CLT-Demo

Enzo Alda

Sunday, December 21, 2014

This exercise is an empirical demonstration of the Central Limit Theorem (CLT)

We will see how starting with a distribution that is nothing like normal, we still end up with an approximately normal distribution of sample means, for sample sizes sufficiently large. The CLT result is very important in statistics because enables us to use what we know about the normal distribution to estimate and compute confidence intervals for the mean of the population from the mean of a sample.

We start by sampling values from an exponential distribution:

```r
# Generate exponential sample

set.seed(7139);

N <- 1000;

lambda <- 0.2;

exp.sample <- rexp(N, lambda);
```

Now we simulate taking multiple samples, for different sample sizes, to see how the mean of these samples is distributed. We extract 1000 samples for each of our chosen sample sizes of 2, 4, 10, 20, and 40. Note that the project only required the last one (40), but showing the progression towards normality as the sample size increases is a great way to get a sense for the CLT.

```r
means <- NULL;
ks <- NULL;

# Compute sample means (i.e. sampling distibution) samples for different sample sizes

S <- c(2, 4, 10, 20, 40); # Sample sizes

for (k in S) {
    for (i in 1:N) {
        means <- c(means, mean(rexp(k, lambda)));
        ks <- c(ks, k);
    }
}

df <- data.frame(x=means, k=ks);
```

We are ready to visualize the convergence of the sample means distribution towards a normal distribution. The pink histogram shows the original exponential distribution. The blue plots show the sampling distributions of the mean for different sample sizes. Note tat a sampling distribution for sample size 1 would be equivalent to the underlying exponential distribution. As the sample size increases, the distribution of the means quickly resembles a normal distribution more than an exponential one.

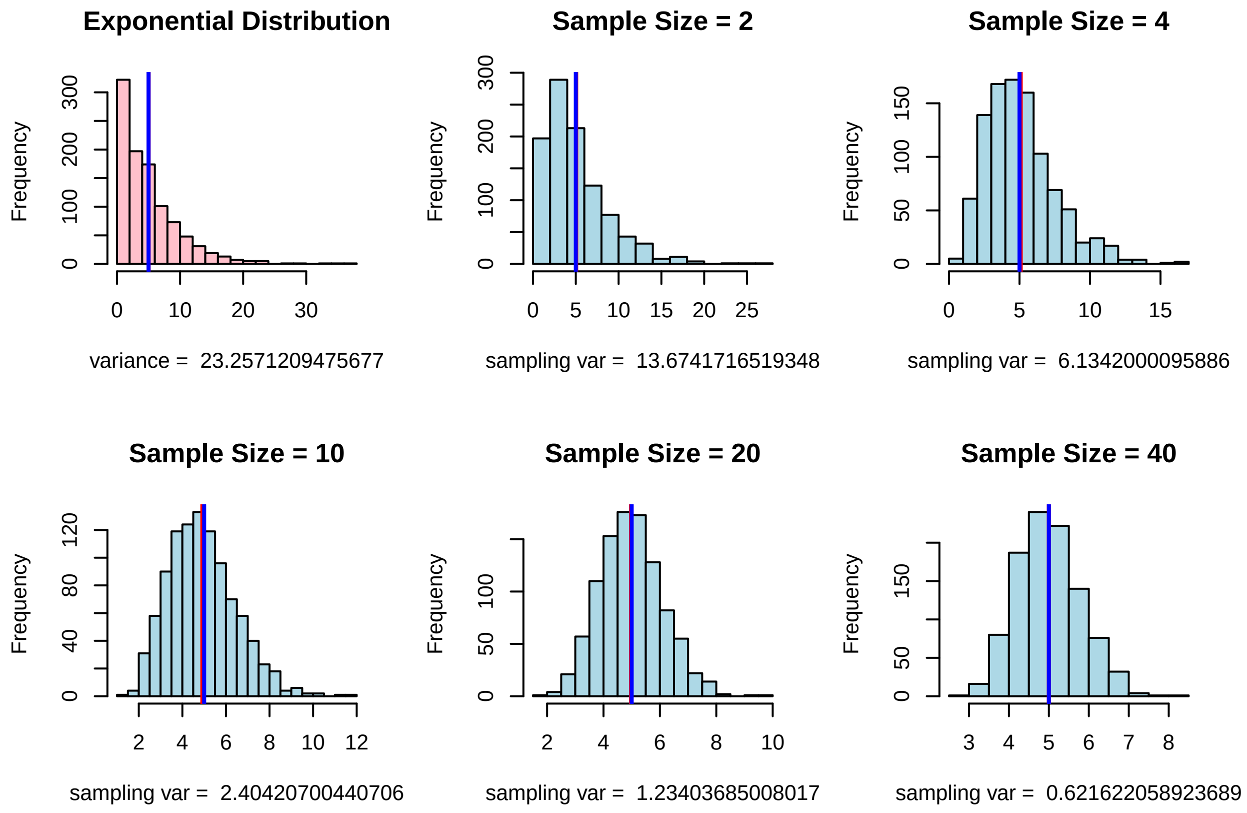
In all plots, the empirical mean is shown in red and the theoretical one in blue. Note how well the empirical mean matches the theoretical one. Indeed , the CLT tells us that the mean of the sample is an unbiased estimator of the mean of the original distribution. Also note how well the empirical variance of each sampling distribution matches the theoretical one: i.e. the variance of the population (use the variance of our sampled exponential as reference) divided by the size of the sample.

```r
# Generate plots

par(mfrow=c(2,3));

title <- paste("Exponential Distribution");
xvar <- paste("variance = ", toString(var(exp.sample)));
hist(exp.sample, breaks=16, main=title, xlab=xvar, col="pink");
abline(v=mean(exp.sample), lwd=2, col="red");
abline(v=1/lambda, lwd=2, col="blue");

for (k in S) {
    title = paste("Sample Size =", toString(k));
    sample <- df[df$k==k,]$x;
    xvar <- paste("sampling var = ", toString(var(sample)));
    hist(sample, breaks=16, main=title, xlab=xvar, col="lightblue");
    abline(v=mean(sample), lwd=2, col="red");
    abline(v=1/lambda, lwd=2, col="blue");
}
```


Exponential Distribution
Frequency
variance = 23.2571209475677
Sample Size = 2
Frequency
sampling var = 13.6741716519348
Sample Size = 4
Frequency
sampling var = 6.1342000095886
Sample Size = 10
Frequency
sampling var = 2.40420700440706
Sample Size = 20
Frequency
sampling var = 1.23403685008017
Sample Size = 40
Frequency
sampling var = 0.621622058923689