Introductory Thoughts on Stats

Héctor Corrada Bravo

University of Maryland, College Park, USA

CMSC320: 2018-03-13

Why Stats?

In this class we learn *Statistical and Machine Learning* techniques for data analysis.

By the time we are done, you should

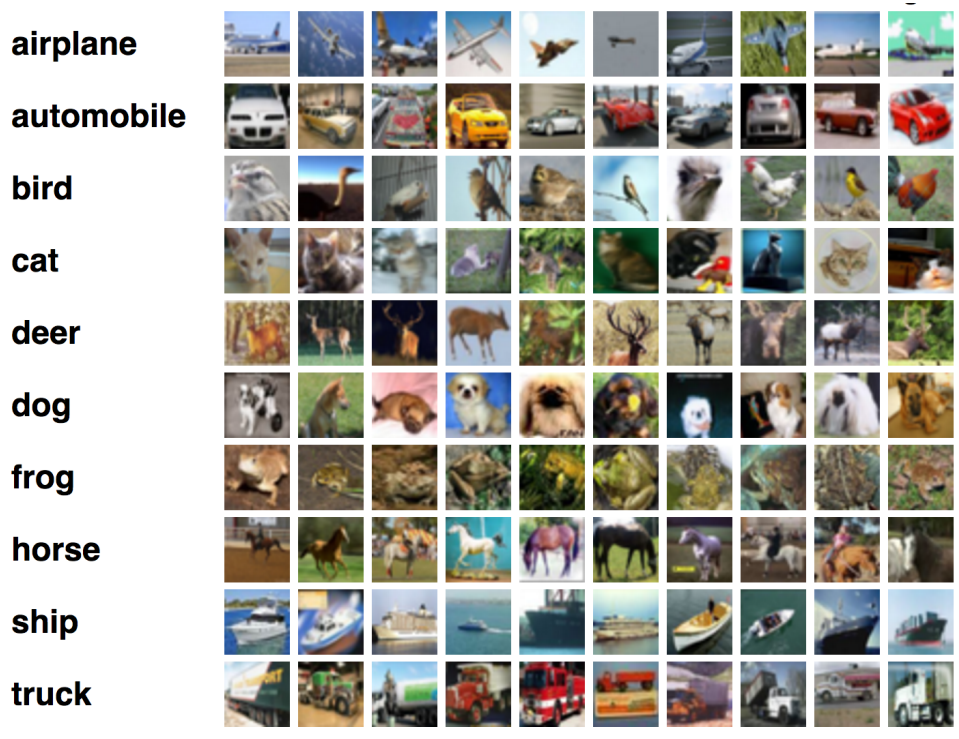
- be able to read **critically** papers or reports that use these methods.
- be able to use these methods for daata analysis

Why Stats?

In either case, you will need to ask yourself if findings are **statistically significant**.

Why Stats?

- Use a classification algorithm to distinguish images
- Accurate 70 out of 100 cases.
- Could this happen by chance alone?



Why Stats?

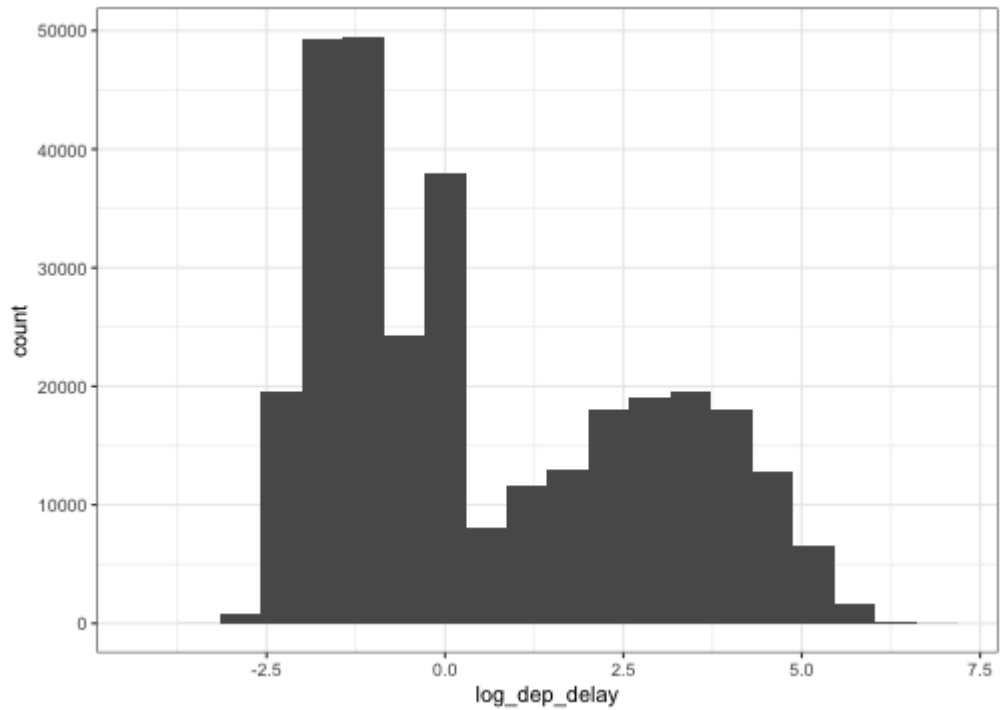
To be able to answer these question, we need to understand some basic probabilistic and statistical principles.

In this course unit we will review some of these principles.

Variation, randomness and stochasticity

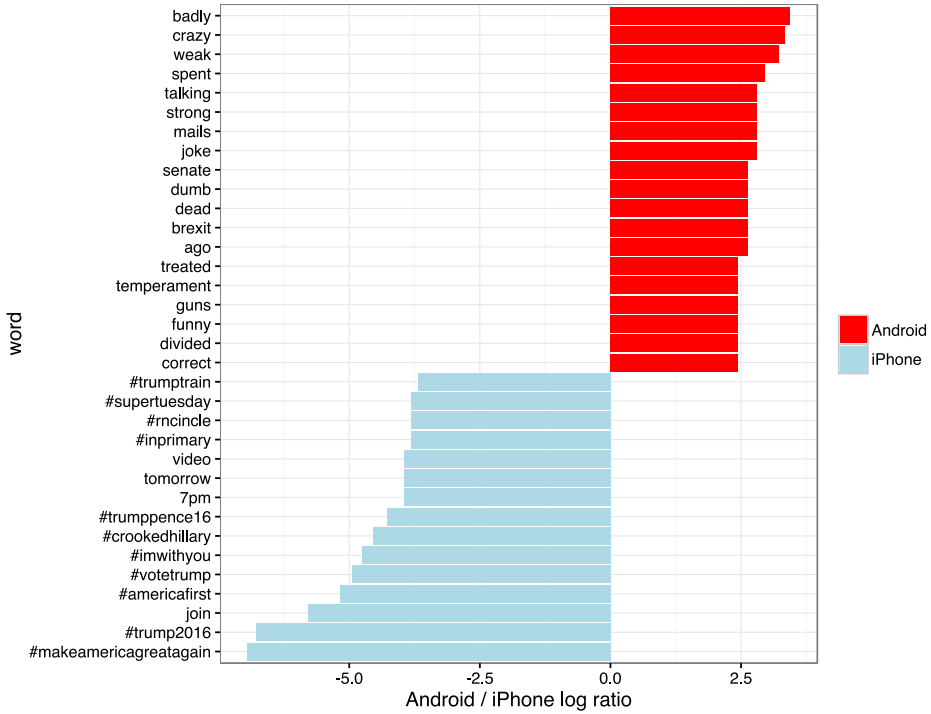
So far, we have not spoken about *randomness* and *stochasticity*. We have, however, spoken about *variation*.

spread in a dataset refers to the fact that in a population of entities there is naturally occuring variation in measurements



Variation, randomness and stochasticity

Another example: in sets of tweets there is natural variation in the frequency of word usage.



Variation, randomness and stochasticity

In summary, we can discuss the notion of *variation* without referring to any randomness, stochasticity or noise.

Why Probability?

Because, we **do** want to distinguish, when possible:

- natural occuring variation, vs.
- randomness or stochasticity

Why Probability?

- Find loan debt for **all** 19-30 year old Maryland residents, and calculate mean and standard deviation.

Why Probability?

- Find loan debt for **all** 19-30 year old Maryland residents, and calculate mean and standard deviation.
- That's difficult to do for all residents.

Why Probability?

- Find loan debt for **all** 19-30 year old Maryland residents, and calculate mean and standard deviation.
- That's difficult to do for all residents.
- Instead we sample (say by randomly sending Twitter surveys), and *estimate* the average and standard deviation of debt in this population from the sample.

Why Probability?

Now, this presents an issue since we could do the same from a different random sample and get a different set of estimates. Why?

Why Probability?

Now, this presents an issue since we could do the same from a different random sample and get a different set of estimates. Why?

Because there is naturally-occuring variation in this population.

Why Probability?

So, a simple question to ask is:

> How good are our *estimates* of debt mean and standard deviation from sample of 19-30 year old Marylanders?

Why Probability?

Another example: suppose we build a predictive model of loan debt for 19-30 year old Marylanders based on other variables (e.g., sex, income, education, wages, etc.) from our sample.

Why Probability?

Another example: suppose we build a predictive model of loan debt for 19-30 year old Marylanders based on other variables (e.g., sex, income, education, wages, etc.) from our sample.

> How good will this model perform when predicting debt in general?

Why Probability?

We use probability and statistics to answer these questions.

Why Probability?

We use probability and statistics to answer these questions.

- Probability captures stochasticity in the sampling process, while

Why Probability?

We use probability and statistics to answer these questions.

- Probability captures stochasticity in the sampling process, while
- we *model* naturally occuring variation in measurements in a population of interest.

One final word

The term *population* means

> **the entire** collection of entities we want to model

This could include people, but also images, text, chess positions, etc.