Coursera Data Science Project: Statistical Inference (Part 1)

Rob Rolleston

August 23, 2015

Introduction

This is the project for the statistical inference class. In it, I will use simulation to explore inference and do some simple inferential data analysis. The project consists of two parts:

1. A simulation exercise. (this report)
2. Basic inferential data analysis.

Simulation Excercise

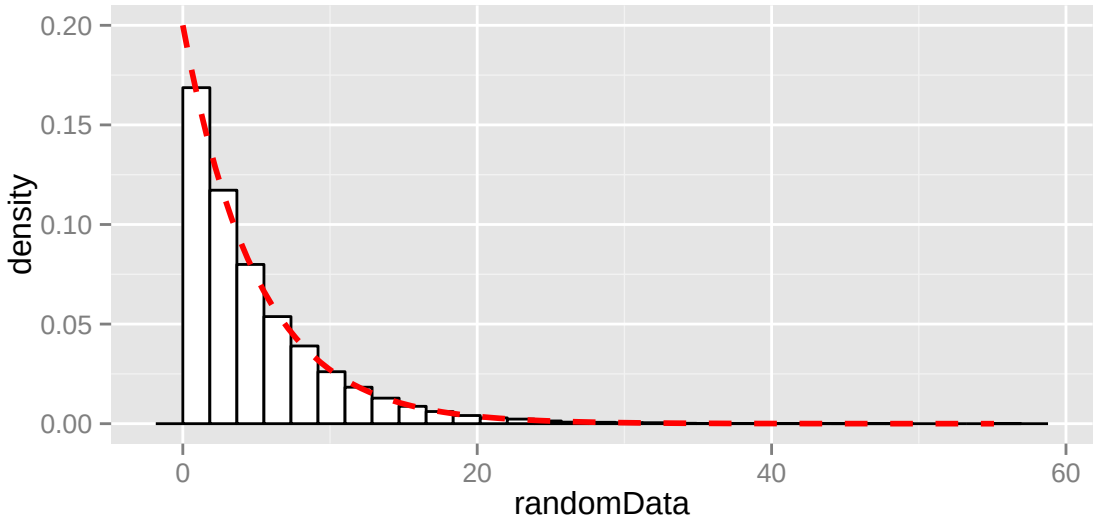
We will generate a large number of iid variables, and then use them to sample without replacement. The parameters used are:

```
nosim <- 1000
n <- 40
lambda <- 0.2
populationMean <- 1./lambda
populationSD <- 1./lambda
```

Simulations

Generate 40000 exponentially iid samples with rate=0.2. To visually inspect these samples, a histogram of the sample is plotted, along with the theoretical density curve for the population (red dashed line).

```
randomData <- rexp(nosim * n, rate=lambda)
tbl_df(data.frame(randomData)) %>%
ggplot(aes(x=randomData)) +
  geom_histogram(aes(y=..density..), color="black", fill="white") +
  stat_function(fun=dexp, args=list(rate=lambda), color="red", linetype="dashed", size=1)
```



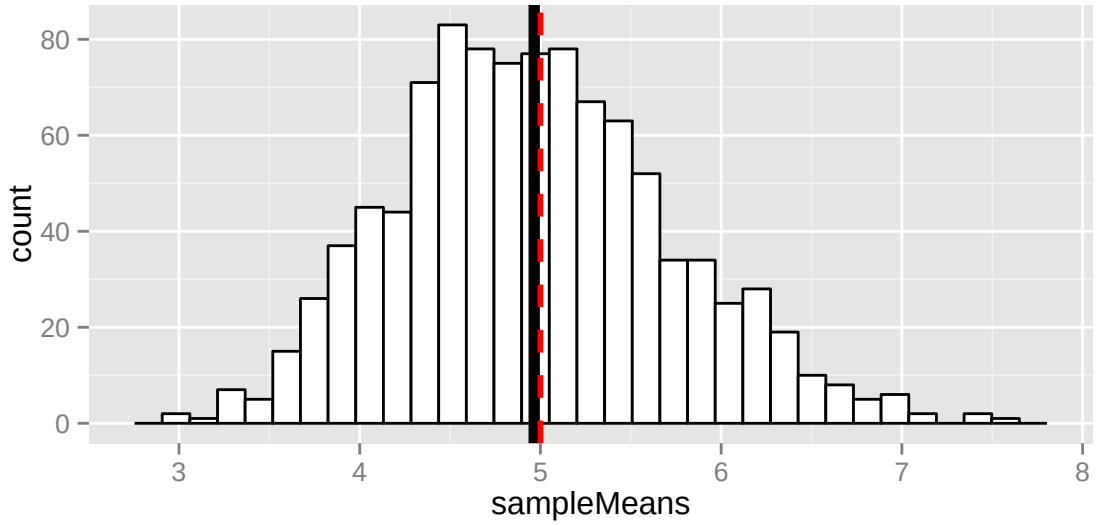
The iid samples have very good agreement to the theoretical distribution.

Sample Mean versus Theoretical Mean

These 40000 samples are now divided into 1000 sets, each with 40 samples. The mean of each sample set is calculated. *It is the propoerties of these sample means which is the focuse of this excercise.*

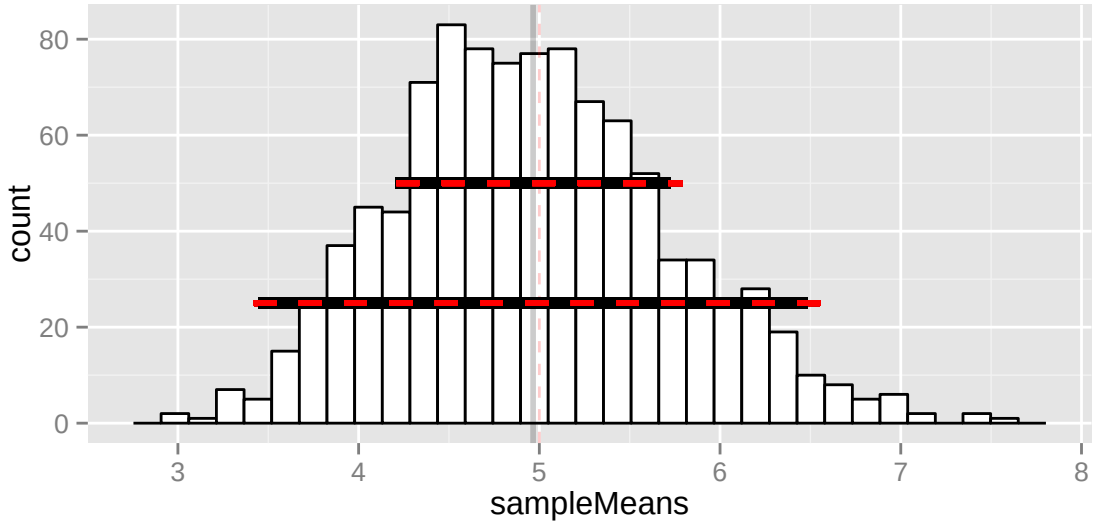
```
sampleData <- matrix(randomData, nosim)
sampleMeans <- apply(sampleData, 1, mean)
sampleMean <- mean(sampleMeans)
```

The mean of the samples is 4.965564 which is very close to the population mean of 5. This can be shown graphically, by plotting a histogram of the means of the samples. In the figure below, the solid black line is the mean of the samples, and the red dashed line is the mean of population.



Sample Variance versus Theoretical Variance

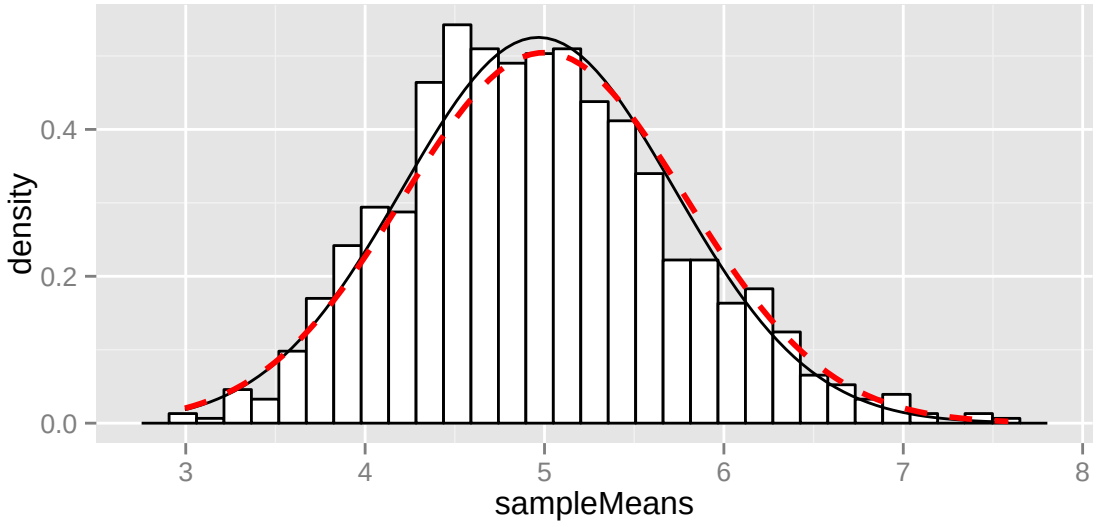
The sd of the distributions of the means of the sample sets is calculated, and we can now compare the inferred variation of the means [= s/sqrt(n))] with the actual sd of the distribution of sample means. This can be shown graphically, by plotting a histogram of the means of the samples. In the figure below, the upper solid black line is +/- 1 sd of the samples, and the upper ed dashed line is +/- 1 sd of population. The lower lines are +/-



2 sd.

Distribution

In theory, the distribution of the means should be normally distributed with a mean of 5 and a sd of 0.7905694. The figure below is a histogram of the means; the solid black line is normal(sampleMean, sampleSD), and the red dashed line is normal;(populationMean, populationSD/sqrt(n).



It can be seen that indeed, the distribution of the sample means is normally distributed.