Statistical Inference Course Project

Part 1: simulation

audrey

25 October, 2015

Overview:

This report investigates how the exponential distribution relates to the Central Limit Theorem (CLT). The CLT states that if we repeatedly take samples from a population, and calculate the averages of each one, the collection of those averages will be normally distributed regardless of the underlying distribution. The CLT applies to more than the *mean* statistic though in the current report we deal with the means.

Distribution properties

Here, we investigate the distribution of averages of 40 exponentials. We take a sample of size 40 from the exponential distribution and compute its average. We then repeat this process 1000 times. We first focus on the sample mean of the distribution and compare it to the theoretical value of $1/\lambda$ where λ is the rate parameter. Note that we fix $\lambda = 0.2$ throughout this report. We also set the seed for reproducible results.

```
set.seed(233)
lambda<-0.2
sims<-1000
n<-40
rexps <- replicate(sims, rexp(n, lambda))
colmean <- apply(rexps, 2, mean)
```

The theoretical mean is calculated as $1/\lambda = 5$ while the sample mean (assigned the variable name samplemean in the R code below) obtained from a thousand simulations is found to be:

```
samplemean<- mean(colmean)
samplemean
```

```
## [1] 4.965749
```

The standard deviation of the exponential distribution is also $1/\lambda$; the variance is the square of the standard deviation hence the theoretical variance is calculated as $1/\lambda = 25$. From the sample distribution, the sample variance is:

```
samplevar<-var(colmean)
samplevar
```

```
## [1] 0.6178026
```

We plot the means in a frequency distribution:

```
library(ggplot2)
g <- ggplot(data.frame(colmean), aes(x = colmean))+geom_histogram(fill = "orange", color = "red")
print(g)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

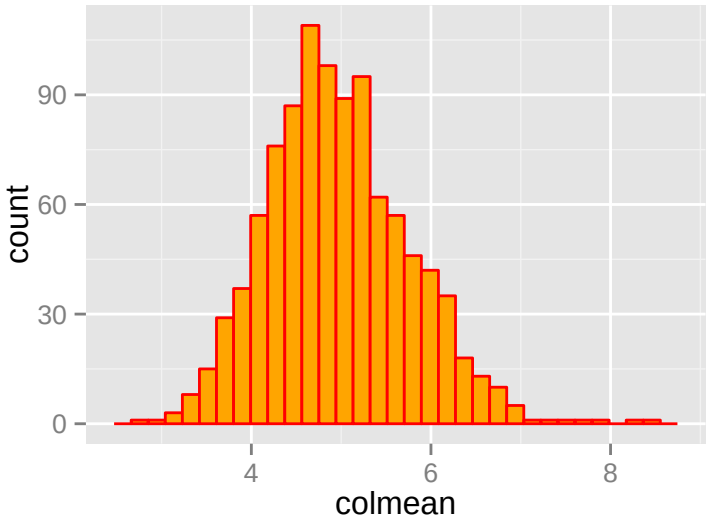

Figure 1: Simulated means distribution

Departure from normality

Here, we aim to show that the distribution is approximately normal by overlapping a density plot generated using simulated data. The blue vertical line indicates the simulated mean while the green, vertical line indicates the theoretical mean. The two means are very close; the result is expected given large sample sizes and the CLT statement.

```
 g <- ggplot(data.frame(colmean), aes(x = colmean))+
  geom_histogram(aes(y=..density..), color="red", fill = "orange")+
geom_density(colour="yellow", size=2)+
  geom_vline(xintercept = mean(colmean), color = "blue")+
  geom_vline(xintercept = 5, color = "green")
print(g)
```

```
## stat_bin: binwidth defaulted to range/30. Use 'binwidth = x' to adjust this.
```

A quantile-quantile (Q-Q) plot shows departure from normality; the closer the data lies on to the diagonal the more normal the data is. We use here our distribution of averages as the argument for qqnorm function:

```
qqnorm(colmean)
qqline(colmean)
```

We observe that the deviations from the diagonal are minimal and thus indicating our means to be normally distributed as expected from the CLT.

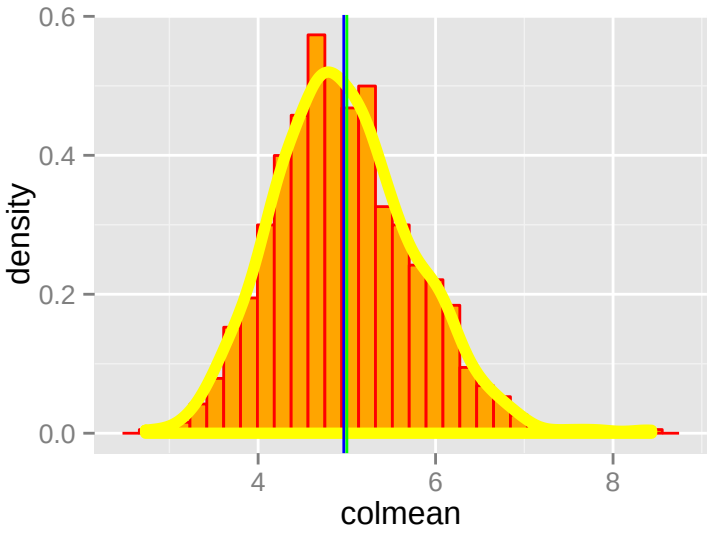

Figure 2: Simulated means distribution with density plot

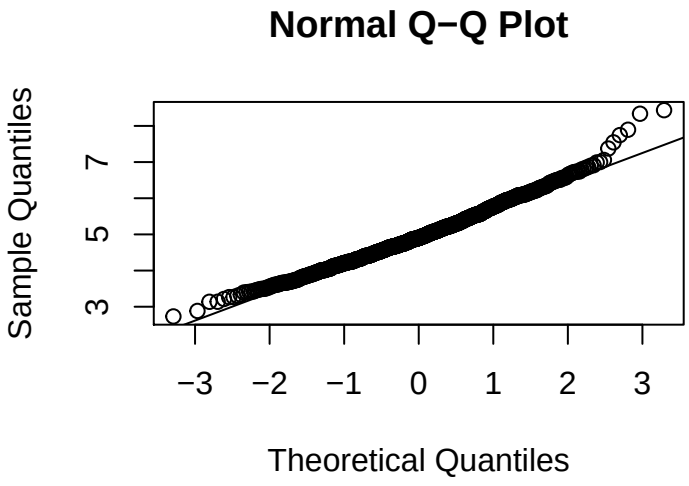

Figure 3: Normal Q-Q plot for simulated data