Illustrating the Central Limit Theorem

Student

August 13, 2015

Introduction

In this assignment we'll use **R** to illustrate the Central Limit Theorem. We consider the sampling distribution of the mean of a random variable Y which follows an exponential distribution with $\lambda = 0.2$. Such a random variable has (by definition) identical mean and standard deviation $\mu_Y = \sigma_Y = \lambda^{-1} = 5$.

We consider y_n to be a sample of n observations of Y, which has empirical sample mean

$$\bar{y}_n = \frac{1}{n}\sum_{i=1}^{n} y_i$$

We may consider $\bar{y}_n$ a particular observation of the random variable $\bar{Y}_n$, which is the mean of a size n sample of Y. $\bar{Y}_n$ follows a some distribution, which we call the *sampling distribution of the mean*. The Central Limit Theorem makes the following statements about $\bar{Y}_n$:

1. $E[\bar{Y}_n] = \mu_Y$
2. $SD(\bar{Y}_n) = \frac{1}{\sqrt{n}}\sigma_Y$
3. $\bar{Y}_n$ is approximately normally distributed.

Simulate

We simulate $\bar{Y}_{40}$ by generating 1000 observations of $n = 40$ sized samples of Y, and considering the mean of each one. Here, the collection of 1000 sample mean observations will be denoted `y.means`, and the 1000 40-member samples themselves will be denoted `y`.

```
# Simulate 1000 observations of a size n = 40 random exponential sample.
set.seed(1)
y <- replicate(1000, rexp(40, rate = 0.2))

# Store the sample mean of each (of the 1000) simulations.
y.means <- colMeans(y)
```

`y.means` is now a vector of observed sample means of Y, each of which may or may not be similar to the population mean of Y. We compare parameters of the observed `y.means` to the theoretical claims about $\bar{Y}_{40}$ made by the Central Limit Theorem.

Mean of the observed sample means

The first result of the Central Limit Theorem is that the **sampling distribution of the mean** has mean

$$E[\bar{Y}_n] = \mu_Y$$

Let's check the mean of our observation of $\bar{Y}_{40}$. It should be very close to $\mu_Y = 5$.

```r
# Observed mean.
mean(y.means)
```

```
## [1] 4.990025
```

Variance of the observed sample means

The second result of the Central Limit Theorem is that the **variance of the sampling distribution of the mean** is

$$Var(\bar{Y}_n) = SD^2(\bar{Y}_n) = \frac{\sigma_Y^2}{n}$$

In our case, theoretical variance $\frac{\sigma_Y^2}{n} = \frac{25}{40} = 0.625$. Let's compare to observation once again:

```r
# Observed variance.
var(y.means)
```

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

Normality of the observed sample means

We'll now discuss the final result of the Central Limit Theorem: that the distribution of sample means is approximately normal.

First, we'll quantify `y.mean` (our $\bar{Y}_{40}$ observation)'s approximation to normal by evaluating its ample skew and kurtosis using **R**'s `moments` package:

```r
library(moments)
skewness(y.means)
```

```
## [1] 0.2788758
```

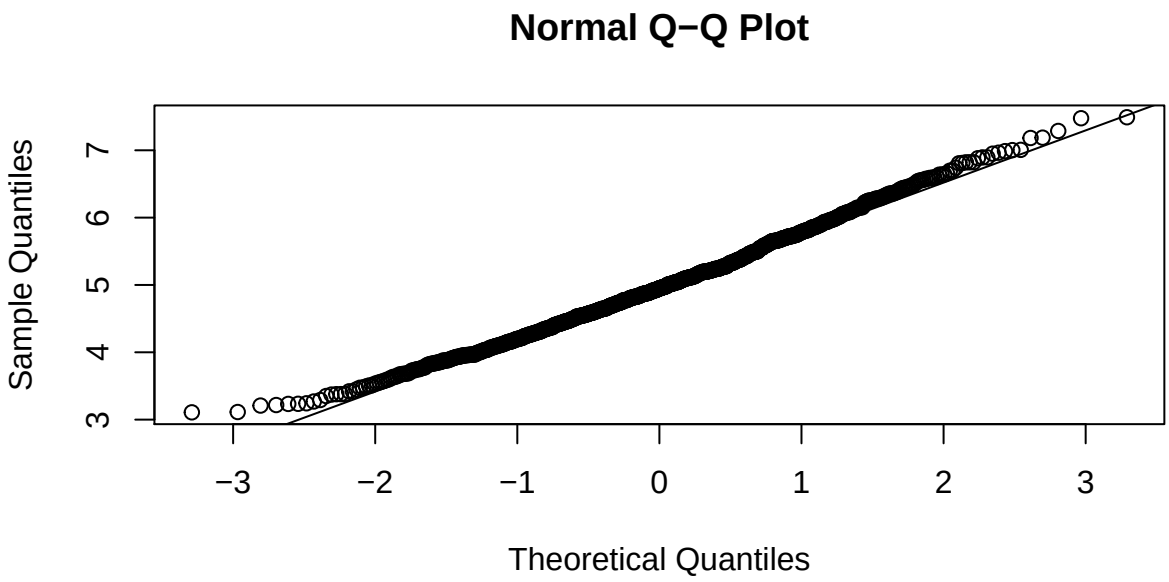
```r
kurtosis(y.means)
```

```
## [1] 2.867473
```

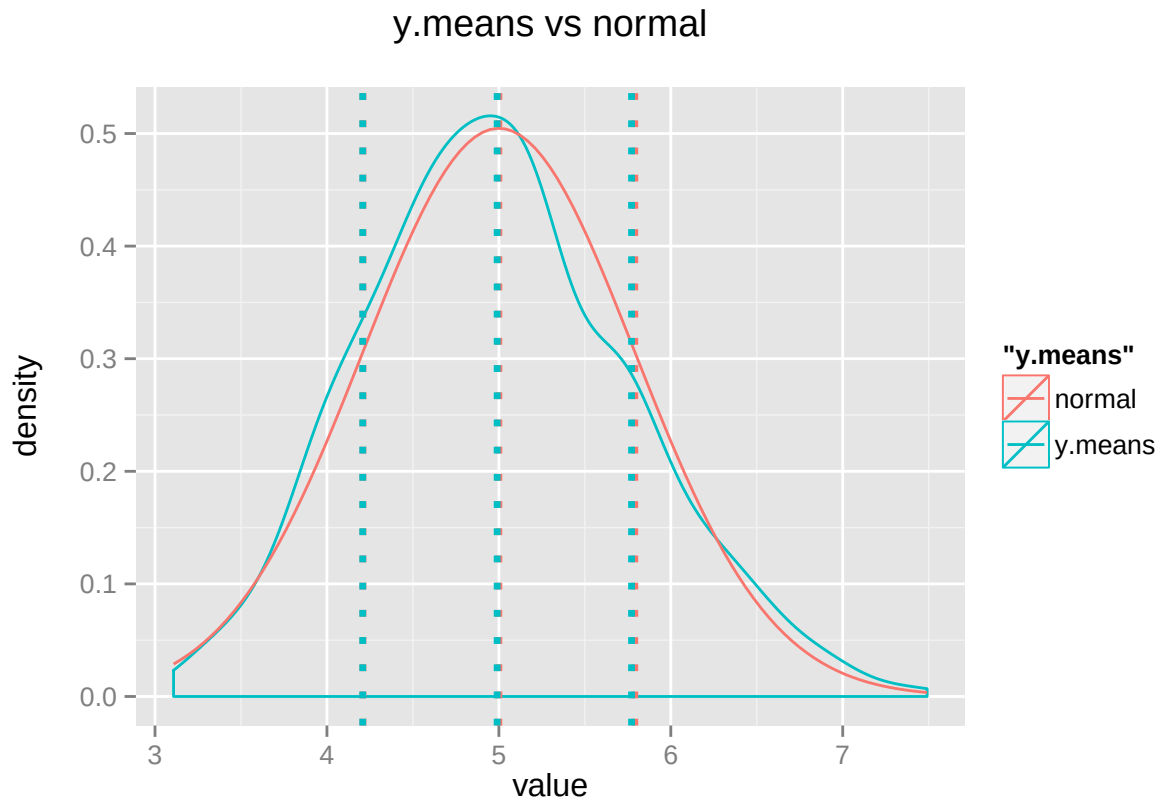
A true normal distribution has skew `0` (being symmetric) and kurtosis `3` by definition. Further discussion of these parameters is beyond the scope of this article. In both cases, `y.means` comes close.

A visual way of comparing a distribution to normal is possible by considering a **QQ Plot**, which plots quantiles against one another. A linear normal QQ plot indicates a distribution that is approximately normal.

```r
# Normal QQ plot.
qqnorm(y.means); qqline(y.means)
```



Finally, let's simply look at the distribution itself by considering a density plot of `y.means`, with the theoretical (and normal) sampling distribution of the mean overlayed[1]. Mean and sd lines for both are drawn.



We observe that our large sample observation of $\bar{Y}_{40}$ is quite normal, as the Central Limit Theorem predicted.

[1] Code for this plot uses **R**'s `ggplot2` library and is included in the appendix.

Appendix

Terminology

It's important to not be confused between the two different types of *samples* discussed here.

- y_{40} is a particular size 40 sample of the exponential random variable Y.
- $\bar{y}_{40}$ is the mean of a particular size 40 sample of Y.
- `y.means` is a size 1000 sample of the random variable $\bar{Y}_{40}$.

Because different size 40 samples of Y have different means, we consider $\bar{y}_{40}$ to be an observation of a random variable $\bar{Y}_{40}$. $\bar{Y}_{40}$ may be called "the sample mean of a size 40 sample of Y"; *its* distribution is the one about which the Central Limit Theorem offers such interesting claims.

Law of Large Numbers

Note that the Law of Large Numbers states that the larger the number of samples, the closer the sample mean will approach the population. As applied here, we have `y.means` being 1000 samples of $\bar{Y}_{40}$, and would therefore expect `mean(y.means)` to be extremely close to $E[\bar{Y}_{40}]$, and indeed, it is.

Each element of `y.means`, however, is the mean of a sample of Y of only size 40. While the sample mean is always an unbiased estimator of the population mean (that is, its expected value is the population mean itself), we might expect much more variance about the population mean of 5 among the elements of `y.means`.

Exploratory Code

Recall that `y` is the collection of 1000 size 40 samples of Y. It's stored as a 40 row x 1000 column matrix. Let's explore it briefly:

```
  # Dimensions in [# rows, # columns]
  dim(y)
```

```
## [1]   40 1000
```

```
   # Quick peek at part of the first four (of 1000) simulations.
  y[1:6, 1:4]
```

```
##            [,1]     [,2]      [,3]      [,4]
## [1,]  3.7759092 5.399406  1.930968 1.1782781
## [2,]  5.9082139 5.141235  5.041282 3.7152181
## [3,]  0.7285336 6.461308  4.092571 8.5496291
## [4,]  0.6989763 6.265527  0.296306 4.9090970
## [5,]  2.1803431 2.773207 11.419267 0.1012521
## [6,] 14.4748427 1.506415  4.020855 0.8928473
```

Each column consists of 40 samples (of which 6 are shown) of the exponential random variable Y.

`y.means` is the sum of each column of `y`; that is, the mean of each size 40 sample of Y. Let's confirm that its dimensions are as expected and demonstrate a few of the values.

```
  length(y.means)
```

```
## [1] 1000
```

```
  head(y.means)
```

```
## [1] 4.860372 5.961285 4.279204 4.702298 5.196446 4.397114
```

Plot Code

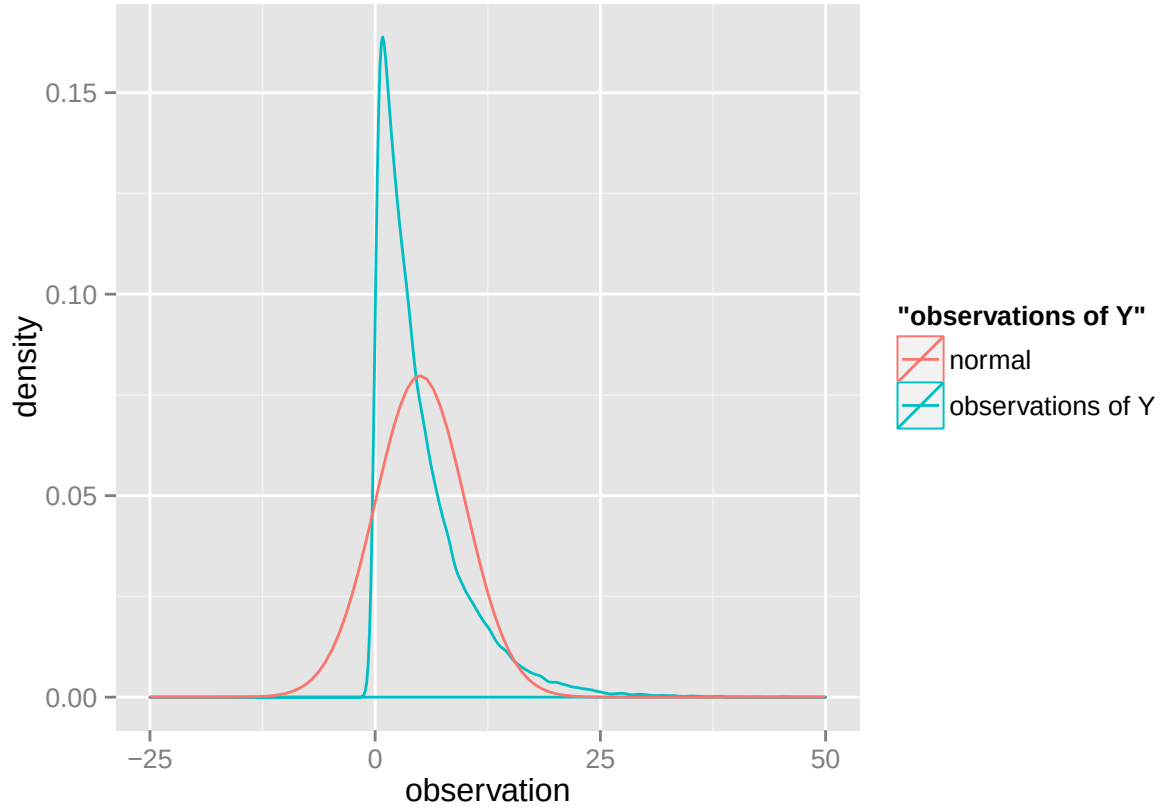
The following code was used to generate our earlier density plot of `y.means`, with normal overlay:

```
  library(ggplot2)
  # Auxiliary lines
  auxlines <- data.frame(
    "distr" = c("normal", "normal", "normal", "y.means", "y.means", "y.means"),
    "value" = c(
      5, 5 + sqrt(25/40), 5 - sqrt(25/40),
      mean(y.means),
      mean(y.means) + sqrt(var(y.means)),
      mean(y.means) - sqrt(var(y.means))
    )
  )
  # Density plot of y.means
  ggplot(data.frame(mean = y.means), aes(x = mean)) +
    geom_density(aes(colour = "y.means")) +
    stat_function(
      fun = dnorm,
      aes(colour = "normal"),
      args = list(mean = 5, sd = sqrt(25/40))
    ) +
    geom_vline(data = auxlines,
      aes(xintercept = value,
        colour = distr
      ),
      linetype = "dotted",
      size = 1.2
    ) +
    labs (x = "value", y = "density\n", title = "y.means vs normal\n")
```

Plotting the source exponential distribution

As we noted earlier in this Appendix, the numbers contained in each size 40 sample of Y are exponentially, rather than normally distributed. Let's take a look at how the distributions compare using our collection of $1000 \cdot 40$ simulated observations of Y.

```r
# Density plot of y.
ggplot(data.frame(observation = as.vector(y)), aes(x = observation)) +
  geom_density(aes(colour = "observations of Y")) +
  xlim(-25,50) +
  stat_function(
    fun = dnorm,
    aes(colour = "normal"),
    args = list(mean = 5, sd = 5) # matching mean and sd to Y.
  )
```



In fact, this is the powerful observation of the Central Limit Theorem. In this article, we used an exponential distribution for Y, but more generally it does not matter *what*[2] kind of underlying distribution we assign to Y. As long as it is a distribution with mean and finite variance, the sampling distribution of its mean will approximate to normal for sufficiently large n!

Closing Remarks

This PDF was written in R Markdown and generated by knitr in RStudio 0.99.465. It was prepared as part of an assignment for the STATINFERENCE-031 class offered on Coursera by the Johns Hopkins Bloomberg School of Public Health. I appreciate your feedback, hope that you are enjoying the Data Science Specialization!

[2]It should be disclaimed that not all distributions meet the conditions of the CLT. For example, the Cauchy distribution has no mean and undefined variance, and its sampling distribution of the mean is also Cauchy, rather than normal!