INTRO TO STATS

09/07/18

Statistics?

- The what:
  - *A branch of mathematics*
  - *Collection, organization, analysis, interpretation, and presentation of data*
- The why:
  - *Applications broadly to any industry (financial or non-financial)*
  - *Trend towards big data (which is not classical statistics) and predictive analytics*
  - *A method of understanding the world better; perspective is important when understanding statistics that is being presented*
  - *Statistics is as much as an art as it is a science*

Probability

- 3 definitions of probability:
  - ***Classical / Theoretical:*** *what are the odds of rolling a 1 on a fair die?*
  - ***Empirical:*** *a study has shown that a weighted coin has 623 heads out of 1000 flips; what are the chances of the next flip being a heads?*
  - ***Subjective:*** *"I think that Tesla has a 30% chance of actually going private"*
- Komolgorov axioms of probability:

$$
\begin{aligned}
&(1)\ P(A) \geq 0 \text{ for all } A \subset S\\
&(2)\ P(S) = 1\\
&(3)\ \text{If } A \cap B = \emptyset,\\
&\quad\ \text{then } P(A \cup B) = P(A) + P(B)
\end{aligned}
$$

Conditional Probability & Bayes Theory

- Conditional probability:
  - *New universe space defined due to a certain event occurring*
  - $P(A|B) = \frac{P(A\cap B)}{P(B)}$*; with P(B) being the new universe*
- Independence: $P(A) = P(A|B)$
- Bayesian theory:
  - *Conditional probability of evidence occurring provides additional information on the hypothesis itself*
  - $P(H|E) = \frac{P(E|H)}{P(E)} \cdot P(H)$
  - *Note that* $\frac{P(B|A)}{P(B)}$ *is known as the* ***likelihood ratio***

Fundamental statistics

- Mean:
  - $E[X] = \frac{1}{n}\sum n$
  - *Measure of* ***central tendency;*** *also referred to as the (long-run) average*
- Standard deviation / Variance:
  - $Var(X) = \frac{1}{n}\sum(x-\mu)^2; s.d. = \sqrt{Var(X)}$
  - *Measure of dispersion around central tendency*
  - *Variance reflects the sum of squared deviations (sum of deviations from mean itself is always 0, i.e.* $E[X-\mu] = E[X] - \mu = 0$*)*
  - *Standard deviation is in the same units as the underlying data set*

Standard deviation of samples

- Usage of samples:
  - *When entire population is infinite, or finite but too large to be observed in entirety, samples are used to provide information of the population*
  - *Sample selection can be random or non-random*
  - *Sample is supposed to represent a slice of the population*
- Unbiased estimate of population mean / s.d.:
  - *As implied, you are using the statistics from the sample to infer/estimate the statistics of the population*
  - *Unbiased estimate of pop. mean μ is* ***equivalent*** *to sample mean $\bar{x}$*
  - *Unbiased estimate of pop. variance*

Distributions

- Discrete:
  - *Random variable can only take on* ***discrete, finite number of values***
  - *E.g. Bernoulli, Binomial, Geometric, Hypergeometric, Poisson, etc.*
- Continuous:
  - *Random variable can take on an* ***infinite range of values;*** *note this* ***does not*** *mean the range of the distribution itself has to be infinite*
  - *E.g. Gaussian, Exponential, Gamma, Chi-squared, etc.*
- Komolgorov Axioms:
  - *Whether discrete or continuous or a mix, a distribution must satisfy Komolgorov's axioms*
  - *Most importantly, the event space of the distribution (discrete summation for discrete, integral summation for continuous)* ***must equal to 1***

Central Limit Theorem

- Clarifications on definition:
  - *CLT applies to iid. distributions as a sample*
  - *Given sufficient observations of iid. distributions in a sample, the* ***sample mean distribution*** *approximates a normal distribution*
  - *Note that CLT does not provide any information on the original distribution itself; original distribution can be both discrete or continuous*
  - $f_x(x)$ *has mean* $\bar{x}$ *and s.d.* σ^2

Additional tidbits

- Normal approximations:
  - *Under certain circumstances, discrete distributions (e.g. Binomial, Poisson) can be approximated to a Normal distribution*
  - *If so**, continuity correction** is required to account for the differences between a discrete vs. continuous distribution*
- Hypothesis testing:
  - *Hypothesis testing uses a Bayesian approach to obtain a conclusion*
  - *My way of thinking about hypothesis testing is, if the result of the sampling is beyond the critical value, the probability of me randomly obtaining such a result is too small for it to be purely by chance, therefore another factor (e.g. the initial hypothesis being not true) is most likely the cause, and so I reject the null hypothesis*