Estimating Distribution Properties by Sampling

Darren Bishop (mail@darrenbishop.com)

```
library(knitr)
library(dplyr)
library(ggplot2)
library(scales)

opts_chunk$set(message=FALSE, fig.path="figures/", fig.width=10, fig.align = "center")

set.seed(221181)

lambda = 0.2
```

Overview

This report looks at how sampling can be used to estimate properties of a distribution.

The method works regardless of the population's distribution, however this report will use IIDs from the Exponential distribution, $X \sim Exponential(\lambda)$, which has both mean and standard deviation $\frac{1}{\lambda}$.

Specifically, this report shows how population mean, μ, and population variance, σ^2, can be estimated by the sample mean, $\bar{X}$, and the sample variance, S^2, respectively.

Simulations

```
simulations = 1000

n = 40

sim_rexp = data.frame(replicate(simulations, rexp(n, lambda), simplify = "matrix"))

mu = 1/lambda
```

To demonstrate the application of sampling as a method for estimating mean and variance, 1000 simulations of 40 samples are taken from the exponetial distribution.

From this sample data, a distribution of sample means is created by taking the mean of each of the 1000 size-40 sample collections.

Similarly, a distribution of sample variance is created by taking the variance of each of the 1000 size-40 sample collections.

For all simulations, the rate parameter is given as:

$$\lambda = 0.2 = \frac{1}{5}$$

Sample Mean versus Theoretical Mean

The mean of the exponetial distribution is given as:

$$\mu = \frac{1}{\lambda} = 5$$

The mean of a sample from a distribution, $\bar{X}$, is itself an IID and has its own distribution.

The sample mean is a normally distributed IID random variable; its distribution is centred around the population mean, μ, to which it approximates.

```
sim_rexp_means = data.frame(mean = colMeans(sim_rexp))
sim_rexp_means.mean = mean(sim_rexp_means$mean)

ggplot(sim_rexp_means, aes(x = mean, y = ..density..)) +
    geom_histogram(colour = "black", fill = "white") +
    geom_density(alpha=.1, fill="red") +
    scale_x_continuous(name = "Sample Mean") +
    geom_vline(aes(xintercept = sim_rexp_means.mean), color = "blue", size = 1) +
    geom_vline(aes(xintercept = 1/lambda), color = "red", size = 1, linetype = "dashed")
```

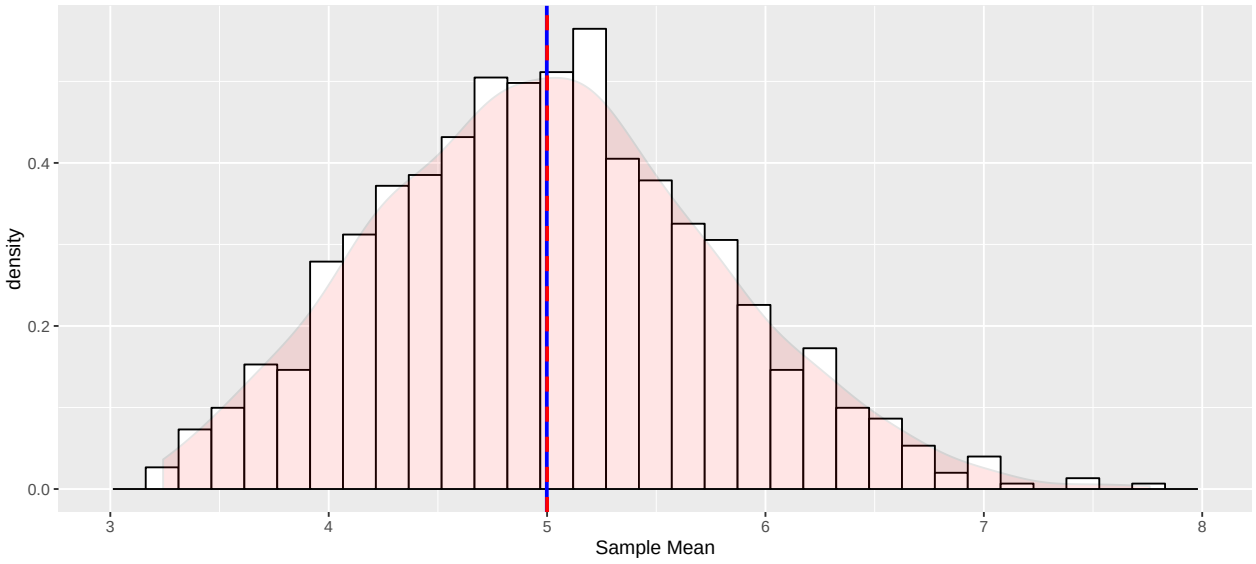

Figure 1: Normal distribution of sample means, with sample size n = 40

Figure 1. shows the sample mean distribution, with the distribution's mean plotted in blue; it has gaussian shape and is centered around the theorectical population mean, $\mu = 5$, plotted in red

The sample mean, $\bar{X} = 4.998667$, is very close to the theoretical population mean, $\mu = 5$.

The sample mean's standard deviation $S = 0.7740452$, is also very close to the distribution's standard error of the mean $\frac{\sigma}{\sqrt{n}} = \frac{1}{\lambda.\sqrt{n}} = 0.7905694$

Sample Variance versus Theoretical Variance

```r
sample_variance <- function(values) {

    x = mean(values)
    n = length(values)
    sum((values - x)^2) / (n - 1)
}

sim_rexp_vars = data.frame(vars = sapply(sim_rexp, sample_variance))
sim_rexp_vars.mean = mean(sim_rexp_vars$vars)

var.theoretical = (1/lambda)^2
```

The variance of a sample from a distribution is itself an IID and has it's own distribution; taking the variance of m simulations of this sampling gives S^2, the sample variance.

The sample variances are normally distributed IID random variables; the sample variance distribution is centred around the population variance, σ^2, to which it approximates.

The theoretical population variance is:

$$\sigma^2 = \left(\frac{1}{\lambda}\right)^2 = 5^2 = 25$$

The sample variance, used as an estimate, is:

$$S^2 = \frac{\sum_{i=1}^{N} \left(X_i - \bar{X}\right)^2}{n-1} = 25.0927$$

```r
ggplot(sim_rexp_vars, aes(x = vars, y = ..density..)) +
    geom_histogram(colour = "black", fill = "white") +
    geom_density(alpha=.1, fill="red") +
    scale_x_continuous(name = "Sample Variance") +
    geom_vline(aes(xintercept = sim_rexp_vars.mean), color = "blue", size = 1) +
    geom_vline(aes(xintercept = (1/lambda)^2), color = "red", size = 1, linetype = "dashed")
```

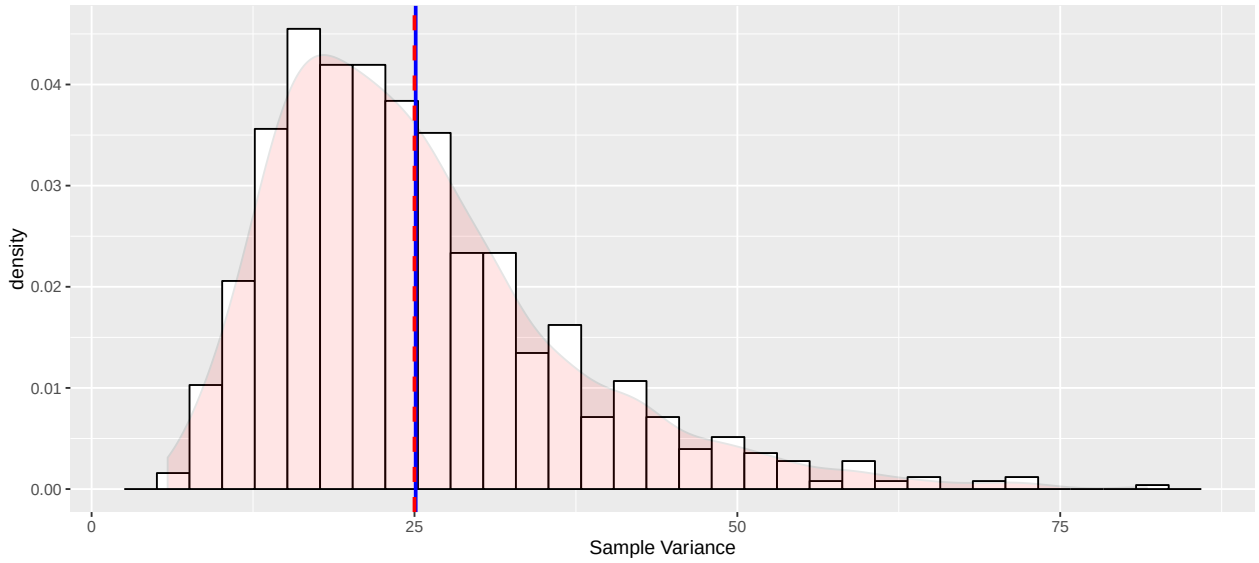

Figure 2: Normal distribution of sample variances, with sample size n = 40

Figure 2. shows the sample variance distribution, with the distribution's mean plotted in blue; it is centered around the theorectical population variance, $\sigma^2 = 25$, plotted in red dashed.

As shown, the sample variance, $S^2 = 25.0927$, is very close to the population variance, $\sigma^2 = 25$.

Distributions

The distribution of the sample mean in Figure 1. shows a distinct gaussian shape, suggesting a normal distribution.

We now compare that to the distribution of 1000 samples from the $X \sim Exponential(0.2)$ distribution.

```
sim_rexp_1000 = data.frame(exp = t(sim_rexp[1,])[,1])

sim_rexp_1000.mean = mean(sim_rexp_1000$exp)

ggplot(sim_rexp_1000, aes(x = exp, y = ..density..)) +
    geom_histogram(colour = "black", fill = "white") +
    geom_density(alpha=.1, fill="red") +
    scale_x_continuous(name = "Exponential") +
    geom_vline(aes(xintercept = sim_rexp_1000.mean), color = "blue", size = 1) +
    geom_vline(aes(xintercept = 1/lambda), color = "red", size = 1, linetype = "dashed")
```

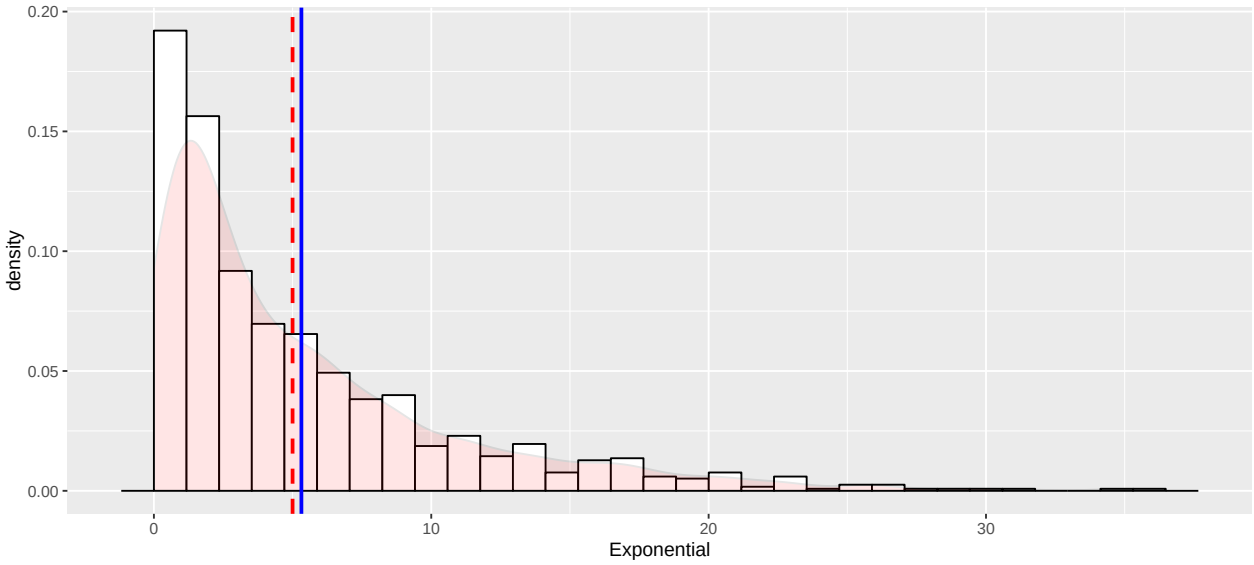

Figure 3: Exponential distribution, with sample size n = 1000

Figure 3. shows the Exponential distribution, with the sample mean plotted in blue and the population mean plotted in red; we see in stark contrast to Figure 1. that this distribution is far from gaussian, with no discernable lower tail.