Gauss-Markov Theorem

Maria Barouti

1/27/2020

Based on `Applied Statistics with R (appliedstats)` by David Dalpiaz (https://github.com/daviddalpiaz/appliedstats)

To verify the results from Gauss-theorem, we will simulate samples of size $n = 100$ from the model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

with $\beta_0 = 3, \beta_1 = 6$, and $\sigma^2 = 4$.

The choice of X_i values is arbitrary. Here we also set a seed for randomization, and calculate s_x.

```
set.seed(42)
sample_size = 100 # this is n
X = seq(-1, 1, length = sample_size)
sx = sum((X - mean(X)) ^ 2)
beta_0 = 3
beta_1 = 6
sigma  = 2
```

The sampling distribution is

```
(var_beta_1_hat = sigma ^ 2 / sx)
```

```
## [1] 0.1176238
```

```
(var_beta_0_hat = sigma ^ 2 * (1 / sample_size + mean(X) ^ 2 / sx))
```

```
## [1] 0.04
```

We now simulate data from this model 10,000 times.

```
num_samples = 10000
beta_0_hats = rep(0, num_samples)
beta_1_hats = rep(0, num_samples)

for (i in 1:num_samples) {
  eps = rnorm(sample_size, mean = 0, sd = sigma)
  y   = beta_0 + beta_1 * X + eps

  sim_model = lm(y ~ X)

  beta_0_hats[i] = coef(sim_model)[1]
  beta_1_hats[i] = coef(sim_model)[2]
}
```

Each time we simulated the data, we obtained values of the estimated coefficiets. The variables `beta_0_hats` and `beta_1_hats` now store 10,000 simulated values of b_0 and b_1 respectively.

We first verify the distribution of b_1.

```
mean(beta_1_hats) # empirical mean
```

```
## [1] 6.001998
```

```r
beta_1 #true mean
```

```
## [1] 6
```

```r
var(beta_1_hats)  # empirical variance
```

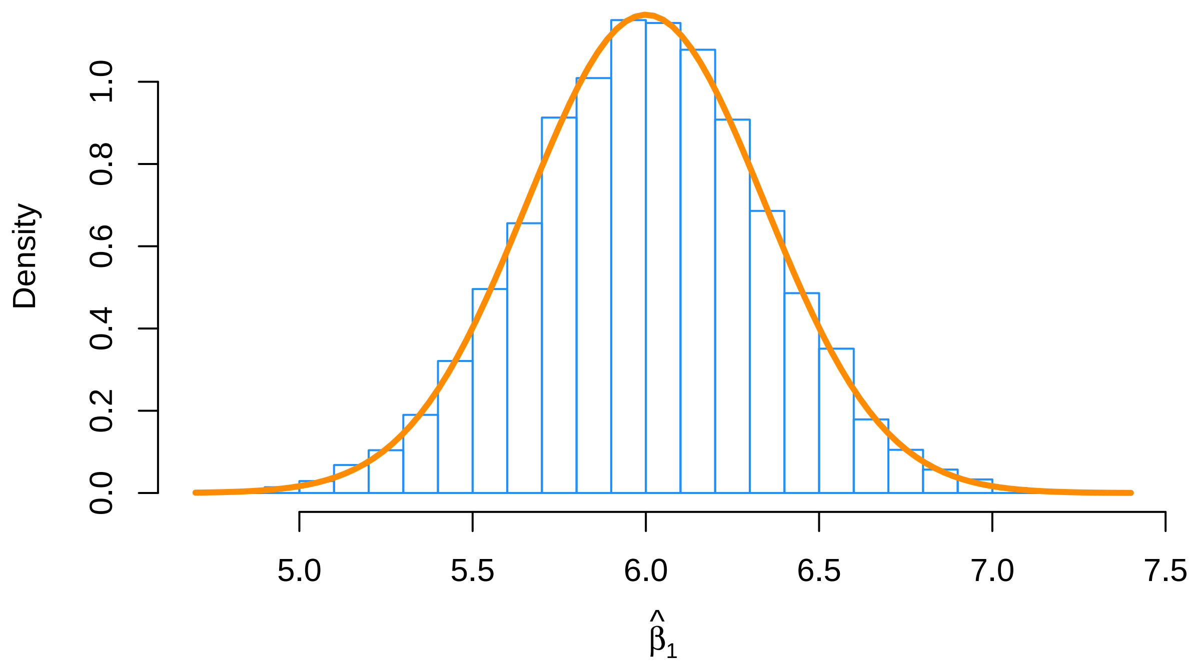
```
## [1] 0.11899
```

```r
var_beta_1_hat    # true variance
```

```
## [1] 0.1176238
```

We see that the empirical and true means and variances are very similar. We also verify that the empirical distribution is normal. To do so, we plot a histogram of the `beta_1_hats`, and add the curve for the true distribution of b_1. We use `prob = TRUE` to put the histogram on the same scale as the normal curve.

```r
# note need to use prob = TRUE
hist(beta_1_hats, prob = TRUE, breaks = 20,
     xlab = expression(hat(beta)[1]), main = "", border = "dodgerblue")
curve(dnorm(x, mean = beta_1, sd = sqrt(var_beta_1_hat)),
      col = "darkorange", add = TRUE, lwd = 3)
```



Similar for b_0.

We first verify the distribution of b_1.

```r
mean(beta_0_hats) # empirical mean
```

```
## [1] 3.001147
```

```r
beta_0 #true mean
```

```
## [1] 3
```

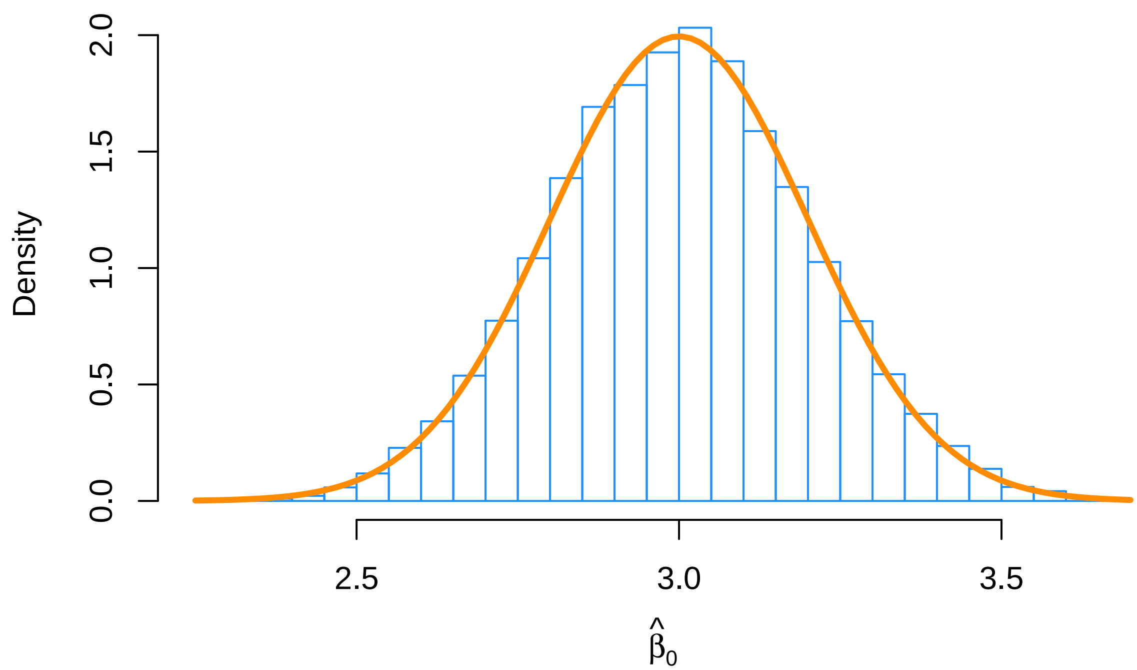
```r
var(beta_0_hats)  # empirical variance
```

```
## [1] 0.04017924
```

```r
var_beta_0_hat    # true variance
```

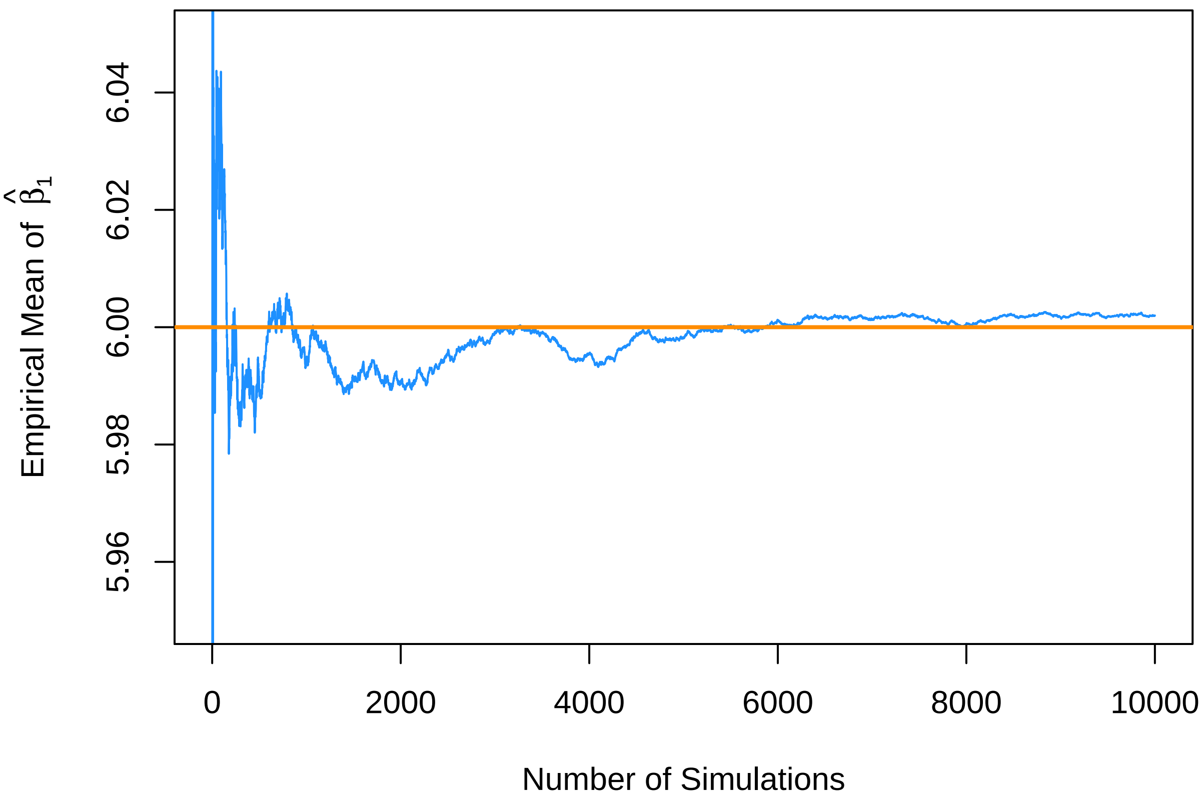
```
## [1] 0.04
```

```r
hist(beta_0_hats, prob = TRUE, breaks = 25,
     xlab = expression(hat(beta)[0]), main = "", border = "dodgerblue")
curve(dnorm(x, mean = beta_0, sd = sqrt(var_beta_0_hat)),
      col = "darkorange", add = TRUE, lwd = 3)
```

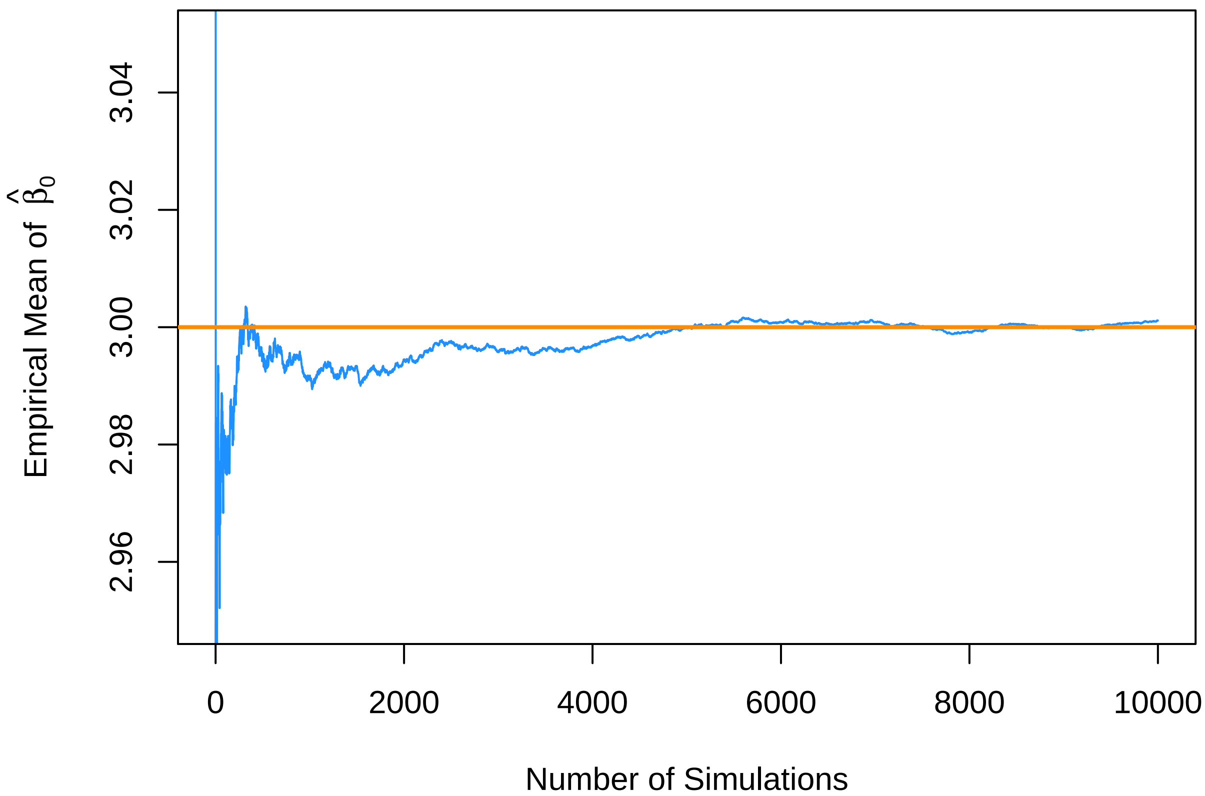


In this simulation study, we have only simulated a finite number of samples. To truly verify the distributional results, we would need to observe an infinite number of samples. However, the following plot should make it clear that if we continued simulating, the empirical results would get closer and closer to what we should expect.

```r
par(mar = c(5, 5, 1, 1)) # adjusted plot margins, otherwise the "hat" does not display
plot(cumsum(beta_1_hats) / (1:length(beta_1_hats)), type = "l", ylim = c(5.95, 6.05),
     xlab = "Number of Simulations",
     ylab = expression("Empirical Mean of " ~ hat(beta)[1]),
     col  = "dodgerblue")
abline(h = 6, col = "darkorange", lwd = 2)
```




```r
par(mar = c(5, 5, 1, 1)) # adjusted plot margins, otherwise the "hat" does not display
plot(cumsum(beta_0_hats) / (1:length(beta_0_hats)), type = "l", ylim = c(2.95, 3.05),
     xlab = "Number of Simulations",
     ylab = expression("Empirical Mean of " ~ hat(beta)[0]),
     col  = "dodgerblue")
abline(h = 3, col = "darkorange", lwd = 2)
```



Empirical Mean of $\hat{\beta}_0$
3.04
3.02
3.00
2.98
2.96
0
2000
4000
6000
8000
10000
Number of Simulations