**What Statistics To Learn For Data Science**

**Statistics:** A statistic (singular) or sample statistic is any quantity computed from values in a sample which is considered for a statistical purpose.

* In other words, a statistic summarizes information about some given data, sample or population.

**Summary Statistics:** Summary statistics generally measure four things: location, spread, shape, and dependence.

* **Mean | Mode | Median**
* **Variance | Standard Deviation | Coefficient of Variation**
* **Skewness | Kurtosis**
* **Percentiles | Quartiles | Interquartile Range**
* **Spearman’s | Pearson’s Correlation Coefficient**

**Visualizations:** Data scientists must be able to visually present their data and findings to find the best way to present your results to stakeholders and colleagues.

* **Bar Chart | Line Graph | Pie Chart | Scatter Plot | Violin Plots | Histograms | Frequency Diagrams | Box and Whisler Plot | Heat-maps | Contours**

**Probability Distributions:** When people think of statistics, they will probably think of distributions. Probability distributions help us describe statistical events and understand the frequency and probability of specific outcomes from these events. Their primary critical use in data science is to help us understand the type of relationship between our target variable and its dependent variables (commonly known as features). This ensures we choose the most suitable model for the task to maximize performance.

* **Normal Distribution**
* **Poisson Distribution**
* **Binomial Distribution**
* **Gamma Distribution**
* **Exponential Distribution**
* **T-Distribution**
* **Chi-Square Distribution**
* **Probability Density Function**
* **Cumulative Distribution Function**

**Probability Theory:** Probability theory encompasses the whole mathematical modeling side of how probability works. This large area is sometimes referenced as separate from statistics due to its size and overlaps with some of the places I will or have discussed in this article thus far.

* **Random Variables**
* **State Spaces**
* **Samples, Populations, and Standard Error**
* **Central Limit Theorem**
* **Law of Large Numbers**
* **Maximum Likelihood Estimation** (very important for machine learning, as it explains where loss functions come from)

**Hypothesis Testing:** How do you know if a result is significant or just some random noise? Statistical testing, commonly known as hypothesis testing, is used to determine this. Perhaps the most famous example in the data world is an AB test, which is used constantly by literally every company nowadays

* To understand hypothesis testing, you need to know the following concepts:
* **Confidence and Prediction Intervals**
* **Significance Levels, Critical Values, and P-values**
* **One vs Two-Tailed Test**
* **Null and Alternative Hypotheses**
* **Test Statistics**
* **Z-Test**
* **T-Test**
* **Chi-Square Tests (both types)**
* **ANOVA Test**
* Just learn the process and intuition behind how hypothesis testing is carried out. You also should understand when to use certain tests over others in different scenarios.

**Regression Analysis:** The first algorithm a data scientist typically learns about is linear regression. Despite machine learning being a relatively new field, linear regression dates back to the early 1800s, so it is quite an old statistical technique.

* Linear regression is part of a broader area called regression analysis, which is all about estimating the relationship between a target variable and a set of features (known as covariates)
* Many models and methods in regression analysis are still used today and are very effective.
* As well as regular linear regression, you should learn the following:
  + **Multivariate Linear Regression**
  + **Polynomial Regression**
  + **Generalized Linear Models**
  + **Generalized Additive Models**
  + **Gauss-Markov Theorem**
  + **Ordinary and Least Squares Estimation**

**Bayesian Statistics:** There are two main branches of thinking in statistics: frequentist and Bayesian. Most people “do” statistics in a frequentist framework, as that’s the way most people are taught in schools. It’s also a more accessible system to work with.

* However, biologically, humans tend to think and work in a Bayesian mindset, and so this theory of statistics has been successfully applied in many experiments in the data science field. Many optimization algorithms in machine learning rely on a Bayesian approach instead of a frequentist one.
  + **Marginal, Joint, and Conditional Probability**
  + **Bayes’ Theorem**
  + **Bayes’ Factor**
  + **Conjugate Priors**
  + **Bayesian Updating**
  + **Credible Intervals**
  + **Bayesian Regression**

**Stochastic Process:** A stochastic process is a sequence of events (typically time-indexed) of random variables. Stochastic processes model many phenomena, such as how water molecules move to the stock market price.

* The following areas are the fundamentals of stochastic processes:
  + **Markov Property**
  + **Markov Chains**
  + **Hidden Markov Models**
  + **Random Walks**
  + **Geometric Brownian Motion**
  + **Ito Calculus (Extremely Advanced)**