The Power of Histograms

Cristian Vava, PhD

The main goal of this white paper is to show how histograms could be used to draw important information about a set of data going sometimes beyond what standard statistical formulas could do. The paper will show common examples of (pseudo) random variables with non-normal distributions and the estimation errors created when deriving their statistical properties based on the assumption of normality, especially for low probability events.

Let's start the analysis with a set of pseudo-random numbers having an almost normal statistical distribution. The numbers used below have been generated in Excel using the internal random number generator with uniform distribution and a Box-Muller transformation. Figure 1 below shows the histogram of both the empirical distribution (ED) derived from the raw data (GAUSS) and its equivalent normal distribution (EN) built from the average and standard deviation parameters based on the maximum likelihood estimation model (unbiased estimator of the population variance with the Bessel correction):

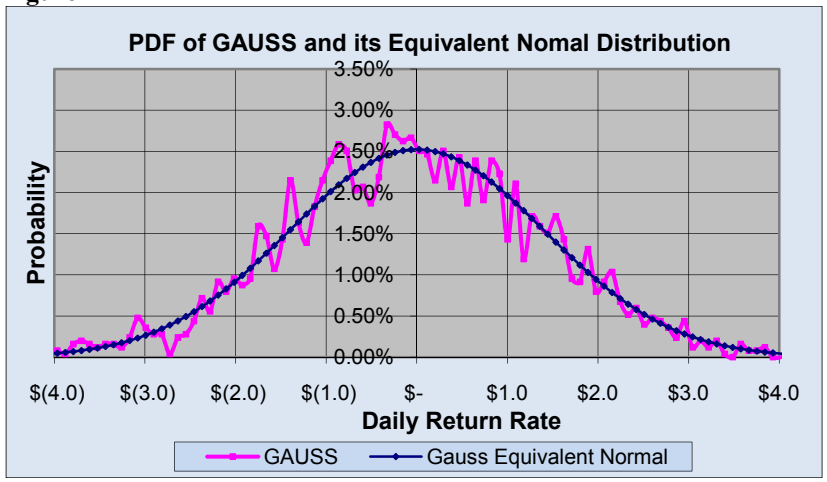
$$S = \sqrt{\frac{1}{N-1} \cdot \sum_{i=1}^{N} (x_i - \bar{x})^2} \qquad (1)$$

Here x_i represent the samples, N the total number of samples, and $\bar{x}$ the average of samples determined as:

$$\bar{x} = \frac{1}{N} \cdot \sum_{i=1}^{N} x_i \qquad (2)$$

Figure 1 below shows that in case of a normal distribution EN (labeled GAUSS in this case) is a good approximation for ED.

Figure 1

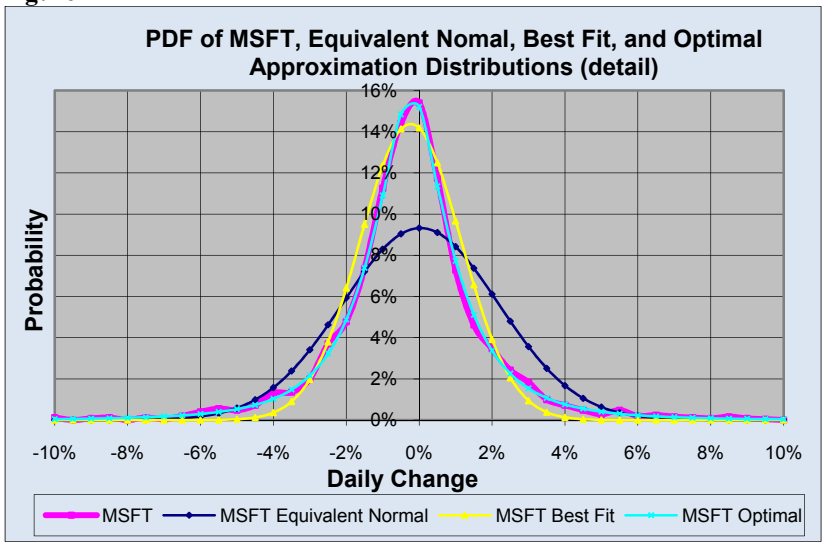


However this is not necessary the case if the empirical distribution is far from normality and Figure 2 below will show such an example. In this last case the raw data represents the daily change of Microsoft (MSFT) stock price over an entire decade (August 2000 to August 2010). The equivalent normal distribution as shown in Figure 2 is obviously inappropriate for many practical purposes. However we could build another normal distribution (MSFT Best Fit) that represents a better fit although not based on the maximum likelihood estimator. Even better we also could build a so called Optimal Approximation estimator which represents the lowest degree exponential distribution matching the original empirical distribution.

Let's now take a more detailed look into the differences among the three representations used in Figure 2. Table 1 below shows

the probability to get a daily change rate of 0%, 1%, 2%, 3%, and 10% as predicted for each representation.

Figure 2



The more complex representations offer a better estimation justifying in most cases the supplementary computational cost.

Table 1

Daily Return / Distribution	0%	1%	2%	3%	10%
Empirical	15.4%	7.3%	3.5%	1.8%	0.04%
Equivalent Normal	9.3%	8.4%	6.1%	3.6%	$2x10^{-4}$ %
Best Fit Normal	14.2%	9.7%	3.9%	0.9%	$2x10^{-11}$ %
Optimal Approximation	15.2%	7.7%	3.4%	1.5%	0.04%

However the real value of a sophisticated histogram analysis could be seen in the case of very low probability events. As Table 1 above indicates for a 10% daily change the empirical probability is 0.04% but the equivalent normal distribution predicts it at $2x10^{-4}$ %. At such a low probability it shouldn't come as a surprise that people start seeing black swans. It is only a common case of a wrongly fitted model and the Optimal Approximation shows that a much more suitable estimator is feasible.

Since we could build such a closely fit estimator a natural next step is to determine the range of possible values. The range limits have an exponential form but obviously can't represent distributions because the area under their curves is different than 100%.

Figure 3

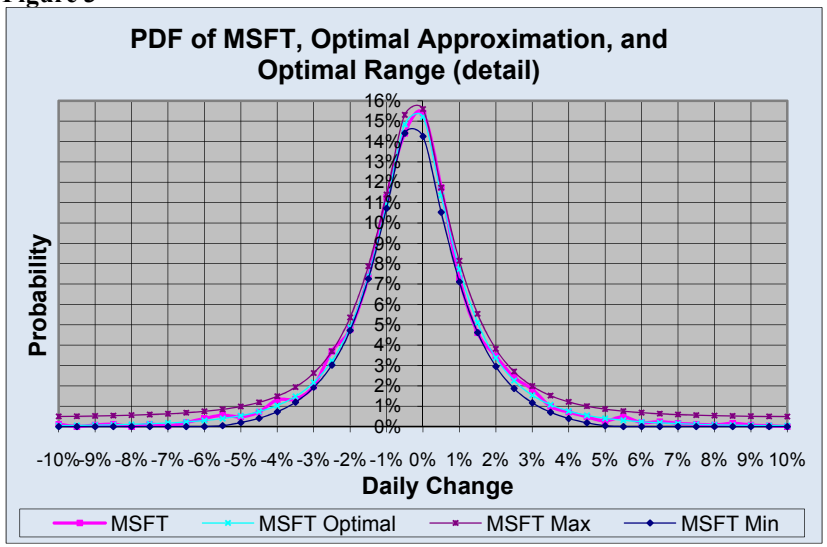


Table 2 shows some details about how accurately the Optimal Approximation and the Optimal Range describe the empirical

distribution. In all cases the range opening was under 1.6% and the Optimal Approximation was accurate within 0.5%.

Table 2

Daily Return / Distribution	0%	1%	2%	3%	10%
Empirical	15.4%	7.3%	3.5%	1.8%	0.04%
Maximum	15.6%	8.1%	3.8%	2.0%	0.5 %
Minimum	14.3%	7.1%	3.0%	0.4%	0.0 %
Optimal Approximation	15.2%	7.7%	3.4%	1.5%	0.04%

When deciding whether to use the assumption of normal distribution a good rule of thumb is to check the statistical moments of higher order starting with the skewness and kurtosis. If all higher order moments are small then the normal distribution could represent a good estimation. The two data sets from our examples had moments as described by Table 3 below.

Table 3

Parameter	GAUSS	MSFT
Average	0.001	0.0003
Standard Deviation	1.4	0.002
Skewness	0.03	0.07
Kurtosis	0.08	7.1

One can use such an optimal range to estimate the financial risk if the original data describes investments. It could be also used in the pharmaceutical industry, the insurance industry, and many other places where there is a need to know either a more accurate description of the distribution function or to predict low or very low probability events.

Conclusions

When applied correctly the histogram gives not only a fast graphical depiction of the whole data set but could also give a more accurate representation of the statistical parameters.

Legal Disclaimer