

Mean Versus Median ©

by

Kevin Lehmann
Department of Chemistry
Princeton University
Princeton, NJ 08544
Lehmann@Princeton.edu

© Copyright Kevin Lehmann, 1997. All rights reserved. You are welcome to use this document in your own classes but commercial use is not allowed without the permission of the author. The author welcomes any constructive criticisms or other comments from either educators or students.

Overview: This worksheet compares the mean and median, two statistics that are often used to characterize the 'central value' of a distribution. They are compared for data drawn from Gaussian and Lorentzian distribution functions. Lastly, the 'maximum likelihood' method is introduced and compared.

Prerequisites: It is assumed in this worksheet that the reader already knows the basic principles of probability and integral calculus, as well as how to use Mathcad. While not required, it would be helpful to the reader if they are already familiar with the Gaussian distribution function, as covered in my Worksheet [GaussianDistribution.mcd](#)

Introduction:

There are many 'statistics' that can be used to characterize a distribution function; which one is best to use depends upon the application and also the form of the distribution itself. In this worksheet, we will focus on the common statistics that are used to define the 'middle' or 'center' of a distribution. These 'central' statistics serve two very different purposes. The first is purely descriptive. Which one is the 'best' descriptor is often a matter of taste or even values. We will not deal with these issues, except to warn the reader that in common usage, the word 'average' can be used for any of these central statistics, and that interested parties will often give only the statistics that is most favorable to their point of view, so caveat emptor!

For many common distribution functions, such as the well known Gaussian or Normal distribution, the different central statistics are in fact equal to one another. However, if one examines a finite number of examples or individuals sampled from that distribution, the central statistics of this data set will not generally be equal. In this case, one is often interested in using the data at hand to provide an estimate for the statistics of the unknown distribution from which the data is drawn or sampled. It is a question of objective statistical analysis which method of estimation will give a lower uncertainty in the properties of the unknown distribution. In this worksheet, we will demonstrate that the best method depends upon what is assumed about the functional form of the unknown distribution.

In order to make the above discussion more concrete, consider the common case where one makes a series of 'identical' measurements, say by titration of an unknown. Common experience is that the measurements will not all exactly agree. How do we combine the set of data to predict the best overall estimate for the unknown concentration? The methods discussed below provide the tools to attack this question, along with the related question, how reliable is the estimate we obtain.

You have likely been taught at sometime to take the mean or arithmetic average of your data. Do you remember WHY you should use this number?

The Mean:

The mean is certainly the best known of the central statistics. For a sample consisting of N points x_i the mean is defined as:

$$\text{mean} = \frac{1}{N} \sum_i x_i$$

This is often called the sample mean, since any real sample consists of a finite number of discrete values. Often, we are also interested in the mean of a distribution function, $P(x)$, that gives the probability density that a value x will occur in any sample. $P(x)dx$ gives the probability that an individual result will be found in the infinitesimal interval $[x, x+dx]$. In terms of $P(x)$, the population mean is defined as:

$$\text{mean} = \int P(x) \cdot x \, dx$$

The integral is over the full range, or domain, of x , i.e. all values of x that are possible results. It is understood that $P(x)$ is normalized, i.e., that the integral of $P(x)$ over the domain of x equals unity. In introductory statistics books, it is usually shown (often heuristically) that the two definitions of the mean will agree (with high probability and to within a small error), for samples of sufficiently large N .

Consider the set of points:

$$\mathbf{x} := (1 \ 2 \ 2 \ 2 \ 5 \ 6 \ 8 \ 10 \ 12)^T$$

x is written as a row vector. The superscript T means transpose. It converts a row vector into a column vector.

Calculate the mean of this sample.

Consider the distribution function:

$$P(x) = k \cdot \exp(-k \cdot x) \quad x, k \geq 0$$

Calculate the mean of this distribution.

The Median:

Another statistic that is often used is the median. The median is the "middle" of the distribution. For a sample population of N points, we put the data in ascending order and take the median to be the value of the $(N+1)/2$ 'th point if N is odd, and the arithmetic average of the $N/2$ and $N/2 + 1$ point if N is even. For a probability density function, $P(x)$, the median is defined as the solution to the equation:

$$\int_{-\infty}^{\text{median}} P(x) dx = \frac{1}{2}$$

Consider the set of points shown immediately below:
Determine the median of this sample.
Which is larger, the mean or median?

$$\mathbf{x} := (1 \ 2 \ 2 \ 2 \ 5 \ 6 \ 8 \ 10 \ 12)^T$$

Consider the distribution function:

$$P(x) = k \cdot \exp(-k \cdot x) \quad x, k \geq 0$$

Determine the median of this distribution.
Which is larger, the mean or median?
If x is time, what would the median commonly be called?

Distributions of family income or assets have a long 'tail' on the high side. In this case, will the mean or the median be larger? Which do you believe represents a 'better' measure of family income or wealth? Does your answer depend upon what you are going to use the measure for?

Mode

A third common central statistic is the mode, which is the most probable value. For a sample population this is the value that appears most often. This can only be applied to a sample that has discrete values, since for a continuous function we do not expect to ever get the same value more than once. However, one can make a continuous distribution discrete by putting the values into a series of 'bins' as when one makes a histogram. The value of the mode, however, can change dramatically if the number or placement of bins is changed, i.e. it is not a robust statistic.

For a probability density, $P(x)$, the mode is the value of x such that $P(x) \geq P(y)$ for all y , i.e., it is the value of x for which $P(x)$ has its absolute maximum. If this equation has more than one solution, then the most probable value is ambiguous. Often, a probability distribution will have two local maxima in which case it is called bimodal. If it has more than two local maxima, it is called polymodal.

Consider the set of points:
Determine the mode of this sample.

$$\mathbf{x} := (1 \ 2 \ 2 \ 2 \ 5 \ 6 \ 8 \ 10 \ 12)^T$$

Consider the distribution function:
Determine the mode of this distribution.

$$P(\mathbf{x}) = k \cdot \exp(-k \cdot \mathbf{x}) \quad \mathbf{x}, k \geq 0$$

Mean Vs. Median for Gaussian Distribution:

Why is the mean is so widely used, almost to the exclusion of the other measures, in statistics and error analysis? There are several reasons for this. One is just that it can be expressed, as a simple integral, instead of a solution of an integral equation, which simplifies the mathematics and allows for a simpler theoretical treatment. Another reason is the widely adopted assumption that one is dealing with a Gaussian distribution of errors. It can be shown that for a Gaussian distribution, the mean has the lowest dispersion of all statistics that return the center of the Gaussian. What this means is that if we estimate the center of the distribution by the mean and another function, such as the median, which returns the same average value, then we will get a smaller root mean squared fluctuation about the correct answer from the mean. Please note that this is not true for any distribution!

In the following pages, we are going to calculate the probability distribution for first the mean and then the median values for data drawn from a known distribution. Please note that this distribution function, in present case a Gaussian, has a well defined value for both the mean and median (and they are equal). However, any particular set of points will give only an estimate for the central value, and in general the sample mean and median values will not be equal. We are naturally lead to ask if we can predict what the probability density is that for any one data set of N_s points, we will calculate a specific value for the sample mean or median.

Consider a 'standard' Gaussian population, with zero mean and unit standard deviation. Let us take a sample of N_s points = $(2m+1)$ (so we get an odd number). The probability density of getting a mean = z is just a Gaussian distribution centered at zero with standard deviation equal to $1/\sqrt{N_s} = 1/\sqrt{2 \cdot m + 1}$. The proof for this can be found in most introductory statistics books.

$$P_{\text{mean}}(\mathbf{z}, N_s) := \frac{N_s}{\sqrt{2 \cdot \pi}} \cdot \exp\left(\frac{-N_s \cdot \mathbf{z}^2}{2}\right)$$

We can 'test' this result by generating a set of pseudo-random numbers using Mathcad.

$N_t := 5000$ N_t is the number of samples of data we will generate by Mathcad

$m := 2$

$N_s := 2 \cdot m + 1$ Size of each 'sample' of data points

$N := N_t \cdot N_s$ Number of pseudo-random data points that must be generated

$\mathbf{p} := \text{rnorm}(N, 0., 1.)$ Generates N pseudo-random numbers from a standard Gaussian distribution with mean zero, and standard deviation equal to one.

$\mathbf{k} := 0..N_t - 1$ Range variable for samples of data points.

$\mathbf{j} := 0..N_s - 1$ Range variable over data elements of each sample

$\mathbf{d}_{j,k} := \mathbf{p}_{k \cdot N_s + j}$ Put data into matrix with N_s rows and N_t columns. Each column contains one sample of N_s data points.

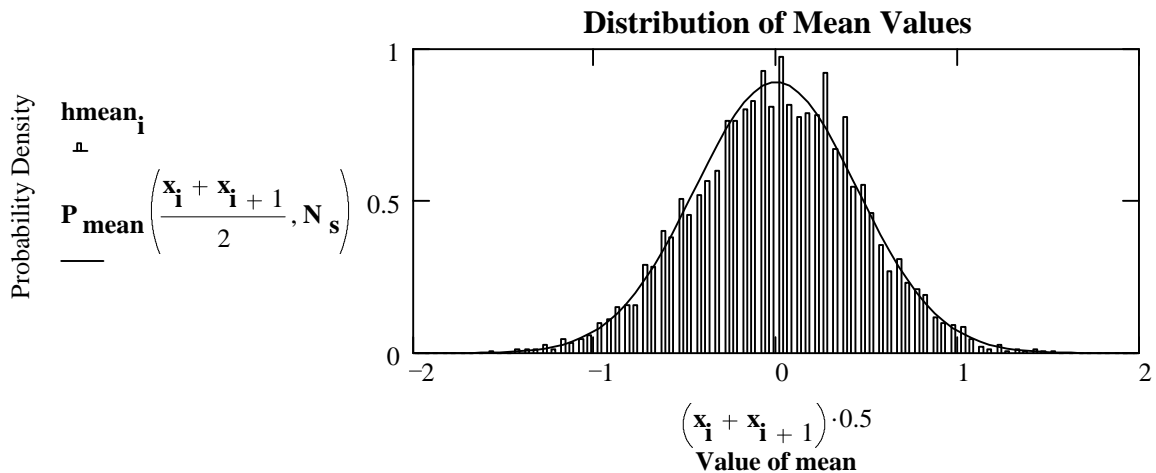
$\mathbf{Gmean}_k := \frac{1}{N_s} \cdot \sum_j \mathbf{d}_{j,k}$ \mathbf{Gmean}_k contains the mean value of the k 'th sample of data.

We can compare the distribution of Gmean values with the prediction $P_{\text{mean}}(z, N_s)$ by using Mathcad's hist function to generate a histogram of observed values

$i := 0..80$ $x_i := -2 + 0.05 \cdot i$ The vector x determines the position of the 'bins' in the histogram

$\text{hmean} := \frac{\text{hist}(x, \text{Gmean})}{0.05 \cdot N_t}$ The factor of $0.05 N_t$ give us a normalized histogram.

$i := 0..79$ We must reduce range of i by 1 because there is one less histogram value than the number of values of x.



Note: $\left(\frac{x_i + x_{i+1}}{2}\right) \cdot 0.5$ is the center value of the i'th bin.

Try other values of N_s and numerically demonstrate that P_{mean} correctly describes the results.

If we had sampled our data from a Gaussian distribution with mean μ and standard deviation σ , how would the above results change?

What is the function for the distribution of the median? I.e., we wish to find a function $P_{\text{median}}(y,m)$ such that for a sample of $(2m + 1)$ data points, the probability of obtaining a median value between y and $y + dy$ is equal to $P_{\text{median}}(y,m)dy$. If the median is y , then we have m data points below y . Each data point has a probability equal to the integral from negative infinity to y to be in that interval. In general, this is known as the cumulative probability. For the case of a standard Gaussian distribution, Mathcad will calculate the cumulative probability with the function $\text{cnorm}(y)$. There are also m points in the interval from y to positive infinity, each with probability $= 1 - \text{cnorm}(y)$, and a data point at y with probability density equal to the Gaussian distribution which Mathcad calculates with the function $\text{dnorm}(y,0,1)$. Lastly, we need to correct for the number of ways of distributing $2m + 1$ points, such that m are in one bin (those below the median), m in another (those above the median), and 1 in the last (at the median). We thus get the following expression for the probability of obtaining a median y in a sample of $2m + 1$ points selected from a normalized Gaussian:

Below you have the function that gives the expected probability density for the median value for a data set of $2m+1$ sample points selected from a standard Gaussian data distribution.

$$P_{\text{median}}(y, m) := \frac{(2 \cdot m + 1)!}{(m!)^2} \cdot [\text{cnorm}(y)^m \cdot (1 - \text{cnorm}(y))^m \cdot \text{dnorm}(y, 0, 1)]$$

Justify the combinatorial prefactor. How many ways are there of putting $2m + 1$ points in a definite order? How many ways of rearranging the first m ? The last m ?

Let us now check if our numerical data agrees with this prediction. What we need to do is find the median of each data set, and then make a histogram plot of the observed distribution of median values and compare that to the analytic expression for P_{median} . We find the median value of any set of data by putting the data in order and taking the central point. Mathcad provides a function $\text{sort}(d)$ which puts the elements of a vector, d , into ascending order. The implementation of the sort in the line below may take a while to execute.

$$d^{<k>} := \text{sort}(d^{<k>})$$

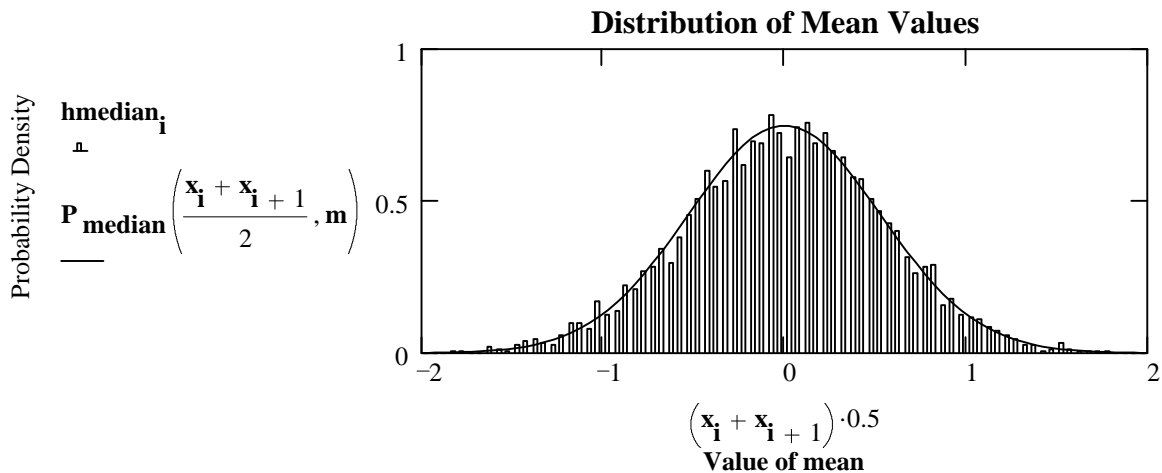
This sorts the elements of the k'th column of matrix d, which is the k'th data set. Since k ranges over all columns, this sorts every data set.

$$G_{\text{median}_k} := d_{m,k}$$

The k'th median value is the m+1 data point, which has index m because the rows start with zero.

$$h_{\text{median}} := \frac{\text{hist}(x, G_{\text{median}})}{N_f \cdot 0.05}$$

Calculate the histogram using the same bins, defined by x, that was used above for the plot of the mean values.

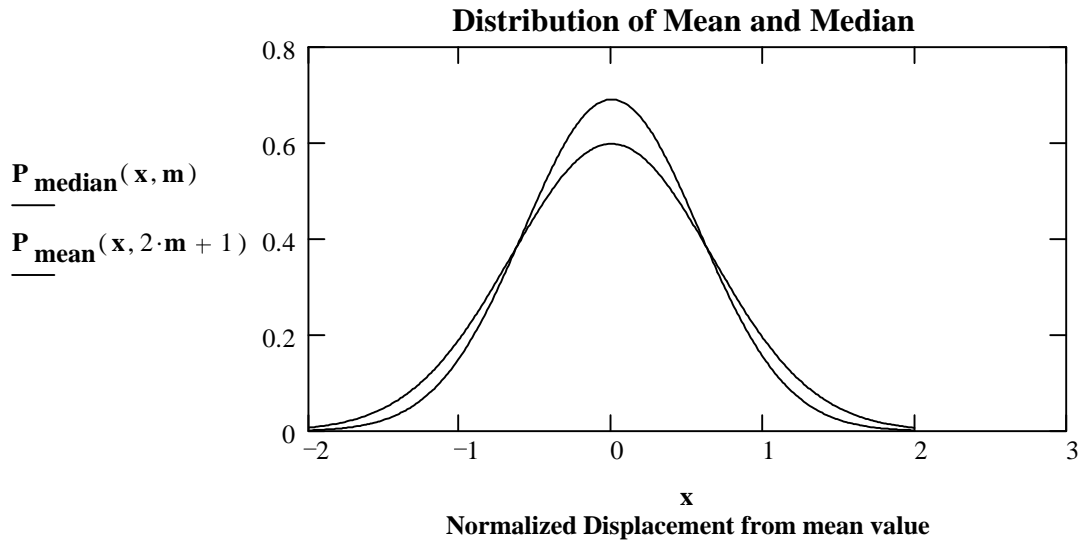


We thus confirm that the function P_{median} predicts the observed distribution of median values. We are now in a position to compare the use of these two central statistics, the mean and the median, to predict the central value of data selected from an unknown Gaussian. These plots which are for a standard Gaussian, can be interpreted as the distribution of the difference between the unknown central value of the Gaussian and the observed value, in units of the standard deviation of the Gaussian. Thus the wider the distribution, the more uncertainty we will have in the value of the predicted center of the Gaussian. Let us compare the predicted distribution of mean values and the distribution of median values. We will compare these distribution both by plotting them on the same graph and by comparing their respective standard deviation since this is widely used to predict confidence intervals. Note, we could have used the distributions of mean and median values calculated from our 'data', but that would be less precise than using the exact distribution function we have derived, since any finite simulation has statistical fluctuations

Let's now plot these two distributions for some selected values:

$x := -2, -1.99..2$ Define range variable for plotting

$m := 1$ How many data points in each sample does this correspond to?



We see that the distribution of the median is somewhat wider than for the mean. Let us compare the values for the standard deviations of these two statistics.

$$\sigma_{\text{mean}}(m) := \frac{1}{\sqrt{2 \cdot m + 1}} \quad \sigma_{\text{mean}}(m) = 0.577$$

$$\sigma_{\text{median}}(m) := \sqrt{\int_{-5}^5 P_{\text{median}}(x, m) \cdot x^2 dx} \quad \sigma_{\text{median}}(m) = 0.67$$

$$\frac{\sigma_{\text{median}}(m)}{\sigma_{\text{mean}}(m)} = 1.16$$

Why did we use limits of ± 5 for the integral instead of $\pm \infty$? Try changing the limits to $\pm \infty$ and see what result is given. Consider how Mathcad does numerical integrals.

Repeat the above plot and calculations of standard deviations for $m = 4, 12, \& 50$. You will need to change the x range of the plot to allow the shapes of the two distribution functions to be compared. Does the ratio of standard deviations appear to be converging?

Mean Vs. Median: The General Case.

Let us assume that we have data selected from some arbitrary distribution function, $P(x)$. Can we make any general predictions about the widths of the expected distribution of sample mean and median values?

It is possible to derive a general expression for the distribution of mean values for samples of N_s points. See problem 3 at the end for this expression. For most cases, it is possible to use the Central Limit Theorem, which states that for any distribution function, $P(x)$, that falls off sufficiently quickly for large x , the distribution of mean values converges to a Gaussian distribution for large N_s . Even for small N_s , it is possible to show that the distribution of sample means has itself a mean value equal to the mean of the distribution from which the data is sampled, and has a standard deviation equal to:

$$\sigma_{\text{mean}}(x, N_s) = \frac{\sigma_x}{\sqrt{N_s}} \quad \text{where:} \quad \sigma_x^2 = \int x^2 \cdot P(x) dx$$

Thus, if we know or can estimate the standard deviation of the function from which the data is sampled, we can predict the standard deviation expected for the distribution of mean values, which is smaller by the well known square root of N factor. Note that an important assumption used in the derivation of the above expression is that the data is statistically independent, which just means that the true probability of getting a particular value for any one measurement is not changed by the fact that we know the results of the previous measurements.

We now consider the distribution of median values. As in the case of the Gaussian, we need to introduce the cumulative probability function, $P_c(y)$ which gives the probability that a data point will be less than y .

$$P_c(y) = \int_{-\infty}^y P(x) dx$$

In terms of this function, the distribution of median values for a sample of $2m+1$ data points can be found by the same argument given previously for the Gaussian distribution:

$$P_{\text{median}}(y, m) = \frac{(2 \cdot m + 1)!}{(m!)^2} \cdot P_c(y)^m \cdot (1 - P_c(y))^m \cdot P(y)$$

It can be shown (see problem 4 at the end) that in the limit of large N_s , the distribution of sample medians of data points selected from a population with continuous distribution $P(x)$ becomes Gaussian centered on the true median of the distribution, μ_m , with a standard deviation given by: .

$$\sigma_{\text{median}} = \frac{1}{\sqrt{4 \cdot N_s \cdot P(\mu_m)^2}}$$

Note that both σ_{mean} and σ_{median} decrease by a factor of the square root of the number of data points per sample (N_s) in the limit of large N_s . The relative merit of the mean and median depends upon the value of $2 \cdot \sigma_x \cdot P(\mu_m)$, which is a dimensionless number. If this number is less than unity, the mean will have a narrower distribution, while if it is greater than unity, the median will have a narrower distribution. A statistic with a narrow distribution of values is said to be 'sharper'.

Consider a normalized Gaussian is given by:

$$P(x) = \frac{1}{\sqrt{2 \cdot \pi}} \cdot \exp\left(\frac{-x^2}{2}\right)$$

What is the median? What is $P(\mu_m)$? What is the ratio of σ_{median} to σ_{mean} for the limit of large N_s ?

We conclude that for a Gaussian distribution, the mean is a 'sharper' statistic for estimating the center of the distribution than the median, though the difference is certainly not dramatic. The expression given earlier predicts that in the limit of large m , the distribution of sample median values will be 1.253 times wider than the distribution of sample mean values.

Consider data sampled from a population with a distribution function: with unknown $k > 0$, and $0 < x < \infty$.

$$P(x) = k \cdot \exp(-k \cdot x)$$

Predict whether you should use the sample mean or median value to determine the best estimate for k . Remember to 'propagate' the error for the statistic to the value of k that is predicted.

Mean Vs. Median for samples drawn from a Lorentzian Distribution:

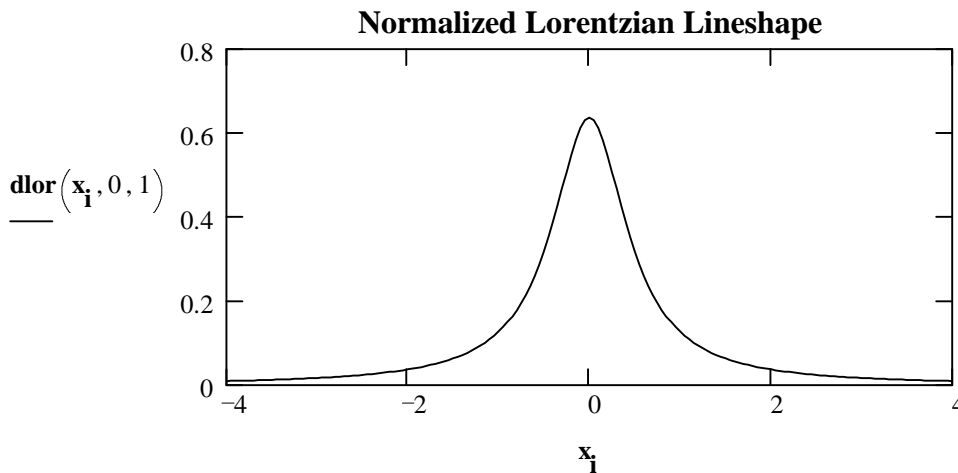
As an example of a distribution function for which the median provides a dramatically sharper statistic, consider a Lorentzian distribution:

$$\text{dlor}(x, x_0, \Gamma) := \frac{1}{\pi} \frac{\left(\frac{\Gamma}{2}\right)}{\left(x - x_0\right)^2 + \left(\frac{\Gamma}{2}\right)^2}$$

In this function, x_0 is both the mean and median value, and Γ is the Full Width at Half Maximum (FWHM), which just means the difference in x values between the points where the curve falls to a value just half as large as the peak. The Lorentzian curve is also often described by the Half Width at Half Maximum (HWHM), which is just half as large as the FWHM. Thus you should be careful to determine which definition an author is using if s/he just refers to the 'width' of a Lorentzian.

The Lorentzian function occurs often in spectroscopy. It describes the distribution of energies of a set of particles that start in a state that undergoes decay with an exponential time dependence. Examples include the energy distribution of states that decay due to spontaneous emission or collisions. A plot of the function is given below:

$i := 0..160$ $x_i := -4 + 0.05 \cdot i$



This plot is in units of Γ , with x_0 set to zero.

In the space provide here make a plot to compare Gaussian and Lorentzian distributions with the same mean value and sample full width at half maximum. Based on your plot, what are the most important differences between the two functions?

Note that the Lorentzian falls off slowly away from the peak. In fact, if one tries to define the standard deviation of the distribution, one gets a divergent integral; the standard deviation is infinite! This has a profound effect on the distribution of means. In fact, it can be shown analytically that the distribution of the mean of N_s points drawn from a Lorentzian distribution function is just the same Lorentzian, i.e. it does not get any narrower! This counter intuitive result is a consequence of the high probability of getting a value far from the mean. We will now demonstrate this fact numerically by sampling points from a Lorentzian distribution function. To do so, we must find a way to generate data points with a Lorentzian distribution. First, we need to define the cumulative distribution.

$$\text{plor}(x, x_0, \Gamma) := \int_{-\infty}^x \text{dlor}(x', x_0, \Gamma) dx'$$

In this case, using the definition of dlors given above, we can do the integral analytically to give:

$$\text{plor}(x, x_0, \Gamma) := \frac{1}{\pi} \cdot \text{atan} \left(2 \cdot \frac{x - x_0}{\Gamma} \right) + \frac{1}{2}$$

Check this result by differentiating with respect to x and demonstrate that you get the Lorentzian function back.

To generate points with the a Lorentzian distribution, we will use the Mathcad function `runif(m,a,b)` which generates `m` pseudo-random numbers with uniform density on the interval `[a,b]`. What we need to do is to find the inverse of the function `plorentzian`, which is easy to do.

$$\mathbf{rlorentzian}(m, \mathbf{x}_0, \Gamma) := \frac{\Gamma}{2} \cdot \overrightarrow{\tan(\pi \cdot (\mathbf{runif}(m, -0.5, 0.5)))} + \mathbf{x}_0$$

The arrow over the top in this definition means that we want Mathcad to calculate the `tan` function for each of the `m` random numbers in the vector returned by the `runif` function.

Let us generate a set of Lorentzian distributed points

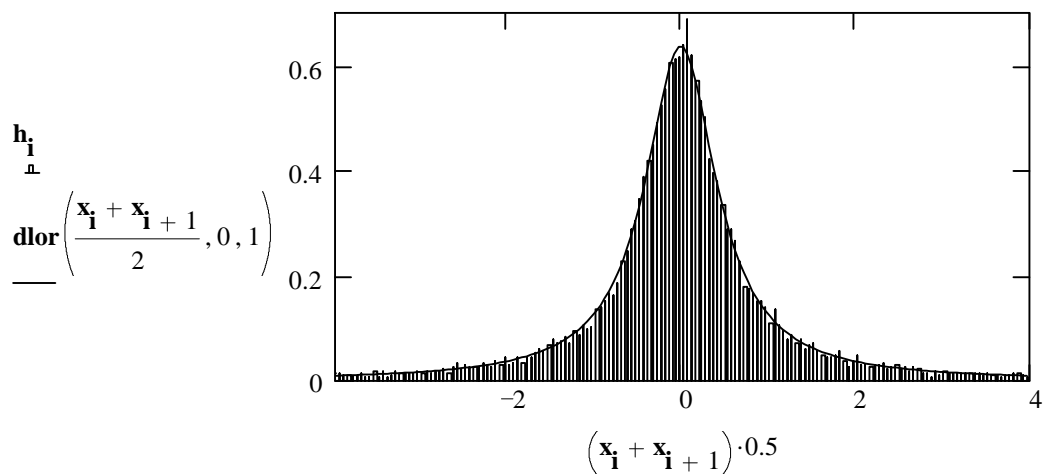
`p := rlorentzian(N,0,1)` What is x_0 and Γ for the distribution from which we sample the data points?
How many points did we generate?

Let's bin this data and compare with the Lorentzian function as a check

`h := hist(x,p)` The factor in denominator is to convert to probability density
 `N·0.05`

`i := 0..159` redefine range variable since h has one less row than x.

Check of Lorentzian Distribution Random Number Generator



What does this graph demonstrate?

Let us now calculate the means for a series of N_t sets of $N = 25$ sample points. In other words we will distribute our N data points into N_t samples of $N_s = 25$ points each.

$$\mathbf{m} := 12$$

$$N_s := 2 \cdot \mathbf{m} + 1$$

Number of data points per sample, must be an odd integer!

$$N_t := \text{floor}\left(\frac{N}{N_s}\right)$$

N_t is the number of samples. Remember that floor(x) returns the largest integer less than or equal to x .

$$N_t = 1 \cdot 10^3$$

$$\mathbf{k} := 0..N_t - 1$$

Range of different data sets

$$\mathbf{j} := 0..N_s - 1$$

Range over data in each sample

$$d_{j,k} := P_{k \cdot N_s + j}$$

Put data into matrix d, with one data set per column

$$d^{<k>} := \text{sort}(d^{<k>})$$

Sort the k'th data sample into ascending order

$$lmean_k := \frac{1}{N_s} \cdot \left(\sum_j d_{j,k} \right)$$

Mean value for the k'th data set

$$lmed_k := (d^{<k>})_m$$

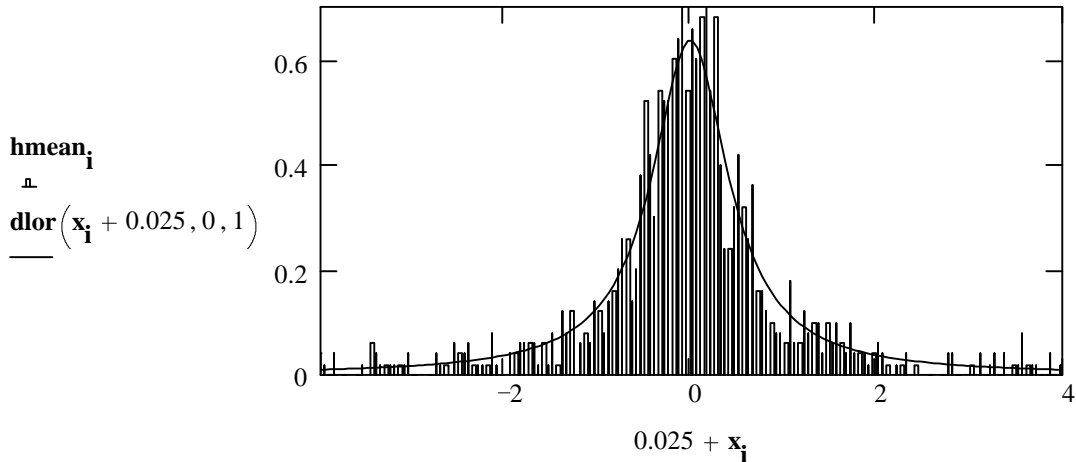
Median value for the k'th data set

Let us now compare the distribution of mean and median values by plotting the histograms of observed values.

hmean := $\frac{\text{hist}(x, \text{lmean})}{N \cdot 0.05}$ Generate histogram of sample mean values

hmedian := $\frac{\text{hist}(x, \text{lmed})}{N \cdot 0.05}$ Generate histogram of sample median values

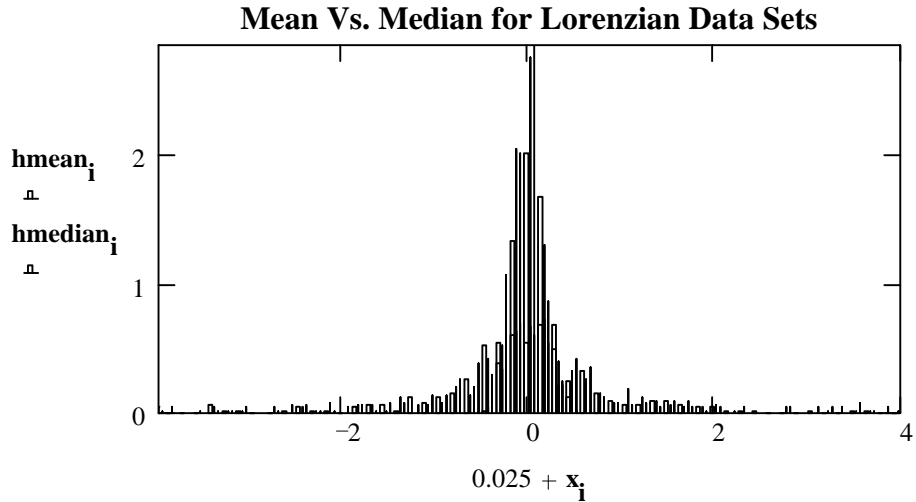
Plot of the distribution function for the mean of 25 points drawn from a Lorentzian distribution function of zero average and unit FWHM



Notice that the $hmean_i$ distribution is, except for increased statistical noise, just the starting Lorentzian distribution! We have gotten zero improvement in the estimate of the center of a Lorentzian distribution if we take one measurement or twenty five. It is possible to demonstrate analytically that the distribution of mean values selected from a Lorentzian distribution is the same Lorentzian as the Lorentzian from which data are selected originally. Clearly, there must be a better statistic!

Why does the above result not violate the general rule that the standard deviation of the mean of N points is just the standard deviation of the sample divided by the square root of N? (Hint: what is the standard deviation for the Lorentzian distribution?)

Here we make a plot to compare the histogram of mean values with the histogram of median values.

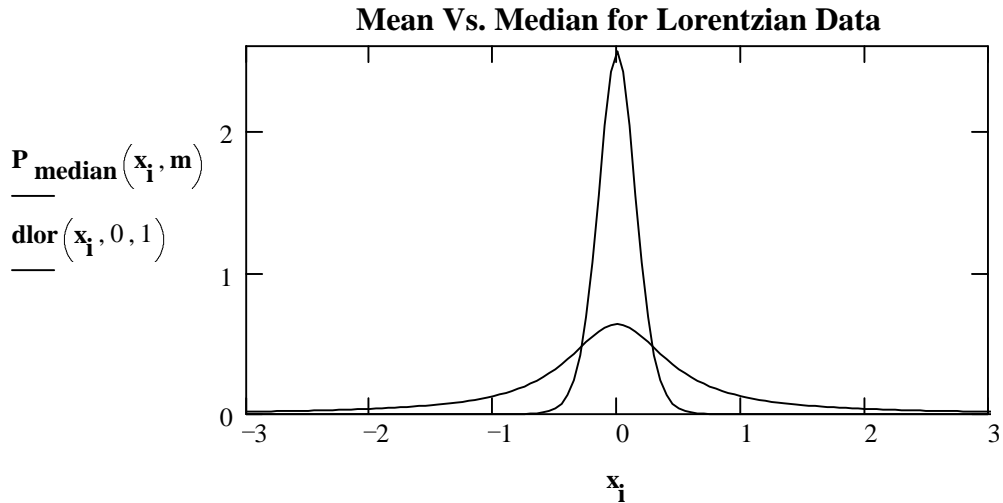


The distribution of sample median values is clearly much more tightly peaked near zero than the distribution of sample mean values. Thus, the median is a much better estimate of the central value than the mean for the case of data sampled from a Lorentzian distribution.

We can analytically determine the distribution of the median values for data sampled from a Lorentzian distribution by the same method we used for the Gaussian above. When we do this we get:

$$P_{\text{median}}(x, m) := \frac{(2 \cdot m + 1)!}{(m!)^2} \cdot \text{plor}(x, 0, 1)^m \cdot (1 - \text{plor}(x, 0, 1))^m \cdot \text{dlor}(x, 0, 1)$$

Let us plot this with the distribution of the mean values on the same scale



This is a very dramatic example of a case where the median is a much better statistic than the mean for estimating the center of a distribution function.

What do you think is the key difference between the Lorentzian and Gaussian distributions that makes the median the best central statistic in the former case, and the mean in the latter case?

Create a numerical way to test your explanation and try it out below.

Make a plot to compare the predicted, P_{median} , and observed distribution of median values.

Maximum Likelihood Method

Is it possible to find an even sharper statistic than the median for estimating the center of a Lorentzian distribution from data points selected from it? Yes, it is. This general method for estimating parameters of a population distribution function from a sample population is known as the **maximum likelihood method**. It is generally considered to provide the best possible estimates for unknown parameters of a distribution function. It also provides an objective way to compare how 'likely' a given set of data is to come from a distribution function of one form or another. As we have seen above, the best way to estimate a central value from a set of data depends upon which form we assume for the distribution function for that data, so it would be valuable to be able to decide if it is more likely that the data was drawn from a Gaussian or Lorentzian distribution function.

The general method is as follows. Imagine we have N points sampled from a distribution function $P(x;p)$, where p is a set of unknown parameters. If $P(x;p)$ is a Gaussian, then the set p would be the mean and standard deviation. If the points are drawn independently, we can calculate the probability density that we would obtain the set of points

$$\vec{x} = (x_1, x_2, \dots, x_N) \text{ as } P(\vec{x}) = \prod_{i=1}^N P(x_i; p). \text{ We seek that set of}$$

parameters p that maximize this probability density. It is usually easier to maximize the $\ln(P)$, since then we get the sum over the \ln of the probability of each point. Many introductory statistics books demonstrate that for an assumed Gaussian distribution, the estimate for the center of the Gaussian that gives the maximum likelihood is the sample mean. We will now use Mathcad's Symbolic Math capabilities to set up the appropriate equations to find the maximum likelihood estimate for an assumed Lorentzian distribution for the data.

$$\ln \left[\frac{1}{\pi} \frac{\frac{\Gamma}{2}}{(x_i - x_0)^2 + \left(\frac{\Gamma}{2}\right)^2} \right] \quad \text{This is } \ln(P(x_i))$$

$$\frac{-1}{\left[(\mathbf{x}_i - \mathbf{x}_0)^2 + \frac{1}{4} \cdot \Gamma^2 \right]} \cdot (-2 \cdot \mathbf{x}_i + 2 \cdot \mathbf{x}_0)$$

Result of differentiating $\ln(P(x_i))$ with respect to x_0 using the Symbolic Math by selecting x_0 in the equation for $\ln(P(x_i))$ and then selecting Symbolics -> Variable -> Differentiate from the menu bar

$$\frac{-\left[-4 \cdot (\mathbf{x}_i)^2 + 8 \cdot \mathbf{x}_i \cdot \mathbf{x}_0 - 4 \cdot \mathbf{x}_0^2 \right] + \Gamma^2}{\Gamma \cdot \left[4 \cdot (\mathbf{x}_i)^2 - 8 \cdot \mathbf{x}_i \cdot \mathbf{x}_0 + 4 \cdot \mathbf{x}_0^2 + \Gamma^2 \right]}$$

Result of differentiating the equation for $\ln(P(x_i))$ with respect to Γ , and then used the Symbolics -> Simplify to get a simpler form.

Using the above derivatives, we find that the values of x_0 and Γ that will satisfy the Maximum Likelihood conditions will satisfy the following two equations (sums are over points in the sample):

$$\sum_i \frac{(\mathbf{x}_i - \mathbf{x}_0)}{\left[(\mathbf{x}_i - \mathbf{x}_0)^2 + \frac{1}{4} \cdot \Gamma^2 \right]} = 0.$$

Derive these equations from the Maximum Likelihood conditions. (Hint, this only requires basic calculus.)

$$\sum_i \frac{4 \cdot (\mathbf{x}_i - \mathbf{x}_0)^2 - \Gamma^2}{4 \cdot (\mathbf{x}_i - \mathbf{x}_0)^2 + \Gamma^2} = 0$$

We can now define a Mathcad function that solves the Maximum Likelihood equations for an input vector of N points. Let's call the function `Lorentian_fit(N,x,x0,Γ)`, We will need to give the Mathcad solver initial values for the center and FWHM. The function defined below passes x_0 and Γ as initial values.

Given

$$\sum_{i=0}^{N-1} \frac{x_i - x_0}{(x_i - x_0)^2 + \frac{1}{4}\Gamma^2} = 0.$$

Equation for derivative of $\ln P$ with respect to changes in x_0 .

$$\sum_{i=0}^{N-1} \frac{4 \cdot (x_i - x_0)^2 - \Gamma^2}{\left[4 \cdot (x_i - x_0)^2 + \Gamma^2\right]} = 0.$$

Equation for derivative of $\ln P$ with respect to changes in Γ .

lorentzian_fit(N, x, x_0, Γ) := **find**(x_0, Γ)

Notice that this solve block contains two equations and two unknowns. N and x are data, x_0 and Γ are the desired statistical parameters.

Note: Mathcad does not yet 'solve' this solve block. Instead, we have defined a function, and Mathcad will attempt to solve these equations when the function `lorentzian_fit` is called. The values of x_0 and Γ that are used as the second and third arguments of the function will act as initial values when Mathcad tries to solve the equations. Mathcad will return the two x_0 and Γ values that satisfy these equations as the result of the function.

Directly solving the above equations does not always lead to a correct solution, because there are singular solutions with x_0 and Γ going to infinity. In cases like this, first solving for one variable at a time and then solving coupled equations can be more stable. We will solve the equation for Γ , for fixed x_0 first because that equation appears to not have any singular solutions for finite x_0 .

Show that if we set $\Gamma = 2x_0$, then for $x_0 \gg$ all x_i , that both of the equations in the solve block will approximately hold.

Show that the first equation in the solve block will be solved, for fixed Γ , when x_0 goes to $\pm\infty$.

Given

$$\sum_{i=0}^{N-1} \frac{x_i - x_0}{(x_i - x_0)^2 + \frac{1}{4} \cdot \Gamma^2} = 0.$$

This block solves for the best x_0 for fixed Γ .

Why do we use this equation to solve for x_0 ?

`lorentzian_mean(N, x, x_0, Γ) := find(x_0)`

Function to find the "best" x_0 value given a fixed Γ and initial guess for x_0 .

Given

$$\sum_{i=0}^{N-1} \frac{4 \cdot (x_i - x_0)^2 - \Gamma^2}{[4 \cdot (x_i - x_0)^2 + \Gamma^2]} = 0.$$

This block solves for the best Γ for fixed x_0

`lorentzian_width(N, x, x_0, Γ) := find(Γ)`

Function to find the "best" Γ value given a fixed x_0 and initial guess for Γ .

Let's try solving the maximum likelihood equations for the different data sets sampled from a Lorentzian above and stored in the matrix d . This will take a while. If your computer is too slow to solve all N_k sets of nonlinear equations, you can restrict the range of the k variable to a smaller value. First find the optimal width for fixed center, using median value as the initial guess for x_0 .

`fitted_width_k := lorentzian_width(N_s, d<k>, lmed_k, 1.)`

Note: for each value of k , the function `lorentzian_width` is called, using a guess of 1.0 for the width and the median, calculated above, for the assumed x_0 value. The function returns the Γ that maximizes the 'likelihood' of obtaining the k 'th data sample from a Lorentzian distribution with known width ($lmed_k$) but unknown full width at half maximum, FWHM.

Now, find the optimal center, x_0 , for fixed width, using the width we just found, and the median for the initial guess for the center.

$$\text{fitted_mean}_k := \text{lorentzian_mean}\left(N_s, d^{<k>}, \text{lmed}_k, \text{fitted_width}_k\right)$$

Note how we are using lmed_k and fitted_width_k for the starting values for the solve block that the function calls.

Now, optimize both the x_0 and Γ parameters at once.

$$\begin{pmatrix} \text{fitted_mean}_k \\ \text{fitted_width}_k \end{pmatrix} := \text{lorentzian_fit}\left(N_s, d^{<k>}, \text{fitted_mean}_k, \text{fitted_width}_k\right)$$

fitted_mean_k and fitted_width_k are now the values of x_0 and Γ that satisfy the maximum likelihood equations for the k 'th data set. These provide the 'best' estimates for these parameters assuming we know that the data was sampled from a Lorentzian distribution for which we did not know the center and width of the distribution.

Warning: If you are running an old version of Mathcad Plus 6 for windows, you may find that Mathcad will fail to 'solve' all N_t equations, i.e. you will get an error when Mathcad tries to evaluate the `lorentzian_fit` function. The solution is to upgrade to at least Patch "e" (the current version when this document was written). You can download upgrade Patches from the MathSoft Web site at <http://www.mathsoft.com/patch.htm>

How does the central value predicted by the maximum likelihood method compare with using the median to estimate the center? Above, we put the sample median values in the vector `lmed`.

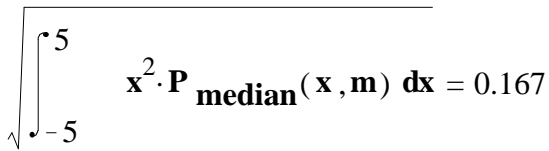
$$\text{mean}(\text{fitted_mean}) = 3.184 \cdot 10^{-3}$$

$$\text{stdev}(\text{fitted_mean}) = 0.143$$

mean(x) and stdev(x)
return the mean and
standard deviation of
the elements of a
vector, x.

$$\text{mean}(\text{lmed}) = 1.791 \cdot 10^{-3}$$

$$\text{stdev}(\text{lmed}) = 0.161$$


$$\int_{-5}^5 x^2 \cdot P_{\text{median}}(x, m) dx = 0.167$$

Exact value for the standard
deviation of a distribution of
median values as calculated from
the exact median distribution
function.

Make an x,y plot with lmedk on the y axis, and fitted_meank on the x axis. Double click the graph and change the plot 'type' to points (under traces). What does this plot demonstrate about the two different estimates of the central value. Are they independent or highly correlated?

We thus find, that the Maximum Likelihood Method returns a statistic that approximates the center of the Lorentzian with smaller dispersion than the median, but the improvement is only modest. Solving the nonlinear equations for the most likely value are much more computationally demanding and may lead to false solutions unless the initial 'guess' for the parameters is sufficiently close to the optimal value for a particular data set. In cases like this, it is often advised to not let the best be the enemy of the good, and give up a little 'theoretical' sharpness for a statistic that is easy to use.

Summary:

In this worksheet, we compared the mean and median values for both theoretical distributions and for data sets sampled from Gaussian and Lorentzian distribution functions. It was found that the mean value provides a moderately better estimate of the central value than the median for the case of a Gaussian. However, in the case of a Lorentzian, due to its slow fall off for large displacements from the central value, the mean is almost useless as a statistic, while the median functions quite well.

We also introduced the idea of finding the optimal estimate by use of the method of maximum likelihood. Application of this method to a Gaussian distribution lead to the expression for the mean value, i.e. the mean value is the best estimate for the central value for a Gaussian distribution. For the Lorentzian distribution, however, the maximum likelihood method leads to a set of coupled nonlinear equations for the parameters. This is the typical situation. We can solve these equations numerically using the built in functions of Mathcad. We find, however, that in this case this optimal estimate gives a standard deviation around the correct value only slightly smaller than that provided by the median value. Thus the median value is almost the optimal estimate of the center of a Lorentzian distribution. Further, we found that in this case, the central value predicted by the sample median and the maximum likelihood methods are highly correlated; they are typically much closer to each other than to the 'true' value of the Lorentzian distribution from which the data was sampled.

Bibliography

Some introductory text books that discuss statistics, as applied to the analysis of experimental data, are as follows, in increasing order of sophistication:

1. John R. Taylor, *An Introduction to Error Analysis, University Science Books*, 1982 (ISBN 0-935702-10-5).
2. Philip R. Bevington and D. Keith Robinson, *Data Reduction and Error Analysis for the Physical Sciences*, McGraw-Hill (1992).
3. E. Bright Wilson, Jr., *An Introduction to Scientific Research*, 1952 (Reprinted by Dover, 1990) (ISBN 0-486-66545-3)

Problems :

1. Generate a large number of points evenly distributed on the unit interval, $[-0.5, 0.5]$. Break the points into a number of data sets, each with N_s points (you select N_s), and calculate the mean and median of each data set. Compare the two statistics as estimates of the central value in this case.
2. Derive an analytic expression for the distribution of median values for a data set of $(2m + 1)$ points sampled from the uniform distribution considered in problem 1. Calculate the standard deviation at large m .
3. For N_s points sampled from an arbitrary distribution function, $P(x)$, it is possible to derive the following expression for the distribution of the sample mean, x_m :

$$P(x_m) = \frac{N_s}{2 \cdot \pi} \int_{-\infty}^{\infty} e^{-i \cdot N_s \cdot k \cdot x_m} \cdot Q^{N_s}(k) dk$$

with:

$$Q(k) = \int_{-\infty}^{\infty} P(x) \cdot e^{i \cdot k \cdot x} dx$$

(see F. Reif, *Fundamentals of Statistical and Thermal physics*, section 1.10 for proof)

$Q(k)$ is the Fourier Transform of the probability density function, and probability density for the mean is (up to a change in scale factor), the inverse Fourier Transform of the N_s 'th power of $Q(k)$.

Use this expression to calculate the distribution of sample mean values for the value of N_s used in your answer to question 1. Plot both the distribution of mean value and distribution of median value for samples of this size. How do they compare? Which do you suggest is the better statistic to use. (You can make this calculation efficient by using Mathcad's FFT functions.)

4. Derive the expression for the standard deviation of the median in the limit of large samples.

$$\sigma_{\text{median}} = \frac{1}{\sqrt{8 \cdot m \cdot P(\mu_m)^2}}$$

Let μ_m be the true median, and x the median of a sample of $(2m+1)$ data points. Let $P(x)$ be the probability density, and $P_c(y)$ the cumulative probability that a data point is less than y .

- What is the value of $P_c(\mu_m)$?
- define $\delta x = x - \mu_m$. Show that $P_c(\delta x) = 0.5 + P(\mu_m) \delta x$ for small δx .
- Use the exact expression for the distribution of the median to show that for sufficiently small δx ,

$$P_{\text{median}}(\delta x, m) = \frac{(2 \cdot m + 1)!}{4^m \cdot (m!)^2} \cdot \left(1 - 4 \cdot P(\mu_m)^2 \cdot \delta x^2\right)^m \cdot P(\mu_m)$$

d. Use the fact that for small a , large m , $(1 + a)^m$ can be approximated by $\exp(a \cdot m)$ to show that in this limit:

$$P_{\text{median}}(\delta x, m) = \frac{(2 \cdot m + 1)!}{4^m \cdot (m!)^2} \cdot P(\mu_m) \cdot \exp\left(-4 \cdot m \cdot P(\mu_m)^2 \cdot \delta x^2\right)$$

e. Comparing to the standard form for a Gaussian distribution:

$$P(\delta x) = \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} \cdot \exp\left[\frac{-(\delta x - \mu)^2}{2 \cdot \sigma^2}\right]$$

show that for small δx , $P_{\text{median}}(\delta x, m)$ follows a Gaussian distribution, centered at zero, (i.e. the mean value for $x = \mu_m$) with a variance given by:

$$\sigma^2 = \frac{1}{8 \cdot m \cdot P(\mu_m)^2}$$

5. Show that for data of N points x_i , that the Gaussian distribution that satisfies the maximum likelihood conditions is the one with mean and standard deviation equal to the sample mean and standard deviation, i.e.

$$\mu = \frac{1}{N} \cdot \sum_i x_i \quad \sigma = \sqrt{\frac{1}{N-1} \cdot \sum_i (x_i - \mu)^2}$$

6. Show that for $(2m+1)$ data points sampled from the probability distribution:

$$P(x) = \frac{1}{2} \cdot \exp(-|x - \mu|) \quad -\infty < x < \infty$$

that the condition for maximum likelihood reduces to μ equals the median of the $(2m+1)$ points.

Hint: $\frac{d}{d\mu} |x - \mu|$ equals -1 for $x > \mu$ and $+1$ for $x < \mu$. Show that the

derivative of $\ln P$ is positive for all μ less than the median, and negative for all μ greater than the median.

7. Given a set of points, how can the maximum likelihood method be used to decide which of two possible forms for the distribution function is the 'most likely'?