Statistical Inference: Assignment 1

ND

15/09/2014

Introduction

This document briefly investigates the distributions of means of an exponential distribution. It demonstrates the central limit theorem, which very roughly states that the mean of a large number of independent random variables will be approximately normally distributed. The R markdown source for this document is available on github.

Simulation

The following code create a matrix of 1000 rows of 40 columns of random samples from an exponential distribution with $\lambda = 0.2$. It then creates a dataframe containing the means and standard deviations of the 40 samples. The `apply` function is used to apply the mean and standard deviation functions over the rows of the matrix to give vectors of means and standard deviations from each of the 1000 simulations.

```
set.seed(1) # set random seed for reproducibility

n <- 40
lambda <- 0.2
nsim <- 1000

temp <- matrix(rexp(n * nsim, lambda), ncol = n)
dat <- data.frame(mean = apply(temp, 1, mean),
                  sd = apply(temp, 1, sd))
```

Distribution centre

The distribution will be centred on its mean value. In other words, if the distribution were balanced on a point, the point would lie at its mean value. The mean of the empirical distribution is 4.99 while the mean of the theoretical distribution is $1/\lambda$, i.e. 5.

Distribution variance

The variance in the empirical distribution is 0.6177 while the variance of the theoretical distribution is $\left(\frac{1/\lambda}{\sqrt{n}}\right)^2$ i.e. 0.625.

Distribution shape

The following code generates a density plot comparing the empirical distribution (solid black line) with the normal distribution (dotted red line). In addition, vertical lines showing the means of the distribution are plotted.

```
d.plot <- density(dat$mean)

plot(d.plot, col = "black", lwd = 2, lty = 1, main = "", ylim = c(0,0.51))
```

```r
abline(v = mean(dat$mean), col = "black", lwd = 2, lty = 1) # mean of empirical dist.

xnorm <- seq(min(dat$mean), max(dat$mean), length = 50)
ynorm <- dnorm(xnorm, mean = 1/lambda, sd = (1/lambda)/sqrt(n))
lines(xnorm, ynorm, col = "red", lwd = 2, lty = 3)  # normal distribution

abline(v = 1/lambda, col = "red", lwd = 2, lty = 3) # mean of normal distribution
```

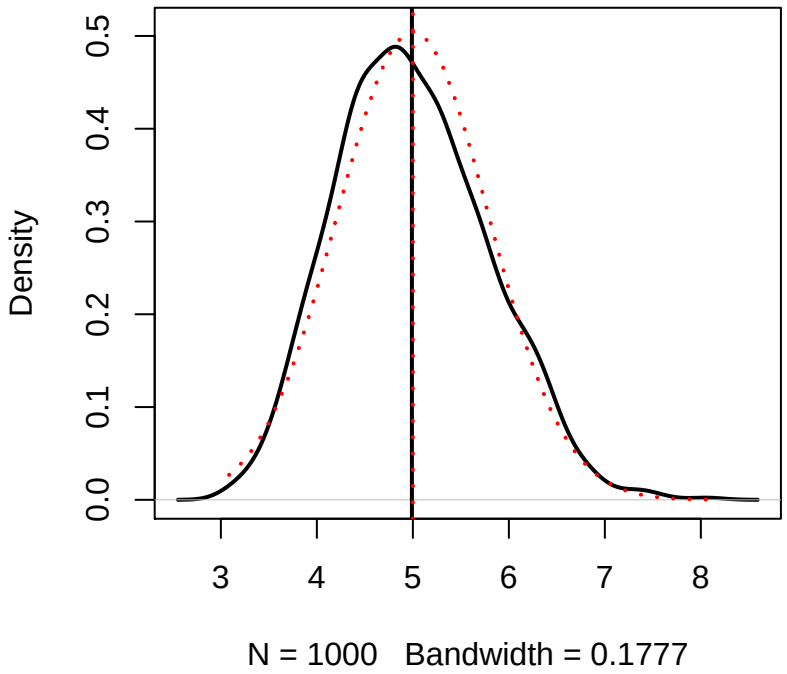

Figure 1: Density of means of 40 samples

A visual comparison of the empirical density curve with the normal density curve shows that the empirical density is approximately normal and the means are virtually identical. Increasing the number of simulations would reduce the difference in shape between the distributions.

Confidence interval

An approximation to the 95% confidence interval for the mean of the exponential distribution is given by $\bar{X} \pm 1.96\frac{S}{\sqrt{n}}$. Coverage can be calculated by generating an empirical confidence interval for each of the 1000 simulations and calculating how frequently this interval contains the true mean $1/\lambda$.

```r
ci.ll <- dat$mean - 1.96 * dat$mean / sqrt(n) # lower CI limit
ci.ul <- dat$mean + 1.96 * dat$mean / sqrt(n) # upper CI limit

coverage <- sum(1/lambda >= ci.ll & 1/lambda <= ci.ul) / nsim
```

In this example, the coverage is 94.0%. This is illustrated by Figure 4, in the appendix.

Appendix

Distribution shape (ggplot2)

Figure 2 simply demonstrates that a similar density plot comparing the empirical and theoretical distributions can be created using ggplot2.

```
ggplot(dat, aes(x = mean)) +
  geom_density(size = 1, aes(y = ..density..), ylim = c(0, 0.51)) +
  xlab("Mean") + ylab("Density") +
  geom_vline(xintercept = mean(dat$mean), size = 1) +
  stat_function(fun = dnorm, colour = "red", size = 1, linetype = 3,
                arg = list(mean = 1/lambda,  sd = (1/lambda)/sqrt(n))) +
  geom_vline(xintercept = 1/lambda, colour = "red", size = 1, linetype = 3) +
  theme_bw()
```

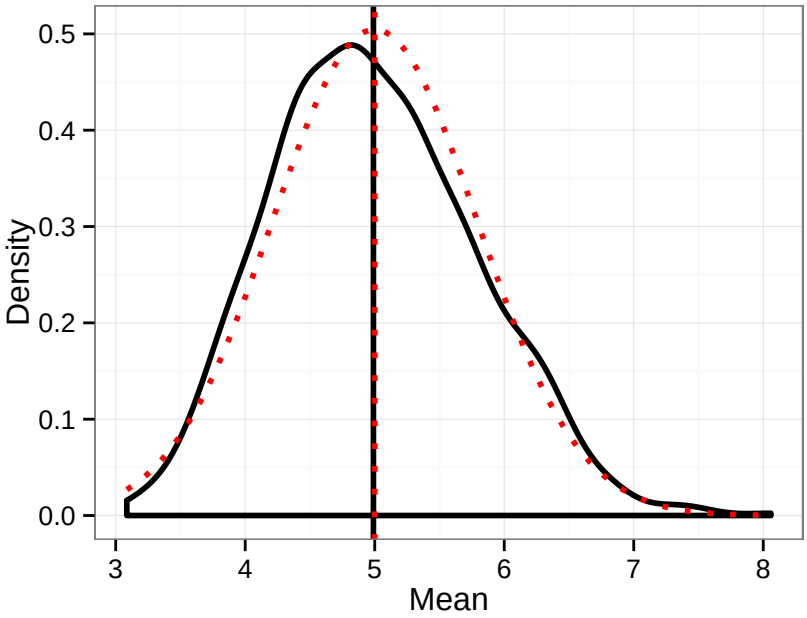

Figure 2: Density of means of 40 samples (ggplot2)

Normal q-q plot

The normal q-q plot provides another way to compare the two distributions. The relationship will be linear if the empirical data is normally distributed.

```
qqnorm(y = (dat$mean - mean(dat$mean)) / sd(dat$mean), ylim = c(-3, 3))
qqline(y = (dat$mean - mean(dat$mean)) / sd(dat$mean))
```

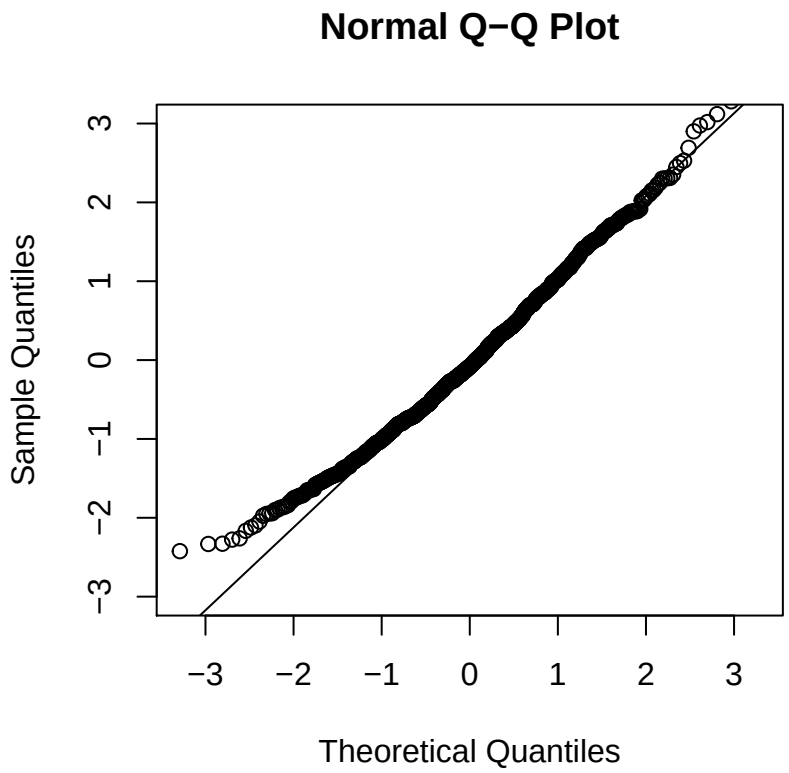

Figure 3: Normal q-q plot

The plot in Figure 3 is fairly linear, which suggests the distribution of means is approximately normal.

Confidence interval and coverage

Coverage can be illustrated graphically by plotting the empirical confidence intervals for each sample and seeing whether each interval contains the true mean. Figure 4 shows the confidence intervals for the first 100 of the 1000 simulations. The true mean, $1/\lambda$ is shown as a dashed horizontal line.

```
ggplot(dat, aes(x=1:100, y=mean[1:100], ymin=ci.ll[1:100], ymax=ci.ul[1:100])) +
    geom_pointrange() +
    geom_hline(yintercept=1/lambda, col = "red", lty=2, size = 1) +
    xlab("Sample Number") + ylab("Mean") +
    theme_bw()
```

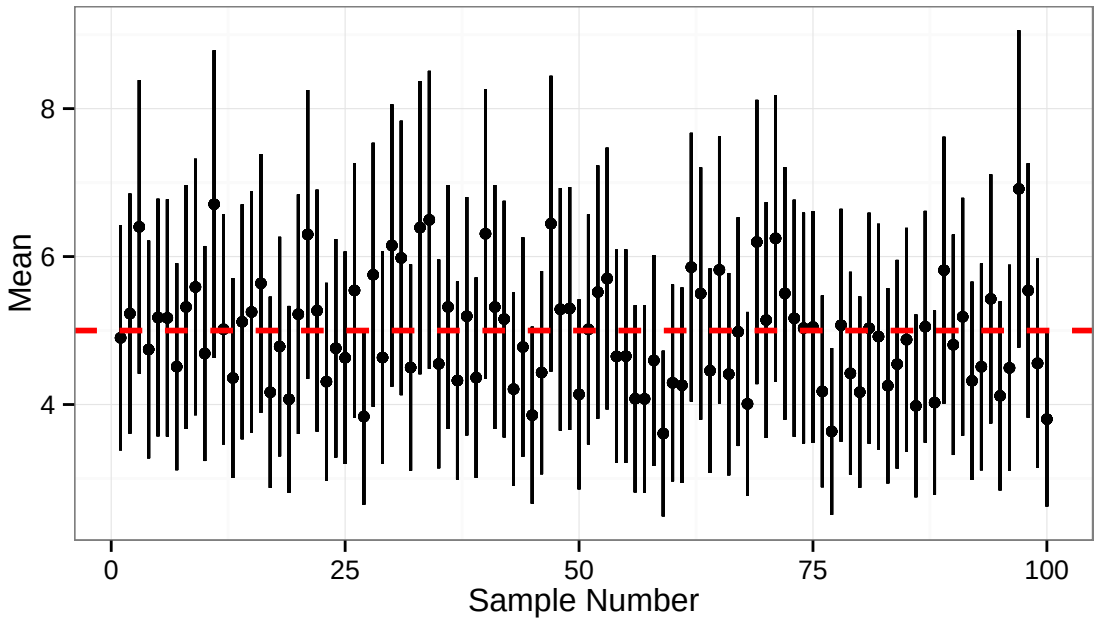

Figure 4: Coverage of empirical confidence intervals

Alternative calculation

It is possible to calculate another form of coverage based on comparing the empirical means with a theoretical confidence interval around the true mean $1/\lambda$.

```
ci <- 1/lambda + c(-1, 1) * 1.96 * (1/lambda)/sqrt(n) # CHECK
coverage <- sum(dat$mean >= ci[1] & dat$mean <= ci[2]) / nsim

ggplot(dat, aes(x=1:100, y=mean[1:100])) +
  geom_ribbon(aes(ymin=ci[1],ymax=ci[2]), fill = "pink", alpha = 0.5) +
  geom_hline(yintercept=1/lambda, col = "red", lty = 2, size = 1) +
  geom_point() +
  xlab("Sample Number") + ylab("Mean") +
  theme_bw()
```

Using the theoretical value $1/\lambda$ for mean and standard deviation results in a confidence interval running from 3.4505 to 6.5495, which gives a coverage of 96.4%. Figure 5 illustrates this approach for the first 100 of the 1000 simulations. The true mean is shown as a dashed horizontal line and its confidence interval is the pink region. This theoretical 95% confidence interval can be plotted on the empirical distribution using the `polygon` function (Figure 6).

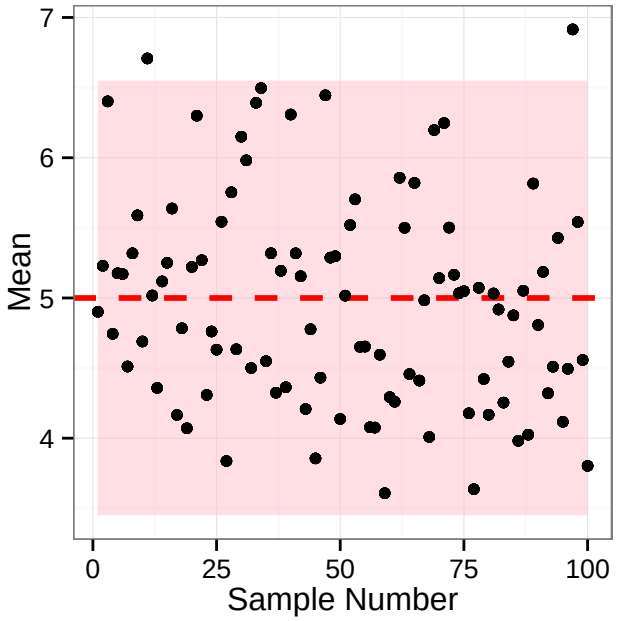

Figure 5: Coverage of theoretical confidence interval

```
x1 <- max(which(d.plot$x <= ci[1])) # lower boundary
x2 <- min(which(d.plot$x >= ci[2])) # upper boundary

plot(d.plot, col = "black", lwd = 2, lty = 1, main ="")
polygon(x = c(d.plot$x[x1], d.plot$x[x1:x2], d.plot$x[x2]),
        y = c(0, d.plot$y[x1:x2], 0), col = "pink")
text(5, 0.25, "95%\n(approx)")
```

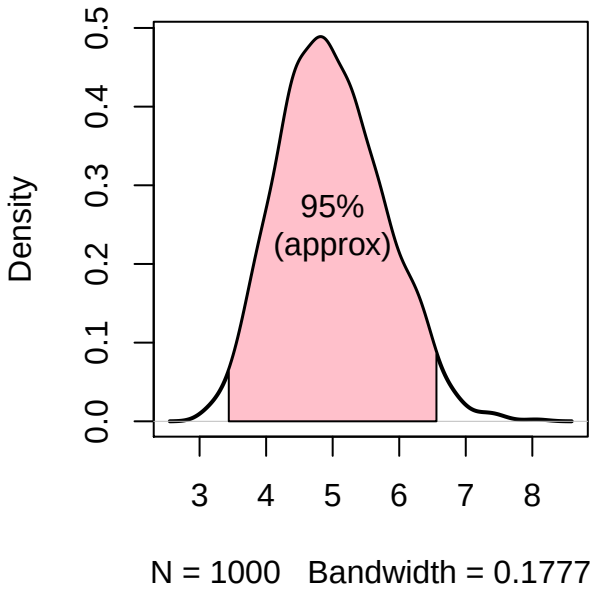

Figure 6: Empirical distribution and theoretical confidence interval