Investigating The Central Limit Theorem

Key to understanding Inferential Statistics is the most popular statistical LAW known as the Central Limit Theorem. In a nutshell, this powerful theorem states three facts about the statistics of sample means.

- The mean of the population and the mean of the samples are EQUAL.
  \[ \mu_x = \mu \]

- The standard deviation of the population (\(\sigma\)) and the standard deviation of the sample means (\(\sigma_x\)) are related by the formula:
  \[ \sigma_x = \frac{\sigma}{\sqrt{n}} \]

- No matter what the distribution of the random variable \(x\) is, the distribution of the sample means is NORMAL.

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!)

\[
\text{> rm(list=ls())} \quad \#\# \text{Removes (rm) all variables (good for saving space)}
\]

Next, lets create a bunch of data. Let’s consider a game like Yahtzee! where you roll 6 dice. We want to keep track of the number of ONES that appear in each roll. Can you figure out what the probability distribution of this random variable (\(X = \) number of ONES in six rolls) is? Is it binomial? If so, what are \(p\), \(q\) and \(n\)?

It is a binomial distribution. To get R to roll 6 dice 10,000 times and count the number of ONES in each roll and store it in variable \(x\), try

\[
\text{> x = rbinom(10000,6,1/6)} \\
\text{> hist(x,breaks=c(-0.5:6.5),prob=T)}
\]

Look at the distribution, you’ve produced. It’s definitely not normal. What should the mean and standard deviation be? Check them using R. Are they close to the theoretical values?

Now we want to try sampling the population data. Suppose we want to take samples of size 40. We can do this in R with the \texttt{sample} command. Lets store one sample of size 40 in \texttt{xsamp} and look at the sample distribution.
> xsamp = sample(x, 40)
> hist(xsamp, breaks=c(-0.5:6.5), prob=T)

How does this sample compare with the population? What is the sample mean?

We can take a number of different samples reusing the `sample` command, looking at the mean each time.

> xsamp = sample(x, 40)
> mean(xsamp)
[1] 0.975
> xsamp = sample(x, 40)
> mean(xsamp)
[1] 1.025
> xsamp = sample(x, 40)
> mean(xsamp)
[1] 1.225

Notice the sample mean is a random variable. It is different for each sample. The Central Limit Theorem is concerned with the statistics of this sample mean.

Suppose we want to look at the mean value of 100 different samples of 40. We can easily create this random variable (let's call it `sampmean`) in R, using a loop. Try this:

> sampmean = numeric(0) # make a place to store the sample means
> for (i in 1:100) sampmean[i] = mean(sample(x, 40)) # find mean of 100 samples of 40

Now let's investigate the three parts of the Central Limit Theorem. First, the mean of x should equal the mean of the sample means. Check it.

> mean(x)
[1] 1.0126
> mean(sampmean)
[1] 1.0175

Not perfect but pretty darn close.

Second, the standard deviation of the sample means and the population are related by $\sigma_{\bar{x}} = \sigma / \sqrt{n}$. In R, we compare

> sd(sampmean) # $\sigma_{\bar{x}}$
[1] 0.1375462

to the population standard deviation divided by $\sqrt{n}$

> sd(x)/sqrt(40)
[1] 0.1445272

Ok, not perfect, but still pretty close.

The last thing the Central Limit Theorem says is that, no matter what the distribution of x, the sample means should be Normally Distributed. Is this the case here? Look at

> hist(sampmean)
Is it 'normal looking'?

**TO DO:**

1. Redo the above analysis for samples of size 50, 100 and 500. 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 $\mu_x$ approach $\mu$? How about the second part of the Central Limit Theorem?

2. Redo the analysis for a different population distribution with sample sizes 50, 100 and 500. 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). Whatever you chose as the population, examine what happens to various sized sample means. Check each part of the Central Limit Theorem.