BPHS 4090 (Fall 2013) - Computer Exercise 6: erf(**w**)

Due Date: Dec. 4, 2013 5:00 PM

Background

Random Numbers and Normal Distributions

Suppose you took N random numbers (during an individual *trial*), each of whom had a value between 0 and 10 and each number independent of the others. Now suppose you lumped all those numbers into integer-wide bins, creating a *histogram*. Since any given number is random, it is equally likely that the number is anywhere between 0 and 10. Thus your histogram might look a bit haphazard depending upon a given trial. Now as N increases (i.e., you have more and more numbers for a given trial), your histogram might appear to start flattening out since there are more chances for any given number to appear during a particular trial (Fig.1.1).
Now what would happen when you took the average of these N numbers? Suppose we define the average (or *mean*) as[1]

$$\mu = \frac{1}{N}\sum_i x_i \tag{1}$$

The average will clearly depend upon the size of N is, but you might reasonably expect a value close to 5. This is because it is just as likely a given random number is between 0 and 5 as it is between 5 and 10. Thus when you take the average, it will be close to 5. But for any given trial, there may be a few more numbers below 5 than above (or vice versa), meaning that the mean μ could be less (or greater) than 5.
Consider now what might happen if you repeated this averaging of N numbers over M trials. What would the distributions of mean values look like? Considering our logic above, we should get values clustered about 5, some above and some below. The remarkable thing is that we actually get a 'bell-shaped' distribution as the number of trial (M) becomes large (Fig.1.1). This sort of distribution is called *normal* (or *Gaussian*) and is the basis for what is known as the *central limit theorem*. This theorem is quite important in scientific and engineering contexts as most uncertainty or random error you will encounter will have statistical properties that are normally distributed (i.e., 'bell-shaped').
A normal distribution can be described by the following function:

$$y(x) = Ae^{-(x-\mu)^2/2\sigma^2} \qquad A, \sigma > 0 \tag{2}$$

[1] As an aside, statistics texts will label the 'average' as

$$\bar{x} = \frac{1}{N}\sum_i x_i$$

and the mean as

$$\mu = \lim_{N\to\infty}\left(\frac{1}{N}\sum_i x_i\right)$$

Thus, the notation μ is intended to indicate the mean for an infinitely large sample pool (approximated by what is known as the 'parent distribution'). For our purposes here and the sake of clarity (especially considering that in data analysis, only a finite number of observations can be made), we will not consider the distinction between finite and infinite N and content ourselves that $\bar{x}$ gives a sufficiently good estimate of μ. The same notion applies to how the standard deviation is defined.

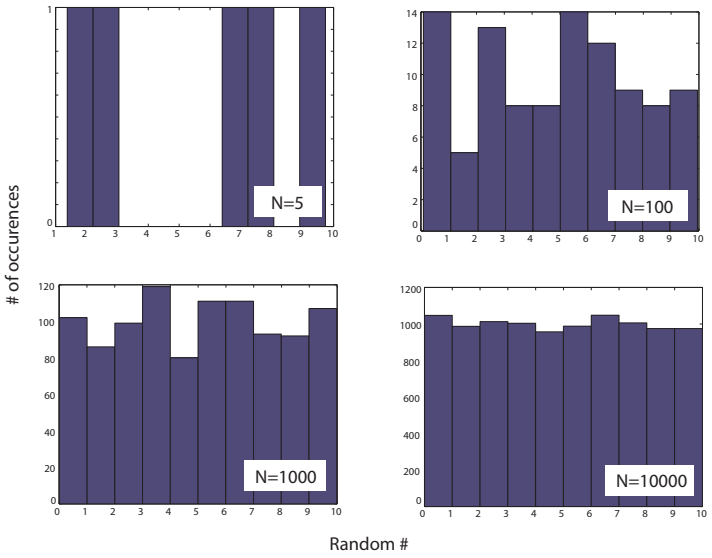

Figure 1: Distributions (or *histograms*) of N random numbers between 0 and 10. Notice that as N gets larger, the distribution appears more and more flat. Note the difference in the vertical scales.

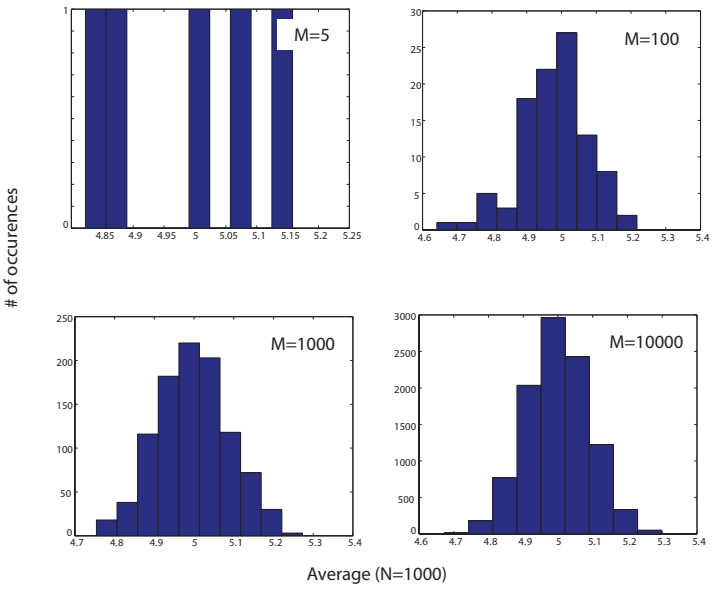

Figure 2: Distributions of the average of $N = 1000$ random numbers between 0 and 10. Notice that as M gets larger, the distribution more centralized about 5 and takes on a more 'bell-curved' shape. Also note the difference in the both horizontal and vertical scales in how they compare across different M values.

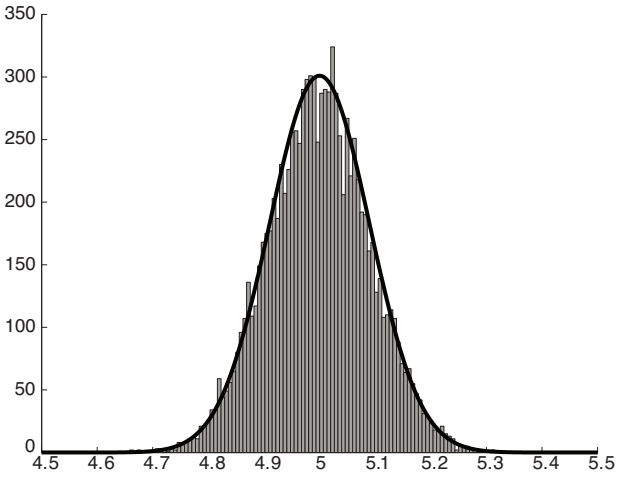

Figure 3: Distributions of the $M = 10000$ averages of $N = 1000$ random numbers between 0 and 10. The black curve indicates the function given by Eqn.2 where $\mu = 5$, $\sigma = 0.1$ and A chosen arbitrarily to fit the data. Note that a different 'bin size' used in the histogram relative to the one shown in Fig.1.1.

Think for a moment what this function will look like. The exponential is always positive, thus the function will be as well. The argument in the exponent will also always be negative no matter what value x is. Thus this function will increase as $|x - \mu| \to 0$ and decrease as x moves away from μ. Because of the square, the sign of x does not matter and the function will be symmetric with respect to $x = \mu$. If we take the derivative of Eqn.2 with respect to x, it is apparent that the slope of the function is zero when $x = \mu$ and approaches zero for large positive or negative values of x (due to the exponential). Additionally, it is clear that the sign of the slope changes from negative to positive about $x = \mu$. Thus, we end up with a curve that is effectively 'bell-shaped' (Fig.1.1).
Note that there are effectively three different parameters in the function given by Eqn.2. A describes how tall the peak is. μ specifies the horizontal location of the peak (i.e., with respect to the x-axis). σ determines how wide the peak is: the smaller σ is, the more narrow the peak is. Based upon our initial example, we wrote down a formula that would allow us to explicitly determine what μ was (Eqn.1). We can similarly do so for σ, such that

$$\sigma^2 = \frac{1}{N-1}\sum_i (x_i - \mu)^2 \tag{3}$$

This value μ is called the *standard deviation* and is a common number used when describing statistical properties of a set of measurements. Note the square in Eqn.3.

Integrating the Bell-Shaped Curve

In a number of useful applications, it is desirable to evaluate an integral where the integrand takes the form of the function given by Eqn.2. This function does not have an explicit anti-derivative, making it unclear at first how to directly integrate an expression of the form

$$w(x) = \int_a^b A e^{-(x-\mu)^2/2\sigma^2}\, dx \tag{4}$$

For simplicity, assume for the moment that $A = 1$, $\mu = 0$ and $\sigma = 1/\sqrt{2}$. Then the integrand simply becomes $y(x) = e^{-x^2}$. The value of this integral will obviously depend upon what the values of a and b are. However, one can 'guess' as to what the solution must look like in general. When x is either a large positive or negative number, the function will be close to zero and there will be little area underneath, thus little contribution to the integral. The function will be largest when $x = 0$ and thus the largest contribution from the integral will come around that point. The function is always positive, thus the integral will always be positive. Suppose we fix a at a large negative number and consider the integral as a function of the upper bound b. Based upon the characteristics described above, if b is also large and negative, the integral will be small. As b is increased, the integral will increase in value. Eventually when you get to $b = 0$, you will get 1/2 the total value of the integral. The rate of increase will reach a maximum (with respect to b) since the largest contribution of the integral comes from that region. As b is further increased, there will be additional contribution to the integral, but it will get progressively smaller with b. Eventually for large positive values of b, there will be very little contribution. Thus as you extend your limits b and a out towards plus and minus ∞ respectively, you will obtain something sigmoidal in shape[2]. A particularly important example is called the *Gaussian integral*, an improper integral (due to the infinite bounds), given as

$$\int_{-\infty}^{\infty} e^{-x^2}\, dx = \sqrt{\pi} \tag{5}$$

Equation 2 takes on many different names, including: the bell-shaped curve, a Gaussian function, and the probability density function. In the latter case, the amplitude A must be chosen in such a way that the total area underneath the curve is unity (since the probability of *something* happening must ultimately be one). This is accomplished by having $A = 1/\sigma\sqrt{2\pi}$ such that

$$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x-\mu)^2/2\sigma^2}\, dx = 1 \tag{6}$$

Note that Eqn.6 is an example of an indefinite integral, where all possible x values are considered.
Another common number you might encounter is the full-width at half maximum (FWHM). This is the width spanned by the function at one half of its maximum value and is given by FWHM=$2\sigma\sqrt{2\ln 2}$, irrespective of whatever value the the amplitude term A takes (Eqn.2).
Suppose now that you want more flexibility in the bounds when integrating Eqn.4. One solution is to either use a means to numerically estimate the integral or look up the needed value in a table (many math books still contain tables of numerical estimates of integrals of the same form as given by Eqn.4!). However, there is another very useful way to think about the area under the curve. By determining the number of standard deviations away you are from the mean, Fig.1.2 shows a straight-forward means to think about what percentage of the total area would be contained within a given region. For example, a region bounded by $\pm 2\sigma$ about the mean contains 95% of the area. Another way to put it, 95% of all points comprising a normal distribution are with two standard deviations of the mean. Thus if you know μ and σ, you have a direct measure of the statistical 'spread' of your data.

The Error Function

Our discussion up to this point has motivated random numbers have this striking 'bell-shaped' symmetry in their statistics and that we can write down an analytic expression for that shape, which upon doing

[2]Note that solving this integral is equivalent to solving the differential equation

$$\frac{dw}{dx} = e^{-x^2}$$

The sigmoidal curve would be the solution curve $w(x)$ to this differential equation.

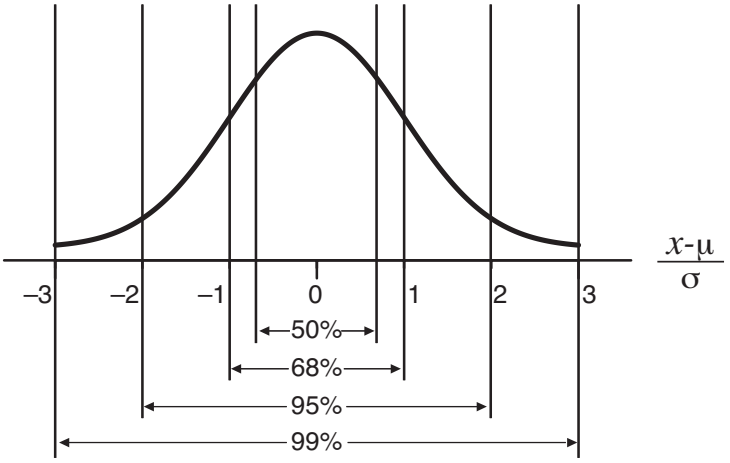

Figure 4: Percentage area underneath the error function as a function of how many standard deviations you are away from the mean (μ). For example, if you set bounds at $\pm x$ such that $|x - \mu| = 2\sigma$ and integrate over that region, you will cover $\sim 95\%$ of the total area that is underneath the curve (i.e. when the bounds extend to $x = \pm\infty$).

some analysis yields some potentially useful tools. We now introduce the error function, which is defined as

$$\mathrm{erf}(w) = \frac{2}{\sqrt{\pi}} \int_0^w e^{-t^2}\, dt \tag{7}$$

The error function is an odd function [i.e. erf(w) = $-\mathrm{erf}(-w)$]. Similar to our reasoning used for Eqn.4, erf(w) will also be sigmoidal in shape. This is shown in Fig.1.3. erf(w) is another way to think about what was shown in Fig.1.2, or in other words, the probability in which a given measurement might be erroneous. Let $t = (x - \mu)/\sqrt{2}\sigma$, then Eqn.7 becomes

$$\mathrm{erf}(w) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}} \int_0^w e^{-(x-\mu)^2/2\sigma^2}\, dx \tag{8}$$

Now suppose we evaluate erf(σ). In this case, the integral on the right-hand side of Eqn.8 will contain 34% of the area underneath the integrand (since we are one standard deviation away from the mean in the positive x-direction). From Eqn.6 (which provided with the appropriate normalization such that the total area underneath the curve is one), the value of the integral must thus be $0.34\sigma\sqrt{2\pi}$. Thus we have

$$\mathrm{erf}(\sigma) = \frac{\sqrt{2}}{\sigma\sqrt{\pi}} 0.34\sigma\sqrt{2\pi} = 0.68 \tag{9}$$

So in this case, $\mathrm{erf}(\sigma)$=0.68 tells us that there is a 68% probability that the error of a given measurement will be within one standard deviation of the mean.
Note that we could have written erf(w) as

$$\mathrm{erf}(w) = \frac{1}{\sqrt{\pi}} \int_{-w}^w e^{-t^2}\, dt \tag{10}$$

assuming that the distribution is symmetric[3]. One might also encounter the *complementary* error function, defined as

$$\mathrm{erfc}(w) = \frac{2}{\sqrt{\pi}} \int_w^\infty e^{-t^2}\, dt = 1 - \mathrm{erf}(w) \tag{11}$$

[3]For our purposes here, we only considered symmetric distributions. However, one can encounter situations where the

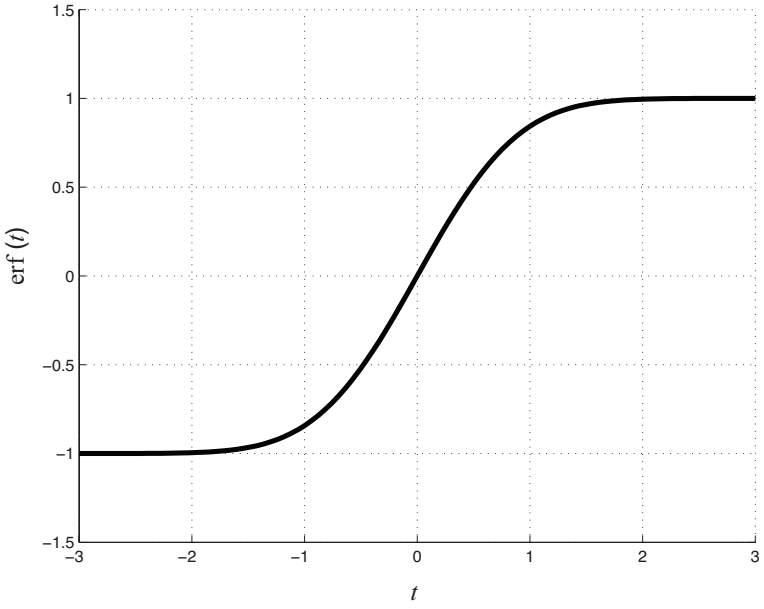

Figure 5: erf(w)

Exercises

- Write a Matlab code to reproduce Figs.1, 2, and 3.
- Use a substitution to verify that Eqns.6 and 5 are equivalent.
- Derive the expression for the FWHM (Hint: remember what the definition of FWHM is!).
- Write a Matlab code to plot $\mathrm{erf}(w)$ by numerically solving Eqn.7. You might want to use the Matlab function trapz to numerically integrate the function.
- What happens to $\mathrm{erf}(w)$ as σ changes? Explain how decreasing σ leads to a change in the probability that one might expect a measurement to be erroneous.
- Use Matlab to numerically integrate Eqn.2 (let $A = 1$, $\mu = 0$ and $\sigma = 1/\sqrt{2}$) in such a way to verify Eqn.5.
- Even though you can not write down an anti-derivative for the integral in Eqn.5, think about other ways you could estimate the value of the integral analytically.
- Briefy describe two different ways that $\mathrm{erf}(w)$ manifests in biophysical problems.

distribution is skewed towards one side and thus a different set of tools than those outlined here are needed. The skewness of a distribution is quantified by what is called the third-moment of the distribution. The fourth-moment of the distribution, or the *kurtosis*, describes the 'peakedness' of the distribution.