Introduction to Data Science: Statistical Principles (Part 2)

Héctor Corrada Bravo

University of Maryland, College Park, USA

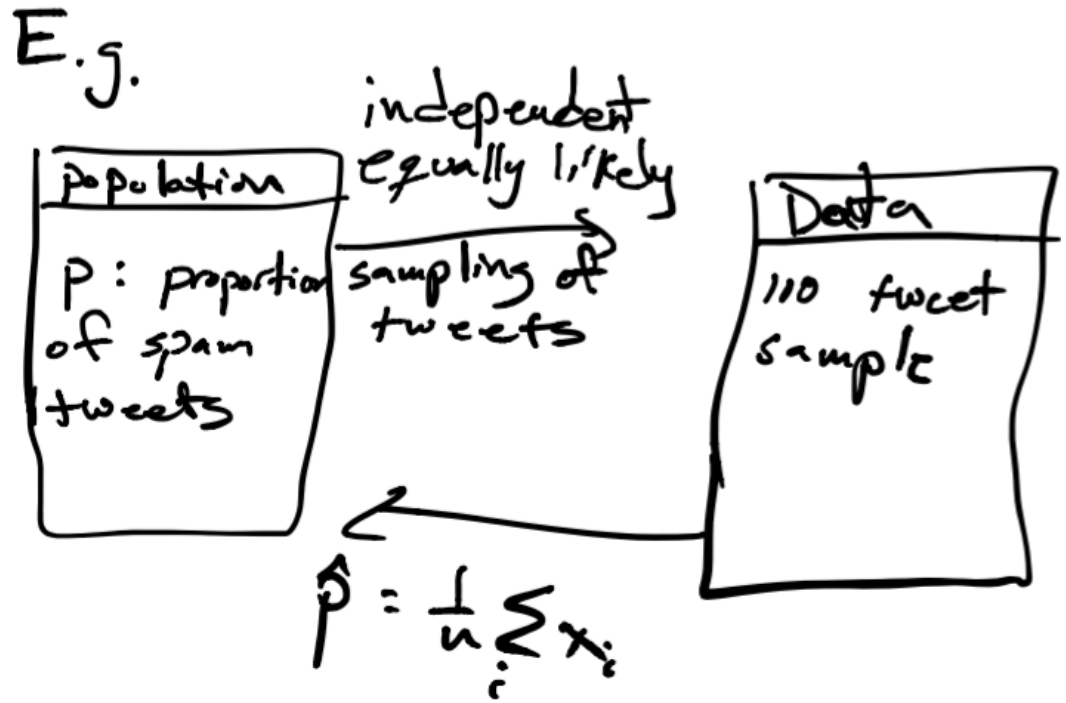
CMSC320: 2020-04-01

Inference

One way to think about how we use probability in data analysis (statistical and machine learning) is like this:

Inference

One way to think about how we use probability in data analysis (statistical and machine learning) is like this:



Inference

Law of Large Numbers (LLN): parameter $\hat{p}$ will be close to p on average,
Central Limit Theorem (CLT): how confident are we that we found p.

Inference

Law of Large Numbers (LLN): parameter $\hat{p}$ will be close to p on average,
Central Limit Theorem (CLT): how confident are we that we found p.

Confidence Interval:

Since $\hat{p} \sim N(p, \frac{\sqrt{p(1-p)}}{\sqrt{n}})$ let's find an interval $[\hat{p}_-, \hat{p}_+]$, with:

- $\hat{p}$ at its center,
- contains 95% of the probability specified by the CLT.

Inference

How do we calculate this interval?

$\hat{p}_-$ will be the value where $N\left(p, \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$ is such that $P(Y \leq \hat{p}_-) = .05/2$.

Inference

How do we calculate this interval?

$\hat{p}_-$ will be the value where $N\left(p, \frac{\sqrt{p(1-p)}}{\sqrt{n}}\right)$ is such that $P(Y \leq \hat{p}_-) = .05/2$.

In R, we calculate with `qnorm`:

$$\begin{aligned}\hat{p}_- &= \texttt{qnorm}(.05/2, \hat{p}, \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}) \\ &= \hat{p} + \texttt{qnorm}(.05/2, 0, \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}})\end{aligned}$$

Inference

The upper value of the interval is computed with probability $1 - (.05/2)$,

By the symmetry of the normal distribution is

$$\hat{p}_+ = \hat{p} + -\texttt{qnorm}(.05/2, 0, \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}})$$

Inference

Let's see how these intervals look for our twitter bot example:

sample_size	phat	se	lower	upper
10	0.700	0.145	0.416	0.984
100	0.690	0.046	0.599	0.781
500	0.702	0.020	0.662	0.742
1000	0.694	0.015	0.665	0.723
10000	0.698	0.005	0.689	0.707

For $n = 500$, our estimate of p is ${}_{0.66}0.7_{0.74}$.

Hypothesis testing

Suppose that **before** I sampled tweets I thought (*hypothesized*) that more than 50% of tweets are bot-generated.

Hypothesis testing

Suppose that **before** I sampled tweets I thought (*hypothesized*) that more than 50% of tweets are bot-generated.

Hypothesis Testing A very popular way of using data to suggest this hypothesis is **true**:

Hypothesis testing

Suppose that **before** I sampled tweets I thought (*hypothesized*) that more than 50% of tweets are bot-generated.

Hypothesis Testing A very popular way of using data to suggest this hypothesis is **true**:

By using inference to **reject** the hypothesis that it is **not true**.

Hypothesis testing

null hypothesis: **50% or less of tweets are bot-generated**

alternative hypothesis (the one we cared about): **more than 50% of tweets are bot-generated**

Hypothesis testing

null hypothesis: **50% or less of tweets are bot-generated**

alternative hypothesis (the one we cared about): **more than 50% of tweets are bot-generated**

You will see this written in statistics textbooks as:

$$H_0 : p \leq .5 \qquad (\text{null})$$
$$H_1 : p > .5 \qquad (\text{alternative})$$

Hypothesis testing

Given sample of n tweets, estimate $\hat{p}$ as we did before.

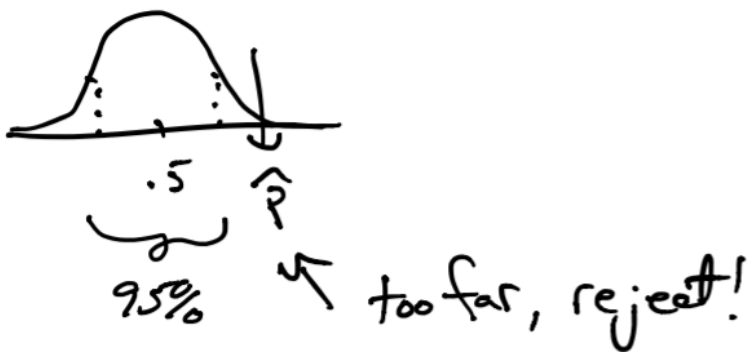
If $\hat{p}$ (sample mean from our sample of tweets) is *too far* from $p = .5$:

then we **reject** the *null* hypothesis: the estimate we derived from the data we have is not statistically consistent with the *null* hypothesis.

Hypothesis testing

How do we say our estimate $\hat{p}$ is too far? Use the probability model given by the CLT.

If $P(Y \geq \hat{p}) \geq .95$ under the null model (of $p = .5$), we say it is too far and we reject.



Hypothesis testing

This 95% threshold is conservative, but somewhat arbitrary.

So we use one more metric, $P(|Y| \geq \hat{p})$ (the infamous p-value) to say:

We could reject the *null* hypothesis for all thresholds greater than this p-value.

Hypothesis testing

Let's see how testing would look like for our tweet example

sample_size	phat	se	lower	upper	p_value
10	0.700	0.145	0.416	0.984	0.084
100	0.690	0.046	0.599	0.781	0.000
500	0.702	0.020	0.662	0.742	0.000
1000	0.694	0.015	0.665	0.723	0.000
10000	0.698	0.005	0.689	0.707	0.000

Hypothesis testing

The t-test

These results hold for n sufficiently large that the normal distribution in the CLT provides a good approximation of the distribution of estimates $\hat{p}$.

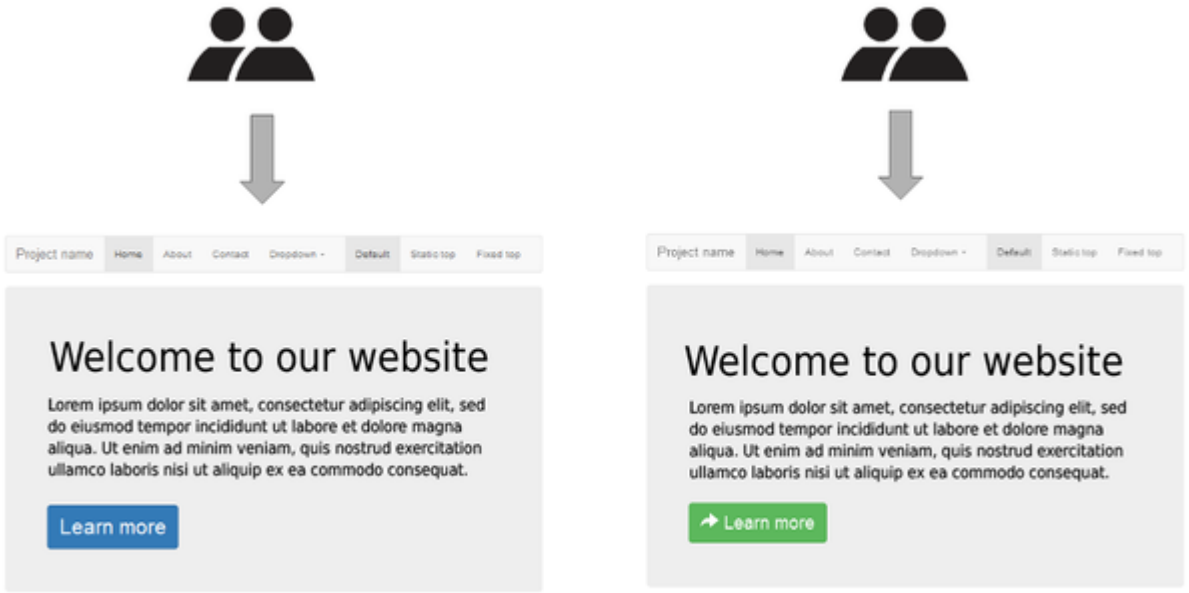
In cases where n is smaller, the t-distribution, as opposed to the normal distribution, provides a better approximation of the distribution of estimates $\hat{p}$.

As n grows, the t-distribution approaches a normal distribution which is why analysts use the t-test regularly.

Hypothesis testing

A/B Testing

A classic experimental design where hypothesis testing is commonly used in A/B testing.



Hypothesis testing

Here we have two estimates $\hat{p}_A$ and $\hat{p}_B$, the proportion of clicks for design A and B respectively.

The null hypothesis we would test is that *there is no difference in proportions* between the two designs.

Mathematically, we would like to know "What is the probability that we observe a difference in proportions this large under the null hypothesis".

We will work this out as a homework exercise (HW4).

Summary

Inference: estimate parameter from data based on assumed probability model

(e.g, matching expectation; we'll see later another method called maximum likelihood).

Summary

Inference: estimate parameter from data based on assumed probability model

(e.g, matching expectation; we'll see later another method called maximum likelihood).

For *averages* the LLN and CLT tells us how to compute probabilities from a single parameter estimate derived from one dataset of samples.

With these probabilities we can construct confidence intervals for our estimate.

Summary

Testing: Use probability *under null hypothesis* to see how statistically consistency of estimates obtained from data,

Reject the null hypothesis if estimates are not statistically consistent enough

(again using probability from CLT when dealing with averages).

Probability Distributions

Check lecture notes for further discussion of the probability distributions we saw in this discussion.

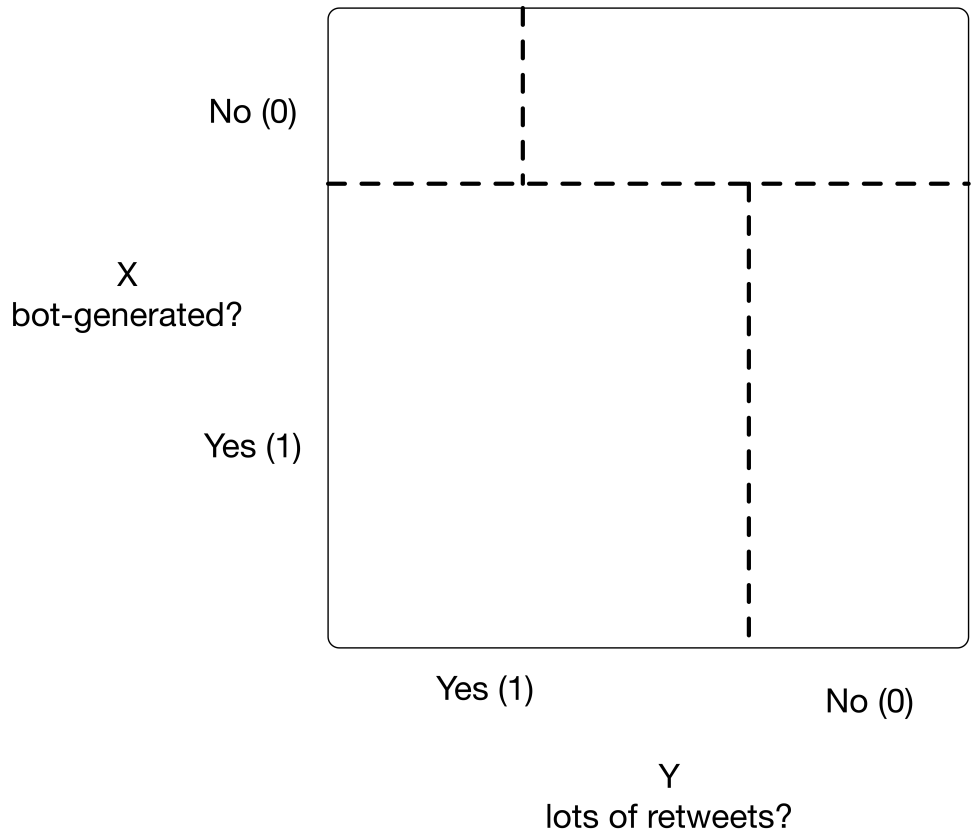
Joint and conditional probability

Suppose that for each tweet I sample I can also say if it has *a lot* of retweets or not.

I have another binary random variable $Y \in \{0, 1\}$ where $Y = 1$ indicates the sampled tweet has a lot of retweets.

Joint and conditional probability

We could illustrate the population of "all" tweets as



Joint and conditional probability

We can talk of the joint probability mass function of X and Y:
$p(X = x, Y = y)$,

where random variables X and Y can take values from domains $\mathcal{D}_X$ and $\mathcal{D}_Y$ respectively.

Here we have the same conditions as we had for univariate distributions:

1. $p(X = x, Y = y) \geq 0$ for all combination of values x and y, and
2. $\sum_{(x,y) \in \mathcal{D}_X \times \mathcal{D}_Y} p(X = x, Y = y) = 1$

Joint and conditional probability

We can also talk about *conditional probability*:

the probability of a tweet being bot-generated or not,
conditioned on whether it has lots of retweets or not:

$$p(X = x|Y = y)$$

which also needs to satisfy the properties of a probability distribution.

Joint and conditional probability

So to make sure

$$\sum_{x \in \mathcal{D}_X} p(X = x | Y = y) = 1$$

we define

$$p(X = x | Y = y) = \frac{p(X = x, Y = y)}{p(Y = y)}$$

marginalization: follows from the properties of joint probability distribution: $\sum_{x \in \mathcal{D}_X} p(X = x, Y = y) = p(Y = y)$.

Joint and conditional probability

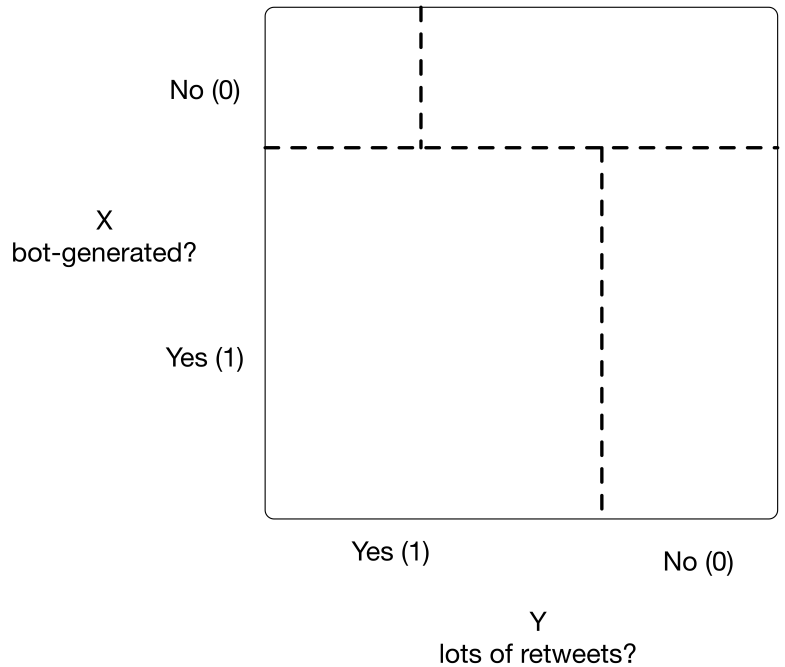
Conditional probability lets us talk about *independence*:

if the probabilty of a tweet being bot-generated *does not* depend on a tweet having lots of retweets

i.e., $p(X = x) = p(X = x|Y = y)$ for all y,

then we say X is *independent* of Y.

Joint and conditional probability



Is X independent of Y? What would the diagram look like if X was independent of Y?

Joint and conditional probability

For independent random variables, the joint probability has an easy form

$$p(X = x, Y = y) = p(X = x)p(Y = y)$$

Generalizes to more than two independent random variables.

Bayes' Rule

An extremely useful and important rule of probability follows from our definitions of conditional and joint probability above.

Bayes' rule is pervasive in Statistics, Machine Learning and Artificial Intelligence.

It is a very powerful tool to talk about uncertainty, beliefs, evidence, and many other technical and philosophical matters. It is however, of extreme simplicity.

Bayes' Rule

Bayes' Rule states that

$$p(X = x|Y = y) = \frac{p(Y = y|X = x)p(X = x)}{p(Y = y)}$$

which follow directly from our definitions above.

Bayes' Rule

One very common usage of Bayes' Rule is that it let's us define one conditional probability distribution based on another probability distribution.

For example, it may be hard to reason about $p(X = x|Y = y)$ in our tweet example.

Bayes' Rule

One very common usage of Bayes' Rule is that it let's us define one conditional probability distribution based on another probability distribution.

For example, it may be hard to reason about $p(X = x|Y = y)$ in our tweet example.

If you know a tweet has a lot retweets $(Y = 1)$, what can you say about the probability that it is bot-generated, i.e., $p(X = 1|Y = 1)$?

Maybe not much, tweets have lots of retweets for many reasons.

Bayes' Rule

However, it may be easier to reason about the reverse: if I tell you a tweet is bot-generated $(X = 1)$, what can you say about the probability that it has a lot of retweets, i.e., $p(Y = 1|X = 1)$?

That may be easier to reason about, at least bot-generated tweets are designed to get lots of retweets.

At minimum, it's easier to estimate because we can get a training set of bot-generated tweets and *estimate* this conditional probability.

Bayes' Rule

Bayes' Rule tells us how to get the hard to reason about (or estimate) conditional probability $p(X = x|Y = y)$

In terms of the conditional probability that is easier to reason about (or estimate) $p(Y = y|X = x)$.

This is the basis of the Naive Bayes prediction method, which we'll revisit briefly later on.

Conditional expectation

With conditional probabilty we can start talking about conditional expectation, which generalizes the concept of expectation we saw before.

The *conditional expected value* (conditional mean) of X given $Y = y$ is

$$\mathbb{E}[X|Y = y] = \sum_{x \in \mathcal{D}_X} xp(X = x|Y = y)$$

Conditional Expectation, which follows from conditional probability, will serve as the basis for our Machine Learning method studies in the next few lectures!

Maximum likelihood

We saw before how we estimated a parameter from matching expectation from a probability model with what we observed in data.

The most popular method of estimation (Maximum Likelihood Estimation) uses a similar idea.

Maximum likelihood

Given data $x_1, x_2, \ldots, x_n$ and an assumed model of their distribution, e.g.,

- $X_i \sim \text{Bernoulli}(p)$ for all i,
- they are iid,

Let's find the value of parameter p that maximizes the likelihood (or probability) of the data we observe under this assumed probability model.

We call the resulting estimate the *maximum likelihood estimate* (MLE).

Maximum likelihood

Here are some fun exercises to try:

1) Given a sample x_1 with $X_1 \sim N(\mu, 1)$, show that the maximum likelihood estimate of μ, $\hat{\mu} = x_1$.

Maximum likelihood

Here are some fun exercises to try:

1) Given a sample x_1 with $X_1 \sim N(\mu, 1)$, show that the maximum likelihood estimate of μ, $\hat{\mu} = x_1$.

It is most often convinient to *minimize negative log-likelihood* instead of maximizing likelihood. So in this case:

$$\begin{aligned} -\mathscr{L}(\mu) &= -\log p(X_1 = x_1) \\ &= \log \sqrt{2\pi} + \frac{1}{2}(x_1 - \mu)^2 \end{aligned}$$

Maximum likelihood

To minimize

$$-\mathscr{L}(\mu) = \log\sqrt{2\pi} + \frac{1}{2}(x_1 - \mu)^2$$

Ignore terms that are independent of μ, and concentrate only on minimizing the last term.

Now, this term is always positive, so the smallest value it can have is 0. So, we minimize it by setting $\hat{\mu} = x_1$.

Maximum likelihood

2) Given a sample $x_1, x_2, \ldots, x_n$ of n iid random variables with $X_i \sim N(\mu, 1)$ for all i,

Show that the maximum likelihood estimate of μ, $\hat{\mu} = \overline{x}$ the sample mean!

Maximum likelihood

Here we would follow a similar approach, write out the negative log likelihood as a function $f(\mu; x_i)$ of μ that depends on data x_i. Two useful properties here are:

1. $p(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) = p(X_1 = x_1)p(X_2 = x_2) \cdots p(X_n$
   ,
2. $\log \prod_i f(\mu; x_i) = \sum_i \log f(\mu; x_i)$

Then find a value of μ that minimizes this function. Hint: we saw this when we showed that the sample mean is the minimizer of total squared distance in our exploratory analysis unit!