INTRODUCTION TO DATA SCIENCE

Elias Gonzalez

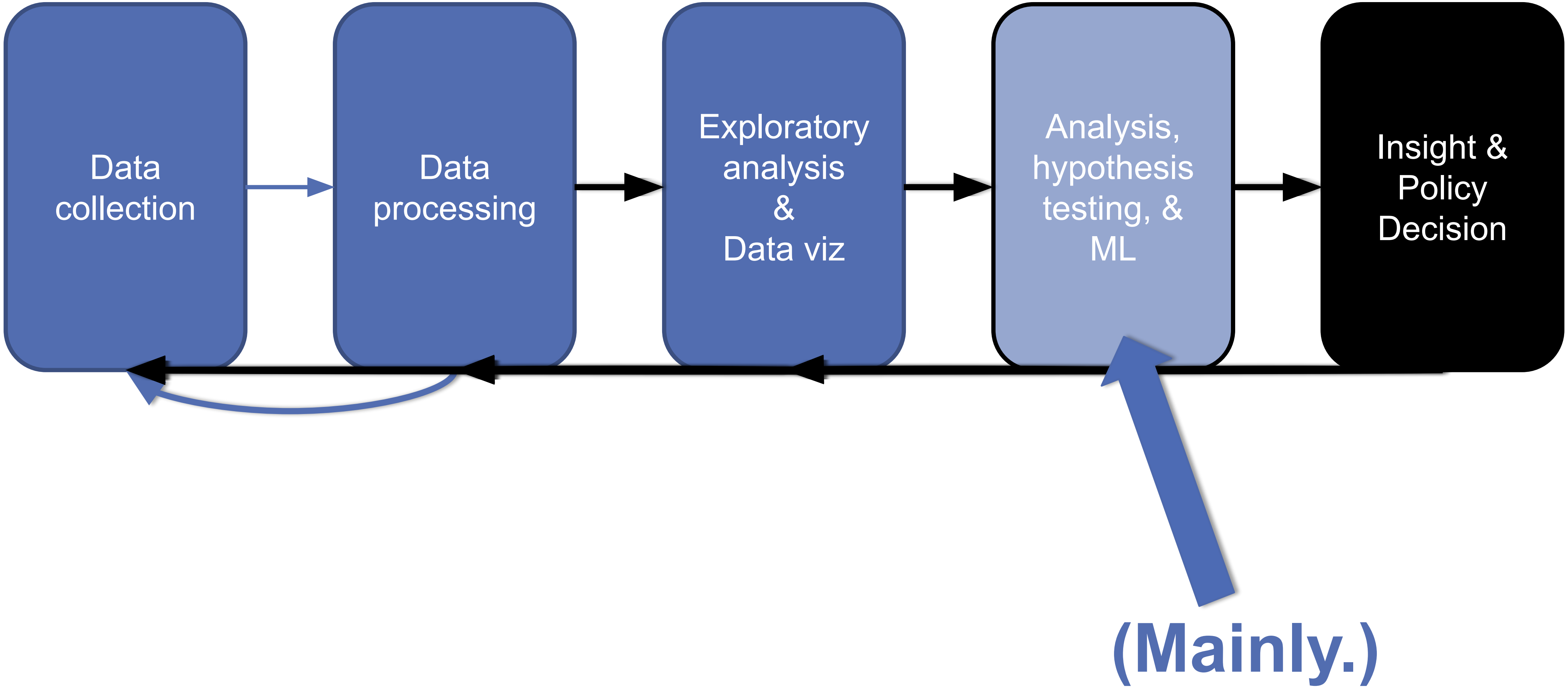
Lecture #14 – 03/01/2022

CMSC320
Mondays & Wednesdays
3:30-4:45

https://cmsc320.github.io/

COMPUTER SCIENCE
UNIVERSITY OF MARYLAND

THE DATA LIFECYCLE



TODAY'S LECTURE

Hypothesis Testing

- What is it?
- What are we looking to prove?
- **How can we use it to justify our hypothesis?**

p-Values

p Hacking

Bayesian Inference

How we use probability in Data Analysis

Probabilistic model of
data generation

<u>Population</u>

has a parameter
of interest

<u>Data</u>

What we observe

Inverse Problem

Infer parameter of
interest from data

How we use probability in Data Analysis

Independent equally likely sampling of tweets

<u>Population</u>

p: Proportion of spam tweets

<u>Data</u>

110 tweet sample

Inverse Problem

$$\hat{p} = \frac{1}{n}\sum_i x_i$$

Confidence Intervals

A confidence interval is way to describe probability.

More mathematically, it is the mean of an estimate minus the variation.

For example, you can construct a confidence interval with a 95% confidence level. In other words you are confident that 95 out of 100 times the estimate will fall between the upper and lower values of the confidence interval.

More concretely

You must know:

- **The point estimate of the confidence interval**
- **Critical values for the test statistic**
- **Standard Deviation**
- **Sample Size**

SampleStatistic ± Margin of Error*

***there are multiple ways of calculating margin of error**

Hypothesis Testing

It's easy to maker claims and hypothesize, but how can we test the likelihood that the hypothesis is true?

H_0- Null Hypothesis that represents the default position
H_1- Alternative Hypothesis that we are comparing to

p-Value: Compute the probability that H_0 is true, that we will see a value at least as extreme as those we observed

p-values

Instead of looking at our data based on some probability cutoff– can we compute the probability assuming that H_0 is true?

Calculated by adding together the probability that random chance generated the data, something else that is equals and something else that is rarer. For normally distributed values, use a density function.

Smaller the p-value the stronger the evidence to reject the null hypothesis.

Typically our p-value cutoff will be 5%

Hypothesis Testing- Flipping a Coin

Let's imagine we want to test to see if a coin flip is fair.

We can assume that the coin has a probability *p* and if we assume that it is fair, it should have a probability of .5

That means:

H_0- p = 0.5

H_1- p ≠ 0.5

Let’s look at this in action

Let’s go back to out Distribution notebook!

p-Hacking

How are p values “hackable” or manipulated?

If we remove enough of the “right” outliers we might be left with data that gets your p-value below the 0.05 threshold.

p-Hacking- Examples

ESP is real?

According to a peer-reviewed study…yes!

How can it be? Start with wanting to show an effect, and reverse engineer a study to get a result without faking anything!

p-Hacking- Examples

How can we avoid p-hacking (or even help us notice it)?

- **Determine your hypothesis *before* looking at the data!**
- **Clean the data without the hypothesis in mind (harder to manage)**
- **p-values are not a substitute for common sense!**
  - Any examples of this?
  - You can’t Trust What You Read About Nutrition

Type of Errors

Type I

- **Incorrectly rejecting the null hypothesis**

Type II

- **Failing to reject a null hypothesis that is false**

Type III

- **Correctly rejecting the null hypothesis but doing so for the wrong reasons**

Examples?

Running an A/B Test

If we have two ads, and we would like to see which is more effective in getting clicks. We can conduct a Z-Test, which will result in a corresponding p-value

Null hypothesis: There is no difference in proportions between the two ads.

```
def estimated_parameters(N, n) -> Tuple[float,float]:
    p = n/N
    sigma = math.sqrt(p*(1-p)/N)
    return p, sigma

def a_b_test_statistic(N-A, n_A, N_B, n_B) ->float: # will return z
    p_A, sigma_A = estimated_parameters(N_A, n_A)
    p_B, sigma_B = estimated_parameters(N_B, n_B)
    return (p_B - p_A)/math.sqrt(sigma_A **2 + sigma_B**2)
```

A/B Test

Ad A had 200 clicks from 1000 views, Ad B had 180 clicks from 1000 views:

```
z = a_b_test_statistic(1000, 200, 1000, 180)
print (z)
```

```
-1.14
```

A/B Test

If we then calculate the resulting p-value:

```
z = a_b_test_statistic(1000, 200, 1000, 180)
p = two_sided_p_value(z)
print (p)
```

```
0.254
```

Large enough that we can conclude there is no difference!

Bayesian Inference

Making a guess based on prior knowledge or data. Use observed data and Bayes theorem for an updated *posterior distribution*.

Rather than making probability judgements about the tests, make probability judgements about the parameters.

Tips for practically using Bayesian Inference:

- **We use beta distribution to model a continuous distribution bounded between 0 and 1**
- **We can then adjust what the distribution is centered around based on the observed data**

Bayes Theorem:

$$P(A|B) = P(B|A)\,P(A)/P(B)$$

Bayesian Inference

Show the examples from the book

Now we can make claims like: “Based on the prior and the observed data there is only a _____ likelihood that the _____ probability is between ____% and ___%.

Which is different from: “If the coin is fair, we would expect to observe data so extreme only 5% of the time”

Likelihood Prior

Posterior

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

Evidence