Info 371 Lab 1

Variance of normal RV-s

Let's compute the sample variance of normal random variables by Monte Carlo (MC) simulations. With sample size $n = 1000$, let's generate n samples to work with.

```
n <- 1000
sample <- rnorm(n)
sample_mean <- mean(sample)
```

The mean of this sample is 0.0442689, which is unsurprisingly very close to 0. Next, let's explore different ways of finding variance of this sample.

```
sample_variance <- mean((sample - sample_mean)^2)
shortcut_variance <- mean((sample^2)) - mean(sample)^2
var_variance <- var(sample)
```

Following the formula **Var X = E[(X - E X)^2]**, we get **0.9320989**. Using the shortcut formula **Var X = E x^2 - (E X)^2**, we get the same value! Computing variance using the canned `var` function gives us **0.9330319**, which is slightly larger than the other variances we computed.

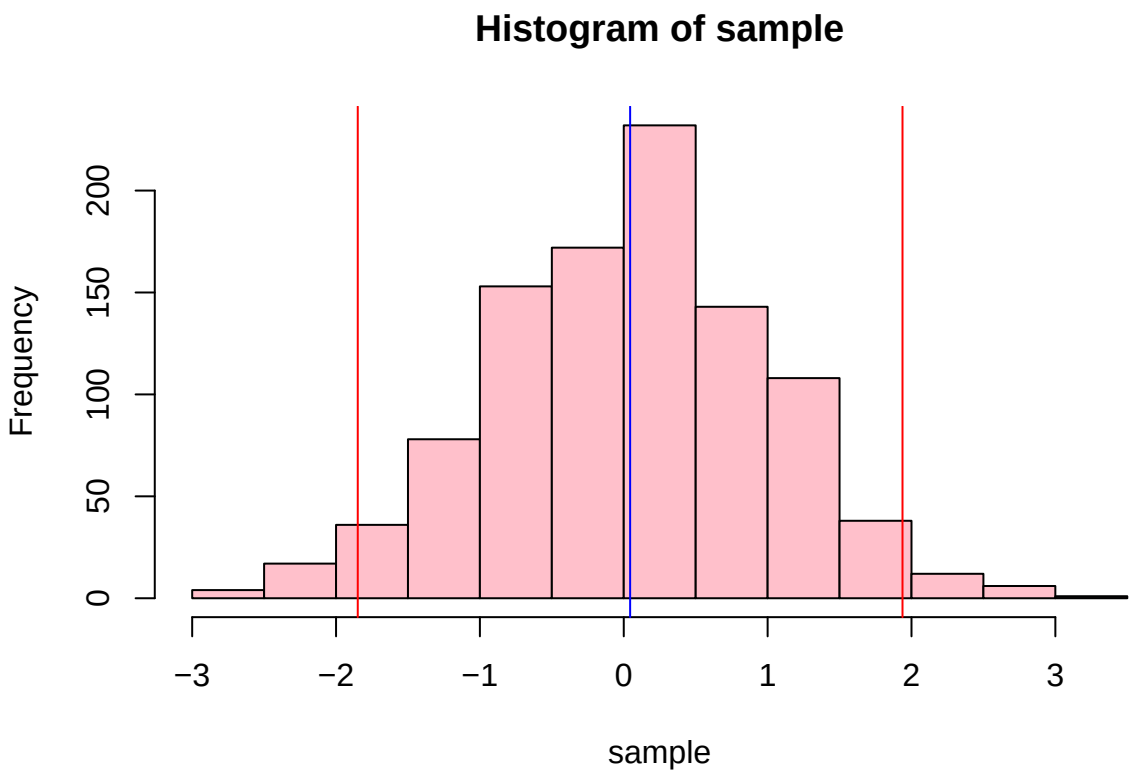
Now let's dive more into our sample, finding the standard deviation and confidence interval bounds.

```
# Finding standard deviation
sample_sd <- sd(sample)

# Finding confidence interval
lower <- sample_mean - 1.96 * sample_sd
upper <- sample_mean + 1.96 * sample_sd
conf_interval <- ((length(sample[sample > upper]) + length(sample[sample < lower])) / n)
```

Using the `sd` function, we get 0.9659358 as the standard deviation of the sample. To find the percentage of our sample that falls outside of the range of the 95% confidence interval, or **[X - 1.96 * sd X, X + 1.96 * sd X]**, we first find the lower and upper bounds of that interval. This happened to have a lower bound of -1.8489652 and upper bound of 1.9375031. The percentage of our values that fall outside of these bounds is about **5.2%**, which is unsurprisingly close to 5%.

We can plot a histogram our our sample, showing the mean and confidence interval. Here we see a binomial distribution of the sample numbers, with a minimal amount (5%) of the data falling outside of the 95% confidence interval. Most values are centered around the mean, while less trickle off and away. The mean of the distribution is nearly zero.



Variance of means

Now we move a step further and compute the variance of means: instead of picking n random numbers, now we pick n random numbers m time, and each calculate their mean. Afterwards, we'll see how the variance of means will change when we change n. For this example, we will have $n = 3$ and $m = 1000$.

```
n <- 3
m <- 1000

# Find means of m samples of n
means <- sapply(replicate(m, sapply(n, rnorm, simplify = FALSE)), mean)
variance <- var(means) # variance of means

# What happens if we make sample size larger?
new_n <- n * 100
new_means <- sapply(replicate(m, sapply(new_n, rnorm, simplify = FALSE)), mean)
new_variance <- var(new_means)
# The variance is much smaller!
```

The variance of our sample of means is **0.3574384**. When sample size increases, variance decreases. For example, when we change our sample size n to be 10x larger (n=300), our new variance is **0.0032451**!

How much smaller will the standard deviation be if the sample size is 10x larger? The first standard deviation of this sample comes out to be 0.5978615, and the new standard deviation when changing the sample size to be 10x larger is 0.178372. The new standard deviation appears to consistently be around 3x smaller than the first one. In this example, the new standard deviation is **3.35x** smaller!