Sampling and Distribution

Ran Ju

Overview

We will estimate the distribution of sample mean and variance of exponential random variables, and compare them to normal distribution. To be able to reproduce the result, the random seed will be set to 344344 and the random variables $\sim Exp(0.2)$.

Simulation

We use sample size $=40$, and run the simulation 1000 times. Here is the code for doing that, and there is also a histogram of all samples combined. Note the scaled exponential distribution PDF and the histogram line match very closely.

Sample Mean

Here is the sample mean histogram. The theoratical mean is computed as $\frac{1}{\lambda}$ where $\lambda = 0.2$.

Sample Variance

Here is the sample variance histogram. The theoratical mean is equal to the mean of each random variable, i.e., $(\frac{1}{\lambda})^2 = 25$. Note this is not the variance of the sample mean.

$\text{var}(\text{sample mean}) = \text{sample variance}/\text{sample size} = 25/40 = 0.625$

Sample Mean Distribution

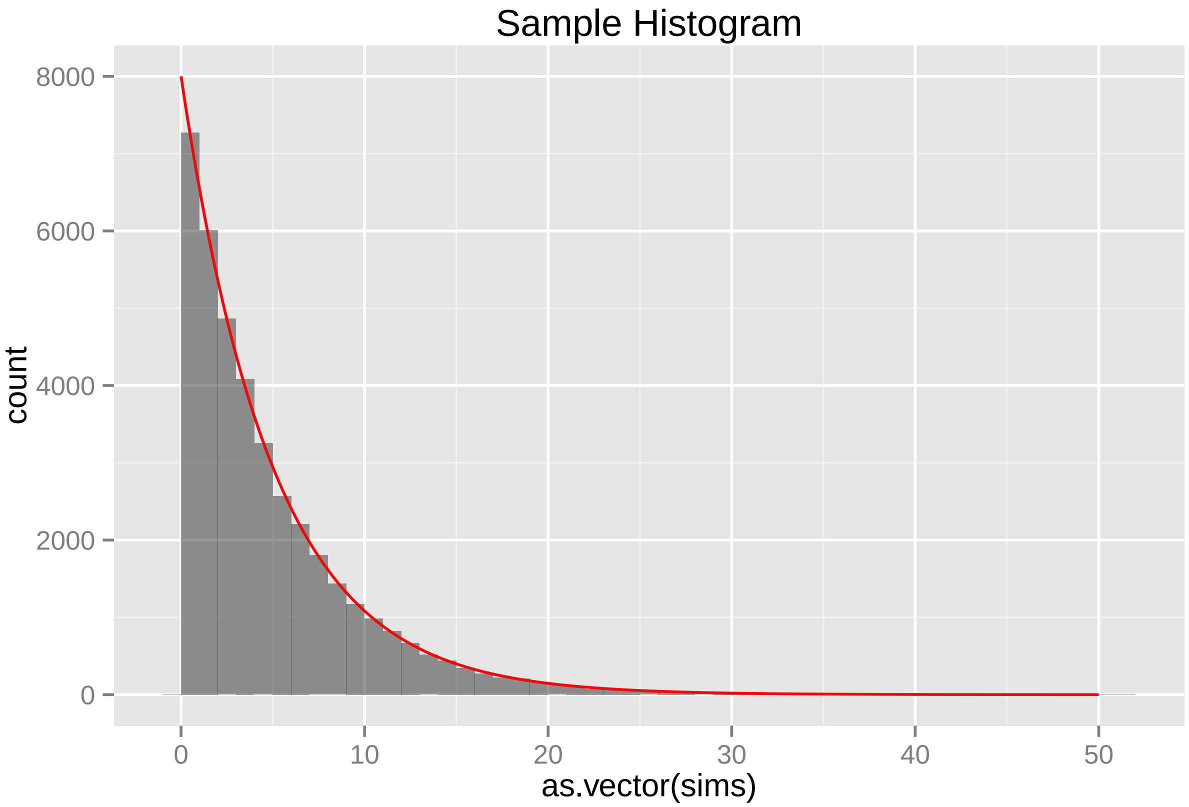
We plot the sample mean again, but overlay the scaled normal distribution $\mathcal{N}(\text{sample mean}, \text{var}(\text{sample mean}))$. Here shows the plot. Clearly the scaled normal distribution (the blue line) and the histogram match up, hence showing the sample mean distribution is approximately normal.

Appendix

Simulation

```r
library(ggplot2)
# Set the random seed
set.seed(344344)
# Simulation
sims <- NULL
for (i in 1:1000) sims <- rbind(sims, rexp(40, 0.2))
```

```
# Store the sample mean and variance into a data frame
mv <- data.frame(mean = apply(sims, 1, mean), var = apply(sims, 1, var))
# Plot histogram and exponential distribution line (scaled by the sample size)
ggplot() +
  geom_histogram(aes(x = as.vector(sims)), binwidth = 1, alpha = 0.5) +
  geom_line(aes(x = seq(0, 50, 0.1), y = 0.2 * exp(-0.2 * seq(0, 50, 0.1)) * 40000), color = "#FF0000")
  ggtitle("Sample Histogram")
```



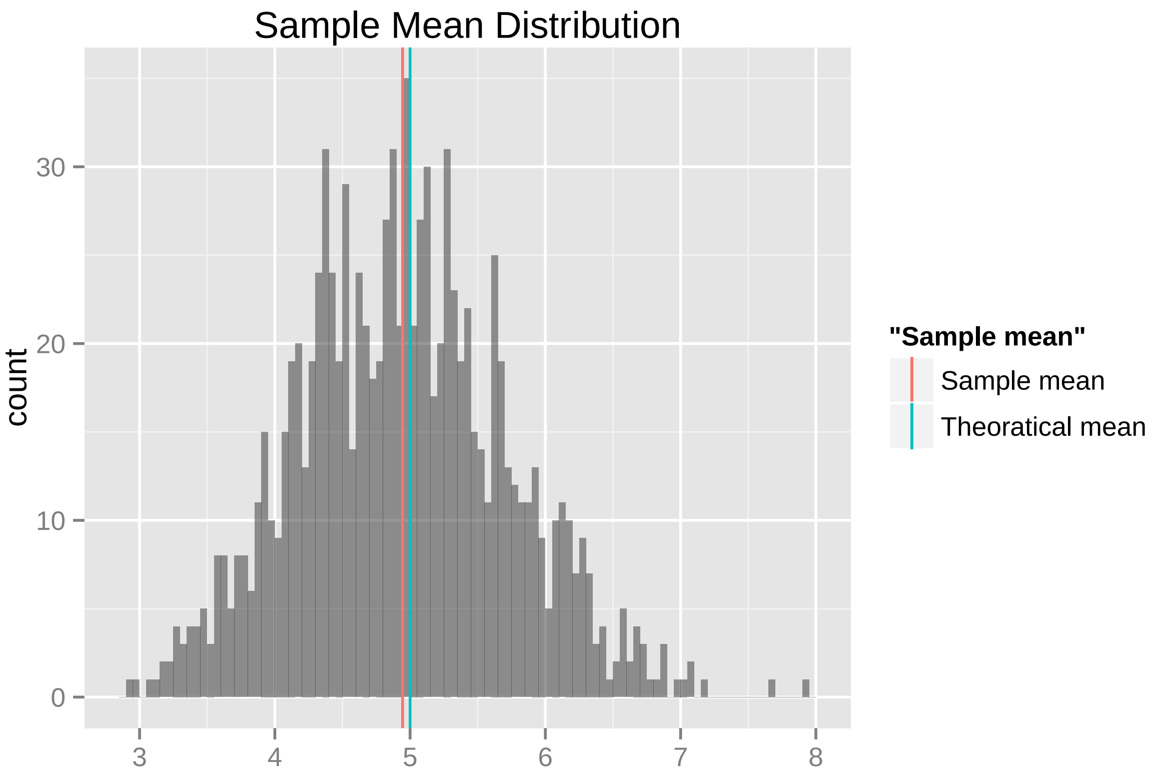
Sample Mean Histogram

```
# Compute the mean of the sample means
mean(mv$mean)
```

```
## [1] 4.944462
```

```
# Plot the histogram, sample mean and theoratical mean
ggplot() +
  geom_histogram(aes(x = mv$mean), alpha = 0.5, binwidth = 0.05) +
  geom_vline(aes(xintercept = mean(mv$mean), color = "Sample mean"), show_guide = T) +
  geom_vline(aes(xintercept = 1 / 0.2, color = "Theoratical mean"), show_guide = T) +
  xlab("") +
  ggtitle("Sample Mean Distribution")
```

```
## Warning in loop_apply(n, do.ply): position_stack requires constant width:
## output may be incorrect
```

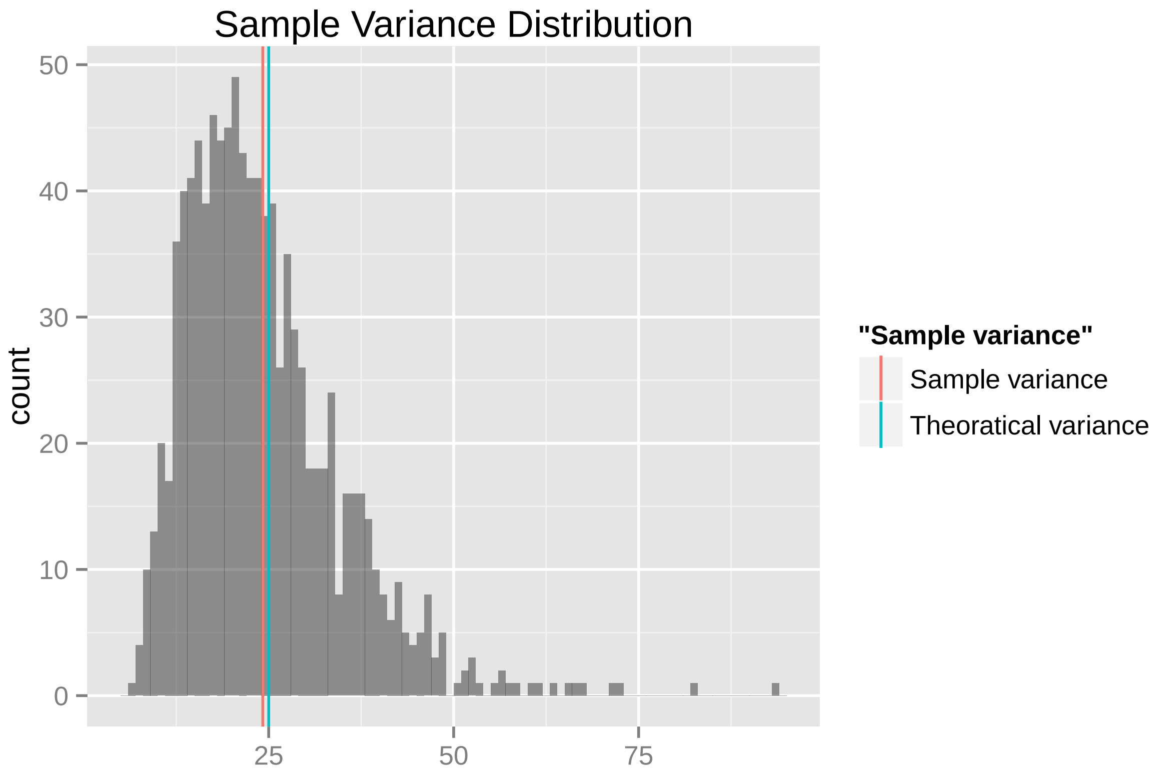


Sample Variance Histogram

```
# Compute the mean of the sample means
mean(mv$var)
```

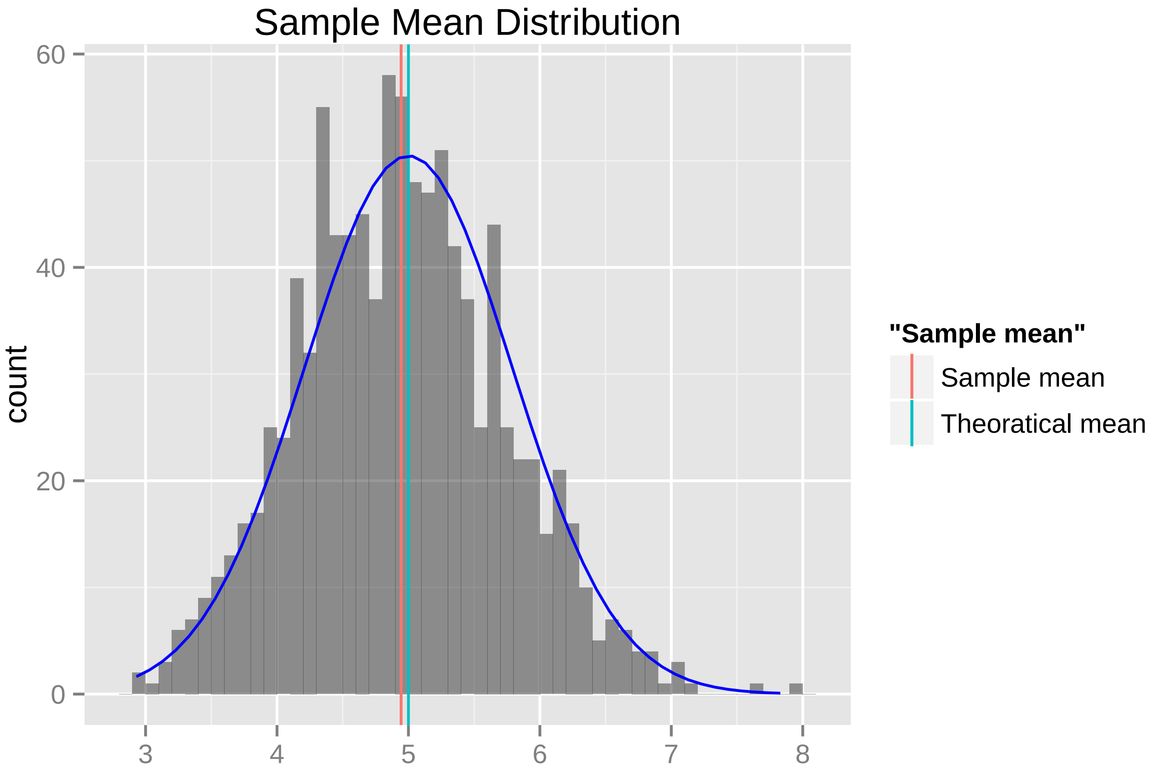
```
## [1] 24.19843
```

```
# Plot the histogram, sample mean and theoratical mean
ggplot() +
  geom_histogram(aes(x = mv$var), alpha = 0.5, binwidth = 1) +
  geom_vline(aes(xintercept = mean(mv$var), color = "Sample variance"), show_guide = T) +
  geom_vline(aes(xintercept = (1 / 0.2) ^ 2, color = "Theoratical variance"), show_guide = T) +
  xlab("") +
  ggtitle("Sample Variance Distribution")
```



Sample Mean Histogram

```
# Create x values for plotting normal PDF
x <- seq(min(mv$mean), max(mv$mean), by = 0.1)
# Theoratical mean
m <- 1 / 0.2
# Theoratical standard deviation
sd <- sqrt((1/ 0.2) ^ 2 / 40)
# Calculate normal PDF, scaled by the number of samples
y <- dnorm(x, mean = m, sd = sd) * 1000 * 0.1
# Plot the histogram, sample mean and theoratical mean
ggplot() +
  geom_histogram(aes(x = mv$mean), alpha = 0.5, binwidth = 0.1) +
  geom_vline(aes(xintercept = mean(mv$mean), color = "Sample mean"), show_guide = T) +
  geom_vline(aes(xintercept = 1 / 0.2, color = "Theoratical mean"), show_guide = T) +
  geom_line(aes(x = x, y = y), color = "blue") +
  xlab("") +
  ggtitle("Sample Mean Distribution")
```


Sample Mean Distribution
count
60
40
20
0
3
4
5
6
7
8
"Sample mean"
Sample mean
Theoratical mean