

### LAB 3

This goal of this lab is to familiarize the student with the Central Limit Theorem, an amazing result from probability theory that explains why the Gaussian distribution (aka "Bell Shaped Curve" or Normal distribution) applies to areas as far ranging as economics and physics. Below are two statements of the Central Limit Theorem (C.L.T.).

I) "If an overall random variable is the sum of very many elementary random variables, each having its own arbitrary distribution law, but all of them being small, then the distribution of the overall random variable is Gaussian".

II) Let  $Y_1, Y_2, \dots, Y_n$  be an infinite sequence of independent random variables each with the same probability distribution. Suppose that the mean ( $\mu$ ) and variance ( $\sigma^2$ ) of this distribution are both finite. Then for any numbers  $a$  and  $b$ :

$$\lim_{n \rightarrow \infty} P \left[ a < \frac{Y_1 + Y_2 + \dots + Y_n - n\mu}{\sigma\sqrt{n}} < b \right] = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-(1/2)y^2} dy$$

Thus the C.L.T. tells us that under a wide range of circumstances the probability distribution that describes the sum of random variables tends towards a Gaussian distribution as the number of terms in the sum  $\rightarrow \infty$ .

Some things to note about the C.L.T. and the above statements:

a) A random *variable* is not the same as a random *number*! Devore in "Probability and Statistics for Engineering and the Sciences" defines a random variable as (page 81):

"A random variable is any rule that associates a number with each outcome in S".

b) If  $y$  is described by a Gaussian distribution with mean ( $\mu$ ) = 0 and variance ( $\sigma^2$ ) = 1 then the probability that  $a < y < b$  is:

$$P(a < y < b) = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-(1/2)y^2} dy$$

c) The C.L.T. is still true even if the  $Y_i$ 's are from different probability distributions! All that is required for the C.L.T. to hold is that the distribution(s) have a finite mean(s) and variance(s) and that no one term in the sum dominates the sum. This is more general than definition II).

1) In this exercise we will see how the mass of trays containing small balls tends towards a Gaussian distribution. There are nine holes in a tray and hence the mass is the sum of nine random variables. As a demonstration of the more general form of the central limit theorem, we intentionally choose to have two different hole diameters, four of small diameter and five of large diameter, to represent the two different kinds of random variables. We should obtain a mass distribution that looks Gaussian regardless of whether the hole diameters are the same or different. The experiment in this lab is rather simple and all the manipulations should be done above a large plastic tray to prevent the lost of balls. Fill up a small plastic beaker with balls

and then pour the balls into the holes in the tray. Use a wood stick to wipe across the tray to remove excess balls. Measure the mass and repeat the process a total of one hundred times.

In this exercise, there are two contributions to the uncertainty in the number of balls in the each hole (and hence the mass). One is the packing of the balls in each hole which is slightly different in each pouring. The other is the slightly different number of balls being removed in each wipe. Due to the two uncertainties, we obtain a slightly different number of balls (mass) each time. The Central Limit Theorem tells us that the total mass should be distributed approximately like a Gaussian probability distribution. To see if this is true make a histogram of the 100 measurements ( $x$  axis = mass,  $y$  axis = # of times a mass within a certain range was measured). A typical bin width for the histogram might be  $\Delta x = 0.8$  gm. Calculate your average mass ( $\mu$ ) and variance ( $\sigma^2$ ) of your 100 measurements. Superimpose a Gaussian p.d.f. on your histogram (this can be done with Kaleidagraph) and comment on how Gaussian-like your measurements are. Remember that for a given bin size ( $\Delta x$ ) a Gaussian p.d.f. function describing  $N$  measurements with mean  $\mu$  and variance  $\sigma^2$  is given by:

$$p(x) = \frac{N\Delta x}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

2) In this exercise we will use the computer and definition II) to illustrate the C.L.T. This exercise uses the properties of the random number generator (RND). The random number generator gives us numbers uniformly distributed in the interval  $[0, 1]$ . This uniform distribution ( $p(x)$ ) can be described by:

$$p(x) = 1 \text{ for } 0 < x < 1$$

$$p(x) = 0 \text{ for all other } x.$$

a) Prove using the integral definitions of the mean and variance that the uniform distribution has  $\mu = 1/2$  and  $\sigma^2 = 1/12$ .

b) According to definition II) if we add together 12 ( $Y_1 + Y_2 + \dots + Y_{12}$ ) numbers taken from our random number generator RND then:

$$P\left[a < \frac{Y_1 + Y_2 + \dots + Y_{12} - 6}{1} < b\right] \approx \frac{1}{\sqrt{2\pi}} \int_a^b e^{-(1/2)y^2} dy$$

This says that just by adding 12 random numbers (each between 0 and 1) together and subtracting off 6 we will get something that very closely approximates a Gaussian distribution for the sum ( $\equiv Z = Y_1 + Y_2 + \dots + Y_{12} - 6$ ) with  $\mu = 0$  variance  $\sigma^2 = 1$ ! Write a program to see if this is true. Generate  $10^4$  values of  $Z$  and make a histogram of your results. I suggest using  $x$  bins of 1 unit, e.g.  $Z < -5$ ,  $-5 \leq Z < -4.0$ ... $Z > 5$ . Superimpose a Gaussian with  $\mu = 0$  and  $\sigma^2 = 1$  on your histogram and comment on how well your histogram reproduces a Gaussian distribution.

NOTE: save this program, we will use it again in LAB 5.