What is an Empirical Cumulative Distribution Function?

An [empirical cumulative distribution function (CDF)](http://en.wikipedia.org/wiki/ECDF) is a [non-parametric](http://en.wikipedia.org/wiki/Nonparametric) estimator of the underlying [CDF](http://en.wikipedia.org/wiki/Cumulative_distribution_function) of a random variable.  It assigns a probability of 1/n to each datum, orders the data from smallest to largest in value, and calculates the sum of the assigned probabilities up to and including each datum.  The result is a [step function](http://en.wikipedia.org/wiki/Step_function) that increases by 1/n at each datum.

The empirical CDF is usually denoted by \hat{F}_n(x) or \hat{P}_n(X \leq x) , and is defined as

\hat{F}_n(x) = \hat{P}_n(X \leq x) = n^{-1}\sum_{i=1}^{n} I(x_i \leq x) 

I() is the [indicator function](http://en.wikipedia.org/wiki/Indicator_function).  It has 2 possible values: 1 if the event inside the brackets occurs, and 0 if not.

I(x_i \leq x) = \begin{cases}  1,&\text{when }x_i \leq x\\  0,&\text{when }x_i > x  \end{cases}

Essentially, to calculate the value of \hat{F}_n(x) at x,

1. count the number of data less than or equal to x
2. divide the number found in Step #1 by the total number of data in the sample

Why is the Empirical Cumulative Distribution Useful in Exploratory Data Analysis?

The empirical CDF is useful because

* it approximates the true CDF well if the sample size (the number of data) is large, and knowing the distribution is helpful for statistical inference

**Fhat is a good estimator for F**

* a plot of the empirical CDF can be visually compared to known CDFs of frequently used distributions to check if the data came from one of those common distributions
* it can visually display “how fast” the CDF increases to 1; plotting key[quantiles](http://en.wikipedia.org/wiki/Quantile) like the [quartiles](http://en.wikipedia.org/wiki/Quartile) can be useful to “get a feel” for the data