Central Limit Theorem

James Crosswell

Monday, May 18, 2015

Overview

In this study we investigate the central limit theorem, which states that $\bar{X}_n$ is approximately $N(\mu, \sigma^2/n)$.

To demonstrate the theorem we calculate the mean and variance of a large number of random samples from the exponential distribution and compare these to the theoretical statistics for this distribution.

Simulations

We start by generating 1000 *samples* (each of size 40) from the exponential distribution and recording both their means and their variances:

```
# Set lambda=0.2 for all our simulations
lambda <- 0.2
nosims <- 1000
sample_size <- 40

# Generate 1000 means of samples from the exponential distribution
sample_means = NULL
sample_vars = NULL
for (i in 1 : nosims)
{
    sample <- rexp(sample_size, lambda)
    sample_means = c(sample_means, mean(sample))
    sample_vars = c(sample_vars, var(sample))
}
```

Sample Mean versus Theoretical Mean

In the case or the exponential distribution we know that $\mu = 1/\lambda$ so our sample mean should approximately equal $\mu = 5$.

```
expected_mean <- 1 / lambda
sample_mean <- mean(sample_means)
c(expected_mean, sample_mean)
```

```
## [1] 5.000000 4.990025
```

And indeed it's pretty damn close.

Sample Variance versus Theoretical Variance

Similarly, we know that the standard deviation for the exponential distribution is $\sigma = 1/\lambda = 5$. so the variance of our sample means should be around $\sigma^2 = 25$

```
expected_variance <- (1 / lambda) ^ 2
sample_variance <- mean(sample_vars)
c(expected_variance, sample_variance)
```

```
## [1] 25.00000 25.06459
```

Once again, this isn't far from what we're expecting.

Distribution

Lastly, we want to see if our sample means are roughly normally distributed, which we can check by laying a standard normal density curve over the top of a plot of the distribution of our sample means.

Since the two densities don't have the same scale however, we need to transform the distribution of our sample means by subtracting the population mean (which should center the distribution around zero) and dividing by the standard error (which should scale the distribution to match the standard normal):

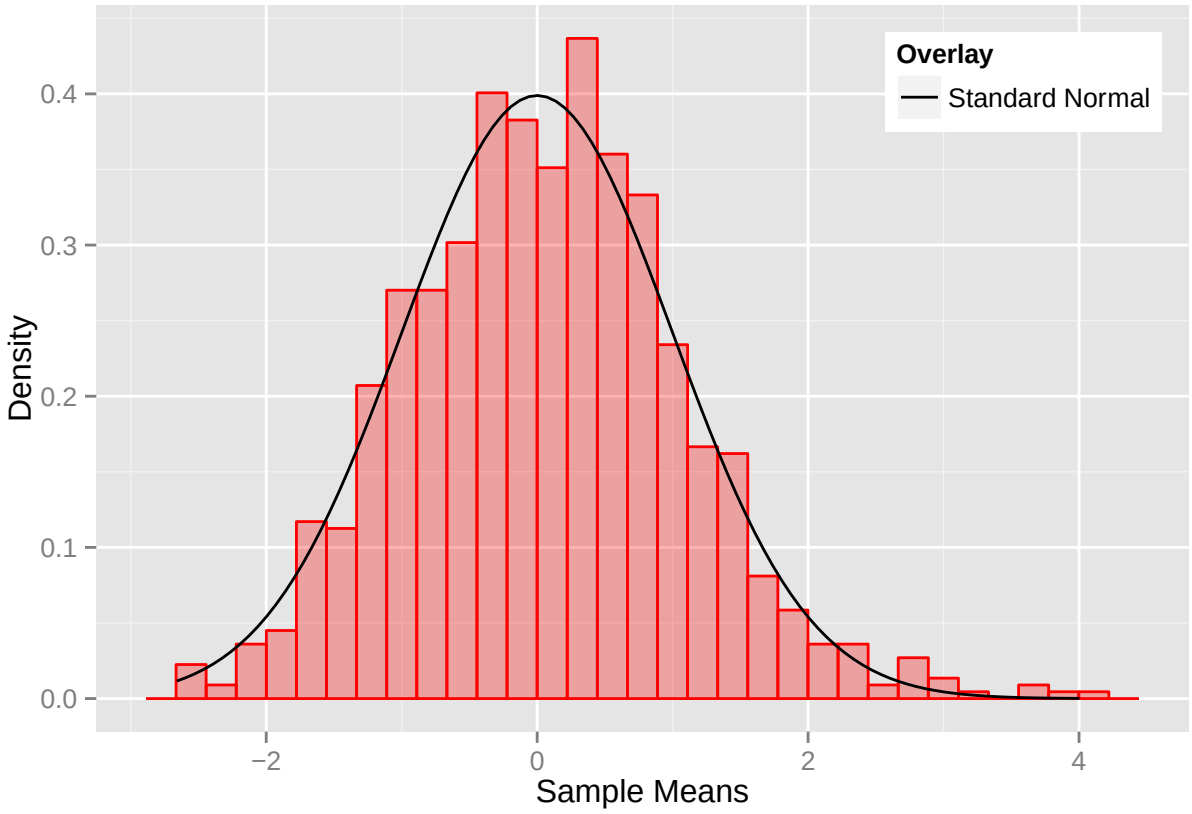
```
expected_sd <- sqrt(expected_variance)

dfunc <- function(x) sqrt(sample_size) * (mean(x) - expected_mean) / expected_sd

sample_distribution = NULL
for (i in 1 : nosims)
{
    sample <- rexp(sample_size, lambda)
    sample_distribution = c(sample_distribution, dfunc(sample))
}
```

We can then plot our transformed sample means alonge with an overlay of the standard normal density:

```
sample_data <- tbl_df(data.frame(data = sample_distribution))
ggplot(sample_data) +
    ylab("Density") +
    xlab("Sample Means") +
    geom_histogram(aes(x=data, y=..density..), color = "red", fill="red", alpha = 0.3) +
    stat_function(fun = dnorm, aes(colour = "black")) +
    scale_colour_identity("Overlay", guide="legend", labels = c("Standard Normal")) +
    theme(legend.position=c(1, 1), legend.justification=c(1,1))
```



And here we can clearly see that our sample means are verly close to normally distributed... which is consistent with the central limit theorem.