How Histogramming and Counting Statistics Affect Peak Position Precision

D. A. Gedcke

Critical Applications

In order to expedite comprehensive data processing with digital computers, most scientific disciplines now acquire spectra as histograms. A histogram breaks up the horizontal axis of the spectrum into small, equal intervals, such that each interval can be assigned to an index in an array. The vertical axis represents the probability of recording an event in each interval, and that probability is represented by the magnitude of the digital word stored at the specific index in the array. Because the histogram quantizes the spectrum by grouping events into small intervals, the question arises: "How does the size of the interval affect the accuracy of measuring the position of a peak in the spectrum?"

Although the results below can be extended to the general case, this study focuses on the application to spectrometers that count stochastic events for the vertical scale in the spectrum. Typical applications are summarized in Table 1.

Table 1. Spectrometers Dominated by Counting Statistics.

- Time Digitizers used in Time-of-Flight Mass Spectrometry (TOF-MS)
- Time Digitizers or Multichannel Scalers (MCS) responding to single photons in LIDAR, or fluorescence and phosphorescence lifetime spectrometry
- Time-of-flight spectrometry with nuclear radiation (neutrons, alpha and beta particles, ions and nuclei)
- Pulse-amplitude or energy spectrometry with nuclear radiation (gamma-rays, X-rays, alpha particles, recoiling nuclei) using multichannel analyzers (MCA)
- Time spectrometry using a time-to-amplitude converter (TAC) followed by an MCA

The above applications involve a variety of species to be measured (photons, molecules, ions, charged particles, etc.). To simplify the discussion, detection of an individual member of any of the above species will be referred to as a detected event, or "event" in the remainder of this application note.

The Impact of Counting Statistics

In the case of time spectrometry, the arrival time of each individual event must be measured to determine the coordinate on the horizontal axis of the histogram. The vertical scale corresponds to the number of events that were counted in each specific time interval. For pulse-amplitude spectrometry, the height of each pulse is measured to determine the location on the horizontal axis of the histogram, ... while the vertical axis records the number of pulses that were counted in each pulse-height interval. In both cases, the arrival rate of events is purposely kept low, so that the arrival time or amplitude of each event can be measured without interference from additional events. As a result, the number of events counted over the duration of the measurement exhibits a statistical fluctuation described by the Poisson Probability Distribution[1,2].

$$P(N) = \frac{\mu^N e^{-\mu}}{N!} \quad (1)$$

P(N) is the probability of counting N events in a single measurement. If the measurement is repeated a large number of times and the values of N are averaged, the average value of N approaches the mean of the distribution, μ, as the number of repeated measurements approaches infinity. Note that the Poisson Distribution has a standard deviation

$$\sigma_N = \sqrt{\mu} \approx \sqrt{N} \quad (2)$$

Substituting N for μ in equation (2) recognizes that the value of N from a single measurement is an adequately accurate estimate[1] of μ.

It is common practice to describe the precision of the measurement of the number of events by expressing σ_N as a percent of N.

$$\sigma_N\% = \frac{\sigma_N}{N} \text{ x } 100\% = \frac{100\%}{\sqrt{N}} \quad (3)$$

Table 2 shows how the precision, σ_N%, depends on N for typical values of N.

Strictly speaking, equations (1) through (3) are applicable only if the dead time losses are negligible, or if the system utilizes an ideal livetime clock to compensate for dead time losses[3–8]. For the purposes of this study, it is presumed that at least one of those conditions is met. Note that equations (1) through (3) can be applied to the counts in any section of the spectrum.

Table 2. The Dependence of Precision, σ_N%, on the Number of Events Counted, N.

N	σ_N%
1	100.0%
100	10.0%
10,000	1.0%
1,000,000	0.1%

How Counting Statistics Limits the Peak Position Precision

Most spectra contain narrow peaks, whose positions define the physical quantity to be measured. For example, the centroid of the peak may define the mass-to-charge ratio of an ionized molecule in mass spectrometry, the energy of a gamma-ray in nuclear spectrometry, or the distance to an object in LIDAR. Unfortunately, counting statistics limits the precision with which the centroid of the peak can be measured. This limitation on precision can be easily calculated as follows.

Figure 1 illustrates a peak in the spectrum that has been acquired with sufficient counts to make the error from counting statistics negligible. Also the width of the quantization interval on the horizontal axis has been chosen to be small enough that it has a negligible effect on the peak shape. These quantization intervals will be referred to as "bins" in the remainder of this application note, although they are also known as "channels" in MCA and MCS applications. The parameter listed on the horizontal axis in Figure 1 is the bin index number, i. This corresponds to the index in the digital array where the data is stored. Normally, the horizontal axis would be calibrated to read in the units that are applicable to the parameter being measured, e.g., nanoseconds for time spectrometry, MeV for gamma-rays, mass per unit charge for

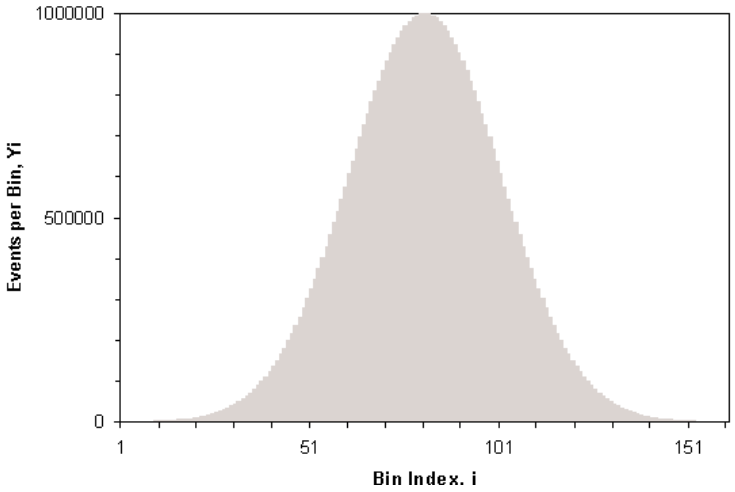

Figure 1. A Gaussian Peak Shape Displayed as a Histogram. The bin width is 0.05 σ_x, where σ_x is the standard deviation of the Gaussian peak.

mass spectrometry, or kilometers for LIDAR. Simplistically, this calibration can be expressed as

$$x_i = k\, i \tag{4}$$

where x_i is the calibrated value for the center of the i^{th} bin, and k is the calibration factor.

When its area is normalized to unity, the peak in Figure 1 represents the empirical probability that a single event will be measured with a particular value of x_i. In other words, it is the probability distribution for x_i. This probability distribution has a mean computed as

$$x_{mean} = \frac{\sum_i y_i x_i}{N} \tag{5}$$

where y_i denotes the number of events counted in bin i. N is the total number of events counted within the peak, i.e., the area of the peak:

$$N = \sum_i y_i \tag{6}$$

The mean is the centroid of the peak, and is considered to represent the position of the peak on the x axis.

The probability distribution in Figure 1 also has a parent population standard deviation σ_x estimated by the sample standard deviation s_x as calculated via the variances σ_x^2 and s_x^2 in equation (7).

$$\sigma_x^2 \approx s_x^2 = \frac{\sum_i y_i (x_i - x_{mean})^2}{N-1} \tag{7}$$

Note that $N-1$ instead of N appears in the denominator of equation (6) because x_{mean} has to be calculated from the measured data, and that removes one degree of freedom[9].

Using conventional probability theorems[1,9], the following conclusion can be stated. *If a single event belonging to the probability distribution is measured, and the measurement yields a value x_i, then x_i is the most likely estimate of the true mean of the distribution, x_{mean}, and a standard deviation σ_x can be assigned to the estimated uncertainty in the single measurement, x_i.* In other words, the uncertainty in estimating the true peak position from a single event is σ_x.

Probability theory also states that the uncertainty of a single measurement can be reduced by combining N measurements to get the average:

$$x_{aver} = \frac{\sum_{j=1}^{N} x_j}{N} \tag{8}$$

The uncertainty or standard deviation in x_{aver} is

$$\sigma_{aver} = \frac{\sigma_x}{\sqrt{N}} \tag{9}$$

This process of averaging the x_j values from N single events is exactly equivalent to accumulating a peak in the spectrum until it includes N events, and then computing the average or mean position of the N events in the peak from:

$$x_c = \frac{\sum_i y_i x_i}{N} \tag{10}$$

where y_i is the number of events counted in bin i and

$$\sum_i y_i = N \tag{11}$$

The value x_c is the most likely estimate of the centroid of the peak, and the uncertainty in this estimate of the centroid is given by the standard deviation

$$\sigma_c = \sigma_{aver} = \frac{\sigma_x}{\sqrt{N}} \tag{12}$$

Equations (10) through (12) apply to any arbitrary peak shape. A further simplification arises in the typical case of a Gaussian peak shape as depicted in Figure 1. The relationship between the full width at half maximum height (FWHM) and the standard deviation of a Gaussian peak is

$$FWHM = 2.35\ \sigma_x \tag{13}$$

Substituting from equation (13) in equation (12) permits the uncertainty in the measured centroid of the Gaussian peak to be expressed as a percent of the FWHM, and as a function of the number of events, N, counted in the peak:

$$\sigma_c\% = \frac{\sigma_c}{FWHM} \times 100\% = \frac{100\%}{2.35\sqrt{N}} \qquad (12)$$

Equation (14) is illustrated in Figure 2. Note that a peak containing $N = 100$ events yields a standard deviation in the centroid, $\sigma_c\% = 4.3\%$ of the FWHM, whereas the standard deviation in the number of counts in the area of the peak is $\sigma_N\% = 10\%$. Both numbers improve slowly in proportion to the square root of the number of events counted in the peak. It takes a factor of 100 more events in the peak to reduce $\sigma_c\%$ to 0.43% and $\sigma_N\%$ to 1%.

Estimating the Centroid Uncertainty from Counting Statistics in a Real Spectrum

Equation (14) provides a practical means of estimating the effects of counting statistics on the uncertainty in the measured peak position. For the typical case of a peak shape that is approximately Gaussian, the FWHM of the peak is measured, and the number of counts, N, in the peak is noted. The value of N is inserted in equation (14) to calculate the estimated centroid standard deviation as a percent of the FWHM. Multiplying this result by the FWHM (in appropriate units) and dividing by 100% yields the uncertainty in the centroid of the peak, σ_c, in absolute units.

If the peak deviates strongly from a Gaussian shape, then the standard deviation of the peak can be computed from equation (7), and σ_c can be calculated from equation (12).

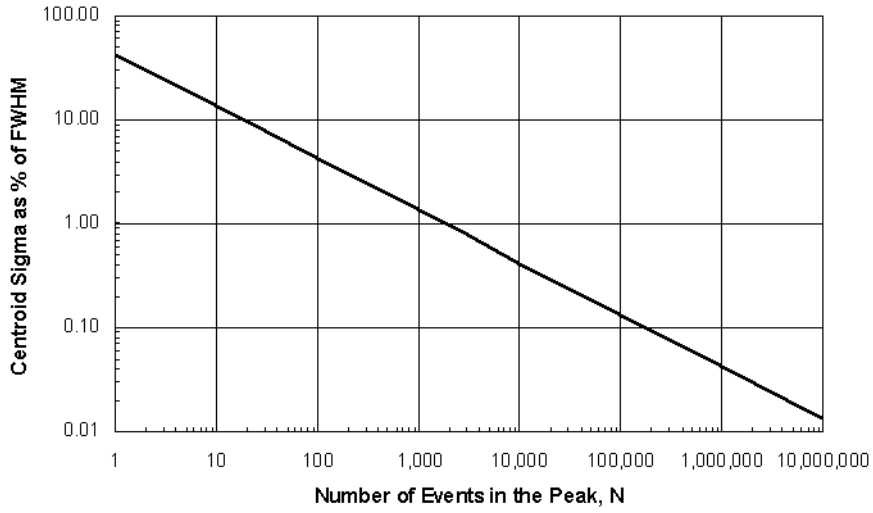

Figure 2. The Dependence of the Percent Standard Deviation of the Peak Centroid on the Number of Events Counted in the Peak. The vertical axis is the left side of equation (14).

The Centroid Error Caused by Large Bin Widths in the Histogram

It is not possible to derive a simple mathematical formula to express the effect of the bin width on the error in measuring the centroid of the peak. Consequently, numerical analysis is the most convenient method for delineating the effect. To study the effects of the bin width, the error caused by counting statistics is considered to be negligible. A Gaussian peak shape of the form

$$y = \frac{exp[-(x-X_c)^2 / 2\sigma_x^2]}{\sqrt{2\pi\sigma_x^2}} \qquad (15)$$

was converted into a histogram and analyzed via an Excel spreadsheet. For convenience, the mean of the Gaussian, X_c, was set to zero, and the x values were computed in units of the standard deviation, σ_x. Note that the area under the peak in equation (15) is exactly 1.

The conversion to a histogram involved choosing a basic bin width $w = 0.05\ \sigma_x$, as illustrated in Figure 1. This is small enough to yield negligible errors in the conversion. The bin overlapping the centroid of the Gaussian was centered at $x = 0$, and the remaining bins extended up to $\pm 8.2\ \sigma_x$ about the centroid. The value of equation (15) at the center of each bin was multiplied by the bin width to compute the value of y_i for each bin. The accuracy of this conversion was verified by the fact that the area under the resulting histogram fell short of 1 by an absolute value of 7×10^{-10}. The x coordinate for each bin was considered to be synonymous with the center of the bin. Wider bin widths were achieved by adding together adjacent bins.

The systematic error caused by the finite bin width can be appreciated by considering the simple example of only two bins spanning the entire Gaussian peak. If the boundary between the two bins exactly coincides with the centroid of the peak, then areas of both bins will be equal and the centroid computed from the two bins will exactly match the centroid of the underlying Gaussian peak. This happens because the Gaussian function is symmetric about its mean. If the boundary between the two bins falls to the left of right of the mean of the Gaussian, the centroid of the two bins will no longer match the mean of the Gaussian.

Now consider one more simple example. If three bins span the area of the Gaussian peak, and the middle bin is exactly centered on the mean of the Gaussian, then the areas on the two outlying bins will be equal. This is another case of symmetry, where the centroid calculated from the three bins will exactly match the mean of the Gaussian. If the middle bin is not exactly centered on the mean of the Gaussian, then the centroid calculated from the three bins will be in error relative to the mean of the Gaussian.

In general, if the bins are not symmetrically positioned relative to the centroid of the Gaussian, then the centroid calculated from the histogram will be in error relative to the true centroid of the Gaussian. The task is to determine the magnitude of that error. This was accomplished by choosing a bin width, w, and sliding the bins across the centroid of the Gaussian in 8 steps. The step size was $w/8$. For each step, the centroid of the resulting histogram was calculated and compared to the true centroid of the Gaussian. From all 8 steps, the maximum error was reported, as displayed in Figure 3. Expressed as fractions of σ_x, bin widths of 0.2, 0.4, 0.8, 1.0, 1.2, 1.6, 2.0, 2.4, and 3.2 were employed. These values have been divided into 2.35 to yield the ratio of the FWHM to the bin width as used on the horizontal axis in Figure 3. The horizontal axis can be considered to represent the number of bins that span the FWHM of the peak.

Figure 3 shows that once the number of bins spanning the FWHM exceeds 1.5, the maximum error in the centroid due to the bin width is less than 0.01% of the FWHM. Furthermore, the error declines steeply with increasing numbers of bins across the FWHM. The error is so miniscule beyond 2 bins across the FWHM that the precision of Excel in dealing with small differences between huge number begins to limit the accuracy of the calculation. For the point on the extreme right side of the graph a further limitation occurred because it was only possible to make 4 steps when shifting the bins.

Common practice is to use at least 5 bins to span the FWHM. As can be seen in Figure 3, this ensures that the maximum error in the centroid as a result of histogramming will be less than 10^{-7}% of the FWHM, ... a truly negligible error!

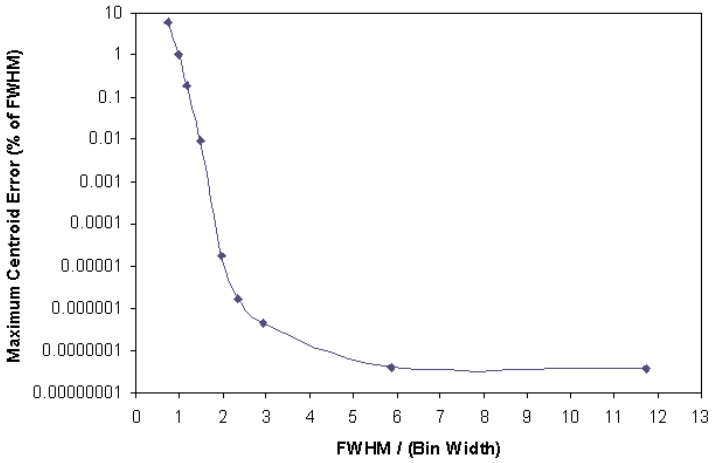

Figure 3. The Maximum Error in the Measured Centroid Due to the Finite Bin Width in the Histogram.

When the Counting-Statistics Error Dominates the Bin-Width Error

The important question is, "Are both sources of error equally important, or does one usually dominate?" That question is most graphically answered by plotting the values of the event counts, N, that yield a random error from counting statistics (equation (14)) that is equal to the maximum systematic error from the bin width (vertical axis in Figure 3). That result is plotted in Figure 4.

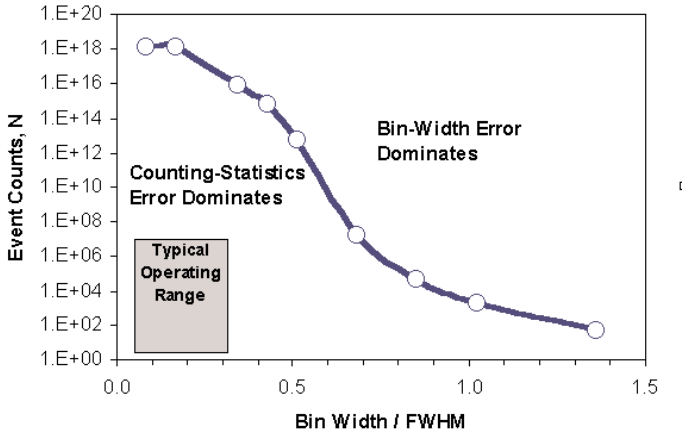

Figure 4. Domains Where the Centroid Error from Event Counting Statistics Dominates (Below and to the Left of the Curve), or the Centroid Error from Bin Width Dominates (Above and to the Right of the Curve). The Grey Rectangle Defines the Typical Operating Range of a Spectrometer.

The curve with the open circles defines where the two sources of error are equal. Below and to the left of this curve the random error from counting statistics dominates the centroid error, while above and to the right of the curve the systematic error from the bin width is dominant. Note that the horizontal axis has been inverted from Figure 3 to Figure 4 in order to define the dominance domains.

It is rare for a spectrometer to acquire less than 5 or more than 10^7 event counts in a peak that must be analyzed. Those numbers set the upper and lower limits of the gray rectangle that defines the typical operating range in Figure 4. The horizontal limits of the rectangle extend from approximately 0.05 FWHM to circa 0.33 FWHM for the bin width. This corresponds to a range from 20 bins across the FWHM down to 3 bins spanning the FWHM. It is obvious from the gray box in Figure 4 that the random error from counting statistics is the dramatically dominant source of error in the centroid measurement for the entire range of normal operation. With reasonable and typical choices for the bin width, the centroid error caused by the finite bin width in the histogram is negligible.

The Effects of Differential Non-Linearity

The results in Figure 3 were based on the presumption that the bin widths in the histogram are all equal. In a practical spectrometer the bin widths can vary. The variation from the average width is called the Differential Non-Linearity (DNL). The percent DNL is defined as

$$\%DNL = \pm \frac{W_{max} - W_{min}}{W_{max} + W_{min}} \times 100\% \qquad (16)$$

where w_{max} is the maximum bin width, and w_{min} is the minimum bin width. High-quality spectrometers typically have a DNL within ±1%, ... i.e., the widest bin is no more than 1% wider than the average bin width, and the narrowest bin is no more than 1% smaller than the average. Naturally, the question arises: "What effect does differential non-linearity have on the accuracy with which the true peak centroid can be measured?"

Using a methodology similar to that employed to investigate the effect of the bin width, the contributions of DNL were evaluated for two commonly encountered patterns of bin width variations a) an odd-even pattern and b) a random pattern. The results are summarized in Table 3. An average bin width equal to 0.4 σ_x was employed in both cases. This corresponds to 5.9 bins spanning the FWHM of the Gaussian peak.

Table 3. The Centroid Error Caused by Differential Non-Linearity.

Pattern	% DNL	Average Bin Width	Centroid Error		Equivalent N
			% of Bin Width	% of FWHM	
Odd-Even	±12.5%	0.4 s_x	4.7×10^{-12} %	8.1×10^{-13} %	2.8×10^{27}
Random	±1.3%	0.4 s_x	0.28 %	0.047 %	820,000

The odd-even pattern set all odd-numbered bins 12.5% narrower than the average bin width, and all even-numbered bins 12.5% wider than the average. As demonstrated by the results in Table 3, this is a rather benign DNL pattern, because the spacing between bin centers is constant for all bins. The contribution to the centroid error is negligible, even though the DNL is an extreme ±12.5%.

The last column in Table 3 is the number of events, N, that would yield a centroid uncertainty from equation (14) that is equal to the centroid error from the differential non-linearity. Below the equivalent value of N the uncertainty due to counting statistics dominates. Above that value of N, the DNL error is dominant.

For the random DNL pattern, a random number generator was used to determine how much each bin boundary should be shifted to the left or the right. The range of the random number was chosen to limit the DNL to ±1.3%. The resulting average for the distribution of bin widths was 0.39996 σ_x, with a standard deviation equal to 0.64% of the average. The average bin width corresponds to 5.9 bins spanning the FWHM of the Gaussian peak.

As listed in Table 3, the centroid error from the random DNL pattern is small, and can be ignored up to circa 820,000 events counted in the peak. Above that value of N, the error from the random DNL pattern will become the limiting centroid error as the percent error from counting statistics declines. It is tempting to diminish the effect of random DNL by doubling or quadrupling the number of bins spanning the peak. That tactic can help, provided the DNL does not double when the number of bins doubles.

Extrapolating to Least-Squares Fitting of Overlapping Peaks

Occasionally, two peaks in the spectrum overlap, and their individual contributions must be separated. Least squares fitting of known peak shapes to the composite peak is the technique normally used to extract the desired parameters for the individual peaks[10–13]. For example, if the peak shapes are Gaussian and have a known width, two mathematical functions such as equation (15) can be fitted to the composite peak in the spectrum. The centroids and heights of each Gaussian function are iteratively adjusted until the sum of the squares of the differences between the actual data in the spectrum and the fitted functions is minimized. Once the iterative minimum is reached, the centroid and area of each Gaussian function are the most probable estimates of peak position and area for each peak.

The information derived in this application note for a single peak can also be applied to overlapping peaks, with some modifications. Obviously, if the two peaks have equal heights and minimal overlap, the statistical errors in each peak have a negligible effect on the other peak. However, as the overlap between the two peaks increases, the statistical uncertainties in establishing the area and centroid of peak 1, magnify the errors in extracting those same parameters for peak 2, and vice versa. The worst case for accuracy is measuring the parameters of a small peak that suffers a domineering overlap from a tall peak.

Practical examples of least squares fitting the extremes of peak overlap from minimal to severe have been thoroughly analyzed by Keyser[13]. His graphical results echo the principles established above for single peaks. For a reasonable choice of the number of bins spanning the FWHM ($\geq$5 bins across the FWHM), it is the random error from counting statistics that normally dominates.

Conclusions

A reasonable operating condition for histogramming peaks implies at least 5 bins spanning the FWHM of the narrowest peak in the spectrum. Under these conditions the error in determining the centroid of the peak is normally dominated by the random error from counting statistics. The error contribution from differential non-linearity becomes important when the number of events counted in the peak exceeds circa 10^6 counts. The error contributed by a finite, constant, bin width does not surpass the error from counting statistics until the number of counted events exceeds approximately 10^{18}.

To improve the accuracy of measuring either the peak position or the area of the peak, the focus must be on increasing the number of events counted in the peak.

References

1. Ron Jenkins, R. W. Gould, and Dale Gedcke, Quantitative X-Ray Spectrometry, Marcel Dekker, New York, First Edition (1981), pp 209 – 229.
2. Phillip R. Bevington, and D. Keith Robinson, Data Reduction and Error Analysis for the Physical Sciences, WCB McGraw-Hill, Boston, Second Edition, (1992), pp 23 – 28.
3. Ibid. ref. 1, pp 229 – 244 and 252 – 276.
4. 1997/1998 EG&G ORTEC Catalog, pp 2.176 – 2.17
5. ORTEC Modular Pulse Processing Electronics Catalog (2001), PerkinElmer Instruments, Oak Ridge, USA, pp 8.3 – 8.4
6. Ibid. ref. 4, pp 2.282 – 2.283.
7. Ibid. ref. 5, pp 10.6 – 10.7.
8. D.A. Gedcke, ORTEC Application Note AN57, Dealing with Dead Time Distortion in a Time Digitizer, (2001).
9. Ibid. ref. 2, pp 1 – 75.
10. Ibid. ref. 1, pp 241 – 270.
11. Ibid. ref. 2, pp 53 – 179.
12. Robert L. Coldwell and Gary J. Bamford, The Theory and Operation of Spectral Analysis Using ROBFIT, American Institute of Physics, New York, 1991.
13. Ronald M. Keyser, Nucl. Instr. and Meth. A286 (1990) pp 403 – 414.

AN58
Application Note

Specifications subject to change
052302

ORTEC®
800-251-9750 • www.ortec-online.com
info@ortec-online.com • Fax (865) 483-0396
801 South Illinois Ave., Oak Ridge, TN 37831-0895 U.S.A. • (865) 482-4411
For International Office Locations, Visit Our Website

AMETEK
ADVANCED
MEASUREMENT
TECHNOLOGY