Statistical Inference

Course Project - Assignment 1

Jim Leach

Synopsis

This report has been created for the the *Statistical Inference* MOOC from Johns Hopkins university on Coursera (https://class.coursera.org/statinference-005).

The first assignment required analysis of the distribution of sample means of random exponentials. Specifically, the following points had to be addressed:

1. Show where the distribution is centered at and compare it to the theoretical center of the distribution.
2. Show how variable it is and compare it to the theoretical variance of the distribution.
3. Show that the distribution is approximately normal.
4. Evaluate the coverage of the 95% confidence interval for $\frac{1}{\lambda}$: $\tilde{X} \pm 1.96 \dfrac{S}{\sqrt{n}}$

See Appendix 1 for the complete, end-to-end R code that generates all the required output.

Data Processing:

Simulating Random Exponentials

Random exponentials are generated in R using `rexp(n, lambda)`, where lambda is the rate parameter.

This analysis generates sets of 40 random exponentials with a rate parameter of 0.2.

In order to assess the distribution of sample mean for these 40 random exponentials, multiple simulations are performed. This analysis has used 1001 simulations.

1 - Distribution Centre

The distribution of the simulated data is centred at a value of **4.9863**.

It is the case that the average of a random sample from a population is itself a random variable with a distribution centred around the population mean, i.e. $E[\tilde{X}] = \mu$.

For the exponential distribution, μ is given by $\frac{1}{\lambda}$, i.e. **5**.

The calculated value of the centre of the distribtion compares well with the theoretical value.

2 - Distribution Variance

The variance of the simulated data (calculated with `var()`) has a value of **0.6197**.

Theoretically, the variance of the sample mean is given by $Var(\tilde{X}) = \frac{\sigma^2}{n}$ where σ is the population standard deviation and n is sample size. For the exponential distribution, σ is *also* given by $\frac{1}{\lambda}$, i.e. **5**. As such, with a sample size of n = 40, the expected variance has a value of **0.625**.

The calculated value of the distribtion variance compares well with the theoretical value.

3 - Distribution Normality

Figure 1 - Distribution of Sample Means for 40 Random Exponentials

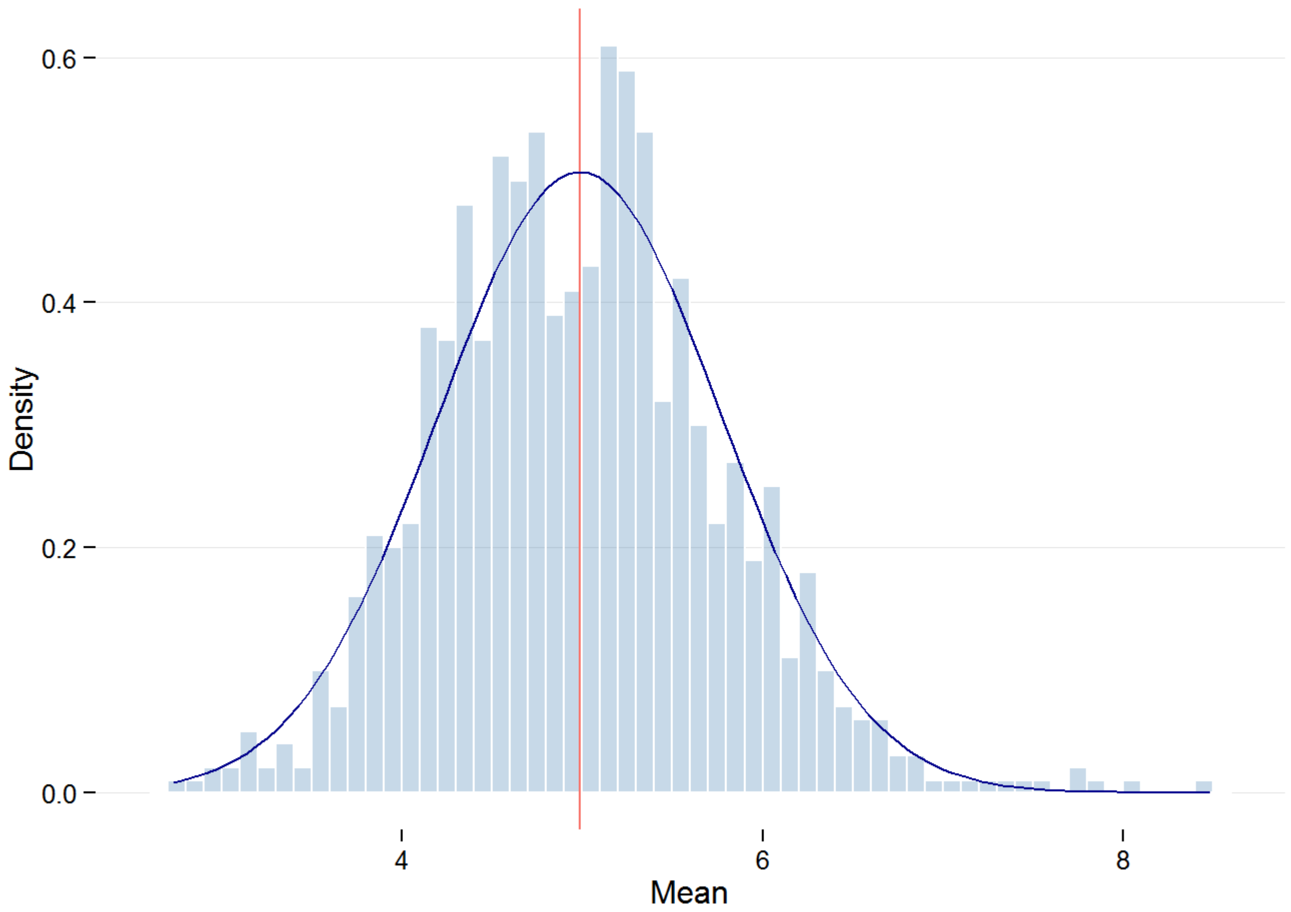

Figure 1 shows a histogram of the data with the vertical red line showing where they are centred (4.9863).

The curved blue line shows the normal distribution curve. It is clear from this figure that the simulated data are distributed approximately normally.

4 - Confidence Intervals

The 95% confidence interval for the exponential distribution is given by $\frac{1}{\lambda}: \tilde{X} \pm 1.96 \frac{S}{\sqrt{n}}$

The coverage of confidence intervals means the percentage of the 1001 computed confidence intervals that contain the theoretical value of the population mean, which is given by $\frac{1}{\lambda}$, i.e. **5**

The confidence interval was calculated for each of the 1001 sample means generated and the proportion of these that contain the true value of the population mean (5) was calculated.

The coverage of the confidence interval for $\frac{1}{\lambda}: \tilde{X} \pm 1.96 \frac{S}{\sqrt{n}}$ was found to be **95.5%**.

References and Contact

A number of packages were used and are therefore required for this analysis. These packages were:

- magrittr (http://cran.r-project.org/web/packages/magrittr/vignettes/magrittr.html)
- ggplot2 (http://ggplot2.org/)

The complete code and documentation for this assignment can be found on GitHub (https://github.com/Jim89/stat_infer) repository.

The author of this report can be contacted on twitter (https://twitter.com/leach_jim)

Appendix 1 - Complete R Code

```
################################################################################
# Step 0 - Prepare the working environment - load packages
################################################################################
# 0i. ggplot2
library(ggplot2,quietly=T)

# 0ii. magrittr
library(magrittr,quietly=T)

################################################################################
# Step 1 - set up the variables to be used in the simulation
################################################################################
# 1i. Set the rate parameter
lambda <- 0.2

# 1ii. Set the number of exponentials to generate
n <- 40

# 1iii. Set the number of simulations
sims <- 1001

################################################################################
# Step 2 - Simulate the data
################################################################################
# 2i. Set a seed for reproducibility
set.seed(8020)
# 2ii. Simulate the data
simulated <- 1:sims %>% lapply(function(i){rexp(n,0.2)})
# 2iii. Calculate the means
means <- simulated %>% sapply(mean)

################################################################################
# Step 3 - Calculate Expected Mean and SD
################################################################################
# 3i. mean
mean_theory <- 1/lambda

# 3ii. SD
sd_theory <- 1/lambda

# 3iii. Variance
var_theory <- sd_expected^2/n

################################################################################
# Step 4 - Calculate actual descriptive statistics of general data
################################################################################
# 4i. mean
mean_actual <- mean(means)

# 4ii. variance
var_actual <- var(means)
```

```
# 4iii. confidence intervals for each mean
intervals <- means %>%
             lapply(function(i){i + c(-1,1) * 1.96 * (sd_theory/sqrt(n))})

# 4iv. function to calcualte if true mean is between confidence intervals
is.between <- function(x, a, b) {
  x > a & x < b
}

# 4v. calculate if true mean is between confidence intervals
between <- 1:length(intervals) %>%
           sapply(function(i){is.between(mean_theory,
                                         intervals[[i]][1],
                                         intervals[[i]][2])})

# 4vi. calculate confidence coverages
confidence_coverage <- round(100*(sum(between)/length(between)),1)

###############################################################################
# Step 5 - Plot the distribution and show it is normally distributed
################################################################################
ggplot(data.frame("value"=means), aes(x= value))+
  geom_histogram(aes(y=..density..),fill="steelblue", alpha= 0.3,
                binwidth = 0.1, colour="white")+
  geom_vline(aes(xintercept=mean(means),color="firebrick"))+
  theme_minimal()+
  xlab("Mean")+
  ylab("Density")+
  #ggtitle("Density plot of sample means for random exponentials")+
  theme(panel.grid.minor.y=element_blank(),
        panel.grid.major.x=element_blank(),
        panel.grid.minor.x=element_blank())+
  stat_function(color="darkblue",
                fun= dnorm , args = list(mean = mean(means) , sd = sd(means)))
```