Math 214
Applied Statistics

Laboratory Project #4
Due: Monday March 19

Investigating The Central Limit Theorem

Key to understanding Inferential Statistics (and most of what follows in Mth 214) is the most popular statistical LAW known as the *Central Limit Thoerem.* In a nutshell, this powerful theorem states three facts about the statistics of *sample means.*

Remember what we are doing: Given a poopulation, we take samples of size n and from these samples we compute a statistic. Lets take our statistic to be the mean of the sample (this is the sample mean, which is a random variable - it changes for each sample)). We redo our sampling many times, and look at the distribution of the sample mean.

Given random samples of size n selected from some population with mean $= \mu$ and standard deviation σ, the following relationships hold:

- The mean of the population and the mean of the sample means are EQUAL.

$$\mu_{\overline{x}} = \mu$$

- The standard deviation of the population (σ) and the standard deviation of the sample means ($\sigma_{\overline{x}}$) are related by the formula:

$$\sigma_{\overline{x}} = \sigma / \sqrt{(}n)$$

- No matter what the distribution of the random variable x is, the distribution of the sample means is approximately NORMAL if the sample size is large.

In this exercise, we will use R to take a look at these 3 facts and see, empirically, how the Central Limit Theorem works.

First start up an R session and lets clear out any old junk that might be lying about in the workspace. (Warning, this command will delete any work you have done!! Use with care!)

```
> rm(list=ls())   ## Removes (rm) all variables (good for saving space)
```

Next, lets create a bunch of data. Let's consider a population of lightbulbs. The 'failure time' of the lightbulbs has a strange distribution. Many fail immediately, but those that dont fail immediately usually last a rather long time. Can you figure out what the probability distribution of this random variable (X = time to failure) should look like?

Lets get specific and assume X is *exponentially distributed.* This is a standard model for distributions which are highly skewed. To get R to take avery large, $n = 100,000$ sample from an exponential distribution, try:

```
> x = rexp(100000,0.05)
> hist(x,,prob=T)
```

Look at the distribution, you've produced. It's definitely not normal. The population mean and the population standard deviation for this example are given by:

$$\mu = \sigma = 20$$

Check these with R.

```
> mean(x)
[1] 20.00177
> sd(x)
[1] 20.01189
```

Pretty close - The population mean and satandard deviation are very close to the sample mean and sample standard deviation for this very large sample.

Now we want to try sampling the population data with more reasonable sample sizes. Suppose we want to take samples of size 100 from the same population and compute the sample mean. This is easy in R

```
> xsamp = rexp(100,0.05);
> mean(xsamp)
[1] 18.60327
```

The sample mean is a random variable that changes with each sample. Try it.

```
> xsamp = rexp(100,0.05);
> mean(xsamp)
[1] 18.60327
> xsamp = rexp(100,0.05)
> mean(xsamp)
[1] 20.37618
> xsamp = rexp(100,0.05)
> mean(xsamp)
[1] 20.92277
```

The Central Limit Theorem is concerned with the distribution of this sample mean.

Suppose we want to look at the mean value of 500 different samples of size $n = 100$. We can easily create this random variable (lets call it `sampmean`) in R, using a loop. Try this:

```
> sampmean = numeric(0)  # make a place to store the sample means
> for (i in 1:500) sampmean[i] = mean(rexp(100,0.05)) #find mean for 500 samples of 100
```

Now lets investigate the three parts of the Central Limit Theorem. First, what does the distribution of sample means look like?

```
> hist(sampmean,prob=T)
```

This is the main statement of the Cenral Limit Theorem. While the population distribution is far from normal, *the distribution of sample means is approximately NORMAL.*

The mean ($\mu_{\overline{x}}$) and the standard deviation ($\sigma_{\overline{x}}$) of the (approximately normally distributed) sample means are related to the mean and standard deviation of the poulation by:

$$\mu_{\overline{x}} = \mu$$

$$\sigma_{\overline{x}} = \sigma/\sqrt{(n)}$$

In R, we compare

```
> mean(sampmean)   # Compare to population mu = 20
[1] 19.98548
> sd(sampmean)    # Compare to sigma/sqrt(100)
[1] 1.952869
> 20/sqrt(100)
[1] 2
```

Ok, not exactly perfect, but pretty darn close.

TO DO:

1. Redo the above analysis for samples of size 50, 400 and 900. Comment on the following:

   (a) How do the histograms of `sampmean` change as the sample size is increased? Does the standard deviation increase or decrease? Is the sample mean looking 'normal'?

   (b) How do the first two predictions of the central limit theorem compare to the actual data as the sample size is increased? Does $\mu_{\overline{x}}$ approach μ? How about the second part of the Central Limit Theorem?

2. Redo the analyis for a different population distribution. You may want to create data using a different binomial distribution or you may try out the R commands `rexp(10000,.1)` (exponential, long-tails) or `rpois(1000,4)` (Poisson Distribution, non-normal) or you may try something else. Whatever you chose as the population, examine what happens to various sized sample means. Check each part of the Central Limit Theorem.

QUESTIONS

1. Consider a population of lightbulbs with mean failure time $\mu = 50$ days and standard deviation of failure times $\sigma = 20 days$. If you take a sample of 100 lightbulbs, what are the chances the mean failure time of your sample will be

   (a) Greater than 53 days?

   (b) Less than 48 days?

   (c) Between 46 and 54 days?