infer

An R package for tidy statistical inference

Dr. Chester Ismay
Data Science Curriculum Lead at DataCamp
GitHub: ismayc
Twitter: @old_man_chester

2018-01-27 (R User Day at Data Day Texas)

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Understanding who you are

- Who uses hypothesis testing/confidence intervals at least once a week?

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Understanding who you are

- Who uses hypothesis testing/confidence intervals at least once a week?
- Who uses the `tidyverse` at least once a week?

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Understanding who you are

- Who uses hypothesis testing/confidence intervals at least once a week?
- Who uses the `tidyverse` at least once a week?
- Who has heard of permutation testing?
    - Randomization-based methods?
    - Resampling methods?
    - Bootstrap methods?

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Pre-requisites for this talk

- Some experience with statistical inference (hypothesis testing / confidence intervals)

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Pre-requisites for this talk

- Some experience with statistical inference (hypothesis testing / confidence intervals)
- A ~~admiration~~, ~~abundance of love~~, ~~won't do anything without it~~ respect for the `tidyverse` and its power to get more users into doing data analysis/visualization quickly
    - The pit of success

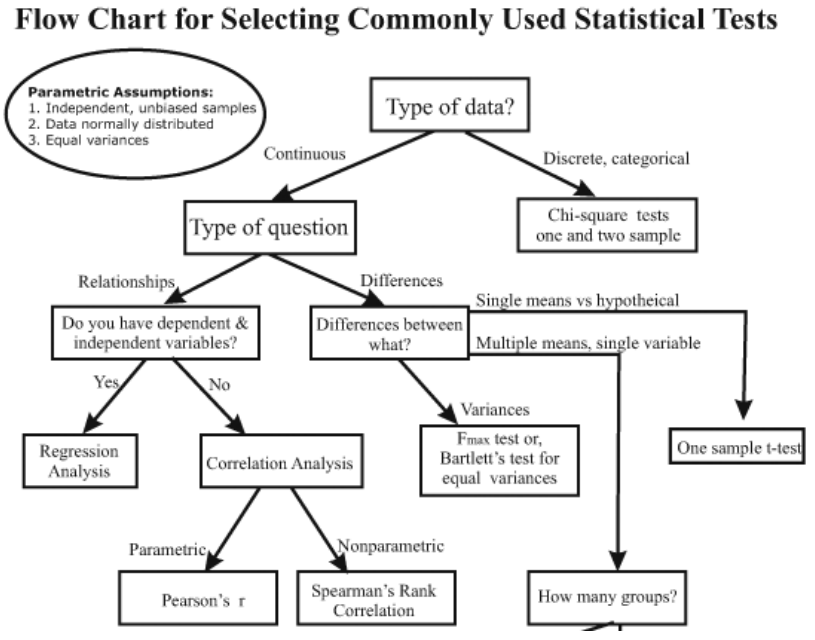
Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Pre-requisites for this talk

- Some experience with statistical inference (hypothesis testing / confidence intervals)
- A ~~admiration~~, ~~abundance of love~~, ~~won't do anything without it~~ respect for the `tidyverse` and its power to get more users into doing data analysis/visualization quickly
    - The pit of success
- ~~Ability~~ Desire to think differently about statistical inference using computational methods as the driver

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Is this statistical inference to you?



Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Students at Virginia Tech studied which vehicles come to a complete stop at an intersection with four-way stop signs, selecting at random the cars to observe. The explanatory variable used here is the arrival position of vehicles approaching an intersection all traveling in the same direction. They classified this arrival pattern into three groups: whether the vehicle arrives alone (`single`), is the `lead` in a group of vehicles, or is a `follow`er in a group of vehicles. Is there an association between arrival pattern and whether a `complete` stop or `not_complete` was made?

- From "Introduction to Statistical Investigations" by Tintle et al.

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Which type of hypothesis test should we conduct here?

- A. Independent samples t-test
- B. One proportion test
- C. Chi-Square test of independence
- D. ANOVA

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

```r
library(tidyverse)
# https://ismayc.github.io/talks/data-day-texas-infer/car_stop.rd
download.file("http://bit.ly/car_stop_rds",
              destfile = "car_stop.rds")
car_stop <- read_rds("car_stop.rds")
car_stop %>% sample_n(10)
```

```
# A tibble: 10 x 2
   stop_type    vehicle_type
   <chr>        <chr>
 1 complete     single
 2 complete     follow
 3 complete     lead
 4 complete     single
 5 not_complete follow
 6 not_complete lead
 7 complete     single
 8 complete     follow
 9 complete     single
10 complete     lead
```

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Which type of hypothesis test should we conduct here?

- A. Independent samples t-test
- B. One proportion test
- C. Chi-Square Test of Independence
- D. ANOVA

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Which type of hypothesis test should we conduct here?

- A. Independent samples t-test
- B. One proportion test
- C. Chi-Square Test of Independence
- D. ANOVA

Answer:

- **C. Chi-Square Test of Independence**

Let's do this in R

- Using a `data` argument

```
chisq.test(data = car_stop, x = stop_type, y = vehicle_type)
```

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Let's do this in R

- Using a `data` argument

```
chisq.test(data = car_stop, x = stop_type, y = vehicle_type)
```

```
Error in chisq.test(data = car_stop, x = stop_type,
  y = vehicle_type)
```

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Let's do this in R

- Using a `data` argument

```
chisq.test(data = car_stop, x = stop_type, y = vehicle_type)
```

```
Error in chisq.test(data = car_stop, x = stop_type,
  y = vehicle_type)
```

- Using a formula

```
chisq.test(data = car_stop,
           formula = vehicle_type ~ stop_type)
```

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Let's do this in R

- Using a `data` argument

```
chisq.test(data = car_stop, x = stop_type, y = vehicle_type)
```

```
Error in chisq.test(data = car_stop, x = stop_type,
  y = vehicle_type)
```

- Using a formula

```
chisq.test(data = car_stop,
           formula = vehicle_type ~ stop_type)
```

```
Error in chisq.test(data = car_stop,
  formula = vehicle_type ~ stop_type)
```

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Finally

```
chisq.test(car_stop$stop_type, car_stop$vehicle_type)
```

```
    Pearson's Chi-squared test

data:  car_stop$stop_type and car_stop$vehicle_type
X-squared = 3.9476, df = 2, p-value = 0.1389
```

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

```
?chisq.test()
```

RDocumentation | Search for packages, functions, etc | R Enterprise Training | R package | Lead

chisq.test

From stats v3.4.3
by R-core R-core@R-project.org

Pearson's Chi-Squared Test For Count Data

`chisq.test` performs chi-squared contingency table tests and goodness-of-fit tests.

Keywords distribution, htest

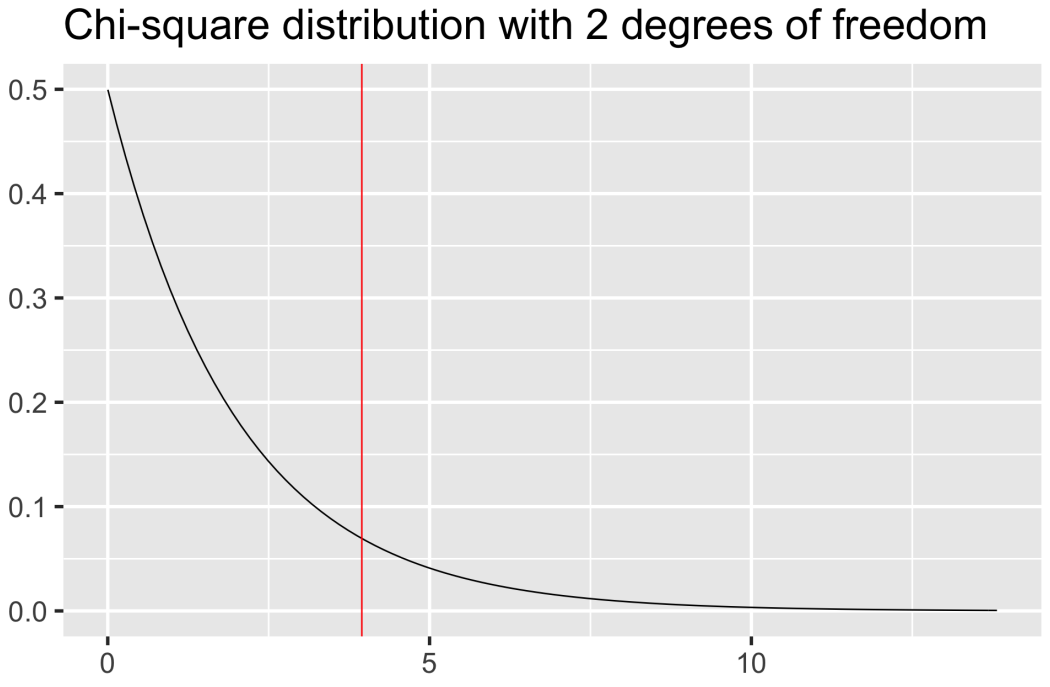
Usage

```
chisq.test(x, y = NULL, correct = TRUE,
           p = rep(1/length(x), length(x)), rescale.p = FALSE,
           simulate.p.value = FALSE, B = 2000)
```

Arguments

x	a numeric vector or matrix. `x` and `y` can also both be factors.
y	a numeric vector; ignored if `x` is a matrix. If `x` is a factor, `y` should be a factor of the same length.

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com



P-value is 0.1389.

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Is there an association between arrival pattern and whether or not a complete stop was made?

The null hypothesis

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Is there an association between arrival pattern and whether or not a complete stop was made?

The null hypothesis

> No association exists between the arrival vehicle's position and whether or not it makes a complete stop.

The alternative hypothesis

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Is there an association between arrival pattern and whether or not a complete stop was made?

The null hypothesis

> No association exists between the arrival vehicle's position and whether or not it makes a complete stop.

The alternative hypothesis

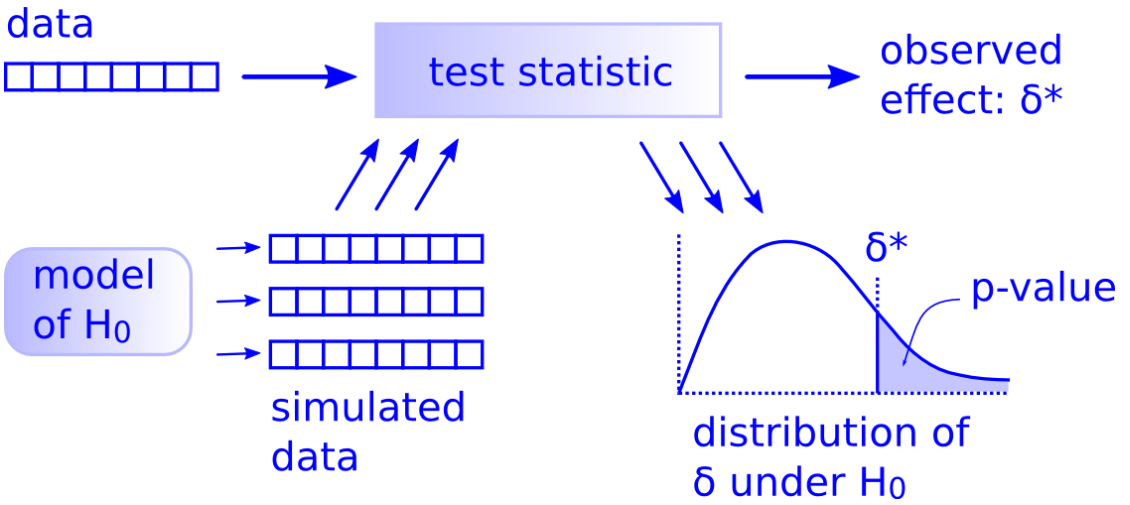
> An association exists between the arrival vehicle's position and whether or not it makes a complete stop.

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

How can computation help us to understand what is going on here?

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

How can computation help us to understand what is going on here?



Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

The tricky step

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

The tricky step

Modeling the null hypothesis

- How do we simulate data assuming the null hypothesis is true in our problem (there is no association between the variables)?

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

The tricky step

Modeling the null hypothesis

- How do we simulate data assuming the null hypothesis is true in our problem (there is no association between the variables)?
- What might the sample data look like if the null was true?

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Properties of the original sample collected

```
car_stop %>% count(stop_type, vehicle_type)
```

```
# A tibble: 6 x 3
  stop_type    vehicle_type     n
  <chr>        <chr>        <int>
1 complete     follow          76
2 complete     lead            38
3 complete     single         151
4 not_complete follow          22
5 not_complete lead             5
6 not_complete single          25
```

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Properties of the original sample collected

```
car_stop %>% count(stop_type, vehicle_type)
```

```
# A tibble: 6 x 3
  stop_type    vehicle_type     n
  <chr>        <chr>        <int>
1 complete     follow          76
2 complete     lead            38
3 complete     single         151
4 not_complete follow          22
5 not_complete lead             5
6 not_complete single          25
```

```
( orig_table <- car_stop %>%
    janitor::tabyl(stop_type, vehicle_type) )
```

```
    stop_type follow lead single
     complete     76   38    151
 not_complete     22    5     25
```

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Permute the sample data

```
# A tibble: 317 x 2
   stop_type    vehicle_type
   <fct>        <fct>
 1 complete     follow
 2 not_complete follow
 3 complete     follow
 4 not_complete single
 5 complete     single
 6 not_complete follow
 7 not_complete single
 8 complete     follow
 9 not_complete lead
10 complete     lead
# ... with 307 more rows
```

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Permute the sample data

```
# A tibble: 317 x 2
   stop_type    vehicle_type
   <fct>        <fct>
 1 complete     follow
 2 not_complete follow
 3 complete     follow
 4 not_complete single
 5 complete     single
 6 not_complete follow
 7 not_complete single
 8 complete     follow
 9 not_complete lead
10 complete     lead
# ... with 307 more rows
```

```
    stop_type follow lead single
     complete     79   35    151
 not_complete     19    8     25
```

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Comparing the original and permuted sample

```
orig_table %>% janitor::adorn_totals(where = c("row", "col"))
```

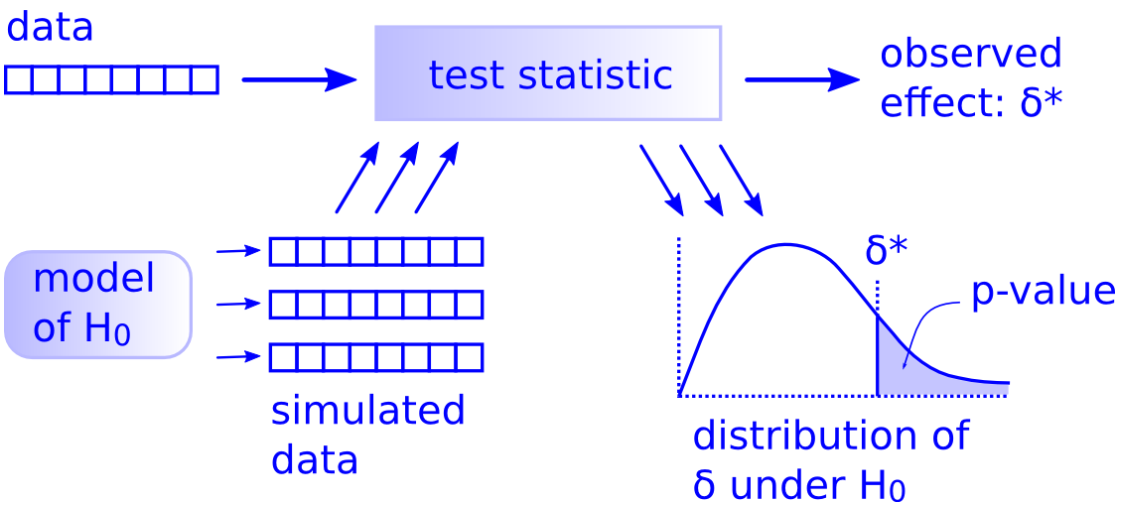
```
    stop_type follow lead single Total
     complete     76   38    151   265
not_complete     22    5     25    52
        Total     98   43    176   317
```

```
new_table %>% janitor::adorn_totals(where = c("row", "col"))
```

```
    stop_type follow lead single Total
     complete     79   35    151   265
not_complete     19    8     25    52
        Total     98   43    176   317
```

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Where are we?



Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Test statistic

- Chi-square test statistic (Wikipedia)
    - Measure of how far what we observed in our sample is from what we would expect if the null hypothesis was true

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Test statistic

- Chi-square test statistic (Wikipedia)
  - Measure of how far what we observed in our sample is from what we would expect if the null hypothesis was true

```
chisq.test(car_stop$stop_type, car_stop$vehicle_type)$statistic
```

```
X-squared
 3.947648
```

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

For the permuted data

```
chisq.test(perm1$stop_type, perm1$vehicle_type)$statistic
```

```
X-squared
 1.408986
```

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

For the permuted data

```
chisq.test(perm1$stop_type, perm1$vehicle_type)$statistic
```

```
X-squared
 1.408986
```

Another permutation

```
chisq.test(perm2$stop_type, perm2$vehicle_type)$statistic
```

```
X-squared
0.3604528
```

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

What does the distribution of multiple repetitions of the permuted data look like?

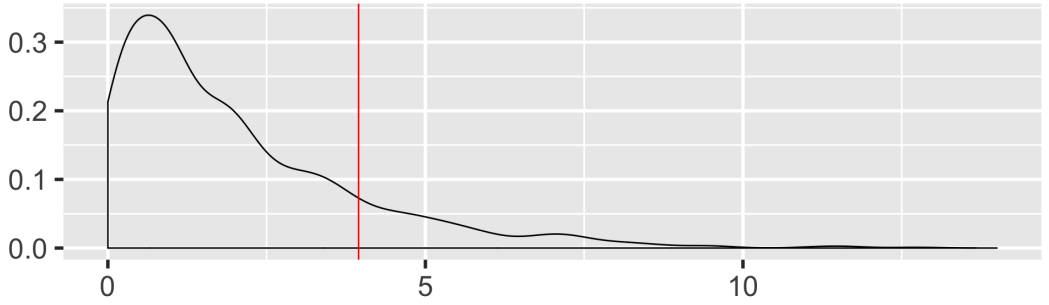
```
# A tibble: 1,000 x 2
   replicate  stat
   <fct>     <dbl>
 1 1         1.05
 2 2         7.24
 3 3         0.253
 4 4         2.16
 5 5         1.60
 6 6         3.95
 7 7         1.94
 8 8         1.68
 9 9         0.242
10 10        3.26
# ... with 990 more rows
```

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

The distribution of multiple repetitions of the permuted data

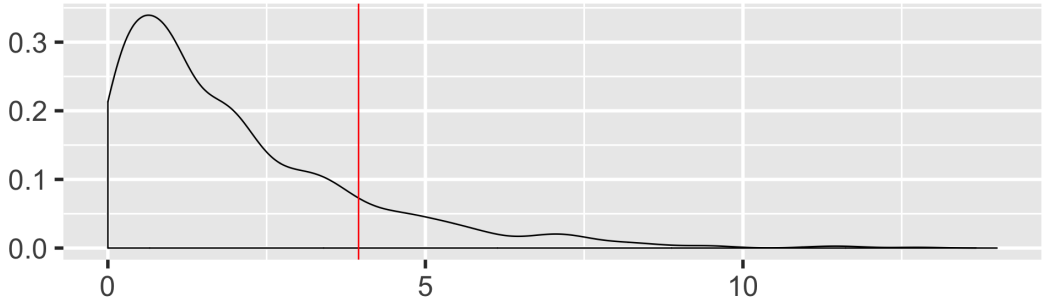
Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

The distribution of multiple repetitions of the permuted data

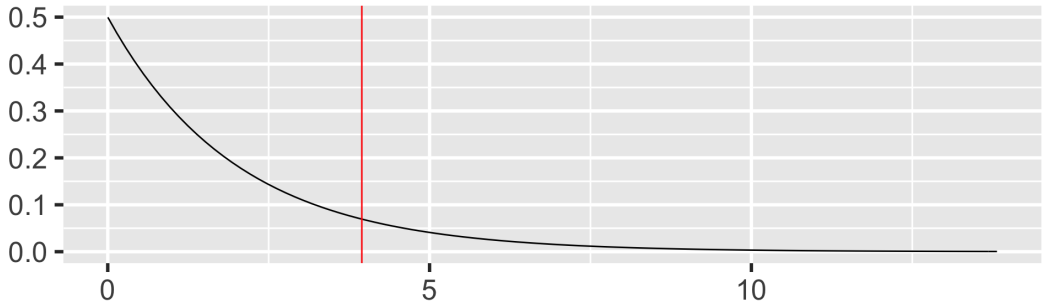


Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

The distribution of multiple repetitions of the permuted data



Recall the traditional method using the Chi-square distribution



Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Objectives of `infer`

- Implement common classical inferential techniques in a `tidyverse`-friendly framework that is expressive of the underlying procedure.

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Objectives of `infer`

- Implement common classical inferential techniques in a `tidyverse`-friendly framework that is expressive of the underlying procedure.
    - Dataframe in, dataframe out
    - Compose tests and intervals with pipes
    - Unite computational and approximation methods
    - Reading a chain of `infer` code should describe the inferential procedure

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

The `infer` verbs

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Hypothesis test

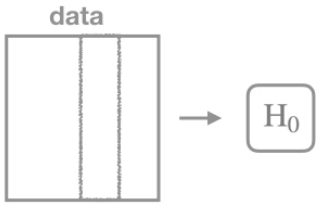
Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Hypothesis test


```
specify()
```

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

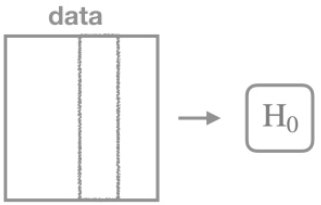
Hypothesis test



`specify()`

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

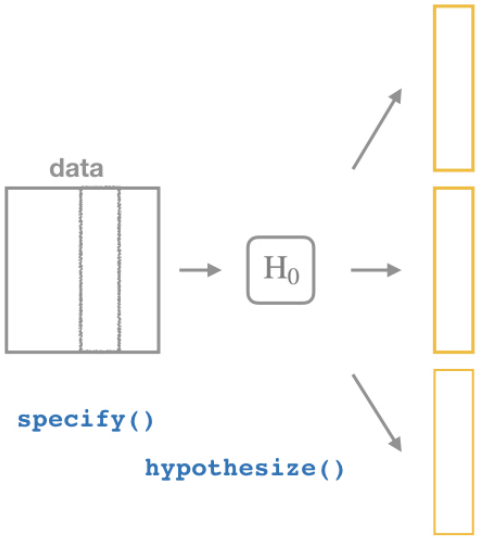
Hypothesis test




```
specify()
        hypothesize()
```

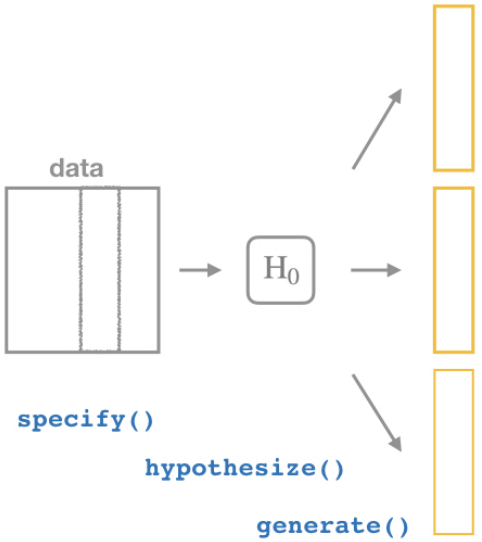
Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Hypothesis test



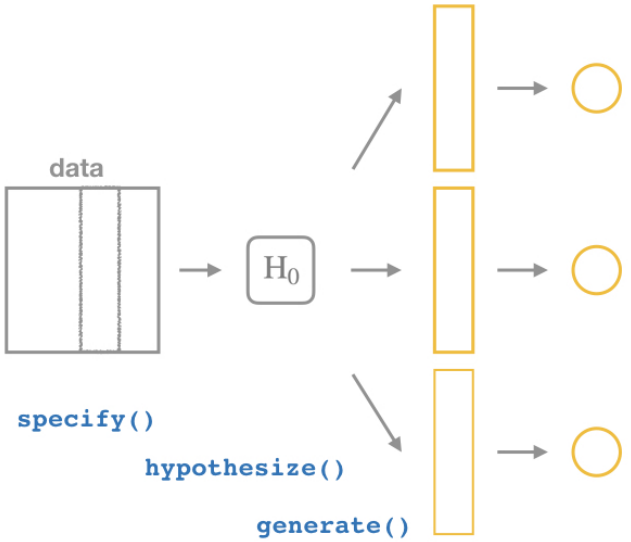
Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Hypothesis test



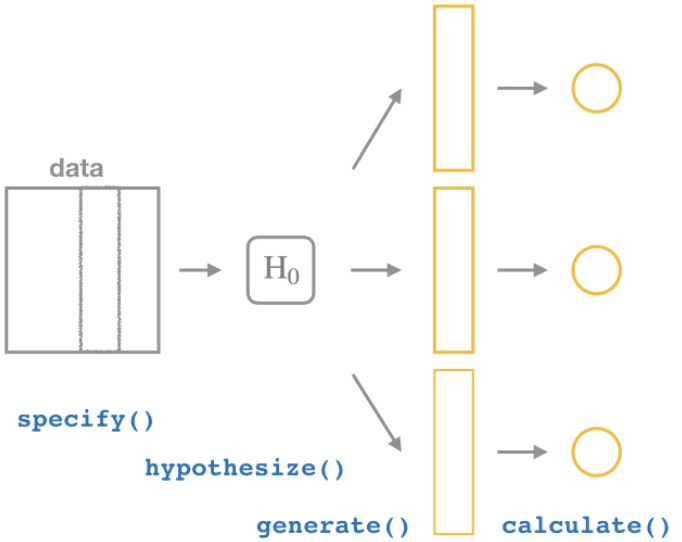
Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Hypothesis test



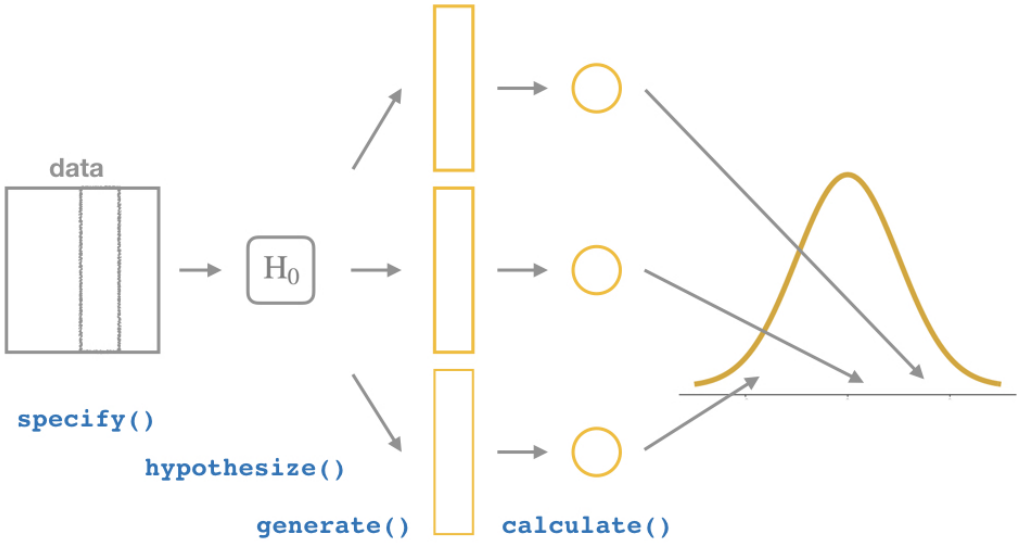
Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Hypothesis test



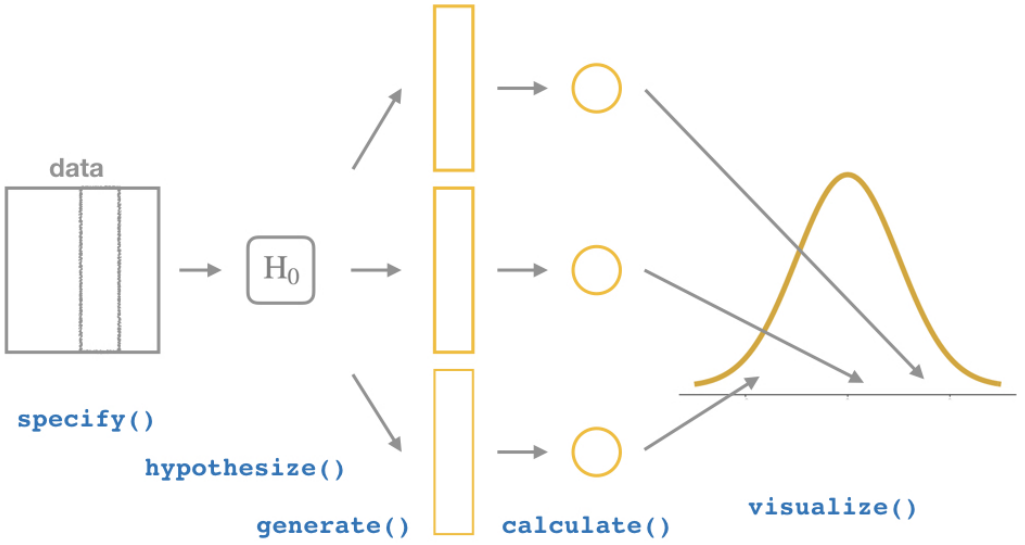
Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Hypothesis test



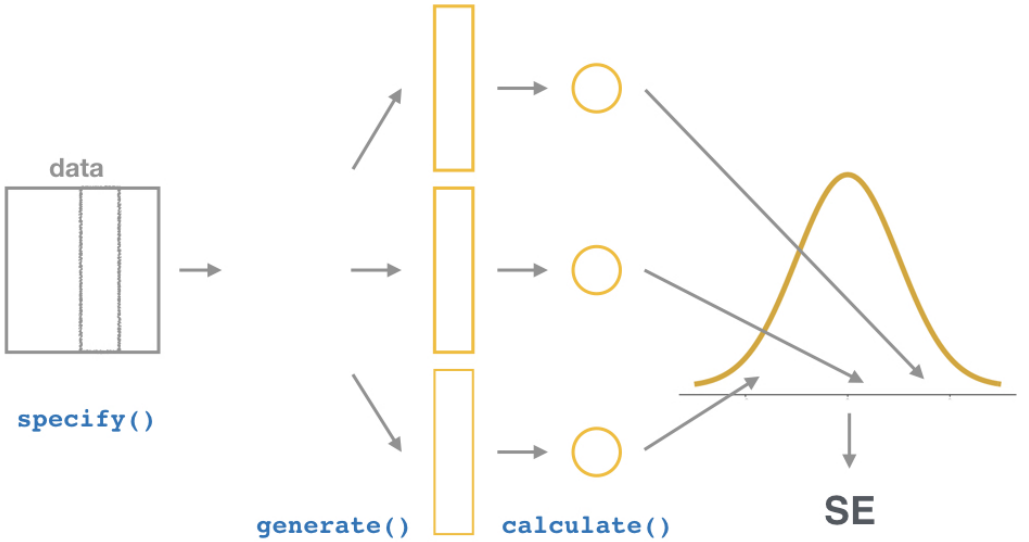
Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Hypothesis test



Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Confidence Interval



Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

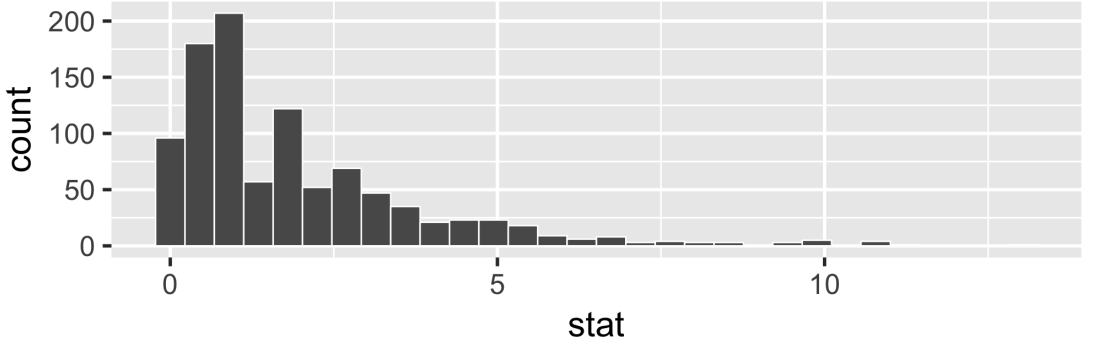
```
car_stop %>%
  specify(stop_type ~ vehicle_type) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "Chisq")
```

```
# A tibble: 1,000 x 2
   replicate  stat
   <fct>     <dbl>
 1 1         0.509
 2 2         0.760
 3 3         1.79
 4 4         1.83
 5 5         0.525
 6 6         4.74
 7 7         1.11
 8 8         2.51
 9 9         1.11
10 10        0.421
# ... with 990 more rows
```

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Back to the example

```
car_stop %>%
  specify(stop_type ~ vehicle_type) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "Chisq") %>%
  visualize()
```



Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

What's to come

- Wrapper functions: `t_test`, `chisq_test`, etc.
- Generalized input to `calculate()`
  - For example, `calculate(trimmed_mean)`
  - Support for more advanced regression models
- Adding features to `visualize()`
  - Show both traditional and computation methods
- Implement list-columns in the `generate()` step

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

Tips and tricks for package development

- Use GitHub and pull requests to the `master` branch
- Create useful vignettes so others know how your pkg works
- Write tests and assertions for your code
    - Buy and read Richie's *Testing R Code* book
- Let travis-ci do the work for you
- Use Hadley's `pkgdown` package to build a pkg website
    - Host it on Netlify.com to be super cool

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

More info

- https://infer.netlify.com
  - Many examples under Articles there with more to come
  - Plans to be implemented in www.ModernDive.com by this summer
    - Sign up for the ModernDive mailing list for details
- Two DataCamp courses currently launched that use `infer`
  - Inference for Numerical Data by Mine Cetinkaya-Rundel
  - Inference for Regression by Jo Hardin
- Two more DataCamp courses to be launched

Slides available at http://bit.ly/infer-austin
Package webpage at https://infer.netlify.com

- Special thanks to Andrew Bray and the other pkg contributors
- Slides created via the R package xaringan by Yihui Xie
- Slides available at http://bit.ly/infer-austin
- Source code for these slides at https://github.com/ismayc/talks/tree/master/data-day-texas-infer