Introduction to R (7): Central Limit Theorem

Central Limit Theorem

Central limit theorem says that if $x \sim D(\mu, \sigma)$ where $D$ is a probability density or mass function (regardless of the form of the distribution), when sample size is large enough: 

$$\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Central Limit Theorem also suggests that when the population distribution is Normal, we can assume $\bar{x} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$.

In this handout, and by using simulation methods I have included a study regarding the sampling distributions of the sampling mean of two populations. A right skewed population and a Normally distributed one.

Figure 1 shows a clearly right skewed population with $\mu = 2.029$ and $\sigma = 1.382$. For 1000 times 1 take samples of size 2 from this population and I will obtain 1000 sample averages associated with those random samples. Then I will repeat this process for samples of sizes 3, 6, 10, 20, 100 and each time I keep track of the mean and the distribution of those 1000 sample averages. I also plot histograms and qq-plots for each scenario. It turns out that as sample sizes increases the distribution of 1000 sample averages converge to normality (figures 2 and 3). Also, the standard deviations of those sampling distributions get closer to $\frac{\sigma}{\sqrt{n}}$ (table 1).

Next, I sample from a normal population with $\mu = 99.602$ and $\sigma = 10.2211$ (figure 4). Then I repeat the same procedure for sample averages. It turns out that regardless of sample sizes, the result associated with central limit theorem hold (figures 5, 6, and table 2).

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>2</th>
<th>3</th>
<th>6</th>
<th>10</th>
<th>20</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>2.0945</td>
<td>2.061667</td>
<td>2.016</td>
<td>2.0301</td>
<td>2.0186</td>
<td>2.02654</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>0.9860487</td>
<td>0.78204</td>
<td>0.5660369</td>
<td>0.453773</td>
<td>0.3010642</td>
<td>0.1300328</td>
</tr>
</tbody>
</table>

Table 1. Results for population 1. The mean and standard deviations of the sample mean with different sample sizes

<table>
<thead>
<tr>
<th>Sample Size</th>
<th>2</th>
<th>3</th>
<th>6</th>
<th>10</th>
<th>20</th>
<th>100</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard Deviation</td>
<td>7.199265</td>
<td>5.83941</td>
<td>4.132356</td>
<td>3.290563</td>
<td>2.240824</td>
<td>0.9427503</td>
</tr>
</tbody>
</table>

Table 2. Results for population 2. The mean and standard deviations of the sample mean with different sample sizes
Figure 1. Case one: Population distribution. The distribution is Skewed to the right.
Figure 2. Sampling Distributions of Sample means for sample sizes 2, 3, 6, 10, 20, 100.
Figure 3. QQ-plots for the sample mean distributions for different sample sizes.
Figure 4. Case two: Population distribution. Normal distribution.
Figure 5. Sampling Distributions of Sample means for sample sizes 2, 3, 6, 10, 20, 100.
Figure 6. QQ-plots for the sample mean distributions for different sample sizes.
R-codes For Simulation

(a) Population 1: Right Skewed Distribution
We can simulate from a Poisson distribution:

```r
> test1<-rpois(1000,2)
> hist(test1)
> mean(test1)
[1] 2.029
> sd(test1)
[1] 1.382777
```

(b) Population 1: Obtaining 1000 Samples With Size 2, 3, 6, 10, 20, 100
Here is the case for Size 2. Others are similar.

```r
> test<-matrix(nrow=1000,ncol=2)
> for(i in 1:1000)
>    test[i,]<-sample(test1,2)

> mean.size2<-apply(test,1,mean)
> mean(mean.size2)
[1] 2.0945
> sd(mean.size2)
[1] 0.9860487
```

(c) Population 2: Obtaining 1000 Samples With Size 2, 3, 6, 10, 20, 100
Again, only the case for size 2 is included.

```r
test<-matrix(nrow=1000,ncol=2)
for(i in 1:1000)
{
    test[i,]<-sample(test2,2)
}
mean.size2.norm<-apply(test,1,mean)

> mean(mean.size2.norm)
[1] 99.95793
```
> sd(mean.size2.norm)
[1] 7.199265

(d) Population 1: Plotting Histograms and QQ-plots

```r
par(mfrow=c(3,2))
hist(mean.size2)
hist(mean.size3)
hist(mean.size6)
hist(mean.size10)
hist(mean.size20)
hist(mean.size100)

par(mfrow=c(3,2))
qqnorm(mean.size2)
qqnorm(mean.size3)
qqnorm(mean.size6)
qqnorm(mean.size10)
qqnorm(mean.size20)
qqnorm(mean.size100)
```

(e) Population 2: Plotting Histograms and QQ-plots

```r
par(mfrow=c(3,2))
hist(mean.size2.norm)
hist(mean.size3.norm)
hist(mean.size6.norm)
hist(mean.size10.norm)
hist(mean.size20.norm)
hist(mean.size100.norm)

par(mfrow=c(3,2))
qqnorm(mean.size2.norm)
qqnorm(mean.size3.norm)
qqnorm(mean.size6.norm)
qqnorm(mean.size10.norm)
qqnorm(mean.size20.norm)
qqnorm(mean.size100.norm)
```