Statistical Inference Course Project

adrianvs

14 Oct 2015

Introduction

This report analyzes the exponential distribution using simulations and the Central Limit Theorem.
The exponential distribution is widely used to model waiting times. A random variable that describes the time that passes until an event occurs has an exponential distribution if the probability that the event occurs in a given time interval is proportional to the length of that time interval. Therefore it describes the time between events in a Poisson process[1] .
For the R code creating the plots see appendix.

Definition

Let X be a continuous random variable defined in the set of positive real numbers.
Then the porbability density funciton of an exponential distribution is:

$f(x) = \lambda e^{-\lambda x}, x \in \mathbb{R}_{\geq 0}$

In this report $\lambda = 0.2$ will be used.

Simulations

The following R code simmulates the drawing of a random sample of size 40 from an exponential distribution with a rate parameter $\lambda = 0.2$ and repeats that process 1000 times. Each time the sample average is calculated and stored (avg).

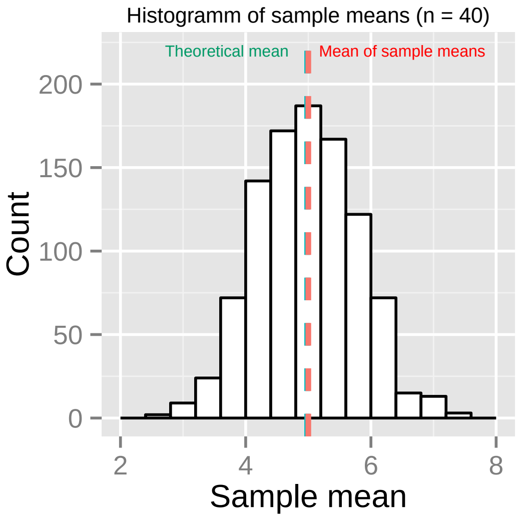
```r
avg <- NULL
nosim <- 1000

for(i in 1:nosim) avg[i] <- mean(rexp(40, rate = 0.2))

## stores avg in a data frame with label 0
hist_data1 <- as.data.frame(cbind(avg,0))

## stores 1000 random exponentials for a comparative histogramm with label 1.
exp_data <- rexp(1000, rate = 0.2)
hist_data2 <- as.data.frame(cbind(exp_data,1))
```

Sample Mean versus Theoretical Mean

To compare the sample mean with the theoretical mean, the following code creates a histogramm of the above generated sample means. Additionally the mean of the sample means and the theoretical mean are drawn. Since $E[\bar{X}]$ is unbiased, the theoretical mean is the population mean as given by: $E[\bar{X}] = \mu = 1/\lambda$

```r
theoretical_mean <- 1/0.2 # = population mean
samp_mean        <- round(mean(avg),2)
```

```
plot0
```

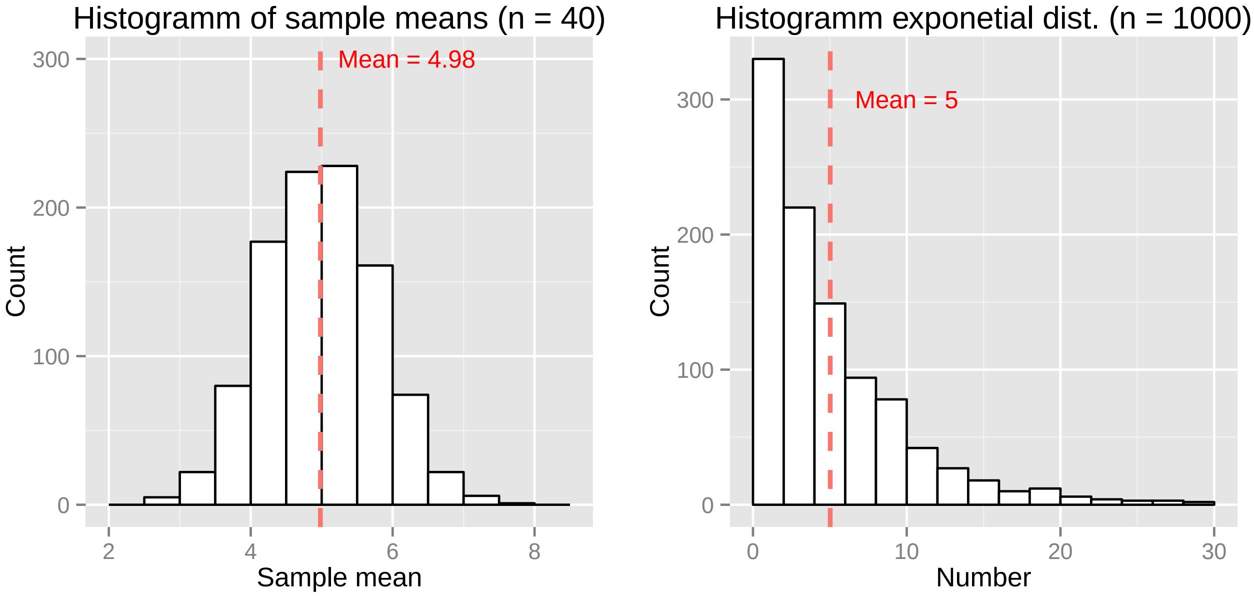


Mean of sample means	Theoretical mean
4.98	5

As can be seen, the mean of sample means estimates the true population mean closely. In this example the accuracy lies within in 99.6%.

The following plot shows, that indeed, though showing completely different distributions, the mean of sample means is a good estimate of the population mean (of 1000 random exponentials).

```
grid.arrange(plot1, plot2, ncol=2)
```



Sample Variance versus Theoretical Variance

The variance of sample means is an unbiased random variable. Its distribution is therefore centred at the populaion variance. Its own variance inversely correlates to the sample size and is theoretically given by the population variance divided by the sample size.
The variance of the exponential random variable is: $Var[X] = 1/\lambda^2$.
The theoretical value of the variance of sample means is therefore: $Var[\bar{X}] = \frac{1}{\lambda^2 n}$.

```
varsample       <- var(avg)
varexp          <- var(exp_data)
vartheoretical  <- 1/(0.2^2 * 40)
```

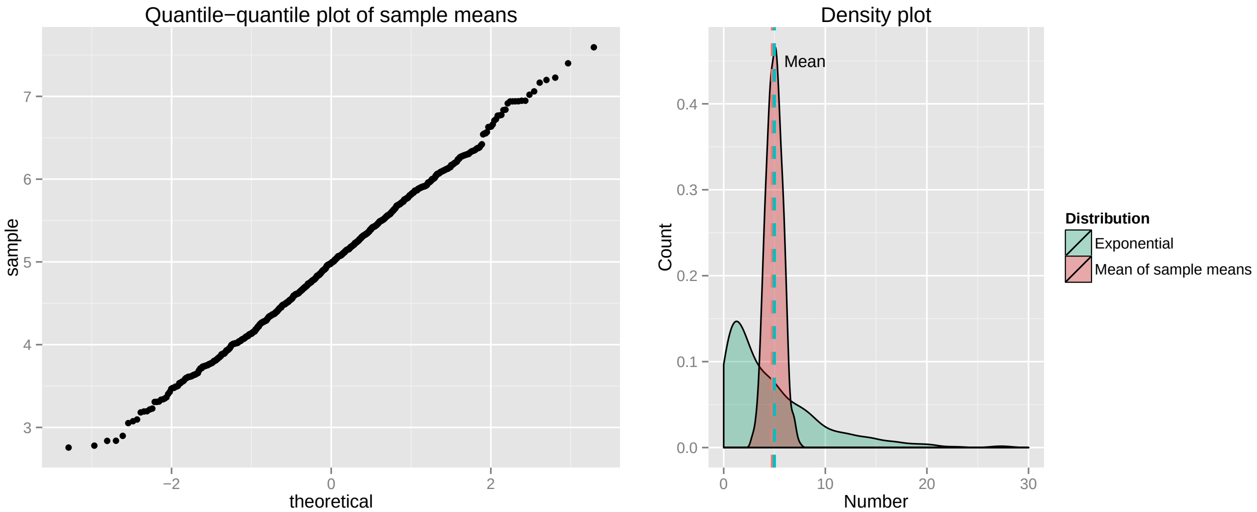
True population variance	Empirical sample variance	Theoretical sample variance
25.34	0.65	0.62

The empirical sample variance lies within 96.6% of the theoretical sample variance. As expected the variance of sample means is narrower than the variance of the population by a factor of sample size.

Distribution

The Central Limit Theorem states that with increasing sample size the distribution of sample averages of an iid variable becomes that of a normal distribution. To see if a sample size of 40 yields a good approximation of a normal distribution, the following plots show a quantile-quantile graph and an overlay of density curves of the mean of sample averages with a random exponential distribution of same rate parameter ($\lambda = 0.2$).

```
grid.arrange(plot3, plot4, ncol=2)
```



The good fit of the qq-plot and symmetry and bell-shape of the sample means density curve both suggest, that the distribution of sample means is a good approximation to a normal distribution.

Conclusion

This report showes that the means of 1000 random draws of size 40 from an exponential distribution approximate a normal distribution with N~$(\frac{1}{0.2}, \frac{1/0.2}{\sqrt{40}})$ in accordance with the Central Limit Theorem.

Footnotes 1: statlect.com

Appendix

R code generating the plots:

```
plot0 <-
ggplot(hist_data1, aes(x=avg)) +
  geom_histogram(binwidth=.4, colour="black", fill="white") +
  labs(title = "Histogramm of sample means (n = 40)",
                   x = "Sample mean", y = "Count") +
  geom_vline(data=hist_data1, aes(xintercept=mean(avg),
                   colour="red"), linetype="dashed", size=1) +
  geom_vline(aes(xintercept=theoretical_mean,  colour="cyan"),
                   linetype="dashed", size=1) +
  annotate("text", label = "Mean of sample means", x = 6.5,
                   y = 220, size = 4, color="red") +
  annotate("text", label = "Theoretical mean",
                   x = 3.7, y = 220, size = 4, color="#009966")
```

```
plot1 <-
ggplot(hist_data1, aes(x=avg)) +
  geom_histogram(binwidth=.5, colour="black", fill="white") +
  labs(title = "Histogramm of sample means (n = 40)", x = "Sample mean", y = "Count") +
  geom_vline(data=hist_data1, aes(xintercept=mean(avg),
        colour="red"), linetype="dashed", size=1) +
  annotate("text", label = "Mean = 4.98", x = 6.2, y = 300, size = 4, color="red")

plot2 <-
ggplot(hist_data2, aes(x=exp_data)) +
  geom_histogram(binwidth=2, colour="black", fill="white") +
  labs(title = "Histogramm exponetial dist. (n = 1000)",
        x = "Number", y = "Count") +
  geom_vline(data=hist_data2, aes(xintercept=mean(exp_data),
        colour="red"), linetype="dashed", size=1) +
  annotate("text", label = "Mean = 5", x = 10, y = 300, size = 4, color="red") +
  xlim(0,30)
```

```
## qq-plot
plot3 <- ggplot(hist_data1, aes(sample=avg)) +
                stat_qq() +
                labs(title="Quantile-quantile plot of sample means")
```

```
## Creates a dataframe containing both data sets.
a <- rexp(1000, rate=.2)
b <- avg

a <- cbind(a,0)
b <- cbind(b,1)
d <- rbind(a,b)
d <- as.data.frame(d)
names(d) <- c("Number","exp")
```

```r
d$exp <- as.factor(d$exp)

# Calculates the mean of both data sets
e <- summarise(group_by(d,exp), means = mean(Number))

# Density curves
plot4 <-
ggplot(d, aes(d$Number, fill=d$exp)) +
  geom_density(alpha=.3) +
  scale_fill_manual(values=c("#009966", "#CC0000"),
        name= "Distribution",
        breaks=c(0, 1),
        labels=c("Exponential", "Mean of sample means")) +
  annotate("text", label = "Mean", x = 8, y = .45, size = 4, color="black") +
  labs(title = "Density plot", x = "Number", y = "Count") +
  geom_vline(data=e, aes(xintercept=e$means,  colour=e$exp), linetype="dashed", size=1) +
  xlim(0,30)
```