# Initial Posts

|  |
| --- |
| **Kernel Density Estimation (KDE)**  Kernel Density Estimation (KDE) is a function that creates a probability area over each data point. Any overlap between each data point’s probability area is then added up. This visual representation of the data is often used to visualize the distribution of some variable.    The area calculated above each data point does not have to be smoothed and can take on multiple different shapes. The area calculated above the points is determined by the Kernel function used.    The bandwidth determines how biased the KDE will be. The higher the bandwidth the higher the bias which in effect will create a smoother line. The lower the bandwidth, the higher the variance will be which in effect will create a line with a lot of noise. The key for bandwidth, much like with any other predictive model, is to find the model with the best bias-variance tradeoff. For KDE, there are multiple different functions to dynamically calculate the optimal bandwidth, the most common being the L2 Loss Function.    *L2 Loss Function*  Reference:  *Intro to Kernel Density Estimation - YouTube. (2022). Retrieved 16 January 2022, from* [*https://www.youtube.com/watch?v=x5zLaWT5KPs&ab\_channel=webelod*](https://www.youtube.com/watch?v=x5zLaWT5KPs&ab_channel=webelod)  *Kernel density estimation. (2022). Retrieved 16 January 2022, from* [*https://en.wikipedia.org/wiki/Kernel\_density\_estimation*](https://en.wikipedia.org/wiki/Kernel_density_estimation) |
| **Exponential Distribution**  An Exponential distribution is usually seen with data that tracks time between events. It is very similar to a Poisson distribution which usually is seen with data that tracks total events between time. For example, in the Emergency Room, if you were to track the number of patients who visit the E.R. per hour, that would be a Poisson distribution but if you were to track the number of hours in between patients arriving, that would be an Exponential distribution. As an example, let’s say that on average, 10 patients visit an E.R. per hour (10 pats/hr). If we were to invert this and look at hours per patient, that would be 1/10 of an hour per patient (.10hrs/pat). You can determine if a distribution is exponential statistically by taking the log of the complimentary of the CDF of the data and then plotting the line. If the line is straight, it would mean that the data is exponentially distributed.  Reference:  Downey, A. B. (2015). *Think Stats: Exploratory Data Analysis*. O'Reilly. |
| **Normal Distribution**  A normal distribution is probably one of the most well known distributions mainly due to the Central Limit Theorem which states that the means of samples taken from any population distribution will follow a normal distribution. "Also known as the Gaussian distribution, it is a probability distribution that is symmetric about the mean, showing that data near the mean are more frequent in occurrence than data far from the mean." (Ivestopedia). A standard normal distribution is a normal distribution with a mean = 0, and a standard deviation = 1. To enforce these parameters on your distribution, you can subtract each values of the distribution by the mean of the distribution and then divide that value by the standard deviation of the distribution.  Reference:  Normal Distribution. (2022). Retrieved 16 January 2022, from <https://www.investopedia.com/terms/n/normaldistribution.asp> |
| **Standard Deviation**  Standard Deviation is a measure of how dispersed the data is around it’s mean. If the standard deviation is low, this would give an assumption that if a new data point were to pop up, that it would more than likely be close to the mean of the data. If the standard deviation is high and a new data point were to pop up, we would be less certain that it would be close to the mean of the data. |
| **Probability Density Function (PDF)**  A Probability Density Function (PDF) is derived from a Cumulative Density Function (CDF) by taking the derivative at each value of CDF(x). By getting the derivative of the CDF, what does this actually show us? Well, if we take the derivative of the CDF at some value of (x), we are getting the rate of change for each value of CDF(x). A higher (steeper) rate of change on a CDF function tells us that the data has a higher frequency around this point. A lower (flatter) rate of change on a CDF function tells us that the data has a lower frequency around this point. As a matter of fact, when looking at the y-axis for a PDF function, you are actually looking at the rate of change for the PDF’s equivalent CDF function. As a note, because taking the derivative of a CDF = PDF, taking the integral of the PDF = CDF.      Reference:  Probability Distribution Functions (PMF, PDF, CDF) - YouTube. (2022). Retrieved 16 January 2022, from <https://www.youtube.com/watch?v=YXLVjCKVP7U&ab_channel=zedstatistics> |

# Replies

|  |
| --- |
| Justin, sounds like a great marketing strategy to me haha. Put a higher weight just within the range where if a scientific experiment were ran that it would be within the 95% confidence range. Now I'm wondering how often something like this occurs? Thinking |
| Good post Saima. A lot of these distributions are familiar to me but am just now learning the technical name of what they are. Also, thanks Chandrasekhar for the additional info. Will have to bookmark both of these! |
| Myranda, I too found the use cases for the Pareto distribution to be very interesting but also make a lot of sense. Good post! |
| Great post and replies. Very helpful information and I now have a better understanding of what a Central Moment is and what they are used for in statistics. |
| Good post Chandrasekhar. Good explanation of what moments calculate different statistics about the distribution. The link you provider is also very helpful as well. |