SESSION 1: THE ESSENCE OF STATISTICS

Session 1

Moneyball

What is data?

- Data is a generic word that covers almost any type of information that you obtain.
- Data can come in different forms.
  - Quantitative versus Qualitative: Quantitative data is numerical. Qualitative data is not, though it can sometimes be converted into numerical form. (Example: Pain scale)
  - Discrete versus Continuous: Discrete data can take only certain values (integer or otherwise). Continuous data take any value, though it can have upper or lower bounds.
- While we have always relied on data for decision making, we have become much better at finding and chronicling the data, collecting and storing it, and making it accessible to a broader audience.

What is statistics?

- The Data Science: Statistics is the discipline that allows us to gather, analyze, depict and make sense of data.
- Data -> Information: Statistics is what helps us convert raw data, noisy and often contradictory, into information that can be used to make better decisions.
- Multifaceted: Defined as such, statistics includes
  - The collection and recording of data
  - The processing of that data
  - The analysis of the data
  - The depiction (presentation) of that data

Why do we need statistics?

1. As data access widens and becomes easier: We have access to more data, quantitative and qualitative, than we have ever had, because of innovations in how we collect data, how we store the data and how we access the data.
2. Data overload is a problem: That access to data is leading to information overload, where we find that the line between facts and opinion is blurred, leaving us more confused than we ever have been.
3. Leading to simplistic rules of thumb: As our brains start to shut down in the face of data overload, we seek out rules from simpler times, many of which have no basis in truth, but make use feel in control.
4. And mental short cuts: We also adopt short cuts, again with no rational basis, just because they have worked in the past.

Statistics drives practice, policy and laws…

- In almost every aspect of our lives, practice and policy is determined by statistics.
  - In fiscal policy, governments decide taxation and spending, based upon statistical assessments of their effects on the economy.
  - In health care, questions of what drugs should be approved and what treatment patients should get is based upon statistics (often in medical research)
  - In our personal lives, our choices of where to work, live and how to save/invest are at least loosely driven by statistics.
- Making good policy and personal decisions requires an understanding of statistics and data. If we misread the statistics or the statistics are unreliable, policy will as well.

Lying with statistics?

- Agenda-driven data: As access to data has increased, the misuse of that data has also gone up, especially when people have agendas and want to further them. These people mislead, without technically lying, as the selectively pick and choose which data they use, and how they present that data.
- Social media as magnifier: Bad data can take on a life of its own, especially with social media operating as a weapon to widen its reach.
- Caveat emptor: As people weaponize data and use selective and slanted statistics, based upon that data, we need to be able to protect ourselves from misinformation. Understanding statistics allows us to:
  - Look for red flags that can be used to detect data manipulation
  - Asking the right questions to separate fact from fiction

Big Data and Data Analytics

- Big Data: Big data is a catch-all that references not just the quantity of data, but also that the data is about behavior that until a few years ago would have been considered private and in real time. (Think location data on your phone, as a simple example). For better or worse, big data has been facilitated by technology platforms, which collect the data, as we use them.
- Data Analytics: All statistics is data analytics, but in the context of big data, data analytics has expanded to cover
  - Very large datasets, creating sample sizes that we have never worked with in the past.
  - Combining quantitative with qualitative data to create composite results.

1. Data Collection and Sampling

- Data collection & storage: The very first exercises in data collection required interactions with subjects, paper documents and physical storage. As technology has evolved, data collection and storage has been digitized, and the risks have shifted.
- Population versus Sample: With large populations, you often have to sample that population, either cross sectionally, across time, or both.
- The Sampling Sins: While sampling, by itself, is part of statistics, there are two fundamental sins that you should try to avoid:
  - *Bias*: The sample has to be representative of the population. If the sampling method creates bias, the results from the sample cannot be extrapolated to the population.
  - *Noise*: Even if the sample is representative, the results that you obtain will have statistical error or noise that can muddy your conclusions.

2. Data Descriptive(Statistic)s

- When faced with sample data, data descriptives try to summarize the data with metrics. Those metrics can include
  - Measures of location: Measures of location try to identify the number or numbers around which the data is centered or is most likely to take.
  - Measures of dispersion: Measures of dispersion measure how much divergence there is on a data item, across a sample.
  - Measures of skewness: Measures of skewness look at whether how symmetric or asymmetric the data is around the central value.
- These data descriptives become short hand for characterizing the data, and the basis for analysis.

3. Data Distributions

- A data distribution is a visual description of the data item in a distribution. In a histogram, for instance, you break the data down into groupings and count the outcomes of each grouping.
- While you can convert any data into a histogram, there are sometimes advantages in replacing this histogram with one of the many standardized descriptions in statistics that fits it.
- These standardized distributions can be
  - Discrete or continuous: A discrete distribution puts the variable into discrete values (0 or 1, numerical scale, bond ratings). A continuous distribution allows the variable to take any value.
  - Symmetric or asymmetric: A symmetric distribution has a center and symmetry on the up and down sides.

4. Data Relationships

- In statistics, we can also examine how two or more variables are related to each other.
  - In some cases, we restrict ourselves to *chronicling whether that co-movement* is in the same direction, opposite directions and that there is no co-movement at all.
  - In others, we try to find *whether there is causation*, where one variable’s movement is the cause of the other variable’s movement.
  - Finally, if there is a link, we can use statistics *to try to predict a variable*, based upon observed values of another variable.
- As an example, consider the linkage between stock prices and interest rates.
  - We can measure whether interest rates and stock prices move together, in opposite directions and are unrelated.
  - We can also examine which direction the causation runs (do higher stock prices cause higher interest rates or vice versa)
  - And if there is causation or a link, we can see if we use changes in interest rates can be use to predict changes in stock prices or vice versa.

5. Probabilities

- Probabilities measure the likelihood of something happening and are at the heart of statistics and data analysis.
- The ”something happening” can be a discrete event or a variable that comes from a continuous distribution.
  - Discrete event: That a company can go bankrupt or stay a going concern.
  - Continuous variable: The likelihood that a company will generate earnings that exceed $1 billion.
- You can estimate the probability of an event (discrete and continuous) happening by
  - Looking at the past or historical data.
  - Assuming a statistical distribution for the data, and that you can estimate the parameters of that distribution.

6. And Probabilistic Tools

- Probit/Logit: A probit/logit measures the likelihood or probability of an event happening, based upon observable variables.
  - Example: Measuring the probability that a company will be acquired, given variables that you believe are correlated.
- Decision Trees: When you have a series of discrete events that are sequential, a decision tree allows you to compute the likelihood of events happening, conditional on events leading up to it.
  - Example: The process by which an FDA drug works its way through the drug approval pipeline, from research to commercial development.
- Monte Carlo Simulations: When you have multiple inputs that determine an output variable, a simulation allows you specify probabilistic distributions for each input variable and get a distribution for the output variable.
  - Example: Valuing a company, given its cash flows, growth and risk characteristics.