5.2.5 Example Histograms: British Petroleum – "Data Pre-Processing"

This Section examines a few months of daily closing prices for British Petroleum (BP) in the context of "basic histograms". The data is examined in its "raw prices" form, and it is examined via two different pre-processing methodologies. One of the pre-processing approaches is a transformation into another (more "convenient[112]") form. Here, two such forms will be reviewed: P&L-basis (by way of price changes or differences called "Diffs"), and a continuously compounding returns-basis. These are closely related but there are subtle distinctions.

Another pre-processing approach applies a "proper de-trending". Here, only a straight-line fitting process will be used to de-trend the data and put it on a "zero drift" basis.

It cannot be emphasised too strongly that these pre-processing matters are very much tied to modelling issues that belong in later Chapters. To preserve the "basic statistical characterisation" perspective, the modelling issues are kept to a minimum, and the two should not be confused with one another.

In all cases, it is the "shape" of the results that describe the statistical measures, and it is possible to extract specific or summary results from that "shape". It will be seen, that the "interpretation" of the results/shape may depend strongly on the pre-processing applied.

Notice that moving to a returns-basis (or a returns proxy basis) may imply additional modelling assumption. For example, taking statistical measures of the returns or "Diffs" histogram, and then reversing out the implied "price" statistics can only be achieved by relating the returns-based measure to at least two point in time for the prices-based measures. Put differently, the price/return relationship is itself a complication in the "characterisation" vs. "modelling" issue[113].

Figure 5.2 – 8 a) and b) illustrate a few months of BP daily closing prices, and a histogram for that history. This histogram characterises the data, and the shape of the histogram indicates that there is a "central tendency" in the data with diminishing likelihood of events (in this case Diffs) far away from this central tendency.

---

[112] Keep in mind that "convenient" can be either from a mathematical/technical perspective, or from a context sensitive/market convention perspective. For example, transforming prices to returns may provide an "easier" analysis problem due to, possibly, "special properties" of returns. In other situations, such as rebalance strategy analysis or VaR, P&L is the "objective" and the transformation may need to reflect that.

[113] As an illustration consider that the price/return relationship can be chosen to be: $F = Pe^{rt}$. Now, suppose that the price data is transformed to a returns-basis and pre-processed in some way (e.g. linear de-trending). If statistical measures are made on this transformed/de-trended dataset, then the reversing of "price-based" results is a bit complex since, for example, the linear de-trending in returns-space requires log-linear de-trending assumptions in price-space (see Chapters 11 - 13, and [1], [2], and [8] for detailed discussions).

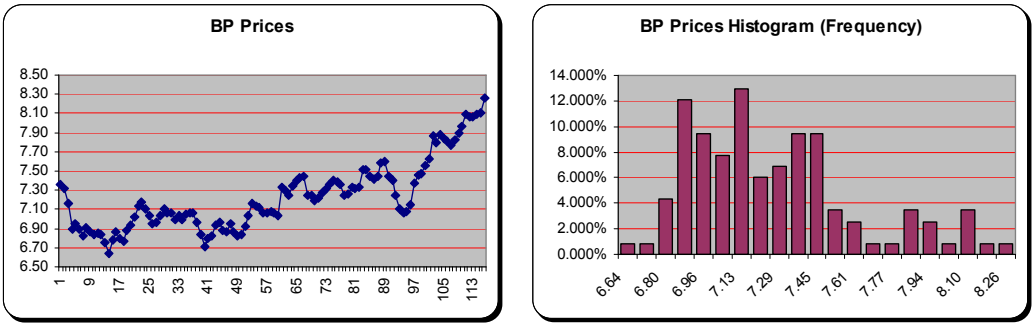

Figure 5.2 – 8 a) Daily closing prices for British Petroleum, b) histogram for same.

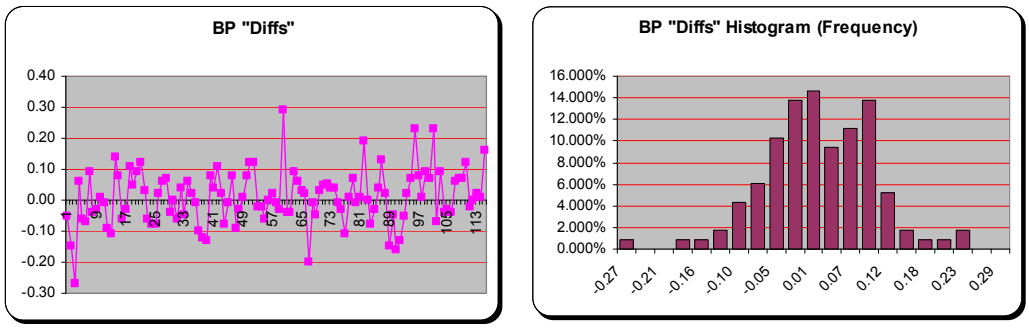

Figure 5.2 – 9. a) Daily closing "Diffs" for British Petroleum, b) histogram for same

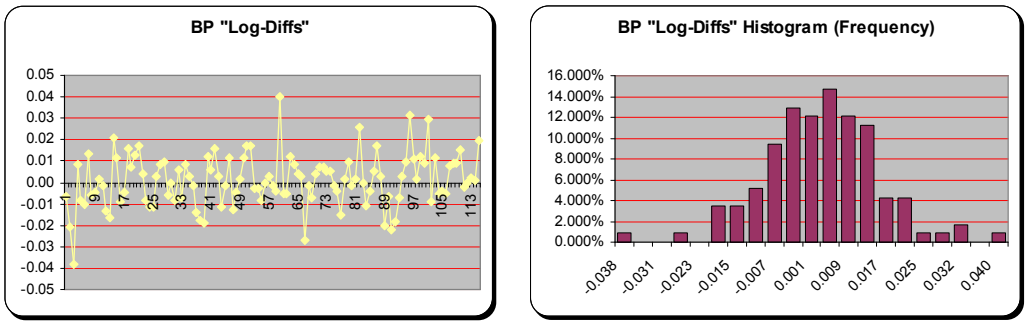

Figure 5.2 – 10 a) Daily closing "Log-Diffs" for British Petroleum, b) histogram for same

Though the data has a "central tendency", it's not a simply symmetric shape. Does this mean that measures of "width" of the histogram may "incorrectly" indicate statistical characteristics such as variability?

In one sense, the width is the width (or any proxy for width), and so it must be a measure of variability of "something". In this case, it is the variability of the dataset.

For instance, the "bin counting" procedure to produce the histogram results with a table as shown to the right. The cumulative frequencies can be used, as before, to make quantifiable estimates of the variability of the data. For instance, the price bin at the frequency of 18.10% is 6.88, while the price bin at the frequency of 80.17% is 7.45. This implies that approximately 62% (i.e. 80.17% - 18.10%) of the data is in the range of the prices (bins) 6.88-7.45.

Bin	Frequency	Cumulative %
6.64	1	.86%
6.72	1	1.72%
6.80	5	6.03%
6.88	14	18.10%
6.96	11	27.59%
7.05	9	35.34%
7.13	15	48.28%
7.21	7	54.31%
7.29	8	61.21%
7.37	11	70.69%
7.45	11	80.17%
7.53	4	83.62%
7.61	3	86.21%
7.69	1	87.07%
7.77	1	87.93%
7.86	4	91.38%
7.94	3	93.97%
8.02	1	94.83%
8.10	4	98.28%
8.18	1	99.14%
8.26	1	100.00%

One implication is that one may expect that additional data points, as they arrive, may (with 60% likelihood) fall in this range.

Some may argue that this is unsatisfactory since the data is decidedly asymmetric, and so a range representing 60% should be chosen to reflect this. For example, one may instead approximate the 60% likelihood interval to be 6.80-7.37 (that's actually a likelihood of 64.66 = 70.69% 6.03%).

Compare these two choices for the quantile ranges[114] visually in Figure 5.2 – 11 a) and b).

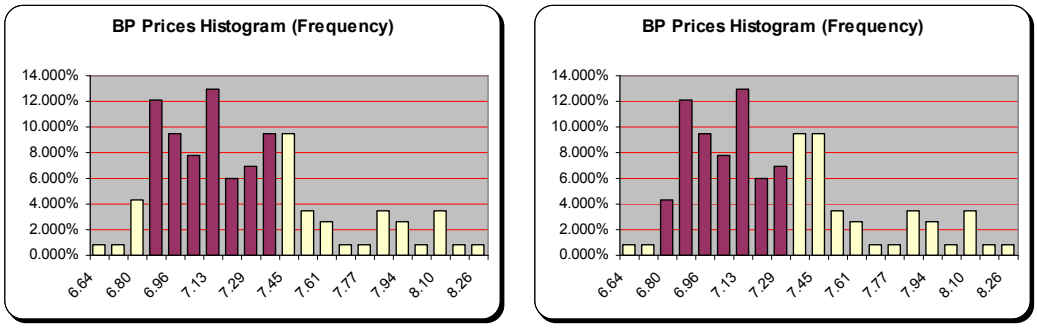

Figure 5.2 – 11. Two different choices for the quantile range each defining (approximately) 60% of the events.

One comment that is worthy at this stage, and one that will repeat itself later, is that even with poor or uncertain choices for quantile ranges, the quantile method still permits a "reasonable" method for dealing with complex and asymmetric histograms. This will not

[114] Keep in mind that these two images do not represent exactly the same "interval" since the area under the curve by one choice is approximately 5% greater than by the other choice. This approximation error is a direct consequence of using rank based methods in conjunction with "low resolution" histograms. For example, the discrepancy might negligible if the histograms were bases on, say, 50 bins. Alternatively, fitting a curve to the histogram data and employing quadrature methods would also reduce this type of approximation error as small as one likes.