Law of Large Numbers and Central Limit Theorem Example

Alexander Alexandrov

Overview

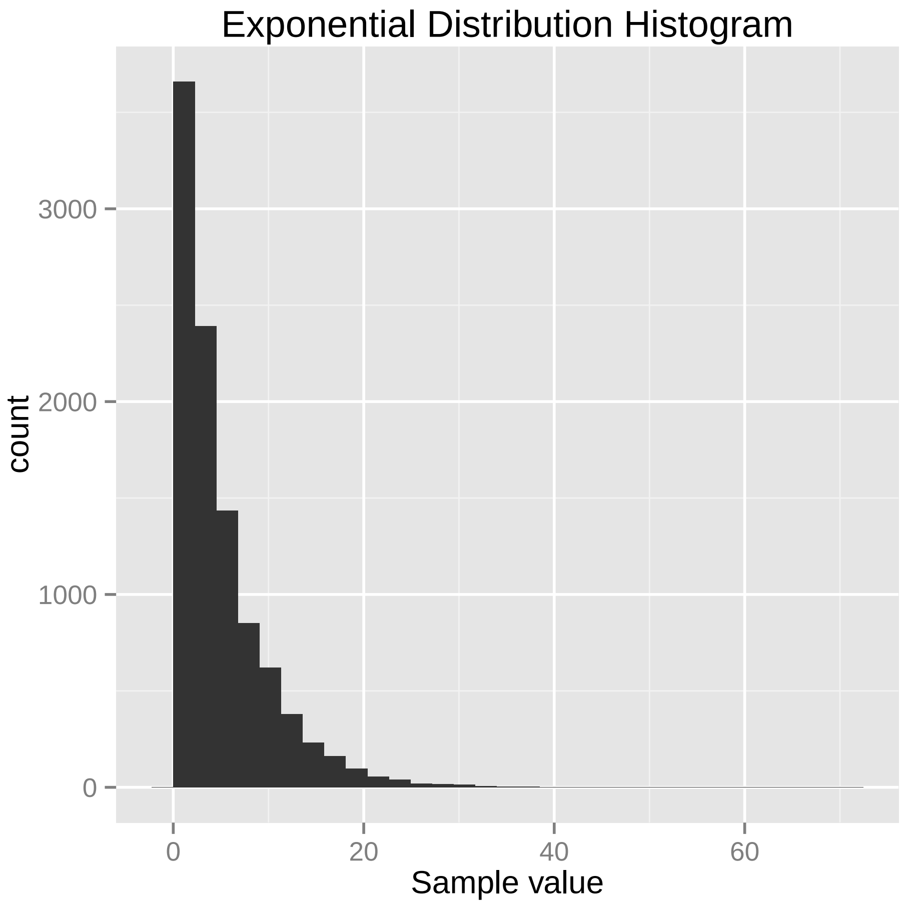
- The LLN states that the average limits to what it's estimating, the population mean.
- The CLT states that the distribution of averages of iid variables (properly normalized) becomes that of a standard normal as the sample size increases.
- This paper illustrates these lwas by the example based on the exponential distribution.
- The mean of exponential distribution is 1/lambda and the standard deviation is also 1/lambda.

Simulations

Lets simulate **exponentially distributed** random variable.

- Let sample size be equal to **40**.
- Number of observations = **10000**.
- Lambda (rate) = **0.2**.

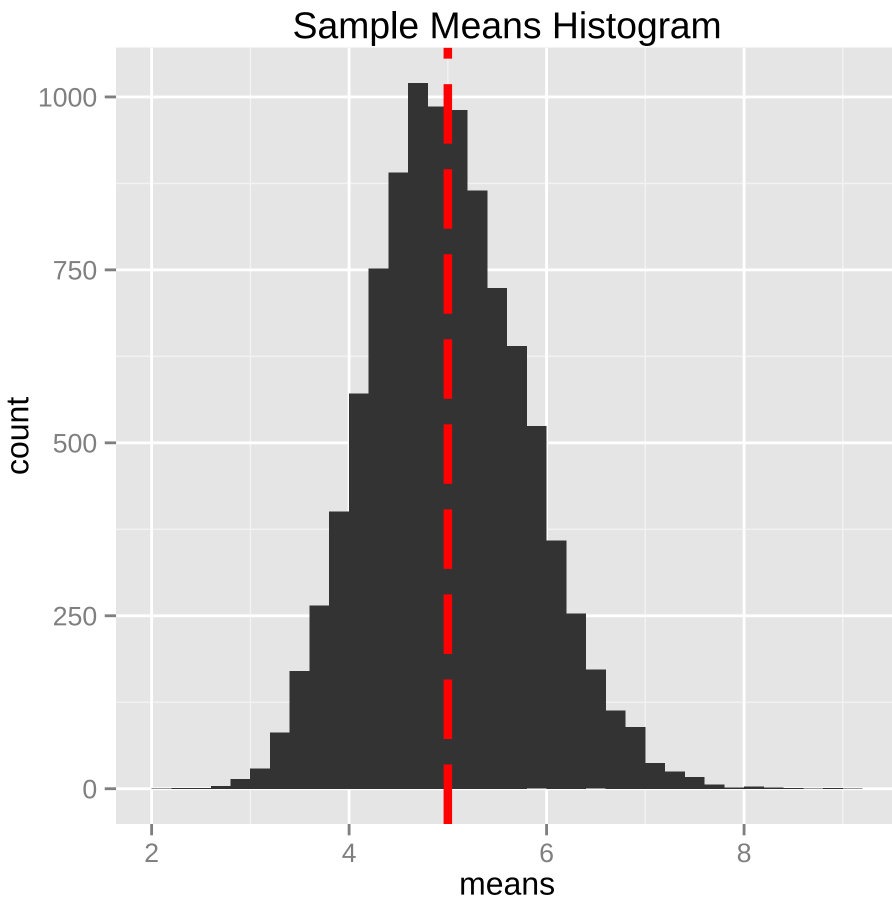
```
n <- 40                          # sample size
nosim <- 10000                   # number of observations
lambda <- 0.2                    # exponential distribution rate
population.mean <- 1 / lambda # theoretical mean
population.sd <- 1 / lambda   # theoretical std. deviation
samples <- matrix(rexp(nosim * n, rate = lambda), nrow = nosim)

library(ggplot2)
qplot(samples[, 1], geom = "histogram", xlab = "Sample value", main = "Exponential Distribution Histogra
```



The **samples** variable contains iid observations (in it's rows). One observation for each row.

Sample Mean versus Theoretical Mean (LLN)

```
means <- rowMeans(samples)
g <- qplot(means, geom = "histogram", main = "Sample Means Histogram", binwidth = 0.2)
g <- g + geom_vline(xintercept = population.mean, colour = "red", linetype = "longdash", size = 1.5)
g
```



According to the LLN, the sample mean limits to the theoretical mean. **Theoretical** mean is showed by the **vertical red line**. And it's evident from this figure that sample mean is distributed around theoretical mean.

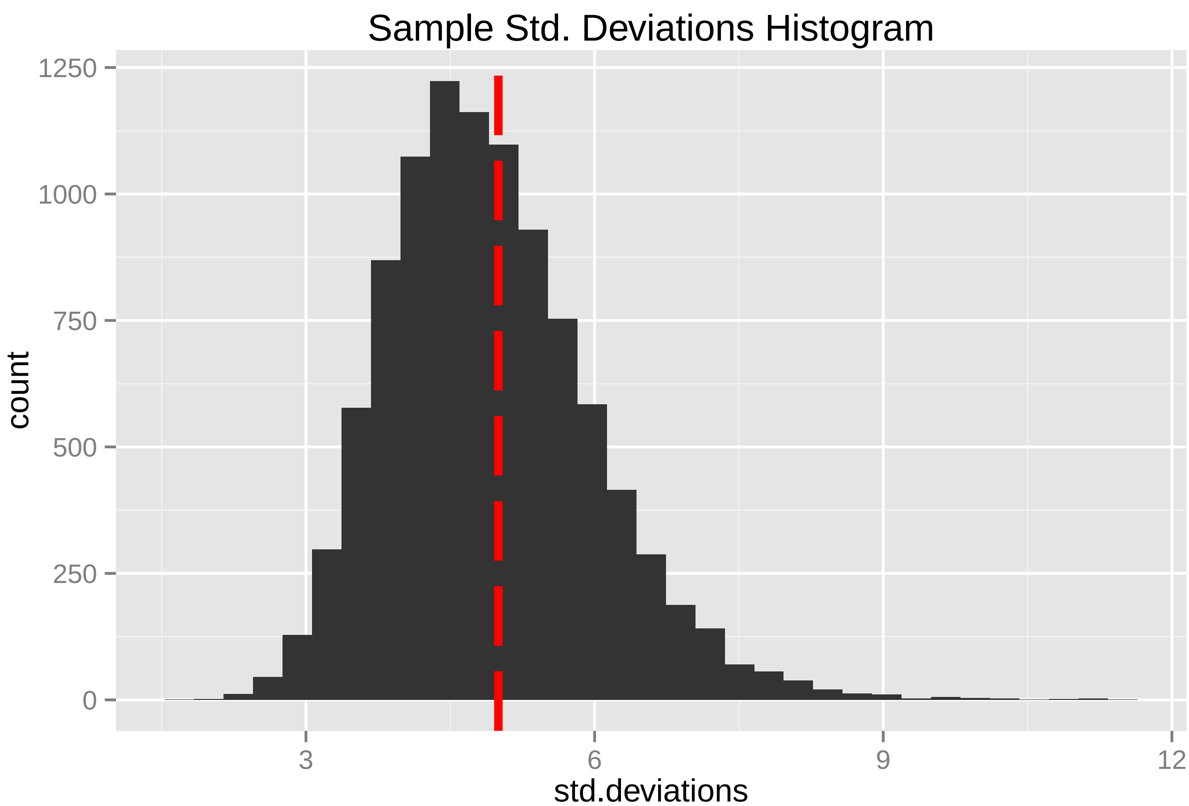
The relative difference between theoretical mean and sample mean:

```
paste(round(abs(mean(means) - population.mean) * 100 / population.mean, 1), "%", sep = "")
```

```
## [1] "0.1%"
```

Sample Std. Deviation versus Theoretical Std. Deviation

```
std.deviations <- apply(samples, 1, sd)
g <- qplot(std.deviations, geom = "histogram", main = "Sample Std. Deviations Histogram")
g <- g + geom_vline(xintercept = population.sd, colour = "red", linetype = "longdash", size = 1.5)
g
```



According to the LLN, the sample std. deviation limits to the theoretical std deviation. **Theoretical** std. deviation is showed by the **vertical red line**.

Distribution (CLT)

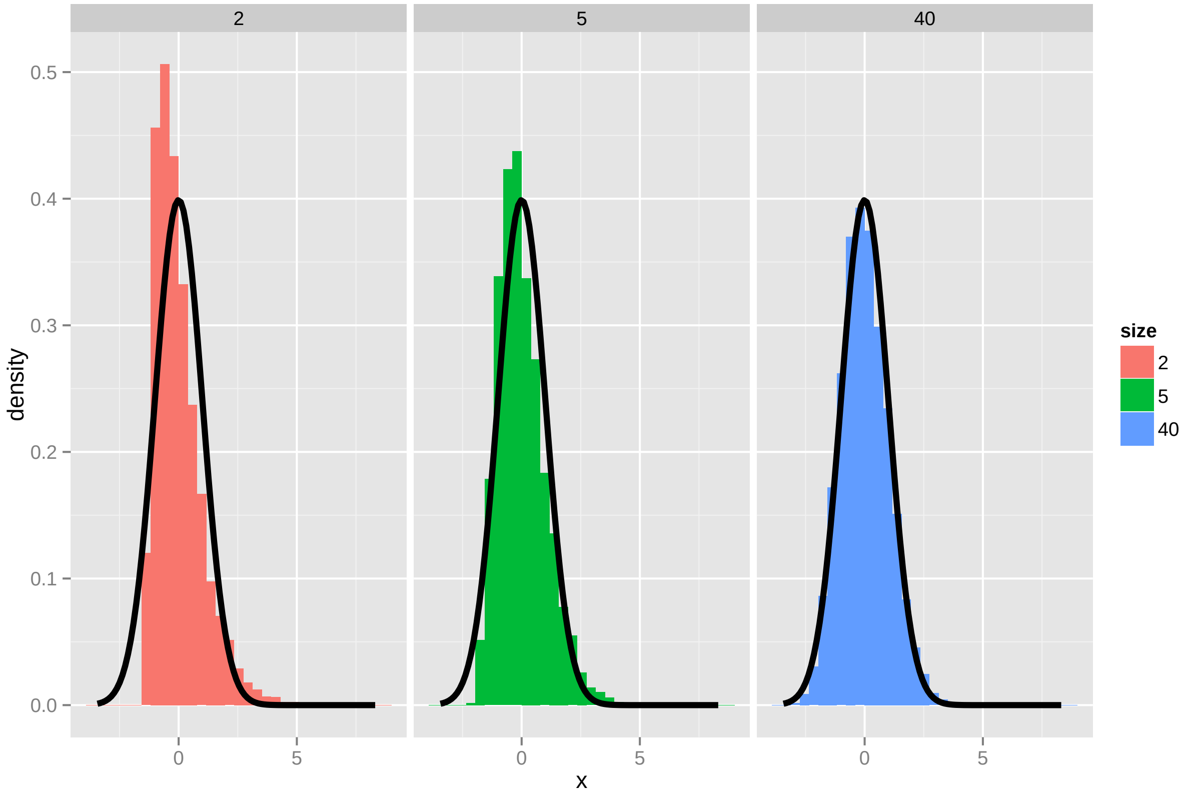
The CLT applies in an endless variety of settings - The result is that

$$\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} = \frac{\text{Estimate} - \text{Mean of estimate}}{\text{Std. Err. of estimate}}$$

has a distribution like that of a standard normal for large n. The useful way to think about the CLT is that $\bar{X}_n$ is approximately $N(\mu, \sigma^2/n)$.

- Let's simulate standart normal random variable.
- Let X_i be the smaple mean from the previous example (exponentially distributed random variable simulation).
- Then note that $\mu = E[X_i] = \frac{1}{\lambda}$.
- $Var(X_i) = \frac{1}{\lambda^2}$.
- SE $= \frac{1}{\lambda\sqrt{n}}$.
- For each observation take mean, subtract off μ, and divide by SE

```
cfunc <- function(x, n) (mean(x) - population.mean) / (population.sd / sqrt(n))
dat <- data.frame(
  x = c(apply(samples[, 1:2], 1, cfunc, 2),
        apply(samples[, 1:5], 1, cfunc, 5),
        apply(samples[, 1:40], 1, cfunc, 40)),
  size = factor(rep(c(2, 5, 40), rep(nosim, 3))))
g <- ggplot(dat, aes(x = x, fill = size)) + geom_histogram(aes(y = ..density..))
g <- g + stat_function(fun = dnorm, colour = "black", size = 1.5)
g + facet_grid(. ~ size)
```



According to the CLT, the larger the sample size, the closer sample param. distribution to the normal distribution.