Simulation and Inferential Data Analysis

Gordon CHAN

19/08/2015

Overview

This report will demonstrate the Central Limit Theorem (CLT) through a simulation exercise and a basic inferential data analysis.

Simulations

In this report, the simulation is done in 2 parts:

First, 1000 random exponentials are generated, with *lambda* equals to 0.2.

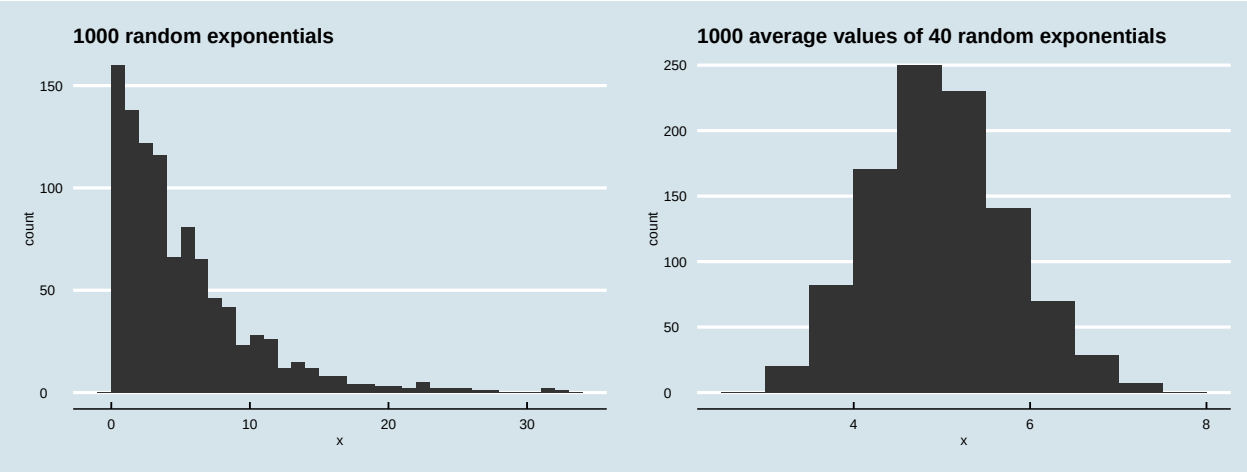
Second, 1000 means of 40 random exponentials are generated, also with *lambda* equals to 0.2.

The generated distributions are plotted. *(Fig.1)*.

```
library("ggplot2")
    library("gridExtra")
    library("ggthemes")

set.seed(1)

    n <- 40
    nosim <- 1000
    lambda <- 0.2



# Generating 1000 random exponentials

ed <- rexp(nosim, lambda)

# Generating 1000 means of 40 random exponentials

f_mnd<- function(nosim, n, lambda, type = "mean"){
    set.seed(1)
    fmnd <- NULL
        if(type=="mean"){
            for (i in 1:nosim){fmnd <- c(fmnd, mean(rexp(n, lambda)))}
        } else if(type=="var"){
            for (i in 1:nosim){fmnd <- c(fmnd, var(rexp(n, lambda)))}
        }
    fmnd
}
mnd <- f_mnd(nosim, n, lambda, "mean")

# Generating 1000 cumulative means of 40 random exponentials
```

```r
means <- cumsum(mnd) / (1   : nosim)
    means_df <- data.frame(x = 1 : nosim, y = means)

# Generating 1000 cumulative variances of 40 random exponentials

vars <- NULL

for (i in 1:nosim){
    vars <- c(vars, var(mnd[1:i]))
}
    vars_df <- data.frame(x = 1 : nosim, y = vars)


# Plot the generated distribution

p <- ggplot()+
    geom_histogram(aes(x=ed), binwidth=1)+
    labs(
        title ="1000 random exponentials",
        x = "x")+
    theme_economist(base_size = 8)+
    scale_fill_economist()

s <- ggplot()+
    geom_histogram(aes(x=mnd), binwidth=0.5)+
    labs(
        title =
        "1000 average values of 40 random exponentials",
        x = "x")+
    theme_economist(base_size = 8)+
    scale_fill_economist()

grid.arrange(p, s, ncol = 2,
             top="Fig.1 - Distribution of random exponentials (lambda = 0.2)")
```

Fig.1 − Distribution of random exponentials (lambda = 0.2)



Sample Mean versus Theoretical Mean

The theoretical mean and the sample mean of the 1000 random exponentials, as well as the 1000 sampling means of 40 exponents are calculated.

```
    t_mn <- 1/lambda
        t_mn
```

```
## [1] 5
```

```
    s_mn <- round(mean(ed), 2)
        s_mn
```

```
## [1] 5.16
```

```
    mn_mn <- round(mean(mnd), 2)
        mn_mn
```

```
## [1] 4.99
```

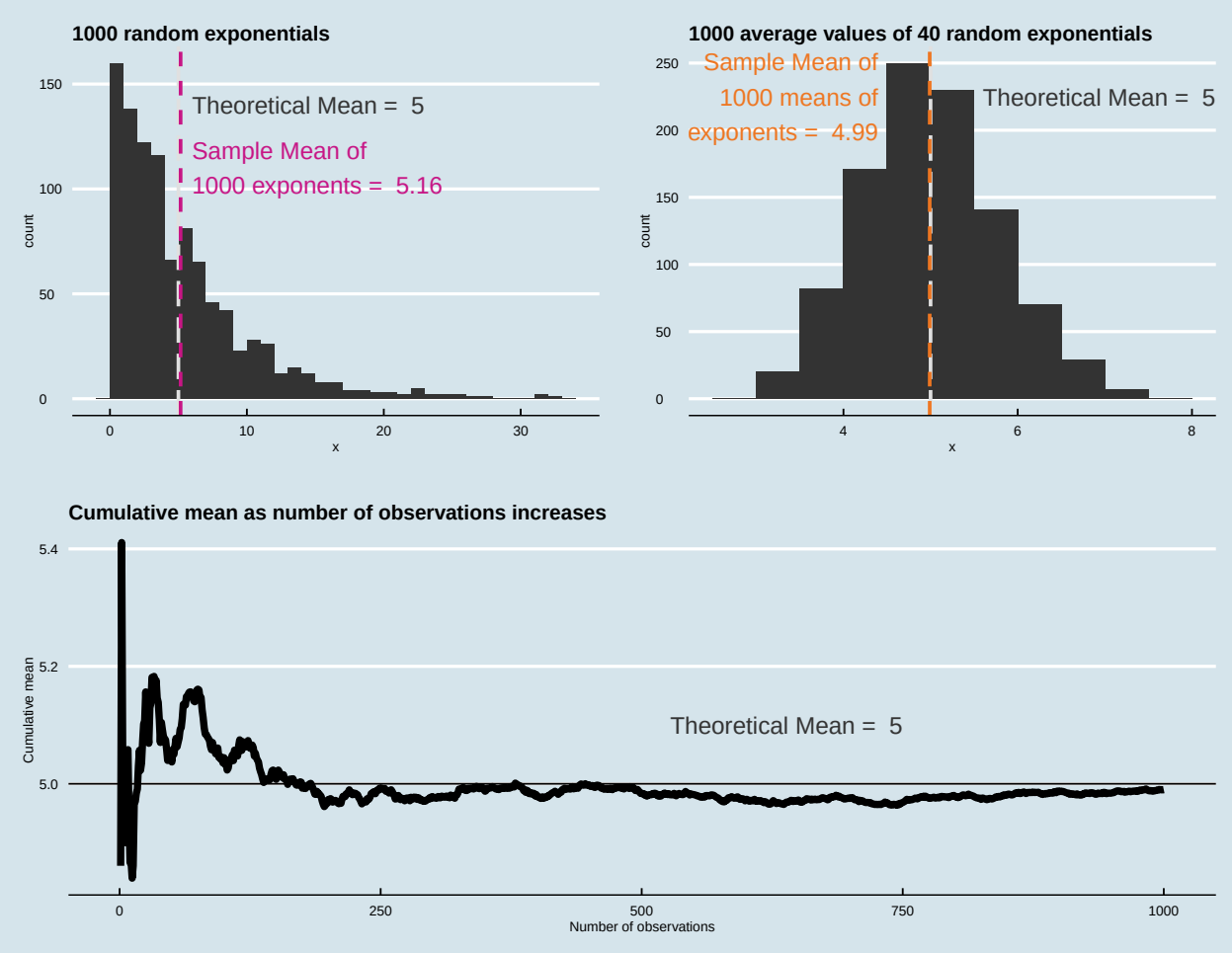
We can see that the theoretical mean being $1/\lambda$ is while the sample mean for 1000 exponents is 5.16, and 1000 means of 40 exonents is .

The theoretical mean and the sample means are plotted to the distribution graphs *(Fig.2)*. From this figure, we can see how the means compared to each other. While naturally we would expect the mean of 1000 random exponents would tend to be the theoretical mean of the exponential distribution, and as the number approaches infinity the sample mean would equals the theoretical mean. The Central Limit Theorem (CLT) predicts that the mean of infinite number of means of population of size n will also equals the theoretical mean, as shown from the bottom panel of *Fig.2*.

```
p <- ggplot()+
    geom_histogram(aes(x=ed), binwidth=1)+
            geom_vline(xintercept=t_mn, linetype=5, size=1, color="gray88")+
                geom_vline(xintercept=s_mn, linetype=2, size=1, color="mediumvioletred")+
    labs(title ="1000 random exponentials", x = "x")+
            annotate("text", x = 6, y = 140,
                     label = paste("Theoretical Mean = ", t_mn),
                     color="gray20", size = 5, hjust = 0)+
                annotate("text", x = 6, y = 110,
                         label = paste("Sample Mean of\n1000 exponents = ", s_mn),
                         color="mediumvioletred", size = 5, hjust = 0)+
    theme_economist(base_size = 8)+
    scale_fill_economist()

s <- ggplot()+
    geom_histogram(aes(x=mnd), binwidth=0.5)+
            geom_vline(xintercept=t_mn, linetype=5, size=1, color="gray88")+
                geom_vline(xintercept=mn_mn, linetype=2, size=1, color="chocolate2")+
    labs(title ="1000 average values of 40 random exponentials", x = "x")+
            annotate("text", x = 5.6, y = 225,
                     label = paste("Theoretical Mean = ", t_mn),
```

```
                    color="gray20", size = 5, hjust = 0)+
                annotate("text", x = 4.4, y = 225,
                         label = paste("Sample Mean of\n1000 means of\n40 exponents = ", mn_mn),
                         color="chocolate2", size = 5, hjust = 1)+
    theme_economist(base_size = 8)+
    scale_fill_economist()

c <- ggplot(means_df, aes(x = x, y = y))+
        geom_line(size=2)+
        geom_hline(yintercept = t_mn)+
            labs(title = "Cumulative mean as number of observations increases",
                 x = "Number of observations", y = "Cumulative mean")+
                annotate("text", x = 750, y = 5.1,
                     label = paste("Theoretical Mean = ", t_mn),
                     color="gray20", size = 5, hjust = 1)+
                theme_economist(base_size = 8)

        layout <- matrix(c(1, 2, 3, 3), nrow = 2, byrow = TRUE)

grid.arrange(p, s, c, layout_matrix = layout,
             top = "Fig.2 - Sample mean v Theoretical mean")
```

Fig.2 − Sample mean v Theoretical mean



Sample Variance versus Theoretical Variance

The theoretical variance and the sampling variance of 1000 means of 40 random exponents are calcualted.

```
t_var <- (1/lambda)^2 / n
    t_var
```

```
## [1] 0.625
```

```
mn_var <- var(mnd)
    mn_var
```

```
## [1] 0.6111165
```

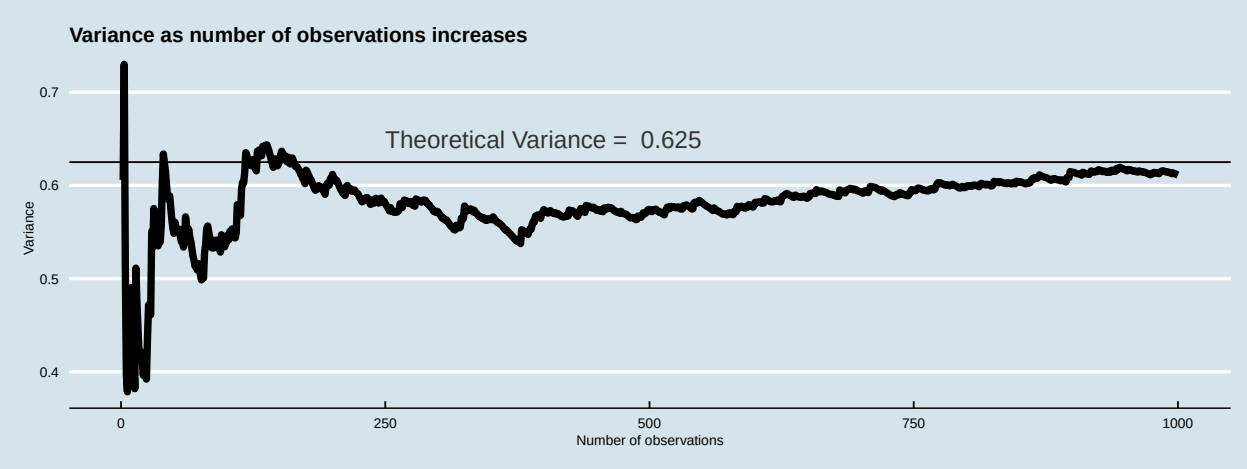
We can see that the theoretical variance being σ^2 / n is while the sampling variance for 1000 means of 40 exponents is .

The theoretical variance and the sampling variance as the number of observation increases is plotted. As predicted from the CLT, as the number of sampling increases, the variance of the sampling will converge to the theoretical variance of the population.

```
cv <- ggplot(vars_df, aes(x = x, y = y))+
        geom_line(size=2)+
        geom_hline(yintercept = t_var)+
            labs(title = "Variance as number of observations increases",
                 x = "Number of observations", y = "Variance")+
                annotate("text", x = 250, y = 0.65,
                     label = paste("Theoretical Variance = ", t_var),
                     color="gray20", size = 5, hjust = 0)+
                theme_economist(base_size = 8)
grid.arrange(cv, ncol = 1,
             top = "Fig.3 - Cumulative variance as number of observations increases")
```

```
## Warning: Removed 1 rows containing missing values (geom_path).
```

Fig.3 − Cumulative variance as number of observations increases



Distribution

As the number of samples increases, the CLT predicts that the sampling distribution would converge to a normal distribution. To show this, sampling distributions with sample size n = 5, 20 and 40 are plotted, with a corresponding normal distribution overlayed on top. *(Fig.4)* We could easily observe this trend from the graph. Note that as teh original population of exponential distribution skewed heavyly, the sampling distribution showed slight skewless also. Although this would improve over a larger sample size.

```
sd <- 1/lambda
cfunc <- function(x, n) sqrt(n) * (mean(x) - 5) / 5

dat <- data.frame(
  x = c(apply(matrix(f_mnd(nosim, 5, lambda), nosim), 1, cfunc, 5),
        apply(matrix(f_mnd(nosim, 20, lambda), nosim), 1, cfunc, 20),
        apply(matrix(f_mnd(nosim, 40, lambda), nosim), 1, cfunc, 40)
        ),
  size = factor(rep(c(5, 20, 40), rep(nosim, 3))))
g <- ggplot(dat, aes(x = x, fill = size))+
    geom_histogram(alpha = .20, binwidth=.5, colour = "black", aes(y = ..density..))+
        theme_economist(base_size = 8)+
            scale_fill_economist()
g <- g + stat_function(fun = dnorm, size = 1)
g <- g + facet_grid(. ~ size) + labs(title = "Sampling distributions")
grid.arrange(g, ncol = 1,
             top = "Fig.4 - Sampling distributions as the number of samples increases")
```

Fig.4 − Sampling distributions as the number of samples increases