

Gaussian Distributions ©

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 introduces the properties of Gaussian distributions, the Mathcad functions used to calculate these properties, and one method to test a distribution to see if it is reasonable to assume the points were drawn from a Gaussian or Normal distribution.

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.

Introduction:

The Gaussian or normal distribution function is widely used in many areas of statistics, including the study of the effects of measurement error, a topic of importance in all areas of experimental science. The form of this distribution is given by:

$$\text{dnorm}(x, \mu, \sigma) := \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} \cdot \exp\left[-\frac{(x - \mu)^2}{2 \cdot \sigma^2}\right]$$

This is one of the predefined functions in Mathcad. This describes the familiar "Bell Curve", with mean μ and standard deviation σ . The special case of $\mu=0$ and $\sigma=1$ is known as the standard Gaussian. Given any variable x that is distributed according to a Gaussian distribution with mean μ and standard deviation σ , it is possible to define a new variable $z = (x-\mu)/\sigma$, which will be distributed according to the standard Gaussian distribution.

The Gaussian function is widely used as a model for random fluctuations or noise. While this is largely because of its convenient mathematical properties, it is often an excellent approximation. Examples include "Johnson" noise which exist in all electronic devices due to thermal motion of the current carriers and shot or counting noise that arise from the fluctuation in the number of carriers of the electrical current, i.e. if in a certain time we expect N electrons to flow through a circuit, then there will be a current noise corresponding to the square root of N electrons in the same time.

In this worksheet, we will numerically demonstrate the properties of a Gaussian distribution and how it can be generated using Mathcad. We begin by showing the predefined Mathcad functions related to the Gaussian distribution, and how they are used. We then generate a large data set of points selected from a Gaussian distribution. We will examine this data set, and then use it to illustrate many of the important general properties of Gaussian distributions, particularly as they are used in error analysis.

Mathcad Functions Related to the Gaussian Distribution:

The probability of finding the independent variable x in the interval [x, x+dx] is given by $dnorm(x,\mu,\sigma)dx$. This is only rigorously correct in the limit that $dx \rightarrow 0$, but it is a good approximation for finite dx as long as $dx \ll \sigma$.

For a standard Gaussian, what is the probability that the variable will be found in the interval between [1,1.01]?

An important 'symmetry' to remember is that $dnorm(x,\mu,\sigma) = dnorm(2\mu-x,\mu,\sigma)$.

Use the definition of dnorm given above to justify this statement.

To find the probability of finding x in a finite interval [a,b], we must integrate the Gaussian distribution function between these points. This integral cannot be 'solved' using the standard methods learned in calculus. In fact, we define the integrated Gaussian probability function by:

$$pnorm(x, \mu, \sigma) = \int_{-\infty}^x \frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} \cdot \exp\left[-\frac{(x' - \mu)^2}{2 \cdot \sigma^2}\right] dx'$$

$\text{pnorm}(a, \mu, \sigma)$ is the probability that a Gaussian distributed x , selected at random, will be $< a$. pnorm is one of the predefined functions in Mathcad. Also predefined is $\text{cnorm}(x)$, which is the cumulative distribution for the standard Gaussian, i.e. $\text{cnorm}(x) = \text{pnorm}(x, 0, 1)$. The Gaussian distribution is normalized, so $\text{pnorm}(\infty, \mu, \sigma) = 1$. The probability of x having some value between $-\infty$ and $+\infty$ must be one! The probability that x will be $> a$ is given by $1 - \text{pnorm}(a, \mu, \sigma)$.

We express the probability of finding x in the interval $[a, b]$ ($a < b$), as:

$$P(a \leq x \leq b) = \text{pnorm}(b, \mu, \sigma) - \text{pnorm}(a, \mu, \sigma)$$

Use the rules of integration to show that this expression is correct. (Hint: remember that one can break an integral from $[-\infty, b]$ into two integrals, between $[-\infty, a]$ and $[a, b]$).

Calculate the probability that a variable distributed according to the standard Gaussian will be found between -1 and +1.

pnorm also has a 'symmetry' that $\text{pnorm}(x, \mu, \sigma) = 1 - \text{pnorm}(2\mu - x, \mu, \sigma)$.

Demonstrate this symmetry numerically.

A widely used variant on the cumulative Gaussian probability distribution is the error function, defined by:

$$\text{erf}(x) = \text{cnorm}(x) - \text{cnorm}(-x)$$

$\text{erf}(x)$ is a predefined Mathcad function.

Compare $\text{erf}(1)$ with the probability calculated above.

Often, we wish to find the inverse of the cumulative distribution function, i.e. the value of 'a' such that the probability that $x < a$ is some specified value. Mathcad has a predefined function, $qnorm(p, \mu, \sigma)$ that gives us this 'a'.

$$qnorm(p, \mu, \sigma) = \text{root}(pnorm(a, \mu, \sigma) - p, a)$$

Use $qnorm$ to find the value of 'a' such that a random point from a standard Gaussian distribution has a 95% probability of being less than this value of 'a'. Repeat for a 95% probability of being greater than 'a'.

Generate Set of Gaussian distributed data points.

There are many situations when we wish to generate points taken from a known distribution, particularly a Gaussian. One important example is when we want to generate 'synthetic' data to test a data analysis method. In such cases, we want to add controlled levels of 'noise' to the data and examine how that effects the results of the analysis. In this worksheet, we will generate a large data set of points from a known Gaussian distribution. Mathcad has a built in function, $rnorm(N, \mu, \sigma)$ which returns a vector of N data points, where the points are drawn from a Gaussian distribution with mean μ and standard deviation σ . More precisely, these numbers are not truly random (they are generated by a reproducible computer algorithm), but they have the same statistical properties as random numbers and are known as pseudo random numbers. See chapter 7 of the text [Numerical Recipes](#) by W. H. Press *et al.* for a thorough discussion of pseudo random numbers and how they are generated.

$\mu := 50$	Average Value of the distribution we will select from
$\sigma := 20$	Standard Deviation of distribution
$N := 100000$	Number of data points we will select
$y := rnorm(N, \mu, \sigma)$	Select points; y is a vector of N elements

We can test how close the statistics of the points we have selected are to those of the distribution function.

$\text{mean}(y) = 1.00018 \cdot \mu$ The function $\text{mean}(y)$ returns the mean or average value of an array of points y .

$\text{stdev}(y) = 0.999895 \cdot \sigma$ The function $\text{stdev}(y)$ returns the standard deviation of an array of points y

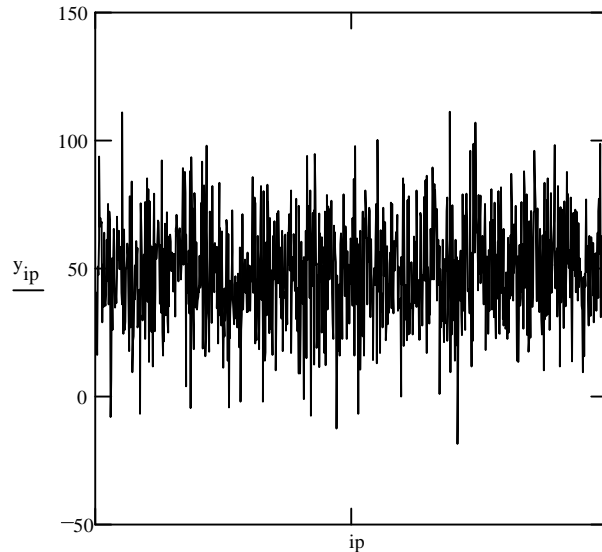
These values are not exactly equal to μ and σ due to fluctuations, which reflect that we only have a finite number of data points. If you select a different distribution and then select "Calculate Worksheet" from the Math menu, these values will change.

Change the specific random numbers that Mathcad calculates by selecting Math->Options->Randomize and then changing the value of the seed. After this, select "Calculate Worksheet" from the Math menu. Mathcad uses the seed value to generate its random numbers.

Next plot a subset of our points:

$N_{\text{plot}} := \min((N - 5000))$ **What value does N_{plot} have?**

$ip := 0, 1 .. N_{\text{plot}}$ Range variable to loop over data points



Vary the maximum value of ip in the graph and describe your observations.

This looks very much like what you would see on an oscilloscope as you looked at the output of a good amplifier with the gain up high. It is what is known as "White Noise". Note that it consists of a "band" centered at μ , with a width of $\sim 5\sigma$.

Verify that the width in the above figure is approximately 5σ .

Comparison of the histogram of data with the Gaussian distribution function

Let us now compare our distribution with the Gaussian function. We do this by 'binning' our data into a histogram. Again, Mathcad makes this easy.

$$i := 0..199 \quad x_0 := -5 \cdot \sigma + \mu$$

$$\delta x := 0.05 \cdot \sigma \quad x_{i+1} := x_i + \delta x$$

$$f := \text{hist}(x, y)$$

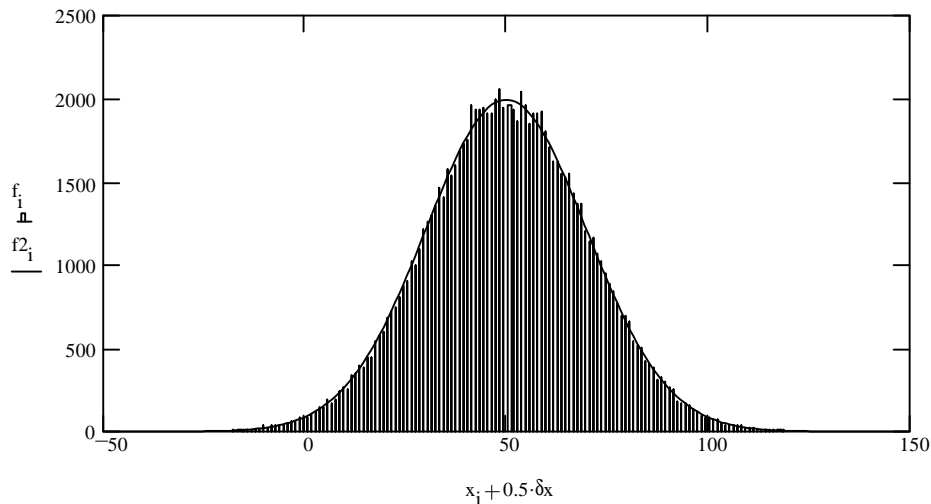
f is a vector with the number of times points of y fall in each bin.

$$f2_i := N \cdot \delta x \cdot \text{dnorm}\left(\frac{x_i + x_{i+1}}{2}, \mu, \sigma\right)$$

Gaussian Distribution for expected number of data points in each bin.

The "bins" are defined between the values in vector x. Recall that μ is the mean. x_0 is the location of the first bin, δx defines the size of each bin. The bins are filled automatically. **How many bins are being generated here? What range of values to the bins cover?**

Why is the Gaussian distribution function multiplied by N (the number of data points) times δx , the size of each bin? Why is the function evaluated at $(x_i + x_{i+1})/2$?



Why does the number of points in each bin not exactly match the predictions of the Gaussian Distribution Function used to generate the data?

Estimating the mean of a distribution function from observed data.

In most applications of statistics to laboratory data, we want to know the true value of the mean for the distribution. We cannot determine this exactly, but we can find an estimate by using a finite number of measurements. Let us now consider splitting our large data set into a set of smaller data sets, so that we can compare how our estimates do.

$N_s := 4$ Number of data points per set; start with 4

$N_t := \text{floor}\left(\frac{N}{N_s}\right)$ Number of data sets (floor(x) is the largest integer $<$ or $=$ to x)

How many data sets of N_s points do we now have?

For a Gaussian distribution, the best estimate of μ from a data set is just the mean of the data. By "best", we mean that this will give on average the lowest root mean squared deviation from the true value. Next we compute the average for each set of 4 points and create the avg vector.

$k := 0..N_t - 1$

$$\text{avg}_k := \frac{1}{N_s} \cdot \left[\sum_{j=0}^{N_s-1} y_{N_t \cdot j + k} \right]$$

What does the k'th element of the avg_k vector contain?

We now compute the statistics of the vector avg.

$\text{mean}(\text{avg}) = 50.009$ $\text{stdev}(\text{avg}) = 10.006$

Express the mean of avg in terms of μ , the standard deviation of the initial distribution, by selecting the result, and then putting μ in where one would type units (the black box after the number). Express the stdev in terms of σ , the standard deviation of initial distribution.

Compare these results for those calculated for the entire distribution above.

Note: the average value of the means is just the mean for the total distribution. The standard deviation of the distribution of means is the standard deviation of the initial distribution divided by the square root of the number of data points averaged. This result will hold for any distribution function as long as the standard deviation is finite.

Let us now look in detail at the distribution of means:

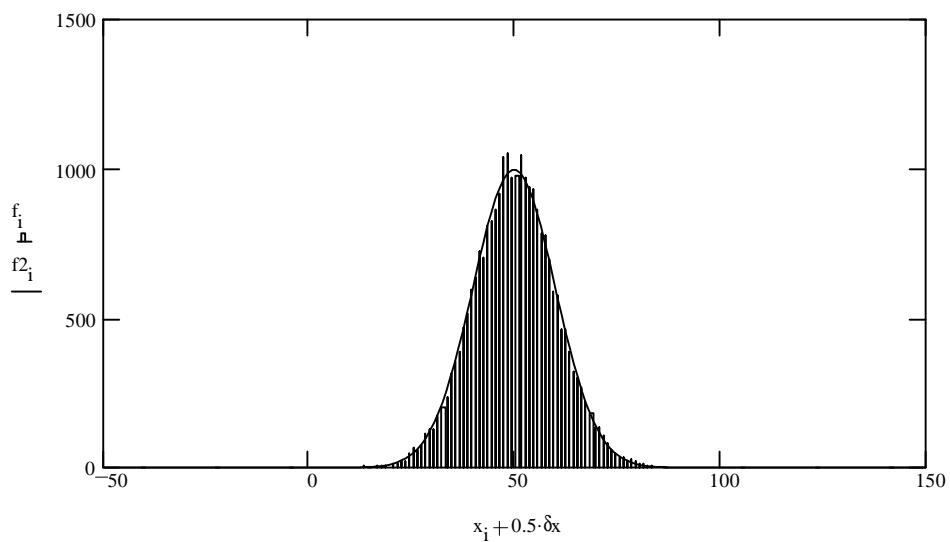
$f := \text{hist}(x, \text{avg})$

Histogram of distribution of mean values

$$f2_i := N_t \cdot \delta x \cdot \text{dnorm}\left(\frac{x_i + x_{i+1}}{2}, \mu, \frac{\sigma}{\sqrt{N_s}}\right)$$

Gaussian with width scaled by $\sqrt{N_s}$

Distribution of Means Compared with Gaussian



Notice that the distribution of means is again a normal distribution, but with the width reduced by the square root of the number of points averaged. For a general distribution, the distribution of the mean of N_s measurements will not be Gaussian, but will have a standard deviation smaller than that of the sample by the factor of square root of the number of data points. Further, for a broad class of distribution functions, the Central Limit Theorem asserts that the distribution of the mean of N_s values approaches arbitrarily close to a Gaussian Distribution for sufficiently large N_s .

Confidence Intervals.

In scientific measurement, no specification of a physical quantity is complete unless it is accompanied by an error estimate. The properties of Gaussian distributions are widely used for this purpose. If we know the true standard deviation of the distribution from which our data was selected, this is easily done using the Mathcad function qnorm, which gives the value of the normal distribution such that there is a given probability that a point at random will be less than that value. We just have to recognize that a normal distribution is symmetric about its mean, so it shows equal probability above or below the mean.

$$\text{Interval}(p, \mu, \sigma) := \left[\begin{array}{l} \text{qnorm}\left(\frac{1-p}{2}, \mu, \sigma\right) \\ \text{qnorm}\left(\frac{1+p}{2}, \mu, \sigma\right) \end{array} \right]$$

Gives the Confidence Interval for a normal distribution. p is the probability level. You must provide a value for p as shown below.

Why are the first arguments of the qnorm function (1-p)/2 and (1+p)/2 for the lower and upper limit of the confidence interval?

Let us calculate a few confidence intervals for the standard Gaussian with unit width and zero mean. The results can be interpreted for a general Gaussian in term of how many σ we must go from the mean

$$\text{Interval}(0.9, 0, 1) = \begin{pmatrix} -1.645 \\ 1.645 \end{pmatrix} \quad \text{Interval}(0.95, 0, 1) = \begin{pmatrix} -1.96 \\ 1.96 \end{pmatrix}$$

$$\text{Interval}(0.99, 0, 1) = \begin{pmatrix} -2.576 \\ 2.576 \end{pmatrix} \quad \text{Interval}(0.999, 0, 1) = \begin{pmatrix} -3.291 \\ 3.291 \end{pmatrix}$$

Thus we see we expect the points to fall within $\sim 3.3\sigma$ of the mean 99.9% of the time. Let us see how well this predicts our selected set of points.

$$\begin{pmatrix} \text{low} \\ \text{high} \end{pmatrix} := \text{Interval}(0.99, \mu, \sigma) \quad \begin{pmatrix} \text{low} \\ \text{high} \end{pmatrix} = \begin{pmatrix} -1.517 \\ 101.517 \end{pmatrix}$$

$$\frac{1}{N} \cdot \left[\sum_{i=0}^{N-1} (y_i < \text{low}) + (y_i > \text{high}) \right] = 0.01044$$

Here we calculate the fraction of points outside of interval(low,high). Not many points lie outside the 99% confidence level. **Change the confidence level to 99.9%. What do you observe?**

Explain why the above expression works to give the fraction of data points outside the confidence interval.

We thus can estimate error (Confidence) intervals **IF** we know the σ for our distribution. How do we estimate this statistic?

Estimating the standard deviation of the distribution from a set of data.

It turns out that we want to first get an estimate for the square of the standard deviation, which is known as the variance. We get an estimate for this by the following expression:

$$\text{var}_k := \frac{\sum_{j=0}^{N_s - 1} (y_{N_t j+k} - \text{avg}_k)^2}{N_s - 1}$$

var_k is the variance of the k'th set of N_s data points. We can compare the mean of var with the variance for the original normal distribution:

Calculate the mean of the vector var and express in terms of σ^2 . The result should be that mean of var is almost equal to σ^2 .

Note that if we had divided by N_s instead of $N_s - 1$, we would get a smaller variance which would be in much worse agreement with the true value for the starting distribution. This is because we are using the avg instead of the true mean of the distribution in the definition of var above. For a particular sample, the mean of the sample, avg, is shifted from the true mean of the distribution, μ , in a way that removes some of the scatter of the data. The Mathcad functions var and stdev compute the variance and stdev by dividing by the N_s , not $N_s - 1$, and thus do not give the correct values for small data sets.

Define a new vector, var2_k , which is the average of the squared deviation of the points in the k'th data set from μ :

$$\text{var2}_k := \frac{\sum_{j=0}^{N_s - 1} (y_{N_t j+k} - \mu)^2}{N_s}$$

What is the mean of var2 in terms of σ^2 ? Note that in the definition of var2 we divide by N_s , not N_s-1 , but we subtract the true mean, μ , not the sample mean, avg_k , as we did for the sample variance, var_k defined above.

Note that we average the variance, and then take the square root to get the standard deviation. Show that for our data set, the average of the sample standard deviations, the square root of the elements of var, does not average to give a good estimate of σ .

The χ^2 function:

We have seen from above that on average the sample variance (with division by $N_s - 1$) gives the variance of the Gaussian distribution from which the data was selected. This is true for most distribution functions. However, to understand how good an estimate it is for a single set of data, we need to know what the distribution of sample variances is. In the theory of statistics, it is shown that the distribution function for sample variance can be expressed in terms of the χ^2 distribution function ('chi-squared function') defined as:

$$\chi^2(x, d) := \frac{e^{-\frac{x}{2}}}{2 \cdot \Gamma\left(\frac{d}{2}\right)} \cdot \left(\frac{x}{2}\right)^{\frac{d}{2} - 1}$$

Note: χ^2 here is a function of real $x \geq 0$ and positive integer d .
This equation is toggled off.

where x is defined on the range $[0, \infty)$, and d is an integer parameter known as the 'degrees of freedom'. $\Gamma(d/2)$ is the Gamma Function. The χ^2 distribution function is predefined in Mathcad as $dchisq(x, d)$.

The probability density function for sample variances (usually written s^2 in statistics) computed from a sample of N_s data points selected from a Gaussian distribution function with variance σ^2 is written as:

$$P(s^2) \cdot ds^2 = \frac{d}{\sigma^2} \cdot \chi^2\left(\frac{s^2 \cdot d}{\sigma^2}, d\right)$$

with: $d = N_s - 1$

Note: The parentheses to the right of χ^2 is the argument of χ^2 . The notation here is due to difficulties with writing some equations in Mathcad.

Based upon this distribution function, what is the probability that a sample of 4 data points will have a sample variance the same as that of the true distribution to within $\pm 1\%$?

Let us now compare the observed distribution of sample variance values, var_k , with that predicted by the χ^2 distribution:

$m := 0..200$ $\delta x := 0.02$ $x_m := \delta x \cdot m \cdot \sigma^2$ Set up bins boundaries spaced by $0.02\sigma^2$

$f := \text{hist}(x, \text{var})$

Put calculated variances into bins

What is contained in the m'th element of vector f?

$m := 0..199$

Now, make m range over bins, which is one less than number of boundary values.

$d := N_s - 1$

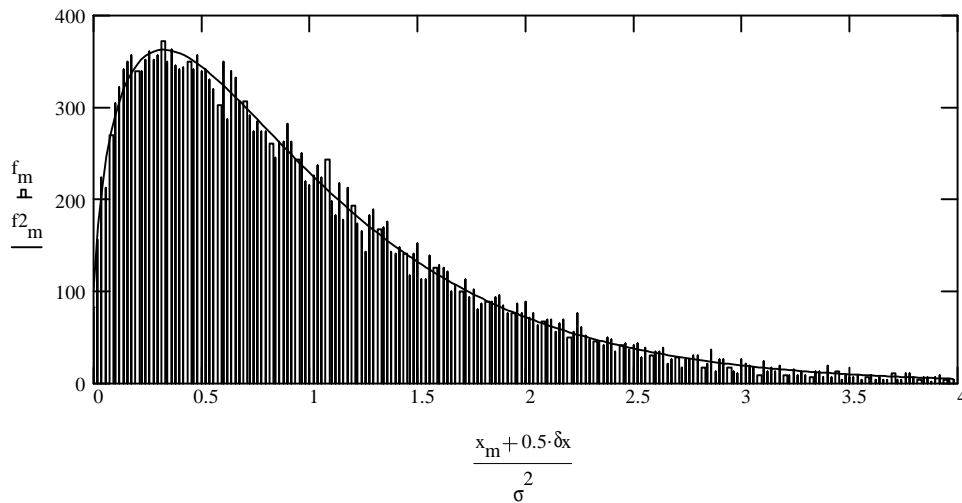
"Degrees of freedom"

$$f2_m := N_t \cdot 0.02 \cdot d \cdot \text{dchisq} \left[\frac{0.5 \cdot (x_m + x_{m+1})}{\sigma^2}, d, d \right]$$

Prediction for number of points in each bin interval, centered on $0.5(x_m + x_{m+1})$

Why does this expression have the given prefactor for the dchisq function?

Distribution of Calculated Variances, compared with expected χ^2 function



The important lesson to learn from the above graph is that for a sample of modest size, the estimate of the standard deviation calculated from the computed variance can be quite different from the true value for the distribution. In particular, we have a high probability of significantly underestimating the real σ . Thus, if we compute a confidence interval using this too small σ , we will get an unrealistically small value for our true uncertainty! This mistake is all too common!

To demonstrate this effect, let us calculate how often our calculated mean value (avg) differs from the true mean by more than "2 σ ", where we use the observed variance to estimate σ . We expect, as shown above, that the standard deviation for the the mean of N_s measurements to be smaller by $\sqrt{N_s}$.

$$\frac{1}{N_t} \cdot \sum_k \left(\left| \text{avg}_k - \mu \right| > 2 \cdot \sqrt{\frac{\text{var}_k}{N_s}} \right) = 0.138$$

This is much larger than the 5% of the time that we would have expected based the normal distribution.

If instead, we use the true value for σ , we of course get a reasonable estimate:

$$\frac{1}{N_t} \cdot \sum_k \left(\left| \text{avg}_k - \mu \right| > 2 \cdot \sqrt{\frac{\sigma^2}{N_s}} \right) = 0.04816$$

This is close to the 5% expected for a Gaussian distribution.

How do we get a reliable error estimate if we must estimate σ from the data itself?

Student's t distribution:

We use a distribution function known as the Student's t distribution, which Mathcad also has as a built in function. We can compute the 95% confidence interval from the observed standard deviation, if we multiply by:

$$\text{range} := \text{qt}(0.975, N_s - 1)$$

$$\text{range} = 3.182$$

The reason we use $p = 0.975$ for the first argument of the qt function instead of 0.95 is that qt returns the value of x (in units of the calculated standard deviation) such that an observed value will occur less than a fraction p of the time. The second argument is d , the degrees of freedom. You will see by changing values of the second argument that the true error confidence interval for a small d is quite a bit larger than when we can assume we know the value for σ . For $d=1$, the 95% interval is greater than six times the traditional " 2σ " estimate, and for $d=10$ it has reduced to only 10% larger

Let us test the fraction of sample means that differ from μ by more than the range predicted by the student's t distribution.

$$\frac{1}{N_t} \cdot \sum_k \left(\left| \text{avg}_k - \mu \right| > \text{range} \cdot \sqrt{\frac{\text{var}_k}{N_s}} \right) = 0.05016 \quad \text{Much closer to the 5\% that we expect!}$$

Explain why the above expression gives the fraction claimed.

An important, often overlooked, point is that the expected square root of N reduction in the uncertainty of the mean is only valid if the points are rigorously independent. This means that the probability of observing a particular value for one point is completely independent of the values observed for the other points. This is often not the case for real laboratory data. The noise in any experiment has its own spectrum. This can often be summarized by giving a time constant, which tells, qualitatively, how long on average it takes the output readings of an instrument to fluctuate on either side of their average values. When data is recorded by a computer, it is possible to make measurements very rapidly, often much faster than the rate at which the output is changing. In these cases, the successive measurements are not independent and one does not expect the uncertainty in the mean to fall as the square root of the number of measurements, since they are highly redundant.

Moments of a Gaussian distribution:

There are many cases where we need to know the average value of a function, $f(x)$, of a variable that follows a Gaussian Distribution. This can be expressed as:

$$\text{mean_of}(f) := \int_{-\infty}^{\infty} f(x) \cdot \left[\frac{1}{\sqrt{2 \cdot \pi \cdot \sigma^2}} \cdot \exp \left[\frac{-(x - \mu)^2}{2 \cdot \sigma^2} \right] \right] dx$$

In favorable cases, we can compute the integral analytically. In most other cases, it can easily be calculated by numerical integration. There exists a very efficient method, known as Gauss- Hermite integration, for the numerical evaluation of the integral of a smooth function times a Gaussian. Another method that we can use is to take N points from the Gaussian distribution, x_i , and compute the average value of $f(x_i)$. This method is known as Monte-Carlo Integration. The average of an infinite set of $f(x_i)$ value is just $\text{mean_of}(f)$, as defined above. We cannot, of course, use an infinite number of points, but if we use a finite number, their average will provide an estimate for $\text{mean_of}(f)$. However, we will still have statistical fluctuations. The standard deviation for the distribution of sample means of f is given by:

$$\text{square_root_of_distribution_mean_of}(f) := \sqrt{(\text{mean_of}(f^2) - \text{mean_of}(f)^2)}$$

Using our data set, y , calculate the average value for $f(x) = |x-\mu|$. Based upon the expression for the standard deviation of the distribution of $f(x)$ values, estimate the likely 2s uncertainty of this estimate.

By the central limit theorem, we can predict that in the limit of large N , the distribution of the average value of N values of $f(x_i)$ will become a Gaussian distribution with the above standard deviation divided by \sqrt{N} and mean equal to the $\text{mean_of}(f)$. The central limit theorem states that for any distribution function $P(x)$ with a mean μ and finite standard deviation σ and satisfying some other mathematical conditions, the distribution function for the mean of N independent points selected at random $P(x)$ will, in the limit of large N , approach a

Gaussian Distribution with mean μ , and a standard deviation of $\frac{\sigma}{\sqrt{N}}$.

A common set of functions that we want to look are the average and distribution of are called 'moments'. The n'th moment of a distribution P(x) is just the mean of (x^n) . The n=1 moment is just the mean value, μ . For $n > 1$, we often want the central moments, which are the mean of $(x - \mu)^n$. The n=2 central moment is just the variance, σ^2 . For a Gaussian distribution, we can use the symmetry of the distribution to show that the odd central moments ($n = 3, 5, \dots$) vanish. For the even central moments, we can use the definition and integration by parts to show that $2k$ central moment = $(1 \cdot 3 \cdot \dots \cdot (2 \cdot k - 1)) \cdot \sigma^{2k} = \frac{((2 \cdot k)!) \cdot \sigma^{2k}}{2^k \cdot (k)!}$. The prefactor grows rapidly with k, being equal to 1, 3, 15, 105, 945 for $k = 1-5$.

Use integration by parts to derive a relationship between the $2k$ and $2(k+1)$ central moments of a Gaussian distribution function. For $k=0$, the moment equals to one since it is the normalization integral. Derive the above expression for the general even moments by induction.

We often report the normalized n'th central moment, a dimensionless number equal to the n'th central moment, by σ^n . The normalized moments characterize the shape of the distribution. The normalized third moment is called the **skew** and is represented by the symbol γ_1 . A positive value for γ_1 means the distribution has a longer 'tail' on the positive side of the mean, the opposite for a negative γ_1 . Distributions symmetric about the mean, such as the Gaussian, have zero γ_1 .

The normalized fourth central moment is often called the **kurtosis**. It is always positive and has a value of 3 for a normal distribution function. A value below 3 means that the distribution dies faster in the wings than a Gaussian. For example, a uniform distribution ($P(x) = \text{constant}$ if x is in $[a, b]$, zero otherwise) has a kurtosis of 9/5. A kurtosis greater than 3 means that the distribution has a 'longer' tail than a Gaussian.

Derive the skew and kurtosis of the uniform distribution in the interval $[0, 1]$. Use Mathcad's symbolics to evaluate the necessary integrals.

Another 'moment' often used is the mean absolute deviation (or misleadingly mean deviation), which is the average of $|x-\sigma|$. For a Gaussian distribution, this has a value of $\sqrt{\frac{2}{\pi}} \cdot \sigma = 0.798$. The ratio of the mean absolute deviation to the standard deviation is larger for a distribution more compact than a Gaussian (it equals 0.866 for a uniform distribution), and smaller for a distribution with longer tails.

Derive the above expression for the mean absolute deviation of a Gaussian distribution. Note that you can break the necessary integral into equal contributions for the parts above and below μ .

Tests for a Gaussian Distribution.

Since most common statistical tests make the assumption that the sample follows a Gaussian distribution, it is important to test this assumption to detect at least gross deviations. One common use for the low order moments is to provide such a test. Let us consider the samples of Gaussian data points we selected earlier in this worksheet and let us calculate the distribution of the normalized absolute deviation as well as the normalized third and fourth central moments. In order for the test to be at all meaningful, we need at least ~ 10 points per sample, so **go back up to page 8 and change N_s to 10**. For small sample sizes, we cannot directly use the expected Gaussian distribution since often we do not know the true σ of the Gaussian but must instead use the sample variance to estimate σ . This introduces additional error, as we discussed above when we discussed the student t distribution. As an example, let us compare the 'moments' calculated using the estimated quantities and the true values for the initial distribution.

If N_s has not been changed to 10 back on page 8 the following discussion will not appear correct.

$$\frac{1}{N} \sum_{j=0}^{N-1} \frac{|y_j - \mu|}{\sigma} = 0.798$$

Here we computed the normalized mean absolute deviation from the entire sample of N points. It agrees well with the expression give above.

$$\text{avg_abs_dev}_k := \frac{1}{N_s - 1} \cdot \sum_{j=0}^{N_s - 1} \frac{|y_{N_{tj+k}} - \text{avg}_k|}{\sqrt{\text{var}_k}}$$

Here we compute the normalized mean absolute deviation for each of the sets of N_s points, using the sample mean and standard deviations in each case.

$$\text{mean}(\text{avg_abs_dev}) = 1$$

The mean of these estimates of the normalized mean absolute deviation does not give the correct value! This is because of the bias introduced by using the sample avg and var.

$$\text{stdev}(\text{avg_abs_dev}) = 0.075$$

This gives an measure of the range of variation in the computed set of mean absolute deviations. Note that it is relatively small.

Let us now make the same comparisons for the third and fourth central moments. First the third:

$$\frac{1}{N} \cdot \sum_{j=0}^{N-1} \frac{(y_j - \mu)^3}{\sigma^3} = -0.013$$

Near zero, as it should be for a Gaussian

$$\text{skew}_k := \frac{1}{N_s - 1} \cdot \sum_{j=0}^{N_s - 1} \frac{(y_{N_{tj+k}} - \text{avg}_k)^3}{(\text{var}_k)^{1.5}}$$

Third central moments for each set of N_s points computed using sample means and standard deviations.

$$\text{mean}(\text{skew}) = -1.97 \cdot 10^{-3}$$

The average value is almost the same as above, indicating that in this case the bias is not large.

$$\text{stdev}(\text{skew}) = 0.506$$

This show the size of the fluctuations of the computed third moments, which are not small.

Now, the fourth:

$$\frac{1}{N} \cdot \sum_{j=0}^{N-1} \frac{(y_j - \mu)^4}{\sigma^4} = 3.008 \quad \text{Close to the value 3 predicted above for the kurtosis}$$

$$\text{kurtosis}_k := \frac{1}{N_s - 1} \cdot \sum_{j=0}^{N_s - 1} \frac{(y_{N_t j + k} - \text{avg}_k)^4}{(\text{var}_k)^2}$$

$$\text{mean}(\text{kurtosis}) = 1.35$$

$$\text{stdev}(\text{kurtosis}) = 0.261$$

Why is this so different from the value 3 predicted for a Gaussian distribution? Can you justify why it is smaller than predicted?

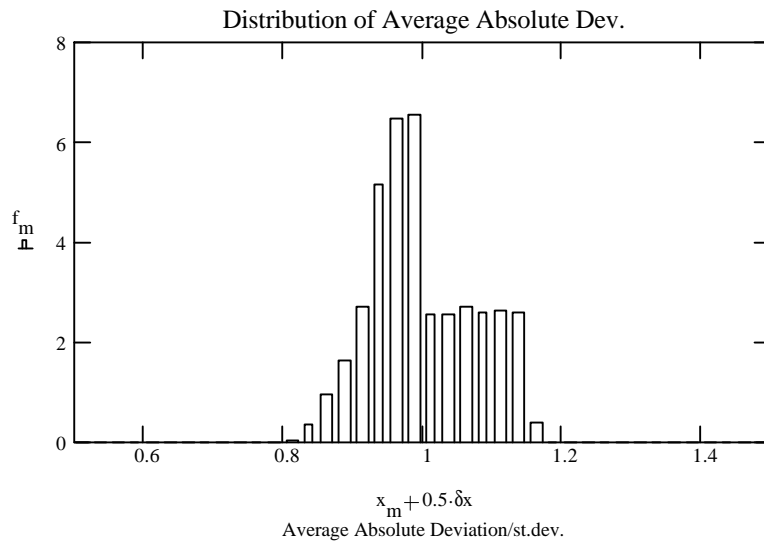
The above results suggest a strategy for testing if a set of points are likely sampled from a Gaussian distribution. We compute the above moments for the data, and then compare the results to those found from sets of the same number of random numbers known to be drawn from a Gaussian distribution. However, because of fluctuations, we will not be able to decide with certainty. Any finite set of points has some probability of being observed when we pull points from a distribution, like the Gaussian, that has nonzero probability density for every value of x .

In order to tell if the moments calculated from the observed points is likely for points from a Gaussian distribution, we need to look at the distribution of values we obtained in our simulation.

Let us make a histogram of the results, first for the mean absolute deviation.

$$m := 0..40 \quad \delta x := 0.025 \quad x_m := 0.5 + \delta x \cdot m$$

$$f := \frac{\text{hist}(x, \text{avg_abs_dev})}{N_t \cdot 0.025}$$



Remember that to see the correct figure you needed to change the value of N_s on page 8.

This distribution is rather 'tight', making for a sensitive test. We see that values of the mean absolute deviation less than 0.6 and more than 1.0 times the sample standard deviation are highly unlikely.

From the above results, we can estimate a 90% confidence interval for this statistic. We do this by putting the set of calculated values, avg_abs_dev , in ascending order by using the Mathcad sort function. Then determine the value of the the elements with arguments closest to $0.05N_t$ and $0.95N_t$.

$avg_abs_dev := sort(avg_abs_dev)$

$avg_abs_dev_{floor}(0.05 \cdot N_t) = 0.886$

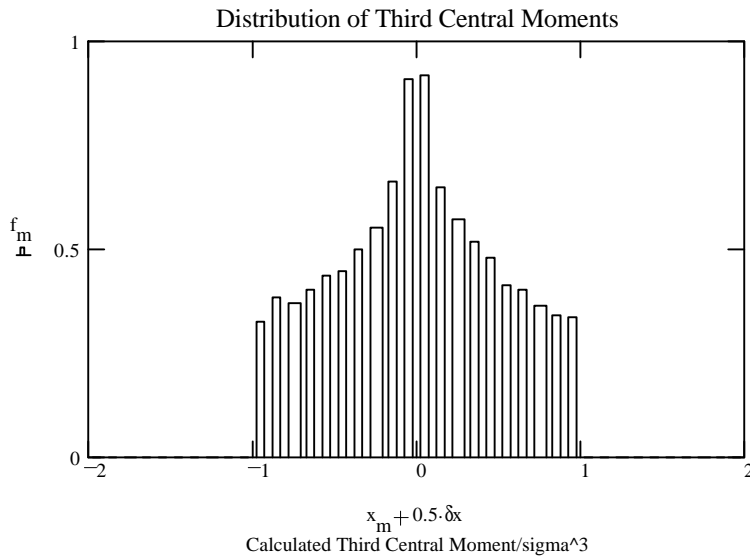
$avg_abs_dev_{ceil}(0.95 \cdot N_t) = 1.135$

If we decided that data sets with mean absolute deviations greater than this value are not drawn from a Gaussian distribution, we would only falsely reject data in fact drawn from a Gaussian distribution 10% of the time.

Determine the 98% confidence interval.

Now, let's look at the distribution of the normalized third moment (skew):

$m := 0..40$ $\delta x := 0.1$ $x_m := -2 + \delta x \cdot m$ $m := 0..39$ $f := \frac{\text{hist}(x, \text{skew})}{N_t \cdot 0.1}$

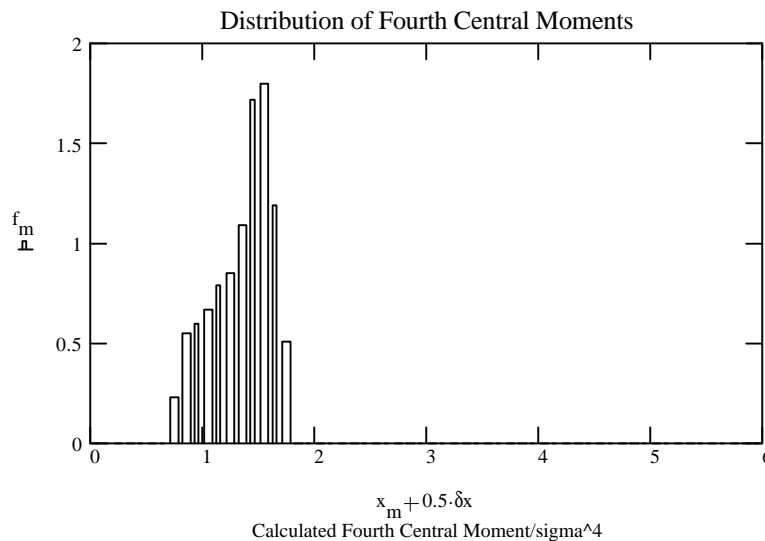


Remember that to see the correct figure you needed to change the value of N_s on page 8.

Determine the 90% confidence interval for the skew. By symmetry, the upper and lower values should be symmetrically placed around zero, but the values you determine are unlikely to be so. Why did you not get symmetric values?

The distribution of the normalized fourth moment (kurtosis):

$$m := 0..60 \quad \delta x := 0.1 \quad x_m := 0. + \delta x \cdot m \quad m := 0..59 \quad f := \frac{\text{hist}(x, \text{kurtosis})}{N_t \cdot 0.1}$$



Remember that to see the correct figure you needed to change the value of N_s on page 8.

Determine the 90% confidence interval for the kurtosis.

Thus, we see that we can use the above three statistics to test a data set of N_s points. Note, that even if the computed statistics are well within the range expected for points drawn from a Gaussian distribution we cannot conclude that our observed data is drawn from Gaussian distribution with any statistical confidence. However, we often make the "Null hypothesis" that the observed data follow a Gaussian distribution and do our analysis of the data accordingly, unless the data indicates that this assumption is unlikely. Note, that if we reject the data if it fails any one of the three tests, with 90% confidence intervals, we will, on average, reject Gaussian data far more often than 10% of the time! If the tests are independent, we would keep the data only $(0.9)^3 = 73\%$ of the time. To be 'safe', we should take the confidence interval for each of the test three times more stringent than we want for the overall test. We rarely test moments higher than the fourth, since the fluctuations of such statistics grows very rapidly

Author's Note: It is hoped that this worksheet demonstrated the basic properties of the Gaussian function and how to generate and work with Gaussian distributions using the Mathcad program.

Problems:

Molecules in a gas or liquid at equilibrium at temperature T have a distribution of each component of velocity described by a Gaussian distribution with mean zero and variance equal to RT/M , where $R = 8.314 \text{ J K}^{-1} \text{ mol}^{-1}$ is the Gas constant, and M is the molar mass.

1. Write down the probability distribution function for v_z ?

Consider below the gas to be made up of Nitrogen molecules ($M = 28 \text{ g mol}^{-1}$) at $T = 298 \text{ K}$.

2. What is the mean value for v_z ? root mean value for v_z ?

3. What fraction of the molecules have $v_z = 100 \pm 1 \text{ m/s}$?

4. What fraction of the molecules have $-100 \text{ m/s} < v_z < +100 \text{ m/s}$?

5. For what value s do 99.99% of the molecules have $|v_z| < s$?

6. Consider samples of N_s points selected from a Gaussian distribution with variance σ^2 . Plot the expected distribution of sample standard divided by σ for $N_s = 5, 10, 25, 50$.

Mathcad has predefined the cumulative χ^2 distribution function, $\text{pchisq}(x,d)$, and the inverse cumulative χ^2 distribution function, $\text{qchisq}(p,d)$. d is the degrees of freedom and p a probability value.

7. Use the pchisq function to determine what fraction of samples of $N_s = 10$ will have a sample variance less than 50% of the variance of the Gaussian population from which the data was selected.

8. Use the qchisq function to determine a 90% confidence interval for the variance of a Gaussian distribution from knowledge of the sample variance of a single sample of $N_s = 10$ data points.

9. For $d = 9$, use Mathcad's Symbolic integration to demonstrate that the $\chi^2(x,d)$ is normalized for $0 < x < \infty$. Then, compute the mean value of x and the variance of x .

The Mathcad function `runif(N,a,b)` returns N pseudo random numbers uniformly distributed on the interval (a,b) . Use this function to generate 10,000 sets of 10 data points on the interval $(0,1)$.

10. What is the mean and variance of this uniform distribution function?

11. Calculate the mean of each data set and plot the distribution of mean values. Compare that plot with a Gaussian of the mean and variance predicted by the central limit theorem.

12. Calculate the variance of each set of data points, and plot the observed distribution. Compare to the χ^2 form predicted for a Gaussian distribution of the same variance.

13. Calculate the distribution of normalized mean absolute deviations for the sets of data points. Plot the results and compare to that found previously for a Gaussian distribution. What fraction of the data sets have values that fall outside the 90% confidence interval for a Gaussian defined above.

14. Repeat question 13, but for the skew. What is the true skew of the uniform distribution?

15. Repeat question 13, but for the kurtosis. What is the kurtosis for the uniform distribution?

16. Based upon the results of questions 13-15, discuss if one can decide with reasonable reliability if a data set of 10 points is drawn from a Gaussian and not a uniform distribution function.