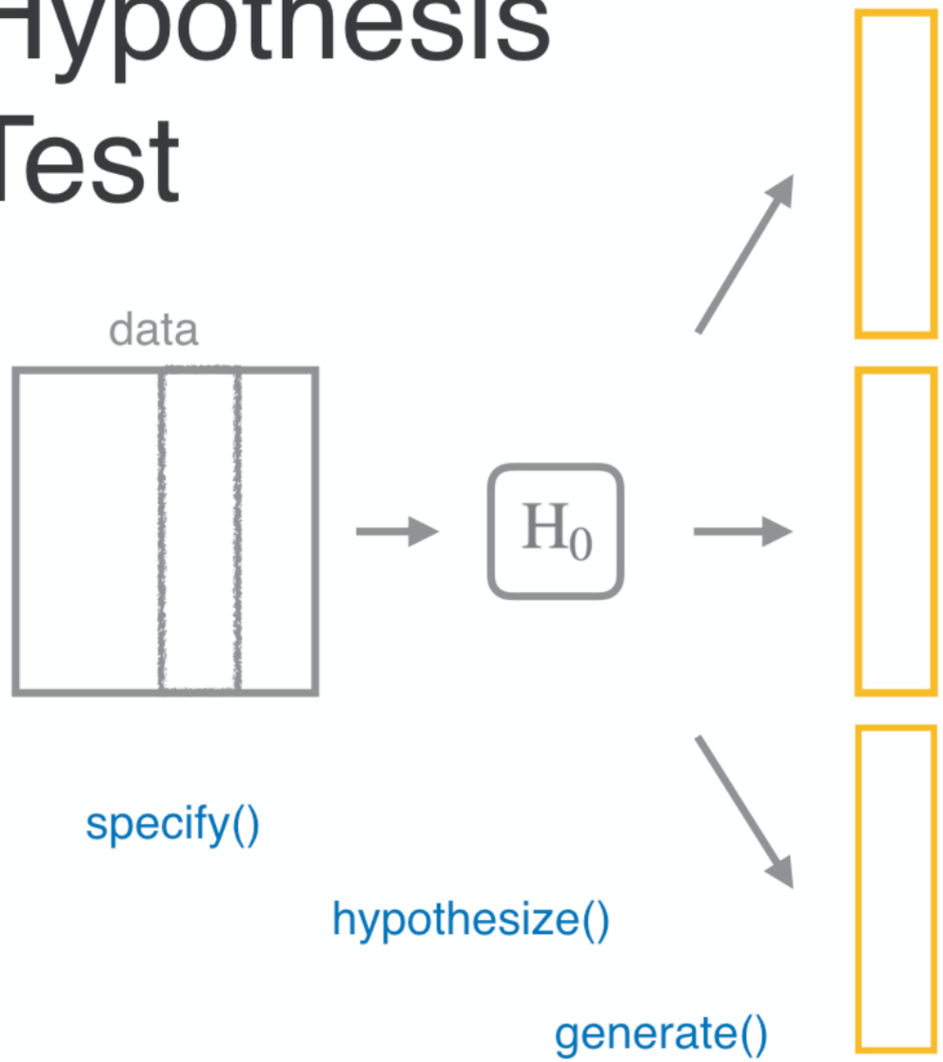
Hypothesis Test



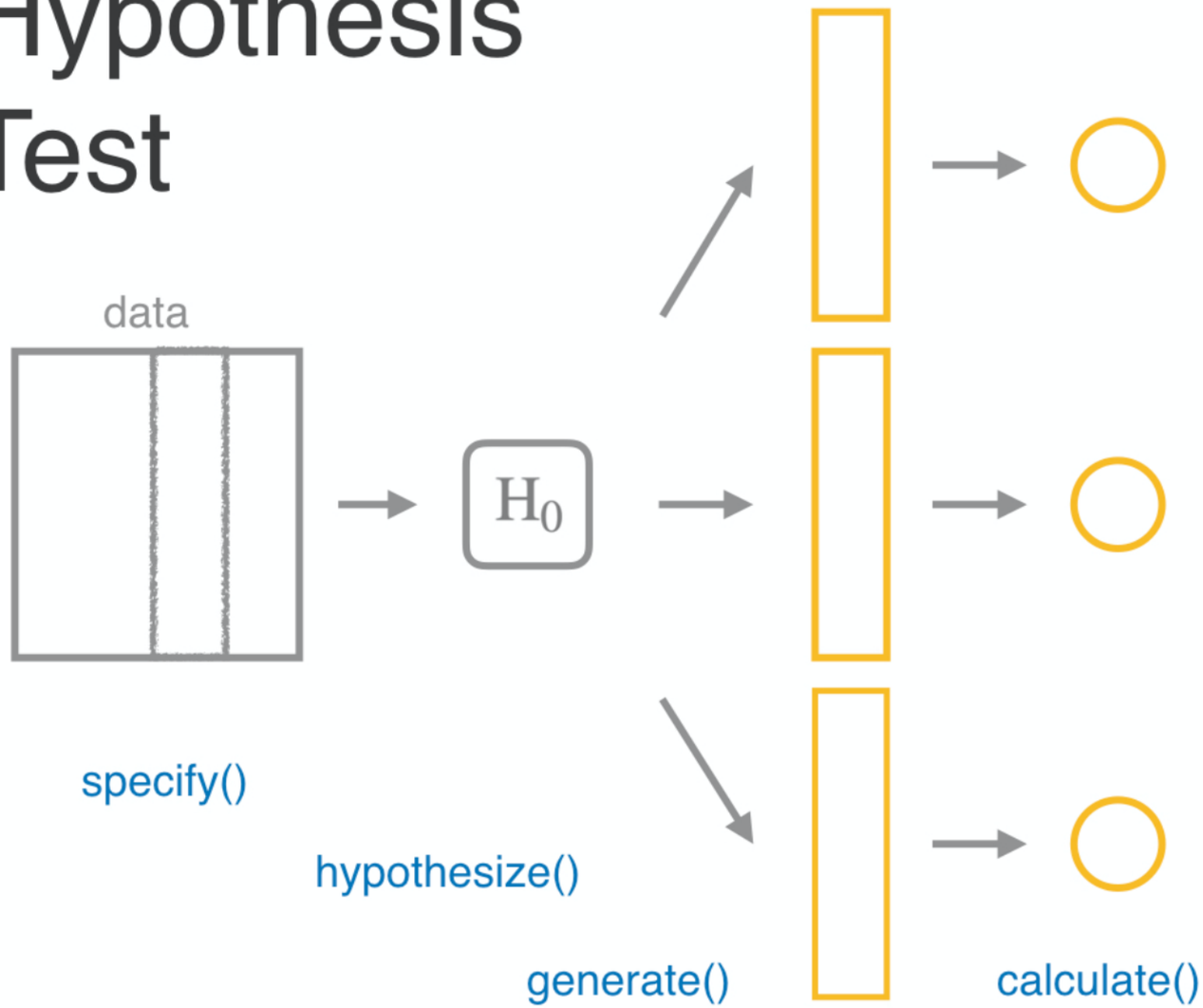
specify()

hypothesize()

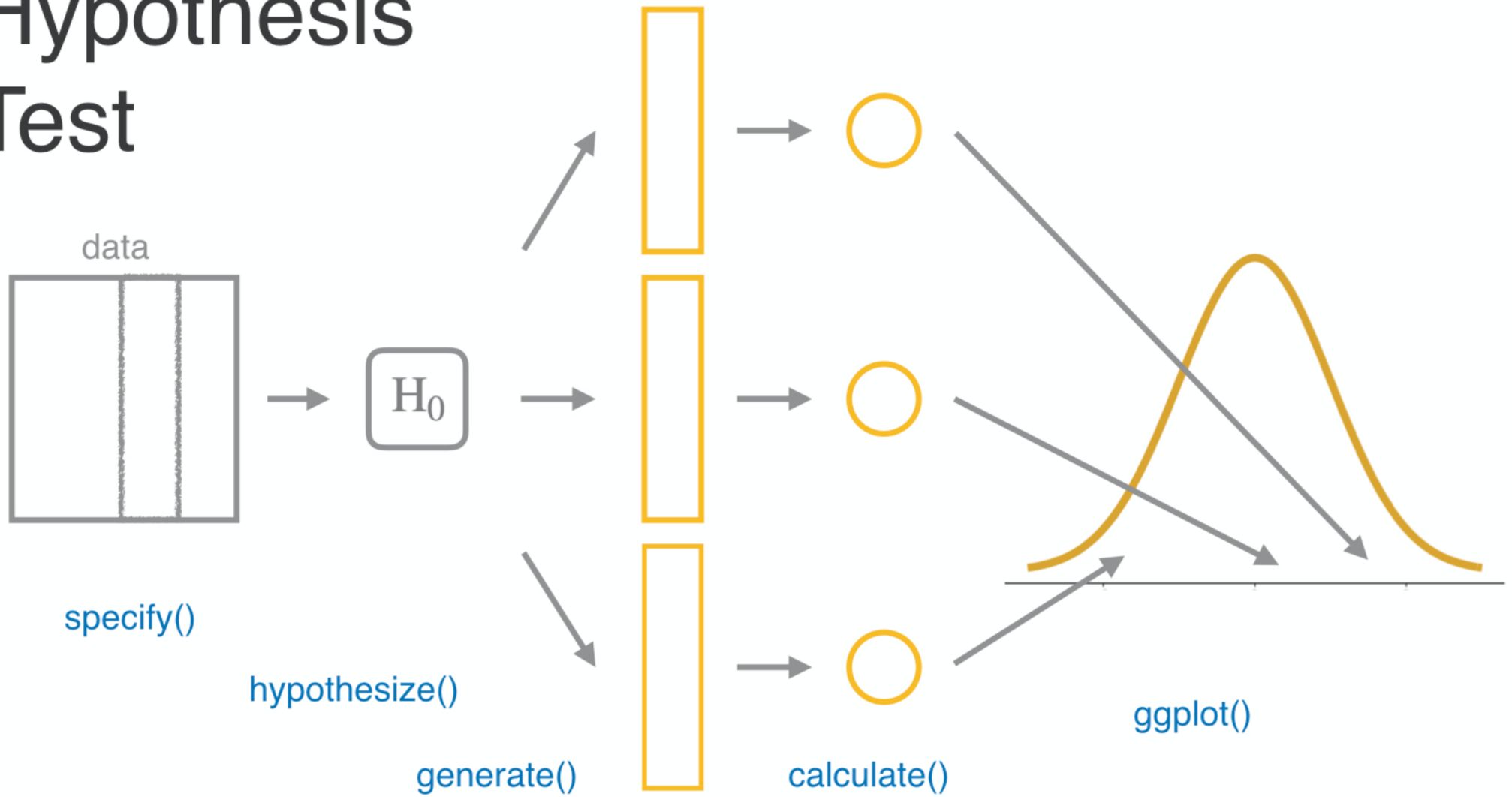
Hypothesis Test



Hypothesis Test



Hypothesis Test



Let's practice!

INFERENCE FOR CATEGORICAL DATA IN R