Homework 1.2

1. Assume that the height of the U.S. adult male population is normally distributed. The height of the entire population has some true mean *mu* and true standard deviation *sigma*. If we draw 100 heights at random (call this set *z*), our sample will have some sample mean *x* and sample standard deviation *s*. The likelihood of our sample mean not being within the population mean +- 1 inch is heavily dependent on *sigma*. Unfortunately, we don't know what *sigma* is. However, we can estimate it to a good degree of accuracy by computing the standard deviation of *z*. Indeed, I found that if I generate a set of 1 million numbers following a Gaussian distribution and take 100 random samples from that, the mean and standard deviation of my sub-sample is very close to the mean and standard deviation of the whole population (at least, more often than not). Now, the probability that our sample mean is less than the population mean minus 1 (assuming the means are in inches) is

$$x - N*s < mu - 1$$

$$x - N*s < x - 1 \; [mu \sim x]$$

$$N*s > 1$$

$$N > 1/s$$

where N is the number of standard deviations. For instance, if we find that s = 0.2, then N = 5 and the probability that our sample mean is less than the population mean minus 1 is 5 sigma. This translates to much less than a fraction of a percent (just compute 1 – erf(N/sqrt(2))). Since the problem asks for the probability outside the range of mu +- 1, we just multiply our end result by two (since the distribution is symmetric). Note this logic relies on *s* being near *sigma*. It follows that, in general, the probability that our sample mean is outside the population mean +- 1 inch is

$$P = 2*[1-erf(1/ (s * sqrt(2)))]$$

I recognize that this doesn't solely rely on techniques from problem 1, but the additional material is just a bit of algebra combined with determining the probability of something being N sigmas away from a mean.

If I just wanted to make a general statement on probability here, I would say something like: "The probability of x being outside of mu +- 1is heavily dependent on sigma. If sigma is small, then the probability of being outside the interval is small. If sigma is large, then the probability of being outside the interval is large." I don't feel that this is quantitative enough though, hence why I did the above calculation.

2. In the above question, I stated that the mean and standard deviation of the sample is "more often than not" very close to the population mean and standard deviation. One program I could write would be computing the above value of P and comparing it to the actual probability. So, I would generate a large sample of heights (~1 million), then pick 100 of those heights at random. I would then calculate the standard deviation of the sample and plug it into P. I would compare that value of P to the fraction of heights within mu +- 1 in my large sample. I can repeat this process a bunch of times to estimate the reliability of my technique described in 1. and plot the agreement in a histogram. In a sense, I would be obtaining a probability distribution of a set of probabilities.

Alternatively, I can write a program that doesn't worry about P. I can generate a million normally distributed heights as my population, then draw 100 random samples from the population and compute their mean. I can repeat the drawing of 100 random samples N times (e.g. 1000 times) and plot their means in a histogram. I can then analyze the resultant histogram to find the fraction of means contained within +- 1 inch of the sample mean (the sample mean will still be a good approximation of the population mean). The probability of our sample mean being outside of the 1 inch interval is then 1 minus that fraction. It might be constructive to compare this approach to the approach discussed above.