## EUROPEAN ORGANISATION FOR NUCLEAR RESEARCH

CERN-PPE/94-72
MAN/HEP/94/3
6 May 1994

# Where to Stick Your Data Points: The Treatment of Measurements Within Wide Bins 

G. D. Lafferty ${ }^{1}$ and T. R. Wyatt ${ }^{1,2}$<br>${ }^{1}$ Department of Physics and Astronomy, The University of Manchester, Manchester M13 9PL, UK<br>${ }^{2}$ PPE Division, CERN, CH-1211 Geneva 23.<br>Email: LAFFERTY@CERNVM.CERN.CH, WYATT@CERNVM.CERN.CH.


#### Abstract

When a data point, measured over a bin of finite width, is to be compared to theoretical or model probability density functions, neither the central value of the bin nor the weighted mean value (the barycentre) of the abscissa within the bin is the appropriate place to plot the data point. It is shown that such data points ought to appear where the value of the predicted function is equal to its mean value over the wide bin. Some consequences of commonly used but incorrect alternative data presentation methods in particle physics are discussed.


## 1 Introduction

It often occurs in high energy particle physics experiments that measurements are made of physical variables whose number density distributions vary rapidly (e.g. exponentially) over the experimentally accessible range. Measurements in regions of small number density are often of considerable interest. One example would be transverse momentum spectra with respect to jet axes, where there is much to be gained in making the measurements out to as large values of transverse momentum as possible. Another example is the measurement of a fragmentation function, the momentum fraction of a quark or gluon jet carried by a particular species of particle. Here too, the measurements at high values of momentum fraction, where the rates for light quark and gluon fragmentation are relatively low, are likely to be the most interesting.

One consequence of a rapid variation in rate with the variable of interest is that the experimenter is often forced to choose wide bins for the data in regions of low number density in order that statistical errors are kept reasonably small. Care in presentation of the data is then needed since the underlying, and presumably unknown, number density distribution in the data will be changing sharply over the wide bins.

## 2 Measuring number densities

In a sample of real events, suppose that the measured number of entries in a bin of some variable $x$, of width $\Delta x$ from $x_{1} \rightarrow x_{2}$, is $n_{\text {meas }}$. The usual approximation for the value of the true density distribution function $g(x)$ measured in the data, at some value of $x$ within the bin, is

$$
\begin{equation*}
g_{\text {meas }}=\frac{n_{\text {meas }}}{\Delta x} \tag{1}
\end{equation*}
$$

It is important to bear in mind that $g_{\text {meas }}$ is not a genuine measurement of the function $g(x)$ at any particular value of $x$, but its expectation value is related to this underlying function by

$$
\begin{equation*}
\left\langle g_{\text {meas }}\right\rangle=\frac{1}{\Delta x} \int_{x_{1}}^{x_{2}} g(x) d x \tag{2}
\end{equation*}
$$

Only in the limit $\Delta x \rightarrow 0$ does $\left\langle g_{\text {meas }}\right\rangle \rightarrow g(x)$; for a bin of non-zero width $\left\langle g_{\text {meas }}\right\rangle$ gives the average value of $g(x)$ within the bin. Nevertheless, it is common practice to present such measurements, $n_{\text {meas }} / \Delta x$, as measurements of differential rates $\mathrm{d} N / \mathrm{d} x$, or after efficiency corrections as differential cross sections $\mathrm{d} \sigma / \mathrm{d} x$, and to plot them at a particular value of $x$.

### 2.1 Plotting the measured densities

At which value of $x$ within the bin should $g_{\text {meas }}$ be regarded as a measurement of $g(x)$ ? A number of possible answers to this question have been used in the presentation of data in the literature, and we give a few examples below. The two most common methods are to plot the data either at the centre of the bin

$$
\begin{equation*}
x_{\mathrm{c}}=x_{1}+\Delta x / 2 \tag{3}
\end{equation*}
$$

as, for example, in reference [1], or at the mean value of $x$ for the entries within the bin

$$
\begin{equation*}
\bar{x}=\int_{x_{1}}^{x_{2}} x g(x) d x / \int_{x_{1}}^{x_{2}} g(x) d x \tag{4}
\end{equation*}
$$

An example of the latter, often called the barycentre, may be found in reference [2]. In one paper this latter quantity is described as "the centre of gravity of the corresponding bin" [3], and in another paper as the "weighted averages" [4]. In general a vertical error bar will be included to represent the measurement error on the function value, but a horizontal 'error bar' across the bin may [5], or may not [6], appear.

It is shown below that none of these methods are strictly correct and that they may give a quite misleading impression of the trends in the data. Before introducing the position $x_{l w}$, which we show to be the correct value of $x$ to plot the data, we give an example.

### 2.2 Example: an exponential density function

As a simple example to illustrate the problem let us assume that the function $g(x)=a e^{-b x}$ represents the true number density distribution in a given data sample. This exponential form for $g(x)$ gives a good representation of hadron fragmentation functions in high energy jets, as well as of transverse momentum distributions of hadrons with respect to jet axes. To be specific let us consider the function $g(x)=10^{4} \times e^{-6 x}$, with the data binned in the ranges of $x$ from $0.0-0.1,0.1-0.3,0.3-0.6$ and $0.6-1.0$. For each bin, Table 1 gives the expected values for: the average number of entries, $\left\langle n_{\text {meas }}\right\rangle$; the corresponding value of $\left\langle g_{\text {meas }}\right\rangle$; the statistical uncertainty $\sigma\left(g_{\text {meas }}\right)$ in $g_{\text {meas }}$; the centre of each bin $\left(x_{c}\right)$; and the mean value of $x$, or the barycentre, for the entries within the bin $(\bar{x})$.

In Figure 1 the function $g(x)$ is compared with the expectation values of $g_{\text {meas }}$. Various possible choices are shown for the $x$ coordinate within the bin at which the data may be plotted. Since the expected average number of entries per bin, and thus the $\left\langle g_{\text {meas }}\right\rangle$ values given in Table 1, have been calculated assuming a precise knowledge of the true $g(x)$ then the plotted points might naïvely be expected to lie exactly on the curve of $g(x)$. However, the figure shows clearly that the choice of the barycentre $\bar{x}$ (the squares) or the bin centre $x_{c}$ (the triangles), to be found frequently in published literature, both give a misleading impression and suggest systematic deviations from the number density distribution to which the data actually correspond. In this particular example the choice of barycentre gives data points lying systematically below the curve, while the choice of bin centre gives points above the predicted curve. The size of these deviations becomes larger as the bins become wider to accommodate increasing statistical errors.

### 2.3 The correct position for plotting measured number densities

For the data properly to represent the true number density distribution in Figure 1, the points should lie on the curve of $g(x)$. They should therefore be plotted at the value of $x$ at which the function value $g(x)$ is equal to the expectation value of $g_{\text {meas }}$; as we observed above, $\left\langle g_{\text {meas }}\right\rangle$ corresponds to the average value of $g(x)$ within the bin. It becomes particularly important to


Figure 1: The function $g(x)=10^{4} \times e^{-6 x}$ with three sets of points showing different possible choices for the abscissa value. The bin ranges, $g(x)$ values and the various $x$ values are given in Table 1.

| bin range | $0.0-0.1$ | $0.1-0.3$ | $0.3-0.6$ | $0.6-1.0$ |
| :---: | ---: | ---: | ---: | ---: |
| $\left\langle n_{\text {meas }}\right\rangle$ | 752.0 | 639.2 | 230.0 | 41.4 |
| $\left\langle g_{\text {meas }}\right\rangle$ | 7519.8 | 3195.9 | 766.5 | 103.5 |
| $\sigma\left(g_{\text {meas }}\right)$ | 274.2 | 126.4 | 50.5 | 16.1 |
| $x_{\mathrm{c}}$ | 0.050 | 0.200 | 0.450 | 0.800 |
| $\bar{x}$ | 0.045 | 0.180 | 0.407 | 0.727 |
| $x_{l w}$ | 0.048 | 0.190 | 0.428 | 0.762 |

Table 1: For the function $g(x)=10^{4} \times e^{-6 x}$ is given the expected values in each bin of: the average number of entries, $\left\langle n_{\text {meas }}\right\rangle$; the corresponding value of $\left\langle g_{\text {meas }}\right\rangle$; the statistical uncertainty $\sigma\left(g_{\text {meas }}\right)$ in $g_{\text {meas }}$; the centre of each bin $\left(x_{c}\right)$; the mean value of $x$ for the entries within the bin $(\bar{x})$, and the value $x_{l w}$ at which the function is equal to its average value over the bin.
use this correct value of $x$, which we will refer to as $x_{l w}$, when the data are measured in bins of $\boldsymbol{l}$ arge $\boldsymbol{w}$ idth. The equation defining $x_{l w}$ is therefore

$$
\begin{equation*}
g\left(x_{l w}\right)=\frac{1}{\Delta x} \int_{x_{1}}^{x_{2}} g(x) d x \tag{5}
\end{equation*}
$$

The appropriate values of $x_{l w}$ for our specific example are given in the last row of Table 1 and are plotted as the circles in Figure 1.

Equation 5 cannot in general be solved, since $g(x)$ is usually unknown and is indeed what one is trying to measure. However a theory or model, for example a Monte Carlo simulation, may be available to predict a distribution $f(x)$ that is a reasonable approximation, at least in shape, to the true number density function $g(x)$. If a linear relation between $g(x)$ and $f(x)$ can be assumed, viz. $g(x)=\alpha f(x)+\beta$, then equation (5) can easily be shown to be equivalent to

$$
\begin{equation*}
f\left(x_{l w}\right)=\frac{1}{\Delta x} \int_{x_{1}}^{x_{2}} f(x) d x \tag{6}
\end{equation*}
$$

and this can now in principle be solved for $x_{l w}$ either analytically or numerically.
A possible alternative to the use of $x_{l w}$ would be to plot the data at $\bar{x}$ or $x_{c}$ and to calculate a correction to the ordinate $g_{\text {meas }}$ such that a point corresponding to the expectation value $\left\langle g_{\text {meas }}\right\rangle$ for a particular bin would lie on the curve $g(x)$. We do not recommend this alternative for a number of reasons. The values $g_{\text {meas }}$ within the bins represent the primary experimental measurements. They can be used for many purposes for which the above problem is irrelevant. For example, they can be integrated over $x$ to give the total number of entries and, as will be seen in Section 2.6, they can be compared directly with model predictions if these are calculated in the same bins as used for the data. It would in any case be rather confusing to have a different set of values $g_{\text {meas }}$ for the graphical representation of the data than for all others purposes. A much more satisfactory solution is to plot the data at the correct value of $x$ (i.e. $x_{l w}$ ) in the first place.

### 2.4 Comparison with the bin centre and barycentre

Let us return to the example of an exponential number density function $f(x)=a e^{-b x}$. Application of equation 6 gives

$$
\begin{equation*}
x_{l w}=x_{1}+\frac{1}{b}\left\{\ln (b \Delta x)-\ln \left(1-e^{-b \Delta x}\right)\right\} \tag{7}
\end{equation*}
$$

From equation 4 the mean value of $x$ for the entries within the bin, or the barycentre, is

$$
\begin{equation*}
\bar{x}=\frac{x_{1}+\frac{1}{b}-\left(x_{2}+\frac{1}{b}\right) e^{-b \Delta x}}{1-e^{-b \Delta x}} \tag{8}
\end{equation*}
$$

The bin centre, $x_{c}=x_{1}+\Delta x / 2$, gives the third possible location for plotting a data point. For a wide bin these three values of $x$ may differ significantly from one another, as was demonstrated in Table 1 and Figure 1.

A particular example (taken as representative from many in the literature) of a wrong choice of $x$ position for plotting data points may be found in Figure 4 of reference [7]. Here a measured fragmentation function for $\eta$-mesons is compared with a Monte Carlo model prediction. The largest $x$ bin covers the range $0.229-0.686$ and the data point is plotted at the bin centre, $x_{c}=0.457$, with a horizontal bar to cover the full bin width. The measured point lies over one standard deviation above the model curve which has a shape proportional to $e^{-7.9 x}$. In this case, the corresponding value of $x_{l w}$ would be 0.395 , and at this value of $x$ the measurement would lie almost exactly on the prediction. From the data given in [7] we have reproduced, in Figure 2a, the data points and the curve. Figure 2a also shows the effect of using $x_{l w}$ as the position of the highest- $x$ data point.

Another example of data points that have been plotted in the wrong place may be found in Figure 9 of reference [8], showing fragmentation functions for $\phi(1020)$ and $\mathrm{K}^{*}(892)^{0}$ mesons in $Z^{0}$ decay. In this case, the bin barycentres have been chosen, with the result that data points at high $x$ appear further to the left of the predicted curve than they ought to be. The impression is then given of a more steeply falling experimental spectrum than is actually the case. For the $\mathrm{K}^{*}$ spectrum, for example, the highest $x$ bin covers the range $0.3-1.0$ with the the barycentre at 0.45 . Since the Monte Carlo curve falls approximately as $e^{-6.2 x}$, the value of $x_{l w}$ in this case is 0.54 . Again we have reproduced, in Figure 2b, the data for the $\mathrm{K}^{*}$ as plotted in [8] together with the curve of the tuned JETSET Monte Carlo prediction and the effect of shifting the highest- $x$ point to its proper position, $x_{l w}$.

In the special case that $g(x)$ varies linearly with $x, g(x)=a x+b$, then $\bar{x}=x_{l w}=x_{\mathrm{c}}$ and there is no ambiguity as to where to plot the data points. It follows for any arbitrary $g(x)$ that if the width of a bin is sufficiently small that $g(x)$ may be taken to vary approximately linearly over the bin then the error caused by plotting the data at $\bar{x}$ or $x_{c}$ may be small. For example, Figure 3 shows the data of Table 1 and Figure 1 plotted with narrower bins in order to reduce the difference between the values of $\bar{x}, x_{l w}$ and $x_{c}$. However, the requirement that the bin widths be kept small in the above sense may then preclude measurements being given in the region of low number density because the observed number of entries per bin is too small. Examples of this approach may be found in references [9].


Figure 2: Examples from the published literature to illustrate the effect of incorrect choice of $x$ value for plotting data: fragmentation functions in $\mathrm{e}^{+} \mathrm{e}^{-}$annihilation for (a) $\eta$-mesons [7] and (b) $\mathrm{K}^{*}(892)^{0}$-mesons [8].


Figure 3: The function $g(x)=10^{4} \times e^{-6 x}$ with three sets of points showing different possible choices for the abscissa value. The data have been rebinned with respect to Figure 1 in order to reduce the difference between the values of $\bar{x}, x_{l w}$ and $x_{c}$.

### 2.5 Errors on the value of the correct position

It has been demonstrated above that a knowledge of the shape of the underlying number density function $g(x)$ is needed in order to calculate the correct position at which to plot a measured data point within a wide bin in $x$. If, as will generally be the case, $g(x)$ is not precisely known then this will introduce an uncertainty in the correct value of $x_{l w}$. As an example let us consider again the values of $g_{\text {meas }}$ given in Table 1. Let us further assume that previous measurements have firmly established the functional form $g(x) \propto e^{-6 x}$ in the range $x<0.6$, but that the value given in the last column of Table 1 represents the first measurement in the range $x>0.6$. If the functional form $g(x) \propto e^{-6 x}$ is assumed to be valid also in the range $x>0.6$ then the appropriate value of $x_{l w}$ can be calculated precisely. However, if a possible variation in the shape of $g(x)$ in this region is considered then an uncertainty in the correct value of $x_{l w}$ is introduced. This is demonstrated in Figure 4. The solid curve shows $g(x) \propto e^{-6 x}$ and the plotted full circle shows the corresponding value of $x_{l w}=0.762$, which has been given in Table 1 and can be calculated from equation 7. The dashed curve shows $g(x) \propto e^{-4.8 x}$ in the range $x>0.6$ and the plotted diamond shows the corresponding value of $x_{l w}=0.769$. The dotted curve shows $g(x) \propto e^{-7.6 x}$ in the range $x>0.6$ and the plotted triangle shows the corresponding value of $x_{l w}=0.753$. The functions $g(x) \propto e^{-4.8 x}$ and $g(x) \propto e^{-7.6 x}$ have been chosen because the associated 'measured' points lie off of the curves by approximately one standard deviation. It would therefore be consistent with the given assumptions to represent the measurement by a point plotted at $x_{l w}=0.762$ with a horizontal error bar extending over the range $0.753<x_{l w}<0.769$. By comparing the plotted circle in Figure 4 with the dashed and dotted curves it can be seen that if the uncertainty in $x_{l w}$ had not been taken into account then these curves would have been excluded at the level of more than one standard deviation.

### 2.6 Presenting data and predictions in the form of histograms

In order to make a quantitative comparison between a set of experimental data and the predictions of a theory or model then an alternative, and perhaps simpler, procedure is often used. The model predictions can be calculated for the same bins in which the data are presented using

$$
\begin{equation*}
n_{\text {pred }}=\int_{x_{1}}^{x_{2}} f(x) d x \tag{9}
\end{equation*}
$$

Data and predictions can then be considered as two histograms, which can be compared directly, perhaps by a fit to free parameters in the theory or model. For the purposes of this quantitative comparison the question of choosing the correct $x$ value within the bin is then irrelevant.

However, for the purposes of a graphical presentation then a histogram of the model predictions within wide bins does not give a very useful representation of the shape of the underlying distribution. For example see Figure 5 a which shows the values of $g_{\text {meas }}$ given in Table 1. Many possible guesses at the underlying number density distribution might be drawn as curves that at first sight roughly follow the shape of the histogram. Of course, the curve representing the true underlying number density distribution $g(x)$ crosses the horizontal line for each bin at the appropriate value of $x_{l w}$. Therefore, a more useful representation of the shape of the underlying distribution may be obtained by adding a point at the appropriate value of $x_{l w}$ within each bin as in Figure 5b. By removing the vertical lines from Figure 5b we arrive at an alternative


Figure 4: Demonstration of the uncertainty in $x_{l w}$ that can result from an uncertainty in the shape of the underlying number density function. The solid curve and the plotted circle correspond to the assumption $g(x) \propto e^{-6 x}$. The dashed curve and the plotted diamond correspond to the assumption $g(x) \propto e^{-4.8 x}$. The dotted curve and the plotted triangle correspond to the assumption $g(x) \propto e^{-7.6 x}$.


Figure 5: Some alternative ways of presenting a histogram of the values of $g_{\text {meas }}$ given in Table 1.
method of presenting the histogram (see Figure 5c). The horizontal lines in Figure 5c are then a purely conventional device to indicate the bin ranges and are not generally to be interpreted as error bars; if they were, then they would render impotent any quantitative comparison between the points and possible model curves. We may therefore note that their presence is not an acceptable excuse for plotting the points at an incorrect value of $x$ (for example $\bar{x}$ or $x_{c}$ )! Indeed, these horizontal bars serve to give the eye a somewhat misleading impression of agreement between data and predicted curves, as was the case in Figure 2a above.

In Figure 6 we propose an alternative way to indicate the bin ranges, by means of short vertical lines at the edges of the plot. This has the advantage of clearly separating the primary information, the data and their genuine errors, from such secondary details as the bin ranges. The potentially misleading effect of the horizontal bars is thus avoided.

The results of two experiments may be compared most easily if they both use the same binning to present their data. Unfortunately this will not normally be the case. If the comparison is to be made on the basis of points in $g(x)$ measured over wide bins then a consistent calculation of the relevant values of $x_{l w}$ for the two datasets must be ensured.

## 3 More complicated situations

We have so far considered the measurement of a number density distribution $g(x)$ as a function of a single variable $x$. We have shown that within a wide bin the observed number density $g_{\text {meas }}$ should be regarded as a measurement of $g(x)$ at the position $x_{l w}$ given by equation 5 . It is important to stress that more complicated situations may arise, in which $x_{l w}$ is not the correct position at which to plot data within a wide bin. However, in such cases some thought along the lines that led to equation 5 should allow the correct position to be deduced.

As an example of a more complicated situation, let us consider that in the presence of the varying number density function $g(x)$ we wish to measure the functional dependence of a second variable $q(x)$ as a function of $x$. A specific example, as in for example [10], is the measurement of the average transverse momentum component of tracks with respect to the jet axis, as a function of the momentum component parallel to the axis. A second example is the measurement in [11] of spin-density matrix elements for quasielastic photoproduction of vector mesons in bins of the four-momentum transfer, $t$, from the photon to the meson. The cross section, $\mathrm{d} \sigma / \mathrm{d} t$, falls exponentially with $t$, necessitating the use of wide bins at large values of $t$; there is in addition a slow variation of the measured spin-density matrix elements with $t$.

The expectation value for the measured $q(x)$ for the entries within a bin in $x$, of width $\Delta x$ from $x_{1} \rightarrow x_{2}$ is given by

$$
\begin{equation*}
\left\langle q_{\text {meas }}\right\rangle=\bar{q}=\int_{x_{1}}^{x_{2}} q(x) g(x) d x / \int_{x_{1}}^{x_{2}} g(x) d x \tag{10}
\end{equation*}
$$

Only in the limit $\Delta x \rightarrow 0$ does $\left\langle q_{\text {meas }}\right\rangle \rightarrow q(x)$; for a bin of non-zero width we have to decide at which value of $x$ within the bin $q_{\text {meas }}$ should be regarded as a measurement of $q(x)$. By analogy with the discussion that led to equation 5 we can easily see that the correct value of $x$ is that point $\left(x_{l w}^{\prime}\right)$ at which the function $q\left(x_{l w}^{\prime}\right)$ is equal to the average value of $q(x)$ for the entries


Figure 6: Our prefered way of presenting the values of $g_{\text {meas }}$ given in Table 1. The bin ranges are indicated by means of the short vertical lines at the edges of the plot. This avoids the potentially misleading effect of using horizontal 'error bars' for this purpose.

| bin range | $0.0-0.1$ | $0.1-0.3$ | $0.3-0.6$ | $0.6-1.0$ |
| :---: | ---: | ---: | ---: | ---: |
| $\left\langle q_{\text {meas }}\right\rangle$ | 7.442 | 3.036 | 0.663 | 0.076 |
| $x_{l w}^{\prime}$ | 0.042 | 0.170 | 0.388 | 0.698 |

Table 2: For the function $q(x)=10 \times e^{-7 x}$ measured in the presence of the varying number density function $g(x)=10^{4} \times e^{-6 x}$ is given in each bin: the expected value of $q_{\text {meas }}$ and the appropriate value of $x_{l w}^{\prime}$. The corresponding values of $\bar{x}, x_{l w}$ or $x_{c}$ may be found in Table 1 .
within the bin. Thus the defining equation for $x_{l w}^{\prime}$ is

$$
\begin{equation*}
q\left(x_{l w}^{\prime}\right)=\int_{x_{1}}^{x_{2}} q(x) g(x) d x / \int_{x_{1}}^{x_{2}} g(x) d x \tag{11}
\end{equation*}
$$

In general, $x_{l w}^{\prime}$ will not be equal to $\bar{x}, x_{l w}$ or $x_{c}$.

### 3.1 Example: exponential variation

For example let us consider an exponential form for $g(x)=a e^{-b x}$ and for $q(x)=c e^{-d x}$. Solution of equation 11 for this case gives

$$
\begin{equation*}
x_{l w}^{\prime}=x_{1}+\frac{1}{d}\left\{\ln \left(\frac{(b+d)}{b}\right)-\ln \left(1-e^{-(b+d) \Delta x}\right)+\ln \left(1-e^{-b \Delta x}\right)\right\} \tag{12}
\end{equation*}
$$

To be specific let us reconsider our previous example of the number density function $g(x)=$ $10^{4} \times e^{-6 x}$ with the data binned in the ranges of $x$ from $0.0-0.1,0.1-0.3,0.3-0.6$ and $0.6-1.0$, as given in Table 1 and Figure 1. In addition let us take $q(x)=10 \times e^{-7 x}$. For each bin, Table 2 gives the expected value of $q_{\text {meas }}$ and the corresponding value of $x_{l w}^{\prime}$. In Figure 7 the function $q(x)$ is compared with the expected values of $q_{\text {meas }}$ plotted as stars at $x_{l w}^{\prime}$. For comparison the data are also plotted as squares at $\bar{x}$, as circles at $x_{l w}$ and as triangles at $x_{\mathrm{c}}$, showing that none of these three choices is appropriate and that in this specific example they all lie systematically above the function. We note that in this case the measurement errors on $q_{\text {meas }}$ will depend on the experimental conditions and so we cannot show any error bars in the figure.

### 3.2 Example: linear variation

In the case where $q(x)$ may be taken to be varying linearly over a bin, as would be appropriate in the examples $[10,11]$ quoted above, the solution of equation 11 is $x_{l w}^{\prime}=\bar{x}$, indicating the choice of the barycentre as the correct position of the measured data point. Thus, while this choice has been shown to be incorrect for plotting the underlying number density, it turns out to be appropriate for a second variable measured in the presence of the varying number density. In fact it is easy to show from equation 11 that this result remains true for any functional form of the density $g(x)$ so long as $q(x)$ may be assumed to be a linear function of $x$ over the width of the bin.


Figure 7: The function $q(x)=10 \times e^{-7 x}$ measured in the presence of the varying number density function $g(x)=10^{4} \times e^{-6 x}$ with four sets of points showing different possible choices for the abscissa value.

## 4 Summary and Conclusions

We now summarize the main conclusions we have reached concerning the measurement of a number density distribution $g(x)$ within wide bins in $x$.

- The expectation value for an experimental measurement $\left\langle g_{\text {meas }}\right\rangle$ corresponds to the average value of $g(x)$ within the bin (equation 2). Therefore the result of a particular experiment, $g_{\text {meas }}$, should be regarded as a measurement of $g(x)$ at the value $x_{l w}$ of $x$ at which the function value $g\left(x_{l w}\right)$ is equal to the average value of $g(x)$ within the bin (equation 5). It is particularly important to use the correct value $x_{l w}$ when the data are measured in bins of $\boldsymbol{l}$ arge $\boldsymbol{w}$ idth.
- Regarding $g_{\text {meas }}$ as a measurement of $g(x)$ at $x_{\mathrm{c}}$, the bin centre, or at $\bar{x}$, the mean value of $x$ for the entries within the bin, is in general incorrect and can only be justified if $g(x)$ can be regarded as varying linearly over the width of the bin. We give examples from the published literature to illustrate this point (see Figure 2).
- The fact that one has to assume a shape for the underlying number density distribution in order to calculate $x_{l w}$ has a number of consequences:
- The assumptions made in a particular analysis should be described in the associated publication.
- An uncertainty in the knowledge of the underlying number density distribution will result in an uncertainty in the calculated values of $x_{l w}$. This must be taken into account when comparing experimental measurements with a curve predicted by a theory or model.
- When comparing the results of two experiments a consistent choice of how to calculate the appropriate values of $x_{l w}$ must be ensured for the two sets of data.
- A quantitative comparison is simplified if the theoretical predictions are calculated for the same bins used to analyse the experimental data. Similarly the comparison of two sets of experimental data is simplified if they are both presented with the same binning. The values within each bin can thus be compared directly and the question of choosing the correct $x$ value within the bin is then irrelevant for this purpose.
- We note that the common practice of using horizontal 'error bars' to indicate bin ranges can sometimes give a misleading impression of consistency in cases where measurements and predictions actually disagree. In Figure 6 we propose an alternative way to indicate the bin ranges - by means of short vertical lines at the edges of the plot - that avoids this problem.
- Although we mainly discuss the measurement of a number density distribution as a function of a single variable we give examples of how more complicated situations may be treated using a similar approach.
- It is clear that choosing bins to be as narrow as possible within the constraints of limited data and Monte Carlo statistics helps to minimize the ambiguity in the interpretation of the resulting data points.


## Acknowledgements

We should like to thank Roger Barlow and Richard Hemingway for their helpful comments on this paper.

## References

[1] I. V. Ajinenko et al., NA22 Collaboration, Z. Phys. C35(1987)7.
[2] P. D. Acton et al., OPAL Collaboration, Phys. Lett. B305(1993)407.
[3] H.-J. Behrend et al., CELLO Collaboration, Z. Phys. C46(1990)397.
[4] M. Bonesi et al., WA70 Collaboration, Z. Phys. C37(1988)535.
[5] M. Arnedo et al., EMC Collaboration, Z. Phys. C33(1986)167.
[6] M. Althoff et al., TASSO Collaboration, Z. Phys. C27(1985)27.
[7] D. D. Pitzl et al., JADE Collaboration, Z. Phys. C46(1990)1.
[8] P. D. Acton et al., OPAL Collaboration, Z. Phys. C56(1992)521.
[9] P. D. Acton et al., OPAL Collaboration, Phys. Lett. B291(1992)503.
P. Abreu et al., DELPHI Collaboration, Phys. Lett. B318(1993)249.
[10] W. Bartel et al., JADE Collaboration, Phys. Lett. 157B(1985)340.
[11] J. Ballam et al., Phys. Rev. D5(1972)545.

