Exploring the Central Limit Theorem

adatum

2015-06-21

Overview

The Central Limit Theorem (CLT) states that independent and identically distributed (iid) random samples will themselves be nearly normally distributed, regardless of their underlying distributions, and will have a mean that approaches the population mean, $\bar{x} \simeq \mu$, and a variance of $\frac{\sigma^2}{n}$ where σ^2 is the population variance and n is the sample size. We will demonstrate this by performing simulations using random samples drawn from an underlying population defined by the exponential distribution, which has the properties that $\mu = \frac{1}{\lambda}$ and $\sigma = \frac{1}{\lambda}$.

Simulations

For these simulations we will set $\lambda = 0.2$ and the sample size $n = 40$, and we will perform 1000 simulations. With this information, we can calculate the theoretical mean, μ, and variance, σ^2, of the population.

```
set.seed(42)
lambda <- 0.2
n <- 40
nsims <- 1000
mu <- 1/lambda
sigma2 <- 1/lambda^2
cat(paste("population mean =", mu, "\npopulation variance =", sigma2))
```

```
## population mean = 5
## population variance = 25
```

Now we can construct our two samples. The first is a single sample of 1000 random values from the exponential distribution.

```
x1 <- rexp(nsims, lambda)
```

The second is 1000 averages of 40 random values from the exponential distribution. This is produced by filling a matrix of 1000 rows and 40 columns with values from the exponential distribution, and then averaging over each of the 40 columns to obtain 1000 values which are the *averages* of 40 random exponential variables.

```
x2 <- apply(matrix(rexp(nsims*n, lambda), nrow = 1000, ncol = 40), 1, mean)
```

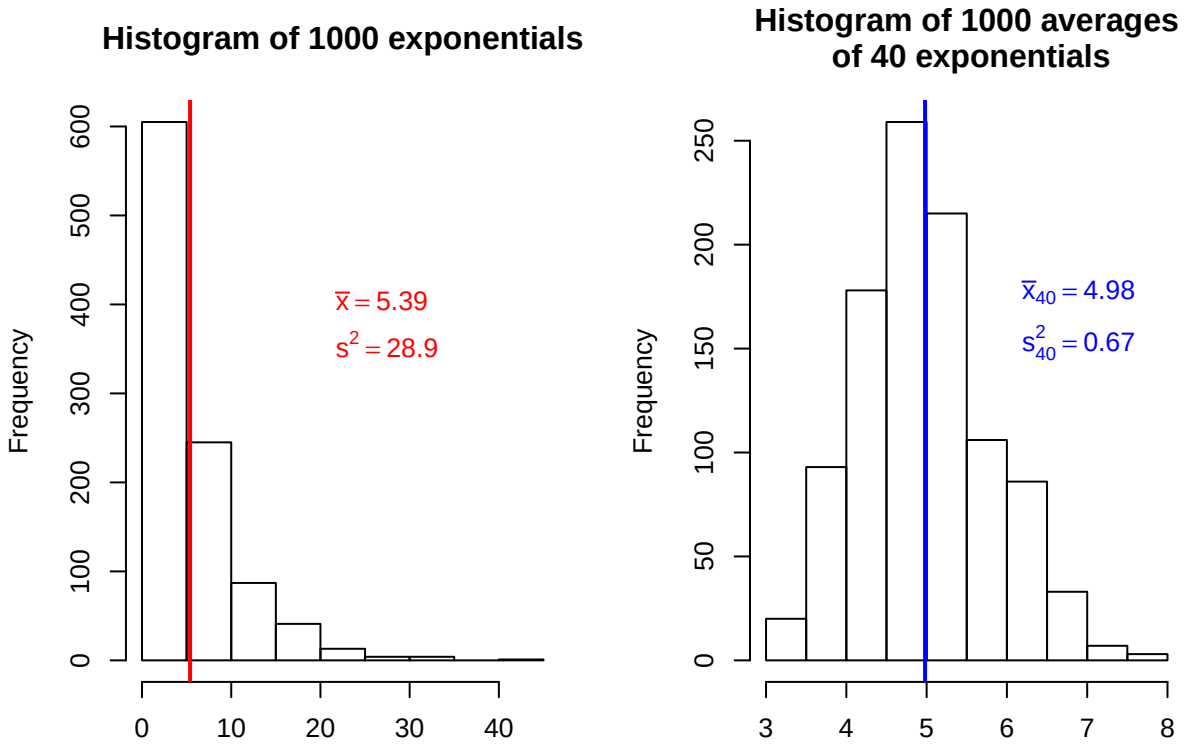
Sample Mean versus Theoretical Mean

We can plot histograms to see what the distributions of the two samples look like:

```
par(mfcol = c(1, 2), cex = 0.8)
hist(x1, main = "Histogram of 1000 exponentials", xlab = "")
abline(v = mean(x1), col = "red", lwd = 2)
```

```
text(20, 400, bquote(bar(x) == .(round(mean(x1), 2))), pos = 4, col = "red")
text(20, 350, bquote(s^2 == .(round(var(x1), 2))), pos = 4, col = "red")

hist(x2, main = "Histogram of 1000 averages\n of 40 exponentials", xlab = "")
abline(v = mean(x2), col = "blue", lwd = 2)
text(6, 175, bquote(bar(x)[40] == .(round(mean(x2), 2))), pos = 4, col = "blue")
text(6, 150, bquote(s[40]^2 == .(round(var(x2), 2))), pos = 4, col = "blue")
```



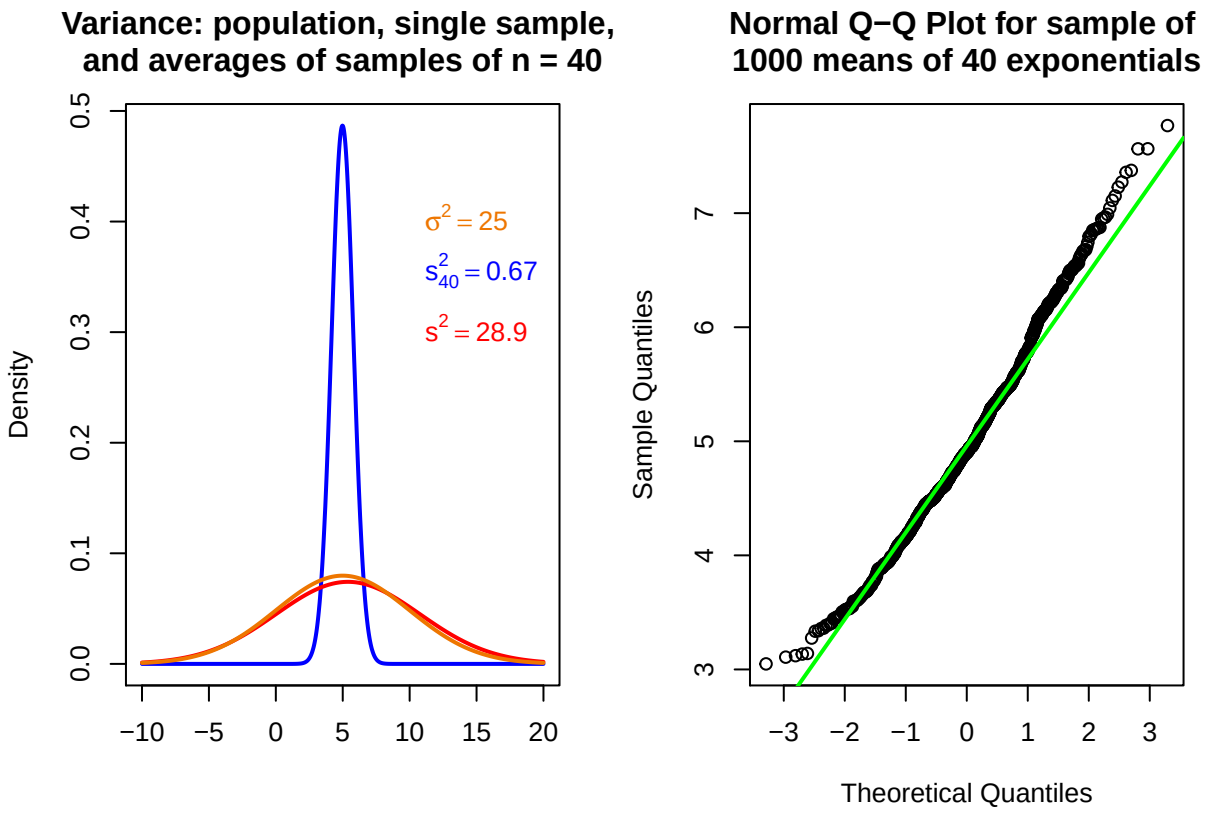
On the left we have the single sample of 1000 exponentials following the shape of the exponential distribution. It has mean $\bar{x} = 5.39$ and variance $s^2 = 28.9$. In contrast the sample of 1000 *averages* of 40 exponentials has a nearly normal distribution with mean $\bar{x}_{40} = 4.98$ and variance $s^2_{40} = 0.67$. Remembering the population parameters of $\mu = 5$ and $\sigma^2 = 25$, we see that although the single sample is relatively close the the population mean, the mean of averages of 40 exponentials is much more closely centered at the true population mean.

Sample Variance versus Theoretical Variance

Similarly, we can plot the variance of the two samples:

```
par(mfcol = c(1, 2), cex = 0.8)
xpts <- seq(-10, 20, length.out = nsims)
plot(xpts, dnorm(xpts, mean = mean(x2), sd = sd(x2)), col = "blue", lwd = 2, type = "l",
     main = "Variance: population, single sample, \nand averages of samples of n = 40",
     xlab = "",
     ylab = "Density")
lines(xpts, dnorm(xpts, mean = mean(x1), sd = sd(x1)), col = "red", lwd = 2)
lines(xpts, dnorm(xpts, mean = mu, sd = sqrt(sigma2)), col = "darkorange2", lwd = 2)
text(10, 0.3, bquote(s^2 == .(round(var(x1), 2))), pos = 4, col = "red")
text(10, 0.35, bquote(s[40]^2 == .(round(var(x2), 2))), pos = 4, col = "blue")
text(10, 0.4, bquote(sigma^2 == .(round(sigma2, 2))), pos = 4, col = "darkorange2")
```

```r
qqnorm(x2, main = "Normal Q-Q Plot for sample of \n1000 means of 40 exponentials")
qqline(x2, col = "green", lwd = 2)
```



Indicated in blue in the plot on the left is the distribution of averages of 40 exponentials with a variance $s_{40}^2 = 0.67$. In orange we have the exponential distribution of the underlying population, with variance $\sigma^2 = 25$. The single sample, in red, has a similar variance to the population from which it is drawn, as would be expected. It is immediately clear that averages of n samples result in a much narrower spread, with a much smaller variance compared to the significantly wider population variance. The CLT states that the variance of averages of samples of size n is given by $\frac{\sigma^2}{n} \approx \frac{25}{40} \approx 0.625$ and indeed we can verify that s_{40}^2 approximates it reasonably well.

Distribution

The CLT also states that averages of sample means will be nearly normally distributed. Earlier we already noted that the histogram is roughly normal. Above to the right, the normal quantile-quantile plot confirms that it is nearly normal since the data fall along the line $y = x$, albeit with some right-skew, just as the histogram shows. As n increases, we can expect a more normal shape.

Conclusion

We have therefore confirmed that the distribution of 1000 samples of size 40 taken from the exponential distribution behave as predicted by the CLT, with the mean approaching the population mean, variance related to the population variance scaled by the sample size, and with the samples distributed nearly normally.