Using R to simulate the Central Limit Theorem

We know that, in the absence of systematic errors, the mean of a sample of measurements, $\bar{x}$, provides us with an estimate of the true value, μ of the quantity we are trying to measure. But, because of random errors, it is most unlikely that the mean of the sample will be exactly equal to the true value. Hence, it is useful for us to give a range of values which is likely to include the true value. The width of this range depends on two factors, i.e.

(1) the precision of the individual measurements, which in turn depends on the standard deviation of the population, and
(2) the number of measurements in the sample.

We know that we have more confidence in the mean of several values rather than in a single value. That means we would expect that the more measurements we make, the more reliable our estimate of μ, the true value, will be.

Hypothetically if we were to make 50 repeated measurements with a test method, we would get a range of values surrounding the true value μ with a mean value $\bar{x}$ and a certain standard deviation. It is however, more likely that we would analyze 10 samples from that population with 5 repeated results in practice. We would see at once that the means of these 10 samples are more closely clustered than the original measurements.

If we were to take still more samples of 5 measurements and calculate their means, those means would have a frequency distribution of their own. The distribution of all possible sample means (in this case, we call it an infinite number) is called the sampling distribution of the mean. Its mean is the same as the mean of the original population. Its standard deviation is called the standard error of the mean (SEM). There is an exact mathematical relationship between SEM and the standard deviation, σ, of the distribution of the individual measurements.

For a sample of n measurements, we have

$$\text{Standard error of the mean} = \sigma/\sqrt{n}$$

As expected, the larger n is, the smaller the value of the SEM and consequently smaller the spread of the sample means about μ.

Do not have the wrong impression that the term "standard error of the mean" is the difference between the μ and $\overline{x}$. This is not so. The SEM actually gives

a measure of the variability of $\overline{x}$. It tends to the normal distribution as n increases. This concept is termed as the central limit theorem.

The central limit theorem states that the sampling distribution of the sample mean approximates the normal distribution, regardless of the distribution of the population from which the samples are drawn, provided the sample size is sufficiently large. This fact enables us to make statistical inferences based on the properties of the normal distribution, even if the sample is drawn from a population that is not normally distributed.

In other words, the central limit theorem can be stated as follows with regard to the sample mean:

Let $X_1, ..., Xn$ be a random sample from some population with mean μ and variance σ^2. Then, for large n, we have

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

even if the underlying distribution of individual observations in the population is not normal.

The ~ symbol represents "approximately distributed", and the formula can be read as "the mean of X is approximately normally distributed with mean μ and variance $\sigma^2/n.$)

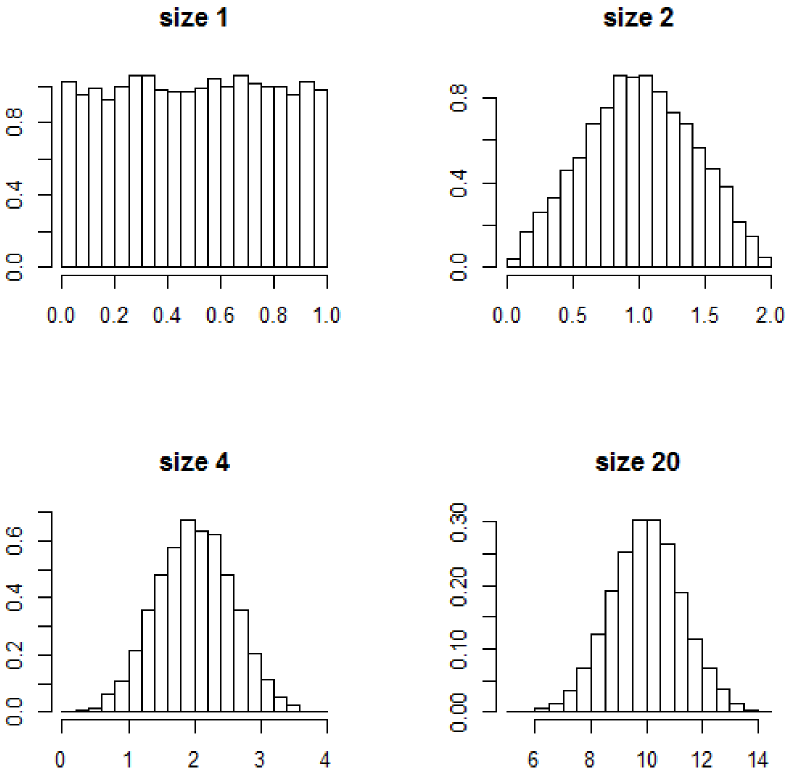
With the understanding of this central limit theorem, we can use a sample to define a range which we may reasonably assume to include the true value (in the absence of systematic errors, of course). Such a range is known as a confidence interval and the extreme values of the interval are called the confidence limits with a given degree of confidence. For example, at 95% confidence, we get 95% of the sample means lying in the range given by:

$$\mu - 1.96(\sigma / \sqrt{n}) < \overline{x} < \mu + 1.96(\sigma / \sqrt{n})$$

The application of the central limit theorem in practice can be seen through computer simulations (such as the R language) that repeatedly draw samples of specified size from a non-normal population, as illustrated below.

```
#=============================================
# Simulation of central limit theorem
#=============================================

layout(matrix(c(1,2,3,4),2,2,byrow=TRUE))

#--------------------------------------------------------------------------
# One uniform random variable simulated 10000 times
#--------------------------------------------------------------------------
size=1                          # No. of random variables in sum.
repeats=10000                   # No. of values to simulate for
                                  # histogram.
v=runif(size*repeats)           # Vector of uniform random
                                  # variables.
w=matrix(v,size, repeats) # Enter v into a matrix
                                  # sizeXrepeats).
y=colSums(w)                    # Sum the columns.
hist(y,freq=FALSE,ann=FALSE)      # Histogram.
title("size 1")

#-----------------------------------------------------------------------------
#Sum of 2 uniform random variables simulated 10000 times
#-----------------------------------------------------------------------------
size=2                       # No. of random variables in sum.
repeats=10000                # No. of values to simulate for
                               # histogram.
v=runif(size*repeats)        # Vector of uniform random
                             # variables.
w=matrix(v,size, repeats) # Enter v into a matrix
                                  # sizeXrepeats).
y=colSums(w)                 # Sum the columns.
hist(y,freq=FALSE,ann=FALSE)      # Histogram.
title("size 2")

#-----------------------------------------------------------------------------
#Sum of 4 uniform random variables simulated 10000 times
#-----------------------------------------------------------------------------
size=4                       # No. of random variables in sum.
repeats=10000                # No. of values to simulate for
                               # histogram.
v=runif(size*repeats)        # Vector of uniform random
                               # variables.
w=matrix(v,size, repeats) # Enter v into a matrix
```

```
                              # sizeXrepeats).
y=colSums(w)                # Sum the columns.
hist(y,freq=FALSE,ann=FALSE)      # Histogram.
title("size 4")

#---------------------------------------------------------------------------
#Sum of 20 uniform random variables simulated 10000 times
#--------------------------------------------------------------------------
size=20                    # No. of random variables in sum.
repeats=10000              # No. of values to simulate for
                             # histogram.
v=runif(size*repeats)      # Vector of uniform random
                             # variables.
w=matrix(v,size, repeats) # Enter v into a matrix
                              # sizeXrepeats).
y=colSums(w)                # Sum the columns.
hist(y,freq=FALSE,ann=FALSE)      # Histogram.
title("size 20")
```


size 1
size 2
size 4
size 20