The Role of Statistics in Big Data

Justin Post

What Do Statisticians Do?

- Understand and account for variability in data
    - Populations & Samples
    - Sampling Distributions and Likelihoods
    - Inferences on the population

Basic Inference Idea

- Statisticians usually consider **populations** and **samples**
- Example:
    - Population - all customers at a bank
    - Parameter - p = proportion of customers willing to open an additional account
    - Sample - Observe 40 *independent* customers
    - Statistic - Sample proportion = $\hat{p} = 8/40 = 0.2$
- Question: Bank makes money if the population proportion is greater than 0.15. Can we conclude that?
- Answer: ?? Is observing $\hat{p} = 8/40 = 0.2$ reasonable if $p = 0.15$ is the true proportion?

Simulating a Sampling Distribution

By simulating this experiment many times, we can understand the sampling distribution of $\hat{p}$

- Assumptions:
  - $p = 0.15$
  - $n = 40$
  - Independent customers

```
import numpy as np
import scipy.stats as stats
import matplotlib.pyplot as plt
```

Simulating a Sampling Distribution

- Where does our value fall in the realm of all possible values?

```
np.random.seed(5)
stats.binom.rvs(n = 40, p = 0.15, size = 1)
```

```
## array([4], dtype=int64)
```

```
stats.binom.rvs(n = 40, p = 0.15, size = 2)
```

```
## array([9, 4], dtype=int64)
```

```
np.random.seed(5)
stats.binom.rvs(n = 40, p = 0.15, size = 1)/40
```

```
## array([0.1])
```

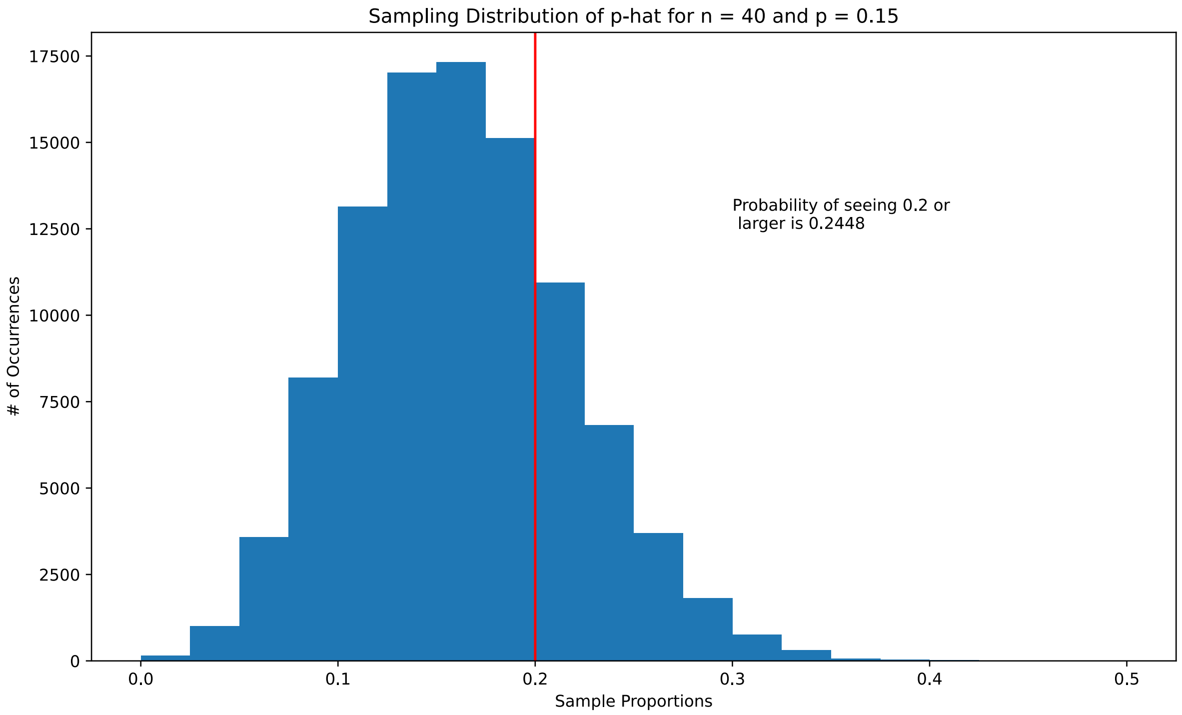
```
stats.binom.rvs(n = 40, p = 0.15, size = 2)/40
```

```
## array([0.225, 0.1  ])
```

Simulating a Sampling Distribution

```
proportion_draws = stats.binom.rvs(n = 40, p = 0.15, size = 100000)/40
plt.figure(figsize = (12, 7))
plt.hist(proportion_draws, bins = [x/40 for x in range(0, 21)])
plt.axvline(x = 8/40, c = "Red")
plt.text(
  x = 0.3,
  y = 12500,
  s = "Probability of seeing 0.2 or \n larger is " + str(round(np.mean(proportion_draws >= 0.2), 4)))
plt.xlabel("Sample Proportions")
plt.ylabel("# of Occurrences")
plt.title("Sampling Distribution of p-hat for n = 40 and p = 0.15")
plt.show()
plt.close()
```

Simulating a Sampling Distribution



Hypothesis Testing

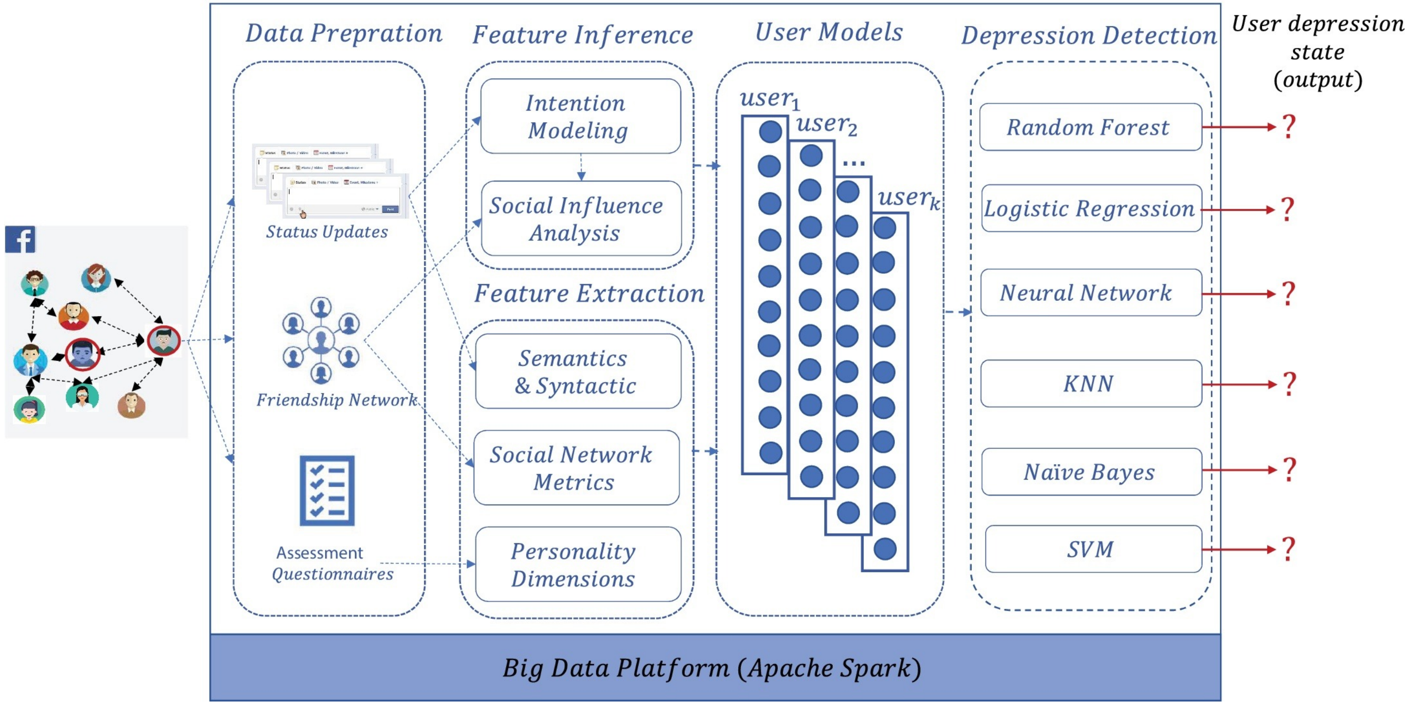
- Logic above is the idea of a hypothesis test
- Assume something about the population
    - Collect data around a quantity of interest
    - Estimate the quantity
    - Use probability to quantify uncertainty in estimate
- If result unlikely to be seen under assumptions, reject assumption

$n =$ all or $n = 1$

- Sometimes we can record every user action… don't we have everything?
  - Is there any variability to consider?
  - Is our sample size the population size? **n = all**

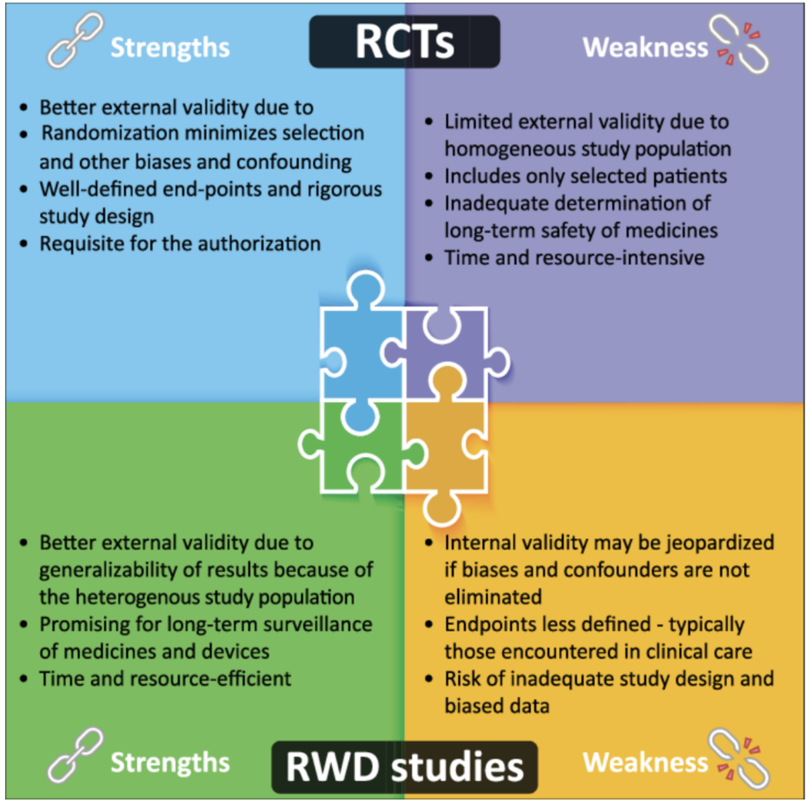
$n = $ all or $n = 1$

- Can now consider **user-level** (or observational unit level) modeling!
  - Example modeling user intention on social media networks to detect depression



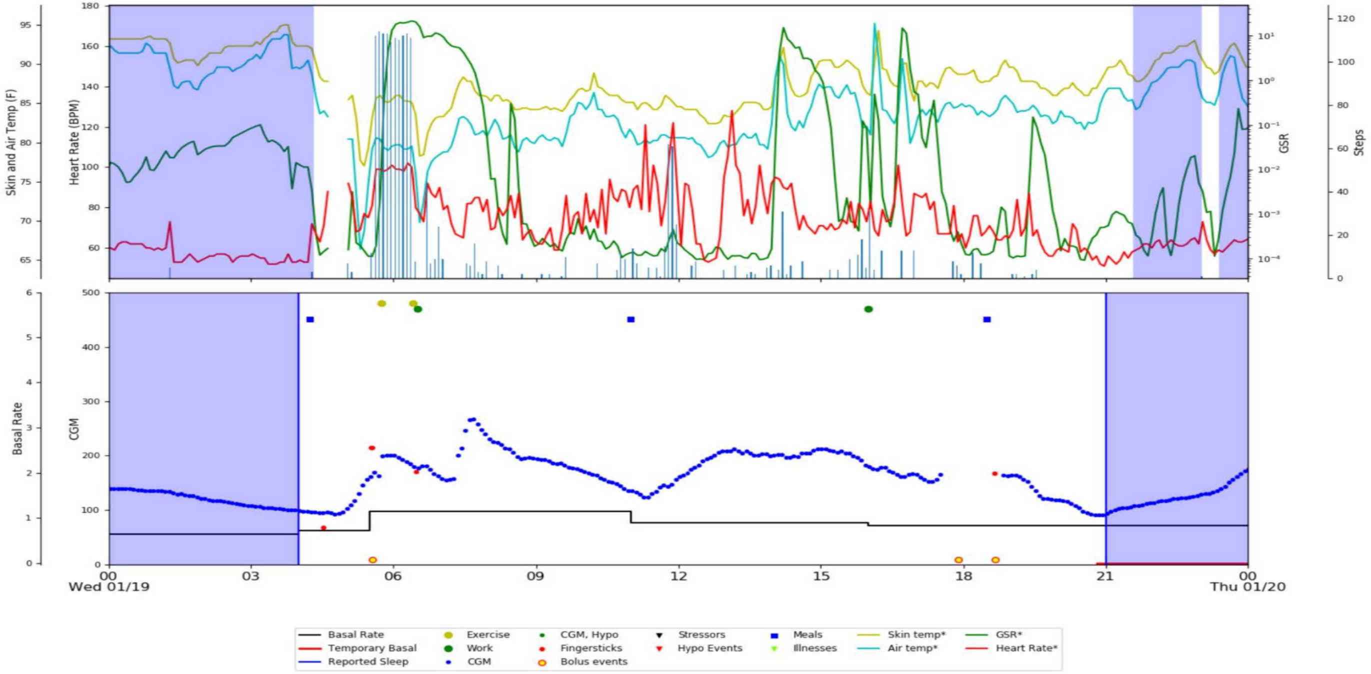
What Do Statisticians Do?

- Carefully consider data sources and **bias**
    - Combining data sets
    - Understanding data quality
    - Causal relationships



What Do Statisticians Do?

- Model data
    - Define assumptions, model structure, and relationships
    - Investigate behavior
    - Provide error measurements



Modeling Big Data

- Explaining variable importance (Random forests, Deep learning)
- Understanding how models relate (Trees as MLR models, a framework for penalized regression)
- Updating models with streaming data

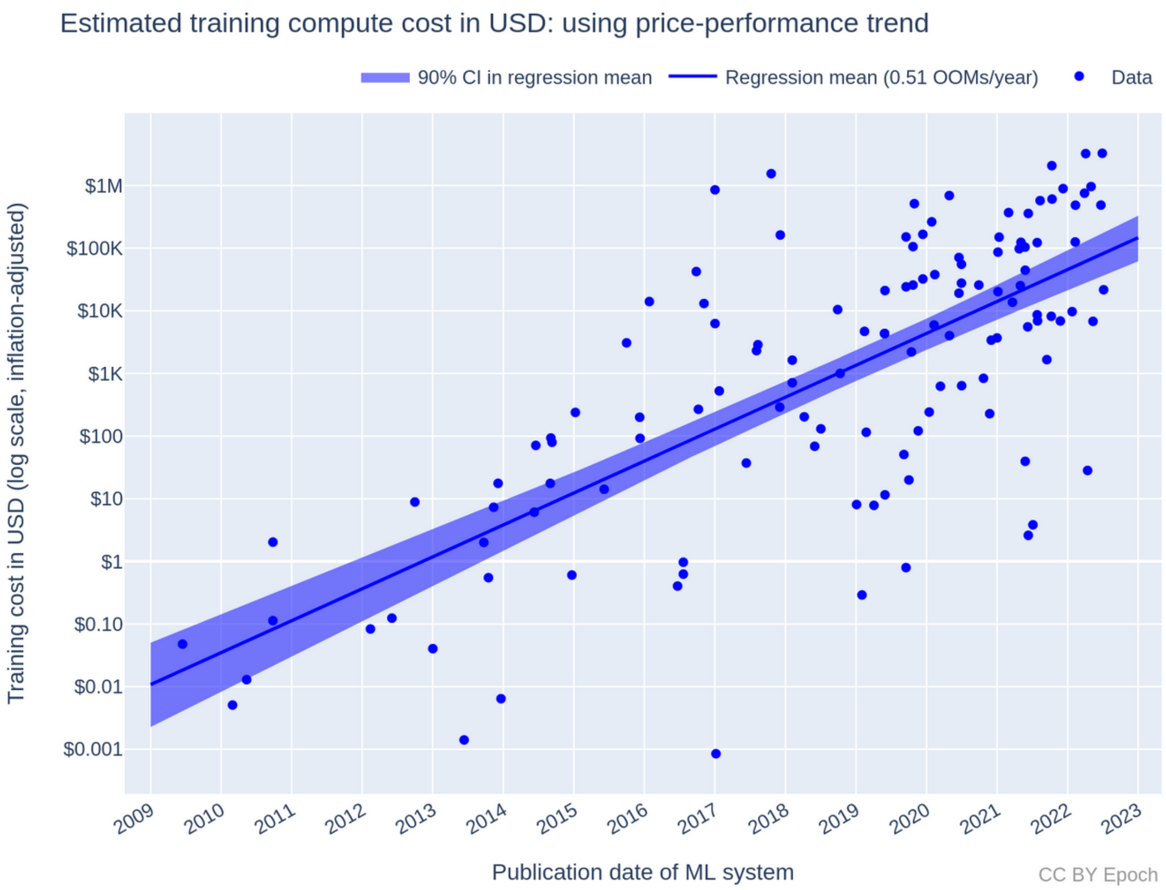
$$\tilde{\beta}_b^{(r+1)} = \tilde{\beta}_b^{(r)} + \{\tilde{S}_b^{(r)\mathrm{T}}(\tilde{V}_b^{(r)})^{-1}\tilde{S}_b^{(r)}\}^{-1}\tilde{S}_b^{(r)\mathrm{T}}(\tilde{V}_b^{(r)})^{-1}\tilde{U}_b^{(r)},$$

where we do not need to access the entire raw dataset except for the observations in the current batch $\mathcal{D}_{ib}$ and the last observation in data batch $\mathcal{D}_{i,b-1}$. Instead, we use

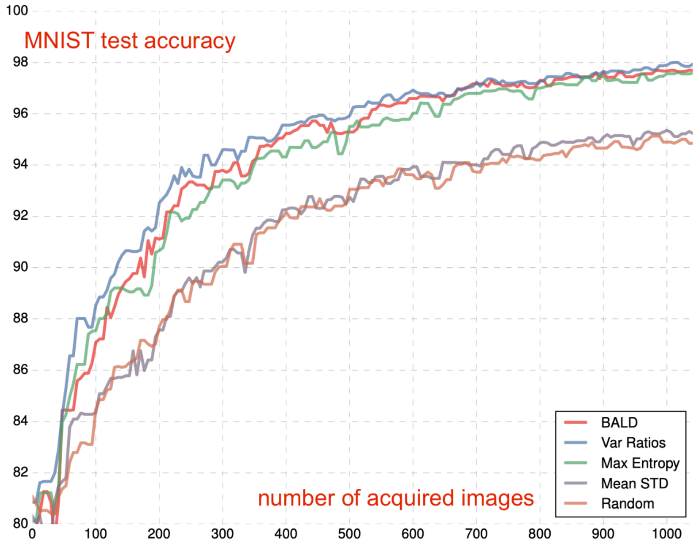
https://academic.oup.com/biomet/article/110/4/841/7048657

What Do Statisticians Do?

- Consider how to be smarter with data



Thinking Critically About Models



- Statistical accuracy and computational cost trade-off
    - Active Learning - which data to acquire (DOE) and causal relationships
    - Coresets - a small, weighted subset of the data, that approximates the full dataset
    - Divide and conquer algorithms

What Do Statisticians Do?

- Understand randomness and rare events
- If you have enough data, you'll eventually see weird things just by chance (similar to multiple testing idea in hypothesis testing)

What Do Statisticians Do?

- Understand randomness and rare events
- If you have enough data, you'll eventually see weird things just by chance (similar to multiple testing idea in hypothesis testing)
- Rare Events & Expected Numbers
  - Suppose we have an event that occurs with probability p
  - We run k different independent experiments

$$P(\text{At least 1 occurrence}) = 1 - (1-p)^k$$

  - We would expect to see the following number of occurrences of the event

$$E(\#\text{ of occurrences}) = k * p$$

Rare Events Example

- Suppose you have an app that screens phone calls for people

$$P(\text{Detected}|\text{Spam}) = 0.99999$$

$$P(\text{Detected}|\text{Non-spam}) = 0.00002$$

And generally, you know that

$$P(\text{Spam}) = 0.2, P(\text{Non-spam}) = 0.8$$

Rare Events Example

- Given a call is detected as spam, what is the probability it wasn't a spam call?

$$P(\text{Non-spam}|\text{Detected}) = \frac{P(\text{Detected}|\text{Non-spam})P(\text{Non-spam})}{P(\text{Det}|\text{Non-spam})P(\text{Non-spam}) + P(\text{Det }|\text{Spam})P(\text{Spam})}$$

$$= \frac{0.00002 * 0.8}{0.00002 * 0.8 + 0.99999 * 0.2} = 0.00008$$

- Our event of interest: Given a call is detected as spam, we were wrong has a tiny probability of happening!

Consider This as a Function of the Number of "Trials"

# of calls flagged as spam	P(At least one mistakenly flagged call)	Expected Number of Mistakes
1	0.00008	0.00008
100	0.007968	0.008
1,000	0.076887	0.08
10,000	0.550685	0.8
100,000	1	8

Recap

Although big data has a lot of info, statisticians help us extract that info in a meaningful way!

Some things statisticians do:

- Understand and account for variability in data
- Carefully consider data sources and bias
- Model data
- Consider how to be smarter with data
- Understand randomness and rare events