Introduction to Data Science: Statistical Principles

Héctor Corrada Bravo

University of Maryland, College Park, USA

CMSC320: 2020-03-30

Why Stats?

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

By the time we are done, you should

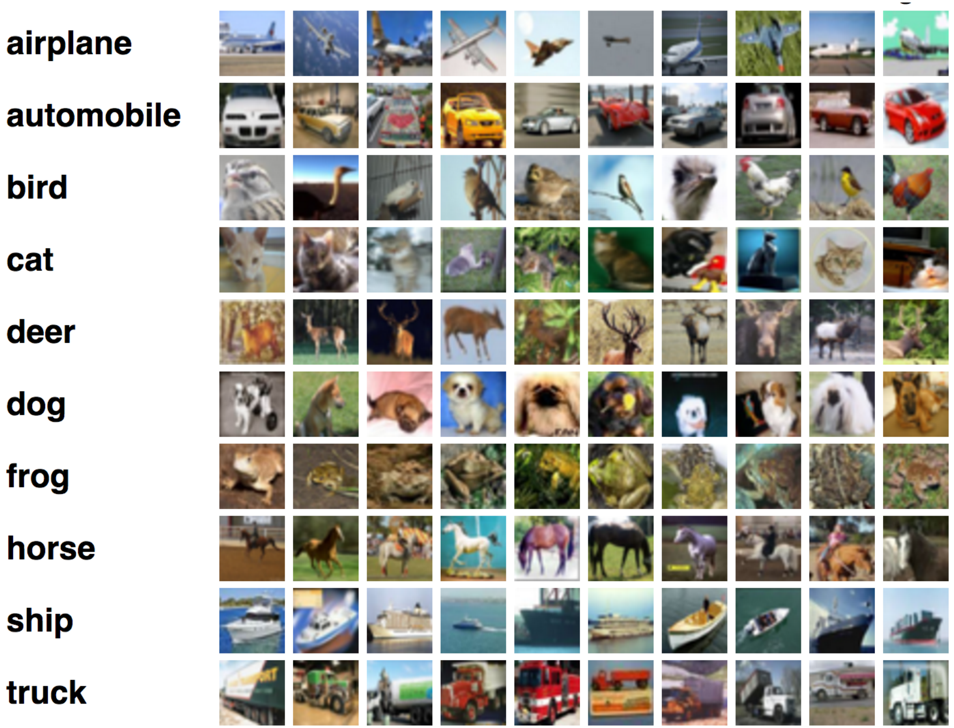
- be able to read **critically** papers or reports that use these methods.
- be able to use these methods for daata analysis

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?

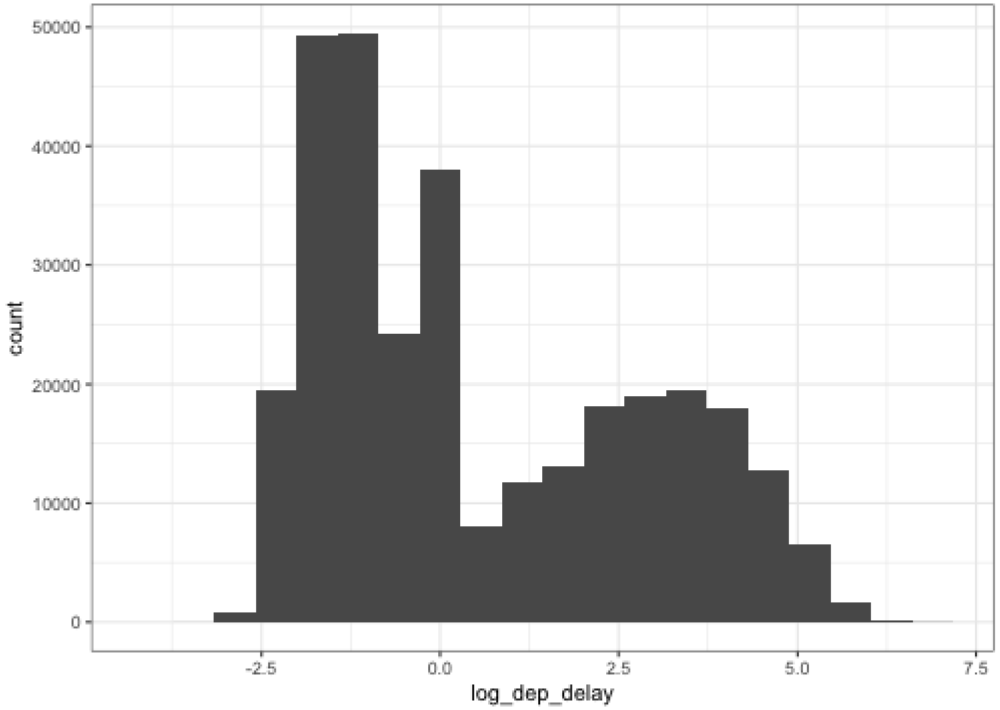
To be able to answer these question, we need to understand some basic probabilistic and statistical principles.

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

Variation, randomness and stochasticity

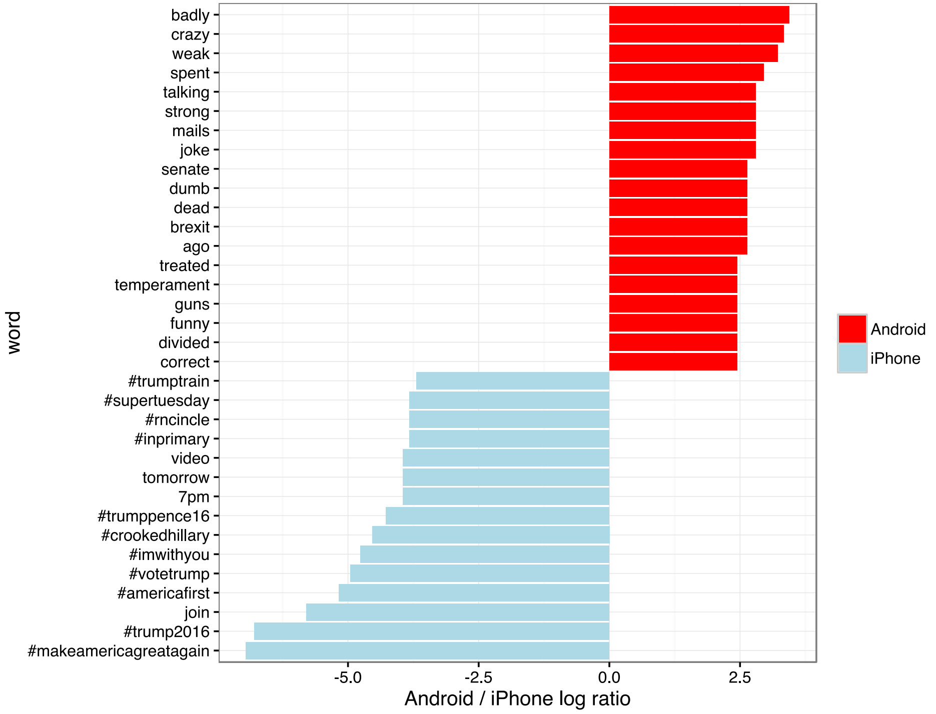
So far, we have not spoken about *randomness* and *stochasticity*. We have, however, spoken about *variation*.

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.

Random variables

The basic concept in our discussion of probability is the *random variable*.

Task: is a given tweet was generated by a bot?

Action: Sample a tweet **at random** from the set of all tweets ever written and have a human expert decide if it was generated by a bot or not.

Principle: Denote this as a *binary* random variable $X \in \{0, 1\}$, with value 1 if the tweet is bot-gerneated and 0 otherwise.

Random variables

The basic concept in our discussion of probability is the *random variable*.

Task: is a given tweet was generated by a bot?

Action: Sample a tweet **at random** from the set of all tweets ever written and have a human expert decide if it was generated by a bot or not.

Principle: Denote this as a *binary* random variable $X \in \{0, 1\}$, with value 1 if the tweet is bot-gerneated and 0 otherwise.

Why is this a random value? Because it depends on the tweet that was *randomly* sampled.

(Discrete) Probability distributions

A *probability distribution* $P : \mathcal{D} \to [0,1]$ over set $\mathcal{D}$ of all values random variable X can take to the interval $[0,1]$.

(Discrete) Probability distributions

A *probability distribution* $P : \mathcal{D} \to [0,1]$ over set $\mathcal{D}$ of all values random variable X can take to the interval $[0,1]$.

We start with a *probability mass function* p:

a. $p(X = x) \geq 0$ for all values $x \in \mathcal{D}$, and
b. $\sum_{x \in \mathcal{D}} p(X = x) = 1$

(Discrete) Probability distributions

How to interpret quantity $p(X = 1)$?

(Discrete) Probability distributions

How to interpret quantity $p(X = 1)$?

a. $p(X = 1)$ is the *probability* that a uniformly random sampled tweet is bot-generated, which implies

(Discrete) Probability distributions

How to interpret quantity $p(X=1)$?

a. $p(X=1)$ is the *probability* that a uniformly random sampled tweet is bot-generated, which implies

b. the proportion of bot-generated tweets in the set of "all" tweets is $p(X=1)$.

(Discrete) Probability distributions

Example The oracle of TWEET

Suppose we have a magical oracle and know for a *fact* that 70% of "all" tweets are bot-generated.

(Discrete) Probability distributions

Example The oracle of TWEET

Suppose we have a magical oracle and know for a *fact* that 70% of "all" tweets are bot-generated.

In that case $p(X=1)=.7$ and $p(X=0)=1-.7=.3$.

(Discrete) Probability distributions

cumulative probability distribution P describes the sum of probability up to a given value:

$$P(x) = \sum_{x' \mathcal{D} \text{ s.t. } x' \leq x} p(X = x')$$

(Discrete) Probability distributions

Expectation

What if I randomly sampled $n = 100$ tweets?

How many of those do I *expect* to be bot-generated?

(Discrete) Probability distributions

Expectation

What if I randomly sampled $n = 100$ tweets?

How many of those do I *expect* to be bot-generated?

Expectation is a formal concept in probability:

$$\mathbb{E}[X] = \sum_{x \in \mathcal{D}} x p(X = x)$$

(Discrete) Probability distributions

What is the expectation of X (a single sample) in our tweet example?

(Discrete) Probability distributions

What is the expectation of X (a single sample) in our tweet example?

$$\mathbb{E}[X] = 0 \times p(X = 0) + 1 \times p(X = 1) = \; 0 \times .3 + 1 \times .7 = .7$$

(Discrete) Probability distributions

What is the expected number of bot-generated tweets in a sample of $n = 100$ tweets.

Define $Y = X_1 + X_2 + \cdots + X_{100}$.

Then we need $\mathbb{E}[Y]$

(Discrete) Probability distributions

We have $X_i = \{0, 1\}$ for each of the $n = 100$ tweets

Each obtained by uniformly and *independently* sampling from the set of all tweets.

Then, random variable Y is *the number of bot-generated tweets in my sample of* $n = 100$ *tweets*.

(Discrete) Probability distributions

$$\begin{aligned}
\mathbb{E}[Y] &= \mathbb{E}[X_1 + X_2 + \cdots + X_{100}] \\
&= \mathbb{E}[X_1] + \mathbb{E}[X_2] + \cdots + \mathbb{E}[X_{100}] \\
&= .7 + .7 + \cdots + .7 \\
&= 100 \times .7 \\
&= 70
\end{aligned}$$

(Discrete) Probability distributions

This uses some facts about expectation you can show in general.

(1) For any pair of random variables X_1 and X_2,
$\mathbb{E}[X_1 + X_2] = \mathbb{E}[X_1] + \mathbb{E}[X_2]$.

(2) For any random variable X and *constant* a, $\mathbb{E}[aX] = a\mathbb{E}[X]$.

Estimation

So far we assume we have access to an oracle that told us $p(X = 1) = .7$.

In reality, we *don't*.

Estimation

So far we assume we have access to an oracle that told us $p(X = 1) = .7$.

In reality, we *don't*.

For our tweet analysis task, we need to *estimate* the proportion of "all" tweets that are bot-generated.

Estimation

So far we assume we have access to an oracle that told us $p(X = 1) = .7$.

In reality, we *don't*.

For our tweet analysis task, we need to *estimate* the proportion of "all" tweets that are bot-generated.

This is where our probability model and the expectation we derive from it comes in.

Estimation

Given *data* $x_1, x_2, x_3, \ldots, x_{100}$,

With 67 of those tweets labeled as bot-generated (i.e., $x_i = 1$ for 67 of them)

Estimation

Given *data* $x_1, x_2, x_3, \ldots, x_{100}$,

With 67 of those tweets labeled as bot-generated (i.e., $x_i = 1$ for 67 of them)

We can say $y = \sum_i x_i = \mathbf{67}$.

Estimation

Given *data* $x_1, x_2, x_3, \ldots, x_{100}$,

With 67 of those tweets labeled as bot-generated (i.e., $x_i = 1$ for 67 of them)

We can say $y = \sum_i x_i = 67$.

We *expect* $y = np$ with $p = p(X = 1)$

Estimation

Given *data* $x_1, x_2, x_3, \ldots, x_{100}$,

With 67 of those tweets labeled as bot-generated (i.e., $x_i = 1$ for 67 of them)

We can say $y = \sum_i x_i = 67$.

We *expect* $y = np$ with $p = p(X = 1)$

Use that observation to *estimate* p!

Estimation

$$
\begin{aligned}
np = 67 &\Rightarrow \\
100p = 67 &\Rightarrow \\
\hat{p} = \frac{67}{100} &\Rightarrow \\
\hat{p} = .67 &
\end{aligned}
$$

Estimation

Our estimate ($\hat{p}=.67$) is wrong, but close.

Can we ever get it right?

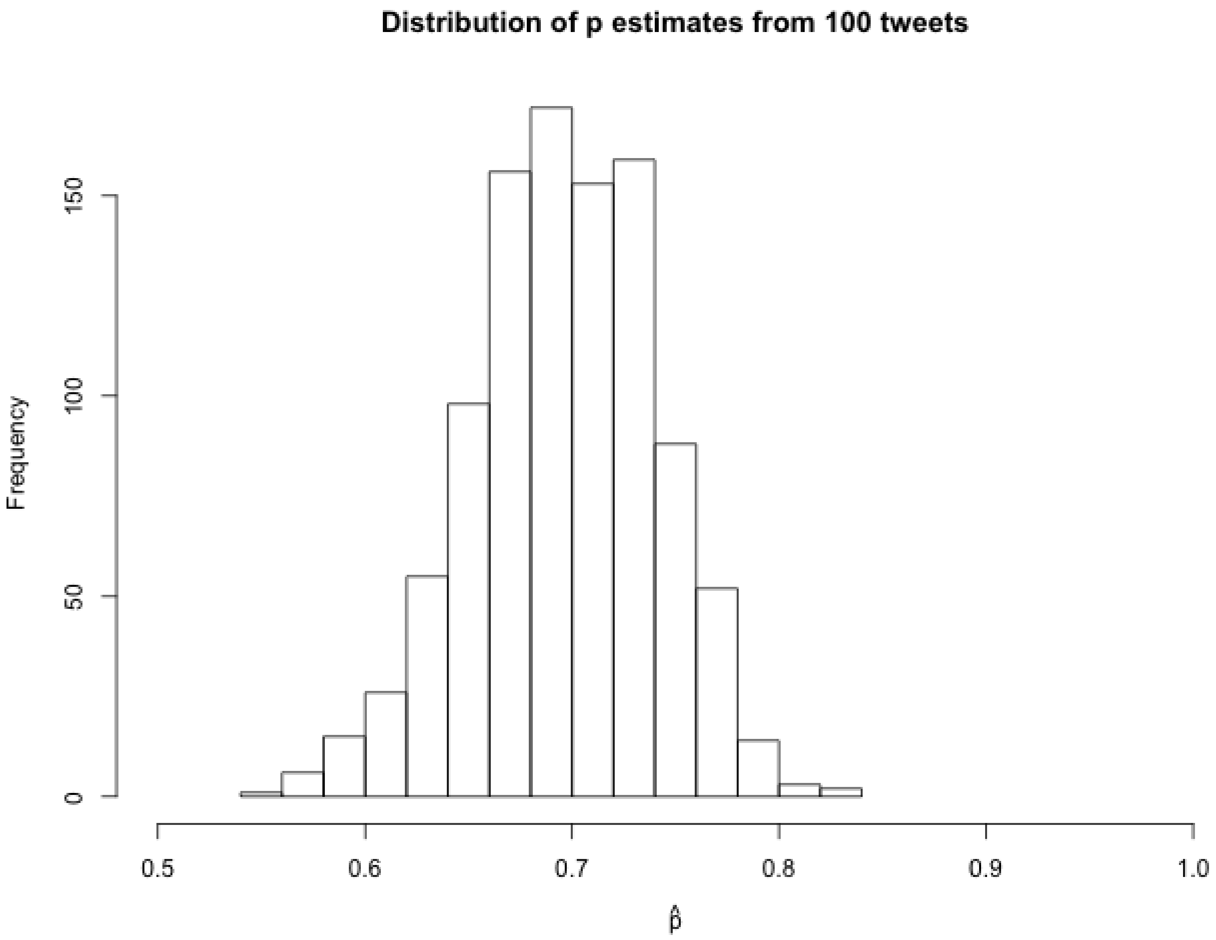
Can I say how wrong I should expect my estimates to be?

Estimation

Notice that our estimate of $\hat{p}$ is the sample *mean* of $x_1, x_2, \ldots, x_n$.

Let's go back to our oracle of tweet to do a thought experiment and replicate how we derived our estimate from 100 tweets a few thousand times.

Estimation

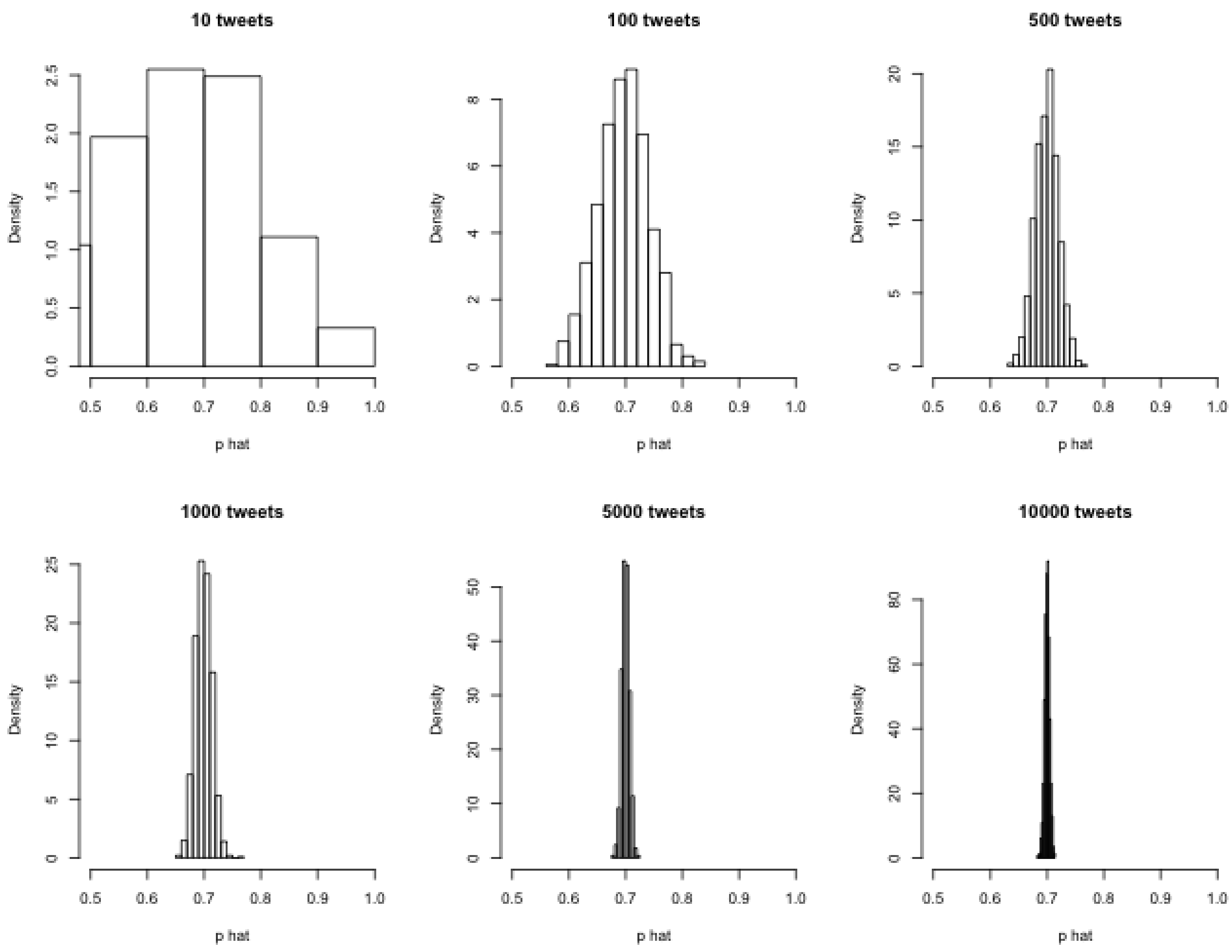

Distribution of p estimates from 100 tweets
Frequency
0
50
100
150
0.5
0.6
0.7
0.8
0.9
1.0
$\hat{p}$

Estimation

What does this say about our estimates of the proportion of bot-generated tweets if we use 100 tweets in our sample?

Now what if instead of sampling $n = 100$ tweets we used other sample sizes?

Estimation



10 tweets
100 tweets
500 tweets
1000 tweets
5000 tweets
10000 tweets
Density
p hat

Estimation

We can make a couple of observations:

1. The distribution of estimate $\hat{p}$ is *centered* at $p = .7$, our unknown *population* proportion, and

2. The *spread* of the distribution **decreases** as the number of samples n **increases**.

Estimation

This was a simulation, we faked the **data generating procedure**.

In reality, we can't.

Estimation

This was a simulation, we faked the **data generating procedure**.

In reality, we can't.

What to do we do then?

(1) Math, or
(2) Resample

Solve with Math

Our simulation is an illustration of two central tenets of statistics:

(a) The law of large numbers (LLN)

(b) The central limit theorem (CLT)

Solve with Math

Law of large numbers (LLN)

Given *independently* sampled random variables $X_1, X_2, \cdots, X_n$ with $\mathbb{E}[X_i] = \mu$ for all i,

$$\frac{1}{n}\sum_i X_i \to \mu, \text{ as } n \to \infty$$

I.E. $\overline{x}$ *tends* to the expected value μ (under some assumptions beyond the scope of this class) regardless of the distribution X_i.

Solve with Math

Solve with Math

Central Limit Theorem (CLT)

The LLN says that estimates built using the sample mean will tend to the correct answer

The CLT describes how these estimates are *spread* around the correct answer.

Solve with Math

Here we will use the concept of *variance* which is expected *spread*, measured in squared distance, from the *expected value* of a random variable:

$$\mathbf{var}[\mathrm{X}] = \mathbb{E}[(X - \mathbb{E}[X])^2]$$

Solve with Math

$$
\begin{aligned}
\mathrm{var}[\mathrm{X}] &= \sum_{\mathcal{D}} (x - \mathbb{E}[X])^2 p(X = x) \\
&= (0 - p)^2 \times (1 - p) + (1 - p)^2 \times p \\
&= p^2(1 - p) + (1 - p)^2 p \\
&= p(1 - p)(p + (1 - p)) \\
&= p(1 - p)(p - p + 1) \\
&= p(1 - p)
\end{aligned}
$$

Solve with Math

$$P\left(\frac{1}{n}\sum_{i=1} X_i\right) \to N\left(\mu, \frac{\sigma}{n}\right), \text{ as } n \to \infty$$

Solve with Math

This says, that as sample size n increases, the distribution of sample means is *well* approximated by a **normal distribution**.

This means we can approximate the *expected error* of our estimates well.

(Continuous) Random Variables

The normal distribution

Random variable $Y = \sum_{i=1}^{n} X_i$ is *continuous*.

The normal distribution describes the distribution of *continuous* random variables over the range $(-\infty, \infty)$ using two parameters:

mean μ and **standard deviation** σ.

(Continuous) Random Variables

The normal distribution

Random variable $Y = \sum_{i=1}^{n} X_i$ is *continuous*.

The normal distribution describes the distribution of *continuous* random variables over the range $(-\infty, \infty)$ using two parameters:

mean μ and **standard deviation** σ.

We write " Y is normally distributed with mean μ and standard deviation σ" as $Y \sim N(\mu, \sigma)$.

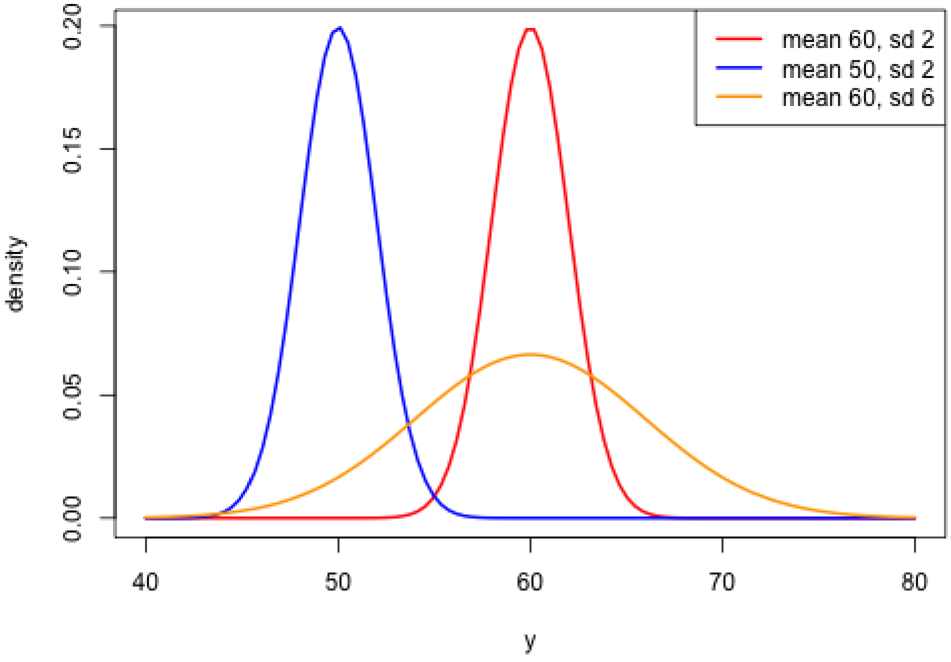
(Continuous) Random Variables

Continuous random variables are described by a *probability density function*. For normally distributed random variables:

$$p(Y = y) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2} \left(\frac{y - \mu}{\sigma} \right)^2 \right\}$$

(Continuous) Random Variables

Three examples of normal probability density functions with mean $\mu = 60, 50, 60$ and standard deviation $\sigma = 2, 2, 6$:



(Continuous) Random Variables

Like the discrete case, probability density functions for continuous random variables need to satisfy certain conditions:

a. $p(Y = y) \geq 0$ for all values $Y \in (-\infty, \infty)$, and
b. $\int_{-\infty}^{\infty} p(Y = y) dy = 1$

(Continuous) Random Variables

One way of interpreting the density function of the normal distribution is that probability decays exponentially with rate σ based on squared distance to the mean μ. (Here is squared distance again!)

$$\boxed{p(Y=y) \propto \exp \left\{ -{\frac{1}{2\sigma^2} (y-\mu)^2} \right\}}$$

(Continuous) Random Variables

Also, notice the term inside the square?

$$z = \left(\frac{y - \mu}{\sigma} \right)$$

this is the *standardization* transformation we saw before.

(Continuous) Random Variables

The name *standardization* comes from the *standard normal distribution* $N(0, 1)$ (mean 0 and standard deviation 1),

Which is very convenient to work with because it's density function is much simpler:

$$p(Z = z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z^2\right\}$$

(Continuous) Random Variables

The name *standardization* comes from the *standard normal distribution* $N(0, 1)$ (mean 0 and standard deviation 1),

Which is very convenient to work with because it's density function is much simpler:

$$p(Z = z) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}z^2\right\}$$

In fact, if random variable $Y \sim N(\mu, \sigma)$ then random variable $Z = \frac{Y-\mu}{\sigma} \sim N(0, 1)$.

(Continuous) Random Variables

One more technicality:

The cumulative probability function for continuous random variables is given by

$$P(Y \leq y) = \int_{\mathcal{D}} p(Y = y) dy$$

where $\mathcal{D}$ is the range of values random variable Y can take (e.g., for normal distribution $\mathcal{D} = (-\infty, \infty)$)

Solve with Math

CLT continued

We need one last bit of terminology to finish the statement of the CLT.

Consider data $X_1, X_2, \cdots, X_n$ with $\mathbb{E}[X_i] = \mu$ for all i, **and** $\mathrm{var}(X_i) = \sigma^2$ for all i,

and sample mean $Y = \frac{1}{n} \sum_i X_i$.

The standard deviation of Y is called the *standard error*:

$$\mathrm{se}(Y) = \frac{\sigma}{\sqrt{n}}$$

Solve with Math

Now we can restate the CLT statement precisely:

the distribution of Y tends *towards* $N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$ as $n \to \infty$.

This says, that as sample size increases the distribution of sample means is well approximated by a normal distribution,

and that the spread of the distribution goes to zero at the rate $\sqrt{n}$.

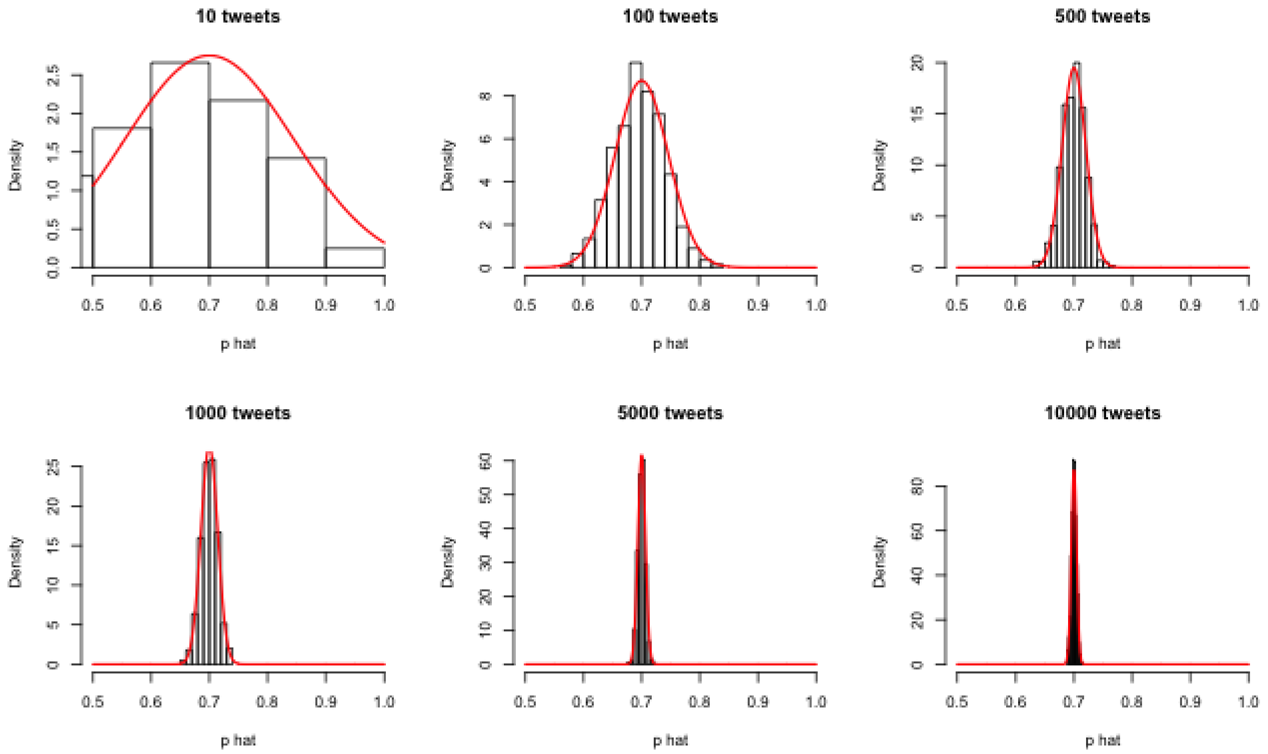
Solve with Math

Disclaimer There a few mathematical subtleties. Two important ones are that

a. $X_1, \ldots, X_n$ are iid (independent, identically distributed) random variables, and
b. $\mathtt{var}[X] < \infty$

Solve with Math

Let's redo our simulated replications of our tweet samples to illustrate the CLT at work:



Solve with Math

Here we see the three main points of the LLN and CLT:

(1) the normal density is centered around $\mu = .7$,
(2) the normal approximation gets better as n increases, and
(3) the standard error goes to 0 as n increases.

Solve with computation

The Bootstrap Procedure

What if the conditions that we used for the CLT don't hold?

For instance, samples X_i may not be independent. What can we do then, how can we say something about the precision of sample mean estimate Y?

Solve with computation

The Bootstrap Procedure

A useful procedure to use in this case is the **bootstrap**.

It is based on using *randomization* to simulate the stochasticity resulting from the population sampling procedure we are trying to capture in our analysis.

Solve with computation

The Bootstrap Procedure

The main idea is the following: given observations $x_1, \ldots, x_n$

and the estimate $y = \frac{1}{n}\sum_{i=1}^{n} x_i$,

what can we say about the standard error of y?

Solve with computation

The Bootstrap Procedure

There are two challenges here:

1) our estimation procedure is deterministic, that is, if I compute the sample mean of a specific dataset, I will always get the same answer; and

2) we should retain whatever properties of estimate y result from obtaining it from n samples.

Solve with computation

The Bootstrap Procedure

The bootstrap is a randomization procedure that measures the variance of estimate y,

using randomization to address challenge (1),

but doing so with randomized samples of size n, addressing challenge (2).

Solve with computation

The Bootstrap Procedure

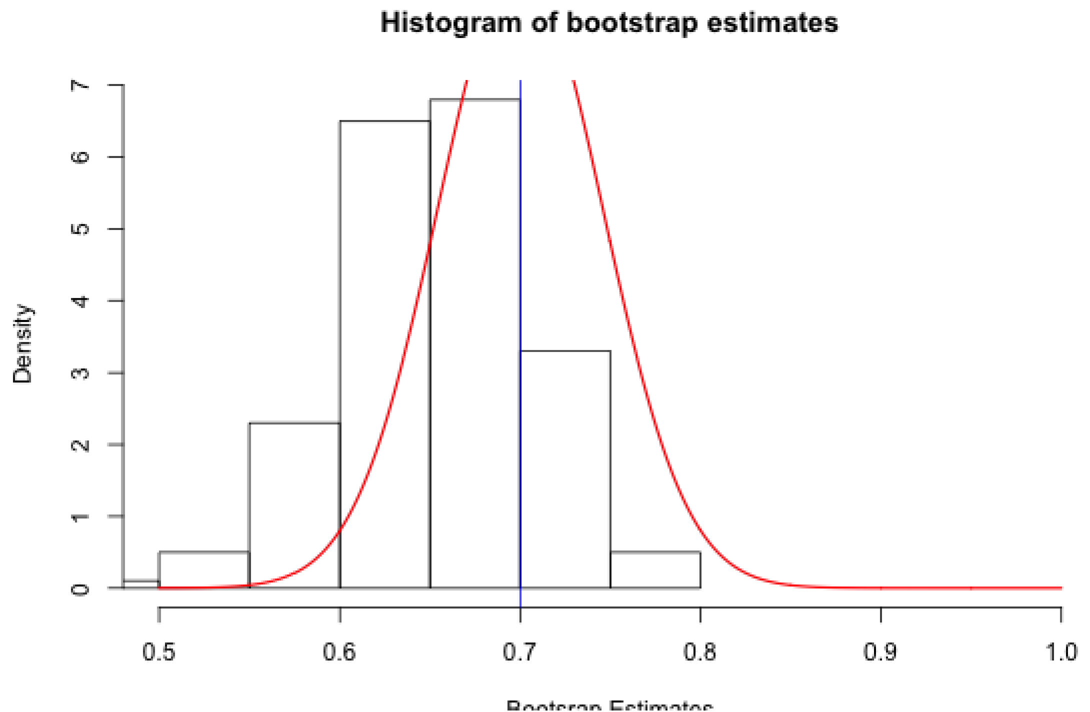
The procedure goes as follows:

1. Generate B random datasets by sampling *with replacement* from dataset $x_1, \ldots, x_n$. Denote randomized dataset b as $x_{1b}, \ldots, x_{nb}$.
2. Construct estimates from *each* dataset, $y_b = \frac{1}{n} \sum_i x_{ib}$
3. Compute center (mean) and spread (variance) of estimates y_b

Solve with computation

The Bootstrap Procedure

Let's see how this works on tweet oracle example



Solve with computation

The Bootstrap Procedure

Not great, math works better when conditions are met.

Solve with computation

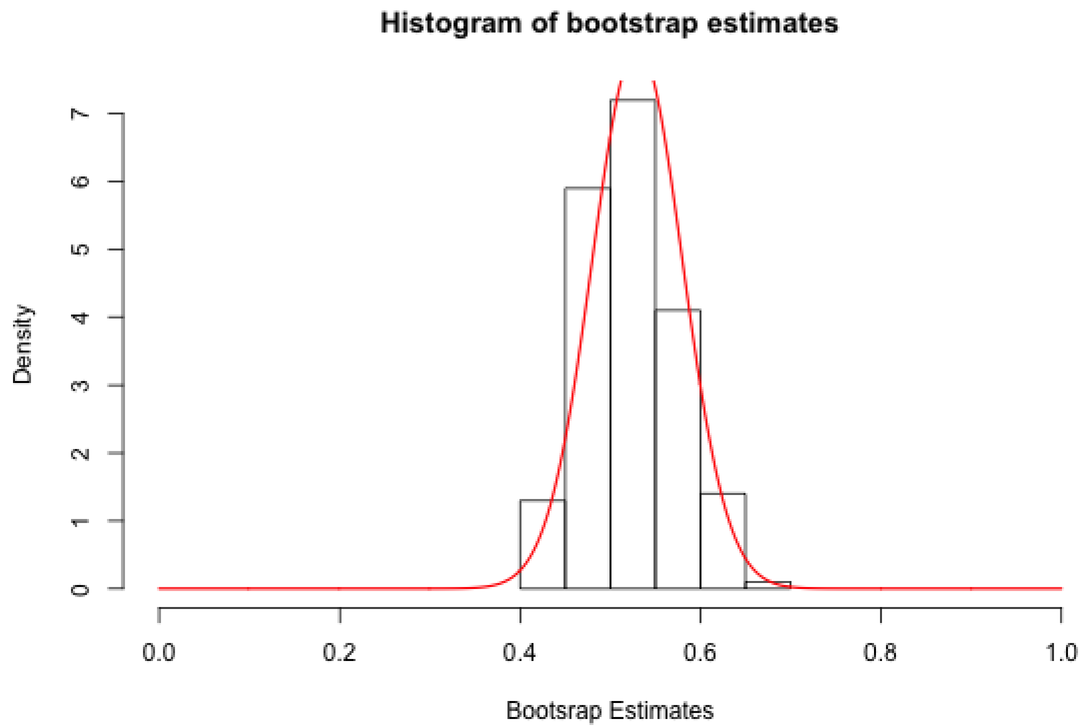
The Bootstrap Procedure

Let's look at a case where we don't expect the normal approximation to not work so well by making samples not identically distributed.

Let's make a new ORACLE of tweet where the probability of a tweet being bot-generated depends on the previous tweet

Solve with computation

The Bootstrap Procedure



Solve with computation

The Bootstrap Procedure

Here, an analysis based on the classical CLT is not appropriate (X_i s are not independent)

But the bootstrap analysis gives some information about the variability of our estimates.