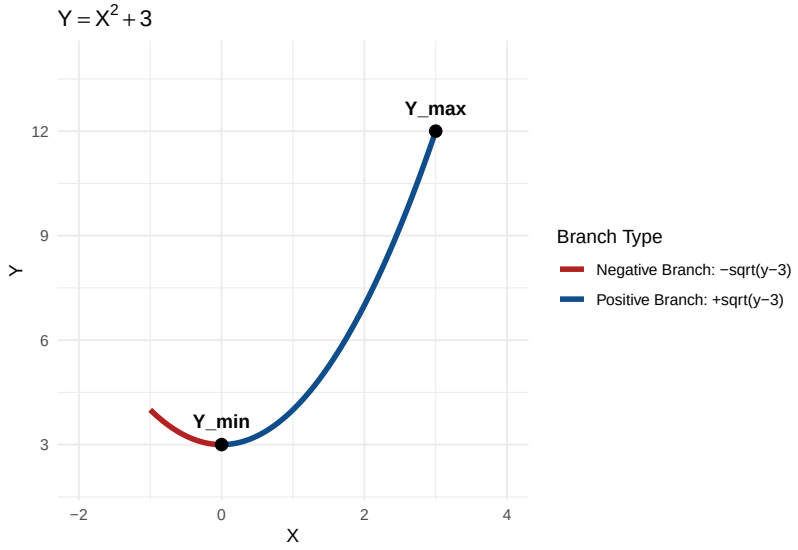
Homework 1

Matthew Pipes

Problem 1: Define $Y = X^2 + 3$ where $X \sim$ Unif$(-1, 3)$.

(a) Derive the cumulative distribution function of Y.



We first note a few features of the graph $Y = X^2 + 3$. We note that Y has a maximum at $Y_{max} = 12$ and a minimum at $Y_{min} = 3$. Thus, $Y \in [3, 12]$.

Furthermore, it is trivial to notice that $F_Y(y) = 0$ when $y < 3$. And similarly, we know that $F_Y(y) = 1$ when $y > 12$.

We now work toward investigating the behavior of the CDF in-between our end-points. We solve our transformation function $Y = X^2 + 3$ for X, which yields the following interval:

$$-\sqrt{y-3} \leq X \leq \sqrt{y-3}, \text{when } y \in [3, 4]$$

and

$$-1 \leq X \leq \sqrt{y-3}, \text{when } y \in (4, 12]$$

continued on next page...

We begin with the interval $y \in [3, 4]$. We solve

$$F_Y(y) = \frac{F_X(b) - F_X(a)}{b - a} = \frac{\sqrt{y-3} - (-)\sqrt{y-3}}{3 - (-1)} = \frac{2\sqrt{y-3}}{4} = \frac{\sqrt{y-3}}{2}$$

Similarly, for the interval $y \in (4, 12]$,

$$F_Y(y) = \frac{F_X(b) - F_X(a)}{b - a} = \frac{\sqrt{y-3} - (-1)}{3 - (-1)} = \frac{\sqrt{y-3}+1}{4}$$

Thus, our CDF is given by:

$$F_Y(y) = \begin{cases} 0 & y < 3 \\ \frac{\sqrt{y-3}}{2} & 3 \leq y \leq 4 \\ \frac{\sqrt{y-3}+1}{4} & 4 < y \leq 12 \\ 1 & y > 12 \end{cases}$$

Finally, it can be verified that $F_Y(y)$ is continuous from the left and the right at $x = 4$, thus $F_Y(y)$ is our valid CDF.

(b) Does Y follow a uniform distribution?

No. Y does not follow a uniform distribution, $F_Y(y)$ would need to be **linear** in y. Since our CDF contains square roots, it cannot be uniform.

(c) Compute $Var(Y)$

To solve for $Var(Y) = E(Y^2) + [E(Y)]^2$, we can use LOTUS

$$E[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

Thus,

$$E(Y) = \frac{1}{4} \int_{-1}^{3} (x^2 + 3) dx = \frac{16}{3}$$

and,

$$E(Y^2) = \frac{1}{4} \int_{-1}^{3} (x^4 + 6x + 9) dx = \frac{176}{5}$$

Finally,

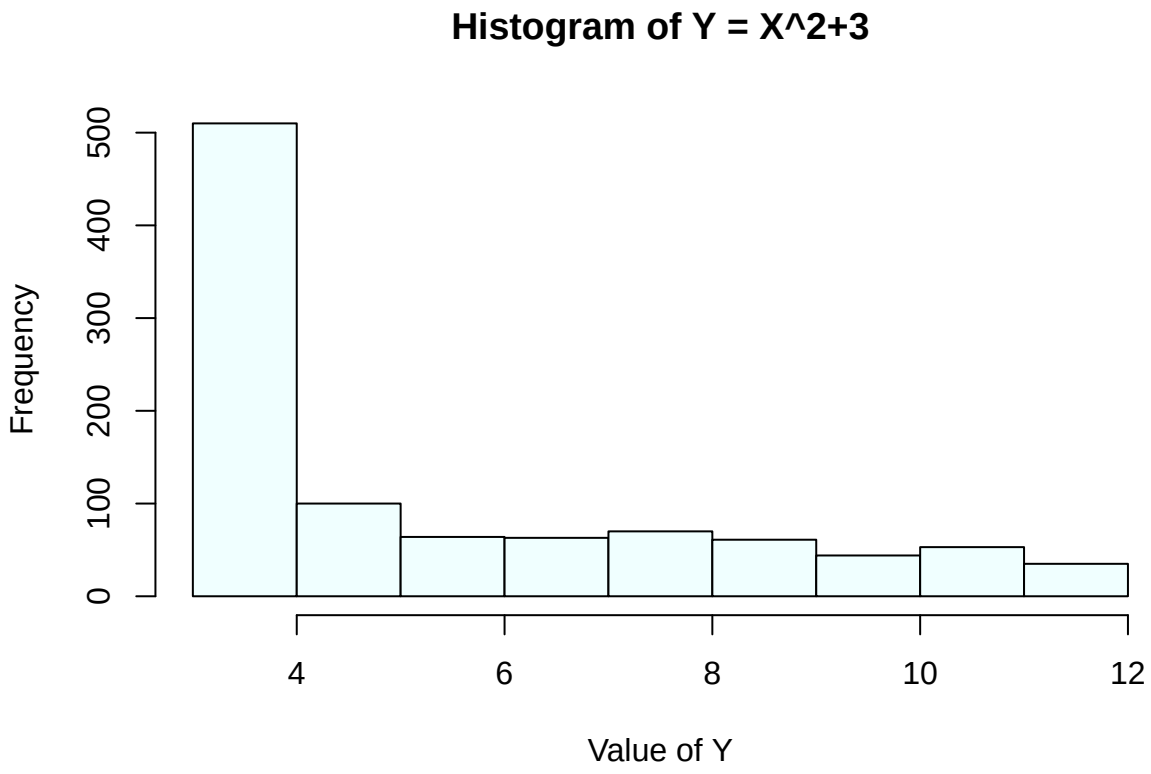
$$Var(Y) = E(Y^2) + [E(Y)]^2 = \frac{176}{5} - \frac{256}{9} = \frac{304}{45} \approx 6.756$$

```
computed_y_var <- 304/45
```

(d) Use R to generate a realization of 1000 independent copies of Y and create a histogram of the realizations.

```
set.seed(3301)
x_vals <- runif(n = 1000, min = -1, max = 3)
y_vals <- (x_vals^2)+3

hist(y_vals, main = "Histogram of Y = X^2+3", xlab = "Value of Y", col="azure")
```



(e) Using the data from (d), compute an estimate of var(Y). Is your estimate close to the actual value from (c)?

I'm assuming the significance of this question is to learn how to do for loops in R, rather than using the built-in R function for variance. So I'll do it that way, using the formula for the sample variance. Continued on next page...

We find the sample mean.

```
sum_of_y_vals <- 0
n <- length(y_vals)

for (i in 1:n) {
  sum_of_y_vals <- sum_of_y_vals + y_vals[i]
}

y_mean <- sum_of_y_vals / n
```

Compare y_mean to built in function mean

```
y_mean
```

```
## [1] 5.325343
```

```
mean(y_vals)
```

```
## [1] 5.325343
```

We find the sample squared difference.

```
sum_of_y_vals_sqd <- 0
for (i in 1:n) {
  sum_of_y_vals_sqd <- sum_of_y_vals_sqd + (y_vals[i] - y_mean)^2 }
```

Find the variance

```
y_variance <- sum_of_y_vals_sqd / (n - 1)
```

Compare the variances between built-in function and created function:

```
y_variance
```

```
## [1] 6.740301
```

```
var(y_vals)
```

```
## [1] 6.740301
```

Thus, the difference between our theoretical variance and measured variance is given by

```
percent_diff <- abs(computed_y_var - y_variance) / ((computed_y_var + y_variance) / 2) * 100
percent_diff
```

```
## [1] 0.226059
```

That is, 0.22%. They agree.

(f) If instead, I had generated a realization of 5000 independent copies of Y, would I expect my estimate from (d) to be closer to that of the true value of var(Y)? Explain why or why not.

I believe so. Fewer values of n means that each occurrence of an outlier can drag the variance away from the theoretical variance. I would expect that with n = 5000, we are closer to the theoretical variance.

(g) Design and perform an expirement in R to verify your claim from (f). Explain your experiment and report your results.

I will do the same thing as I did with the 1000 trials, but with n = 5000 this time. I will then compare the new variance against the old variance and interpret the results by percentage difference as before.

```
set.seed(3302)
x_vals2 <- runif(n=5000, min = -1, max = 3)
y_vals2 <- (x_vals2^2) + 3

y_variance2 <- var(y_vals2)

percent_diff2 <- abs(computed_y_var - y_variance2) / ((computed_y_var + y_variance2) / 2) * 100


percent_diff2
```

```
## [1] 0.9714171
```

That is, 0.97%

So we actually see that we have a higher variance this time. I can't really explain this. Both values seem close to the calculated variance, so perhaps it is just by chance that n=5000 had a larger variance.

Problem 2: A specialized laboratory is testing a new batch of micro-sensors designed for extreme temperatures. Let $X_1, ..., X_n$ represent the testing outcomes for n sensors randomly selected from the production line. In this context, $X_i = 1$ if the i-th sensor functions correctly under stress, and $X_i = 0$ if the i-th sensor fails (for each $i \in 1, ..., n$)

(a) What is the distribution of the random variable $\sum_{i=1}^{n} X_i$? Be as specific as possible; for instance, naming the distribution is insufficient without stating the parameters that define it. What is $E(\sum_{i=1}^{n} X_i)$ and $var(\sum_{i=1}^{n} X_i)$?

Since this experiment consists of n independent Bernoulli trials, each with a θ chance of success. Thus we can say,

$$Y \sim \text{Binom}(n, \theta)$$

with $E(Y) = n\theta$ and $Var(Y) = n\theta(1 - \theta)$.

(b) Perform a simulation study in R to confirm the expected value and variance you claimed in (a) when $n = 20$ and $\theta = 0.4$. To do so, independently repeat the experiment that generates a realization of $X_1, \ldots, X_n$ and the corresponding realization of $\sum_{i=1}^{n} X_i$ a total of reps $= 10000$ times. Report the observed sample mean of the reps realizations of $\sum_{i=1}^{n} X_i$ as a simulation-based estimate of $E(\sum_{i=1}^{n} X_i)$. Report the observed sample variance of the reps realizations of $\sum_{i=1}^{n} X_i$ as a simulation-based estimate of var$(\sum_{i=1}^{n} X_i)$. Are these simulation-based estimates close to their corresponding values from the formulas?

We should expect a sample average of $E(Y)$ of $n\theta = 20(0.4) = 8$

and sample variance $Var(Y) = n\theta(1 - \theta) = 20(0.4)(0.6) = 4.8$

Now, we run the experiment 10,000 times:

```
set.seed(3301)
sensor_trials <- rbinom(n = 10000, size = 20, prob = 0.4)
sensor_mean <- mean(sensor_trials)
sensor_var <- var(sensor_trials)

sensor_mean
```

```
## [1] 8.0045
```

```
sensor_var
```

```
## [1] 4.842364
```

Our computed values of the mean and variance are nearly identical to their theoretical values.

(c) Define $\overline{X} = n^{-1}\sum_{i=1}^{n} X_i$ as the (random) sample mean. Derive formulas for $E(\overline{X})$ and var$(\overline{X})$ in terms of θ and n.

Given $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, we apply the Expected Value operator (E) to both sides, yielding:

$$E(\overline{X}) = \frac{1}{n}E(\sum_{i=1}^{n} X_i) = \frac{1}{n}n\theta = \theta$$

We proceed similarly for variance, noting that $Var(aX) = a^2 Var(X)$

Given Given $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, we apply the Variance operator (Var) to both sides, yielding:

$$Var(\overline{X}) = \frac{1}{n^2}Var(\sum_{i=1}^{n} X_i) = \frac{1}{n^2}n\theta(1-\theta) = \frac{\theta(1-\theta)}{n}$$

(d) Perform a simulation study in R to confirm the formulas derived in part (c) when $n = 20$ and $\theta = 0.4$. To do this, independently repeat the experiment that generates a realization of $X_1, \ldots, X_n$ and the corresponding realization of $\overline{X}$ a total of reps $= 10000$ times. Report the observed sample mean of the realizations as a simulation-based estimate of $E(\overline{X})$. Report the observed sample variance of the realizations as a simulation-based estimate of var$(\overline{X})$. Are these simulation-based estimates close to their corresponding values from the formulas?

We expect a sample

$$E(\overline{X}) = \theta = 0.4$$

We expect a sample

$$Var(\overline{X}) = \frac{\theta(1-\theta)}{n} = \frac{0.4(0.6)}{20} = 0.012$$

Now we run the 10,000 experiments:

```
sensor_trials_xbar <- rbinom(n = 10000, size = 20, prob = 0.4) / 20
sensor_trials_mean <- mean(sensor_trials_xbar)
sensor_trials_var <- var(sensor_trials_xbar)

sensor_trials_mean
```

```
## [1] 0.40084
```

```
sensor_trials_var
```

```
## [1] 0.01211551
```

These values match essentially exactly to those that we calculated.

(e) Consider the random interval $[\overline{X} - \frac{1}{\sqrt{n}}, \overline{X} + \frac{1}{\sqrt{n}}]$. With what probability does this interval contain the true value of θ? To assess this question, perform a simulation study in R with $\theta = 0.4$. Independently repeat the experiment that generates a realization of $X_1, \ldots, X_n$ and the corresponding realization of the random interval a total of reps $= 10000$ times. Report the observed sample proportion of the reps realizations of this random interval that captured $\theta = 0.4$. Perform this simulation study for $n = 10, 30, 50, 100, 200$. Based on the simulation results, what is the probability that the random interval covers the true θ for these n values?

```
set.seed(3301)
n <- 0
repetitions <- 10000
theta <- 0.4
n_values <- c(10, 30, 50, 100, 200)
results <- numeric(length(n_values))

for (i in 1:length(n_values)) {
  n <- n_values[i]

  sensor_xbar <- rbinom(repetitions, size = n, prob = theta) / n

  lower_bound <- sensor_xbar - (1 / sqrt(n))
  upper_bound <- sensor_xbar + (1 / sqrt(n))

  in_bounds <- (theta >= lower_bound) & (theta <= upper_bound)

  results[i] <- mean(in_bounds)
}

data.frame(n = n_values, in_bounds_probability = results)
```

```
##     n in_bounds_probability
## 1  10                0.9813
## 2  30                0.9630
## 3  50                0.9699
## 4 100                0.9666
## 5 200                0.9636
```

Since this question doesn't ask me to interpret the data, just provide it, I won't give commentary. I'm fairly certain this is related to the confidence interval, that is, how sure we can be correct that a value we guessed is within a certain range, but I'm fuzzy on confidence intervals and have been working on this for over 12 hours, so I won't bother.