

Bringing Statistics Education up to Data

Daniel Kaplan

11/15/2019

What's in your intro stats course?

1. Inference?

- What are the settings?
 - “two sample” means and proportions?
 - slope of regression line?
 - ANOVA?
- What is the basic method?
 - Formulas and tables
 - Simulation, resampling, bootstrapping

2. Graphics?

- one variable: histogram, box-and-whisker, ...
- two variables: scatterplot, box-and-whisker, ...

3. Experimental design?

4. Sampling bias?

5. Causation?

6. Confounding and covariates?

StatPREP.org

Faculty and teaching materials to teach statistics in a data-centric way.

Ideally ...

- make every graphic centered on data
- highlight a modern, integrated perspective:
 - not means, instead models
 - multi-variable
- feature accepted contemporary practices
 - large data ($n > 50$, $p > 3$)
 - making responsible causal claims with imperfect data
 - prediction

Realistically ...

- help instructors teach their present topics (e.g. t-tests) but with data.
 - make it easier for instructors to use real data with real questions.
 - help instructors embrace modern computing
-

Three levels of computing

1. Little Apps. Student facing; no programming required.
2. Instructor tutorials. Highly scaffolded computing. Intended mainly for instructors. (But can be used with students.)
3. Writing code starting with a blank page. What most of us think of when we hear “statistical computing.”

All three can be done with browser-based software. (1) can potentially be done with a browser on a smart phone.

This talk will be about (1). But, short digression about ...

Instructor tutorials

- Introduce R commands
- Use a highly consistent command syntax *via* `mosaic` and `ggformula` packages. Examples:

```
gf_point(height ~ sex, data = Galton)
gf_boxplot(height ~ sex, data = Galton)
lm(height ~ sex, data = Galton)
median(height ~ sex, data = Galton)
df_stats(height ~ sex, data = Galton, coverage(0.8))
```

List of currently available StatPREP Instructor Tutorials.

Example

Calculating basic stats

Calculating Statistics **df_stats**

The MOSAIC `df_stats()` function calculates numerical summaries of a variable. It follows the usual MOSAIC template

```
goal(formula, data = data_frame, additional_specifics)
```

The "additional_specifics" list the particular statistics you want calculated.

For example, here is the `df_stats()` command to find the mean height of the adult children in Galton's 1880 collection of height measurements.

Code

```
1 df_stats(~ height, data = Galton, mean)
2
3
```

Note: If you want more information about the `Galton` data frame, you can give the command `help(Galton)` in the command block.

Using a two-sided formula causes groupwise statistics to be calculated. Try this:

Code

```
1 df_stats( height ~ sex, data = Galton, mean)
2
3
```

Structure of a Little App

Do I just need something here?

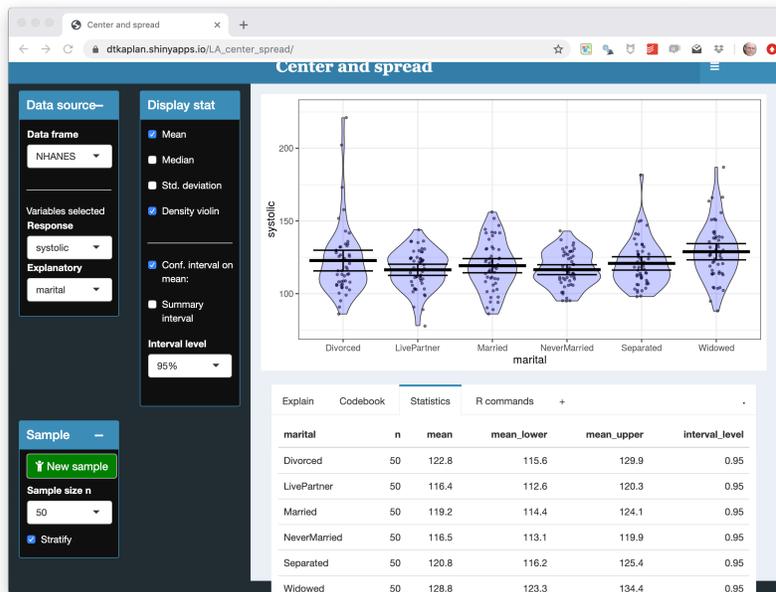
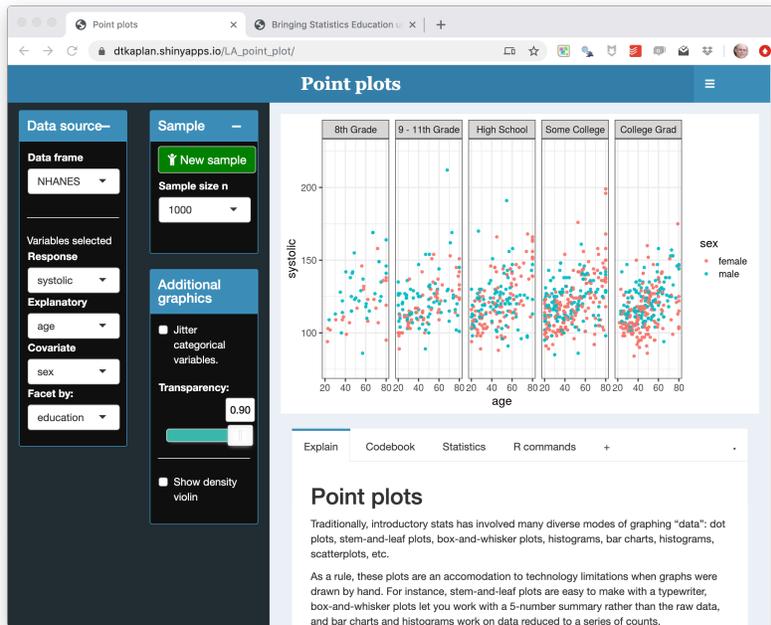
1. Center display is data
 - a. both axes are variables
 - b. each unit of observation is shown
2. Data. Currently, several "large" data sets, e.g.
 - a. National Health and Nutrition Evaluation Survey (NHANES)
 - 10,000 rows
 - 76 variables, including biological, lifestyle, economic indicators (height, age, urine flow, alcohol consumption, smoking, depression, education, income, ...)
 - b. Births in the US
 - 100,000 births
 - 45 variables including baby weight, mother age, gestation length, APGAR, labor induced, ...
3. Operation.
 - Student picks a response variable and an explanatory variable.
 - Optionally picks a covariate
 - Student picks sample size
 - New sample at the push of a button

4. Statistical annotations. Statistics are shown as annotations on the data.
 - point statistics and regression curves
 - confidence intervals and bands
 - densities
 5. Inference:
 - Can see variation when generating a new sample
 - Some apps involve bootstrapping directly
 - Student can vary sample size and see consequences directly.
 6. Statistical tabulation (in a tab)
-

A suggested pedagogy

1. Introduce data: numerical and categorical variables, response and explanatory variables, scatter plot, jittering, sample size. (No stats yet.)
 2. Look for patterns and give simple description and interpretation. (No inference yet.)
 3. Introduce covariates. Does the inclusion of a covariate change the observed pattern? Does it change the story? (This could be later.)
 4. Introduce formal ways to describe patterns: means, proportions, regression models, ... (No inference yet.)
 5. Using small data, raise question of whether the pattern is really there. Do this by observing sampling variation directly. (Press “New Sample”.)
 6. Change sample size and observe how pattern variation changes.
 7. Introduce formal measures of sampling variation in observed patterns.
 8. Introduce formal mechanism for inferring sampling variation from a single, fixed sample.
-

Examples: Plain graphics



Contrasting apps

Dan Adrians Happy Apps

Confidence interval for the population mean μ

Formula: $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$

Click for more info!

Sample mean \bar{x}

10 18 26 34 42 50 58 66 74 82 90

47

Sample standard deviation s

1 5.9 10.8 15.7 20.6 25.5 30.4 35.3 40.2 45.1 50

9

Sample size n

2 12 22 32 42 52 62 72 82 92 100

10

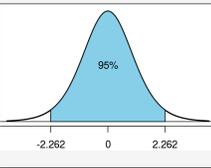
Confidence level

80% 85% 90% 95% 99%

95%

Parts of the result: t^* , se , me

$t^* = 2.262$



-2.262 0 2.262

Standard error

$se = \frac{s}{\sqrt{n}} = 2.846$

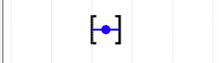
Margin of error

$me = t^*(se) = 6.44$

95% Confidence interval for μ

(40.56, 53.44)

Picture



Range shown

0 10 20 30 40 50 60 70 80 90 100

0 100

Two Sample Apps

Two Sample Apps Simulate 2 samples Two-sample t test Confidence interval for difference in means

The two-sample t test statistic

Investigating how it depends on the summary statistics changed with the sliders in the left panel.

Summary statistics

For simplicity, we will assume the standard deviations and sample sizes of both groups are equal, i.e. $s_1 = s_2$ and $n_1 = n_2$.

Sample 1 mean (\bar{x}_1)

0 1 2 3 4 5 6 7 8 9 10

6

Sample 2 mean (\bar{x}_2)

0 1 2 3 4 5 6 7 8 9 10

9

Sample std devs ($s_1 = s_2$)

1 1.8 2.8 3.7 4.6 5.5 6.4 7.3 8.2 9.1 10

4

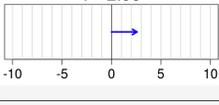
Sample sizes ($n_1 = n_2$)

5 10 20 30 40 50 60 70 80 90 100 200

30

Test statistic

$t = 2.58$



-10 -5 0 5 10

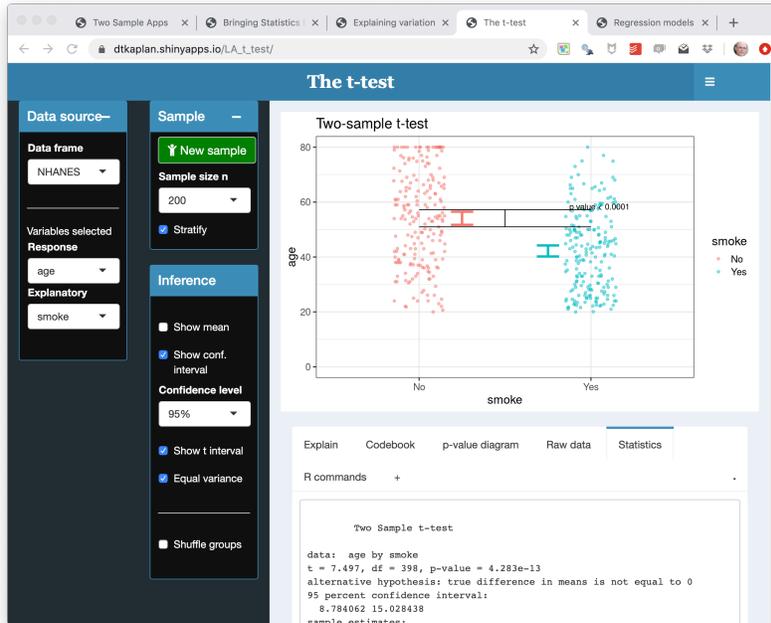
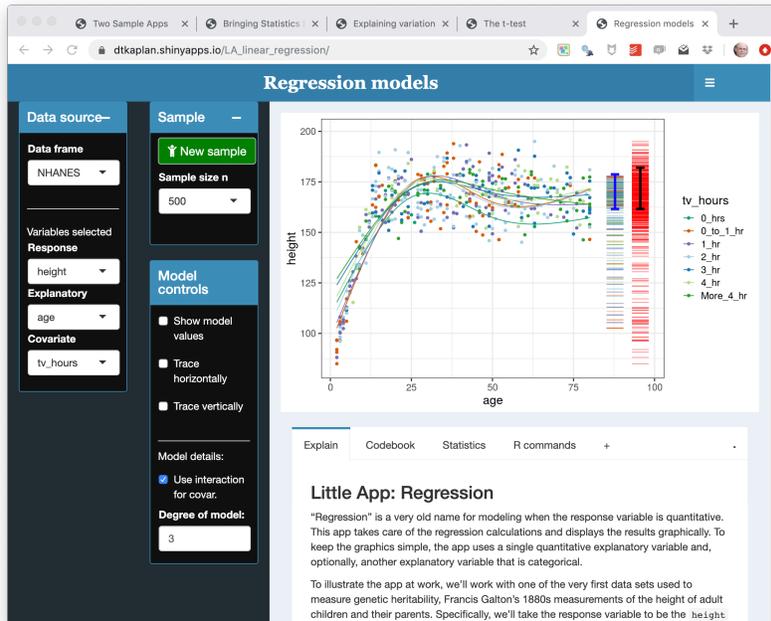
Test statistic formula

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Analysis

Two parts of formula:

- Numerator:** difference in means
- 2
- Which mean is greater determines sign of the test statistic
- Distance between means determines size of test statistic
- Denominator:** standard error
- 0.775
- Inverse relationship with test statistic
- Larger standard deviations -> larger standard error
- Larger sample sizes -> smaller standard error

Example: t-test*Example: Explaining variation*

Summary and discussion

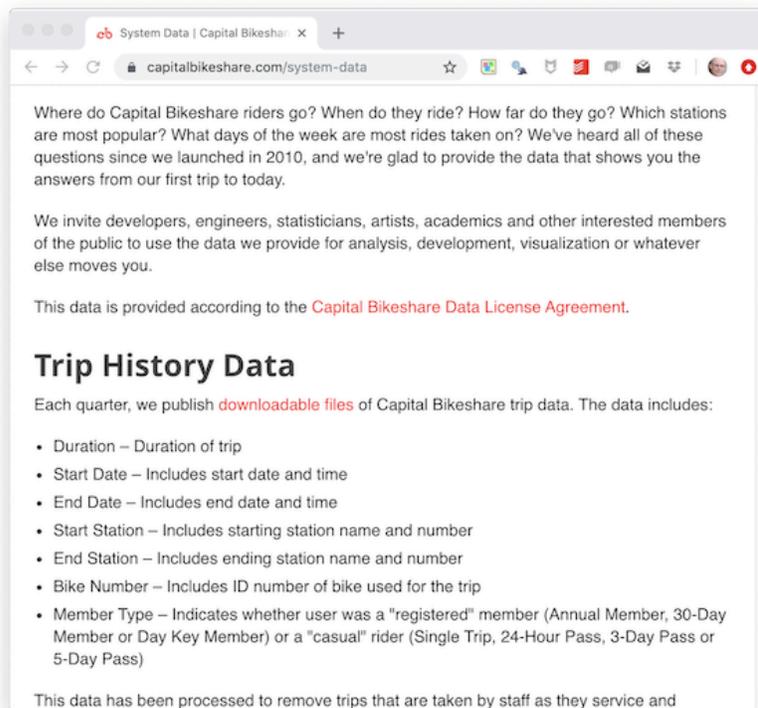
Data can be at the center of your statistics course.

Computing and an associated pedagogy are available.

extra stuff ...

Finding your own data ...

Capital Bikeshare

A screenshot of a web browser displaying the 'System Data | Capital Bikeshare' page. The browser's address bar shows 'capitalbikeshare.com/system-data'. The page content includes an introductory paragraph about the data, an invitation for developers and others to use the data, a link to the 'Capital Bikeshare Data License Agreement', a section titled 'Trip History Data' with a list of data fields, and a note at the bottom stating that staff trips have been removed from the data.

Where do Capital Bikeshare riders go? When do they ride? How far do they go? Which stations are most popular? What days of the week are most rides taken on? We've heard all of these questions since we launched in 2010, and we're glad to provide the data that shows you the answers from our first trip to today.

We invite developers, engineers, statisticians, artists, academics and other interested members of the public to use the data we provide for analysis, development, visualization or whatever else moves you.

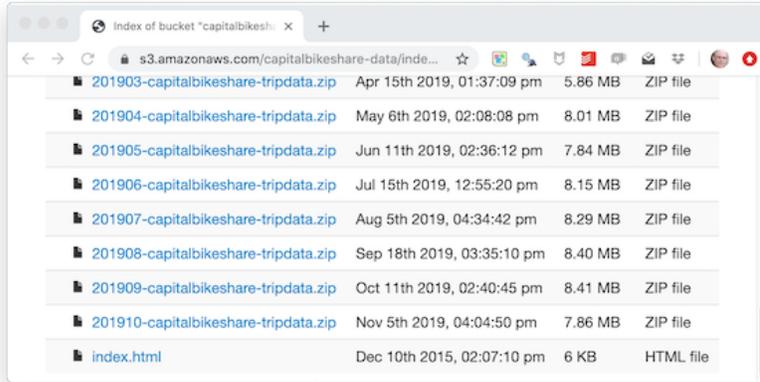
This data is provided according to the [Capital Bikeshare Data License Agreement](#).

Trip History Data

Each quarter, we publish [downloadable files](#) of Capital Bikeshare trip data. The data includes:

- Duration – Duration of trip
- Start Date – Includes start date and time
- End Date – Includes end date and time
- Start Station – Includes starting station name and number
- End Station – Includes ending station name and number
- Bike Number – Includes ID number of bike used for the trip
- Member Type – Indicates whether user was a "registered" member (Annual Member, 30-Day Member or Day Key Member) or a "casual" rider (Single Trip, 24-Hour Pass, 3-Day Pass or 5-Day Pass)

This data has been processed to remove trips that are taken by staff as they service and



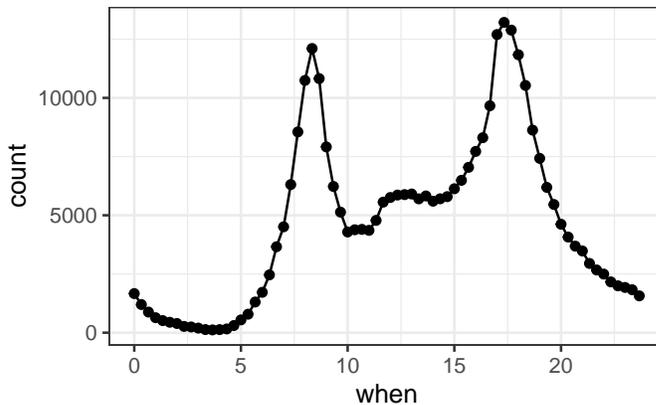
October 2019

Duration	Start date	End date	Start station number	Start station	End station
429	2019-10-01 00:01:59	2019-10-01 00:09:08	31214	17th & Corcoran St NW	
1935	2019-10-01 00:03:07	2019-10-01 00:35:23	31269	3rd St & Pennsylvania Ave SE	
563	2019-10-01 00:03:51	2019-10-01 00:13:14	31214	17th & Corcoran St NW	

... and so on for 337,552 rows altogether.

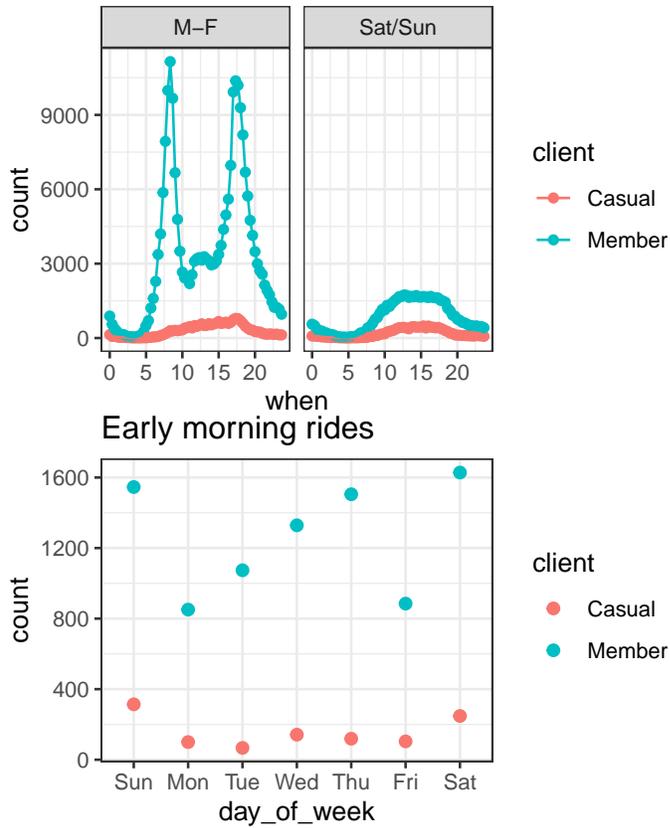
CapitalBikeshare questions

- How many bikes are there? Ans: 4699
- How many are in use at each time? What might this depend on?
- What's a long trip? Are long trips more common at some times than others?



Why does the plot have this shape?

How would you check your hypothesis?



Are tests what we want?

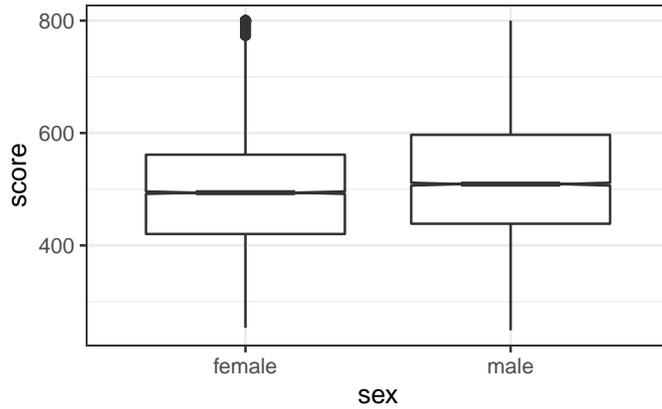
Scores on SAT Math test by sex

t-test: Emphasizes differences, not similarities

```
##
## Welch Two Sample t-test
##
## data:  score by sex
## t = -12.476, df = 19826, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -22.64338 -16.49449
## sample estimates:
## mean in group female    mean in group male
##           498.8845           518.4535
```

What does the p-value tell you, really?

Box-and-whiskers plot



Show individuals!

What's the take-home impression here?

