Statistical Analysis, MSCA 31007, Lecture 1

Hope Foster-Reyes

September 29, 2016

Explore the probabilities associated with the experiment of tossing a simulated fair coin multiple times.

1. Convergence of Probability of Tail to 0.5

a. We will check that the frequency of "Tails" (outcome equals 1) converges to 0.5 as the number of tosses grows. What does this say about the fairness of the coin?

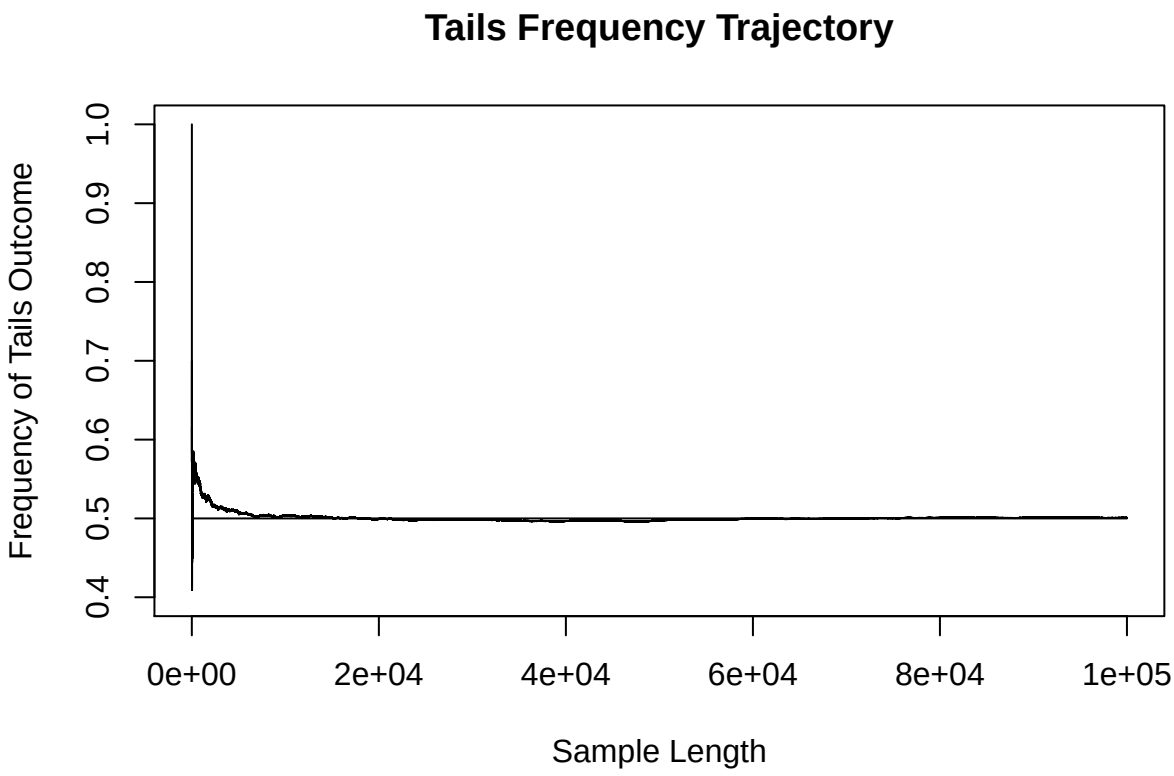
Through the axioms of probability and the Equally Likely Rule, we know that the probability of two equally likely mutually exclusive (disjoint) are equivalent and add to 1, therefore each have a probability of 0.5.

This experiment demonstrates this phenomenon, with our computer-generated pseudo-random simulation of a fair coin. In this case the two equally-likely events are a result of Heads and a result of Tails, each of whose probability is 0.5.

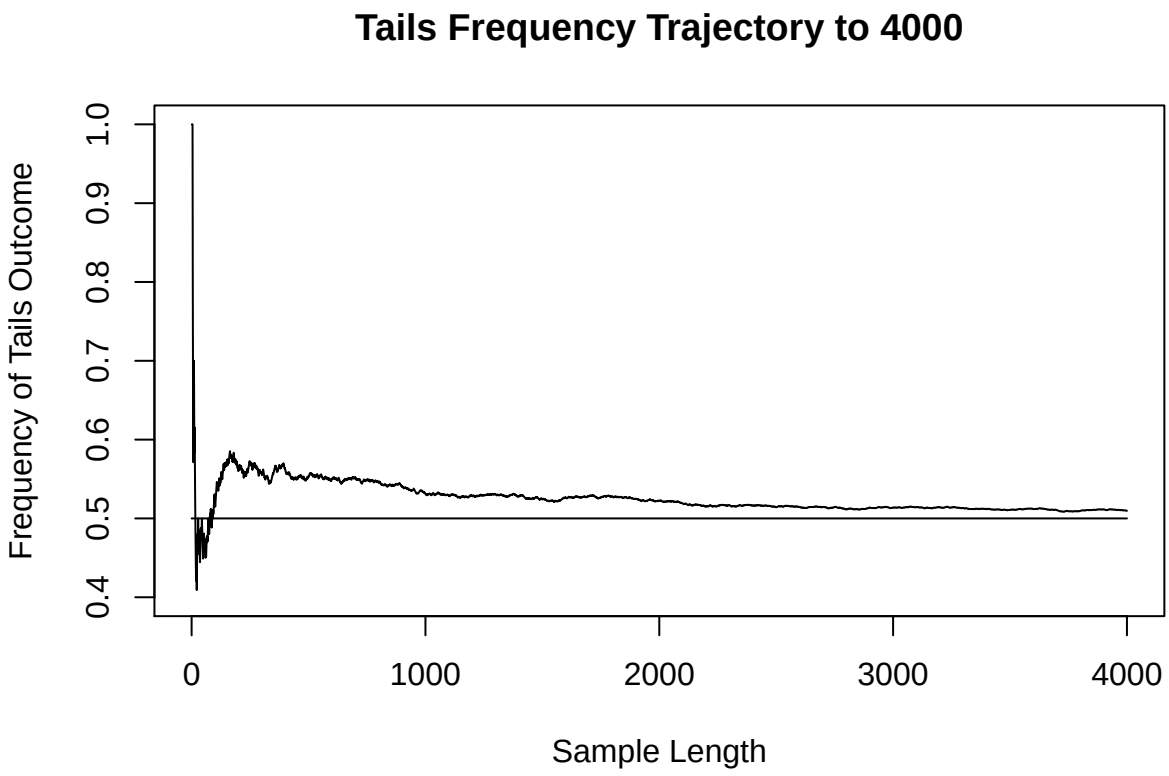
Per the definition of probability and randomness, we also know that in the long run the frequency of the outcome Tails in n empirical trials of this fair coin should converge at 0.5 as n gets larger. Thus we expect the empirical frequency to converge on the theoretical probability of 0.5.

The definition of a fair coin is one one in which both outcomes, Heads and Tails, are equallly likely. Hence by checking that the frequency of Tails converges on 0.5, we are also testing whether our computer-generated coin is fair.

```
# Seed for reproducibility
set.seed(12345)

num.flips <- 100000
#num.flips <- 200000

# Create a random sample of num.flips coin flips, represented by 0 = Heads and 1 = Tails
# Store this sample in the vector flips
flips <- sample(0:1, num.flips, replace = T)

# Create a vector of length num.flips containing the cumulative sum of the flips,
#   incrementing by 1 for each outcome of Tails
trajectory <- cumsum(flips)

# Create a vector of length num.flips containing the running frequency of tails,
#   with each entry representing the calculated frequency after the nth flip
freq.tails <- trajectory / (1:num.flips)

# Create a line graph of all data in the tails frequency vector
plot(1:length(freq.tails), freq.tails, ylim = c(0.4, 1), type = "l",
     ylab = "Frequency of Tails Outcome", xlab = "Sample Length",
     main = "Tails Frequency Trajectory")
lines(c(0, num.flips), c(0.5, 0.5))
```




```
# Create a line graph honing in on the first 4000 entries in the tails frequency vector
plot(1:4000, freq.tails[1:4000], ylim = c(0.4, 1), type = "l",
     ylab = "Frequency of Tails Outcome", xlab = "Sample Length",
     main = "Tails Frequency Trajectory to 4000")
lines(c(0,4000), c(0.5, 0.5))
```



b. Interpret what you see on the graphs.

What we see in the graphs is a behavior in which the trajectory of frequencies jumps up and down erratically as the number of trials is small. The first toss is Tails, so we see a big jump early on.

```
(head(freq.tails, 20))
```

```
##  [1] 1.0000000 1.0000000 1.0000000 1.0000000 0.8000000 0.6666667 0.5714286
##  [8] 0.6250000 0.6666667 0.7000000 0.6363636 0.5833333 0.6153846 0.5714286
## [15] 0.5333333 0.5000000 0.4705882 0.4444444 0.4210526 0.4500000
```

```
(max(freq.tails[0:6]))
```

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

Then the trajectory demonstrates a slightly larger tendency toward Heads, later again a more pronounced tendency toward Tails.

```
(max(freq.tails[20:1000]))
```

```
## [1] 0.5853659
```

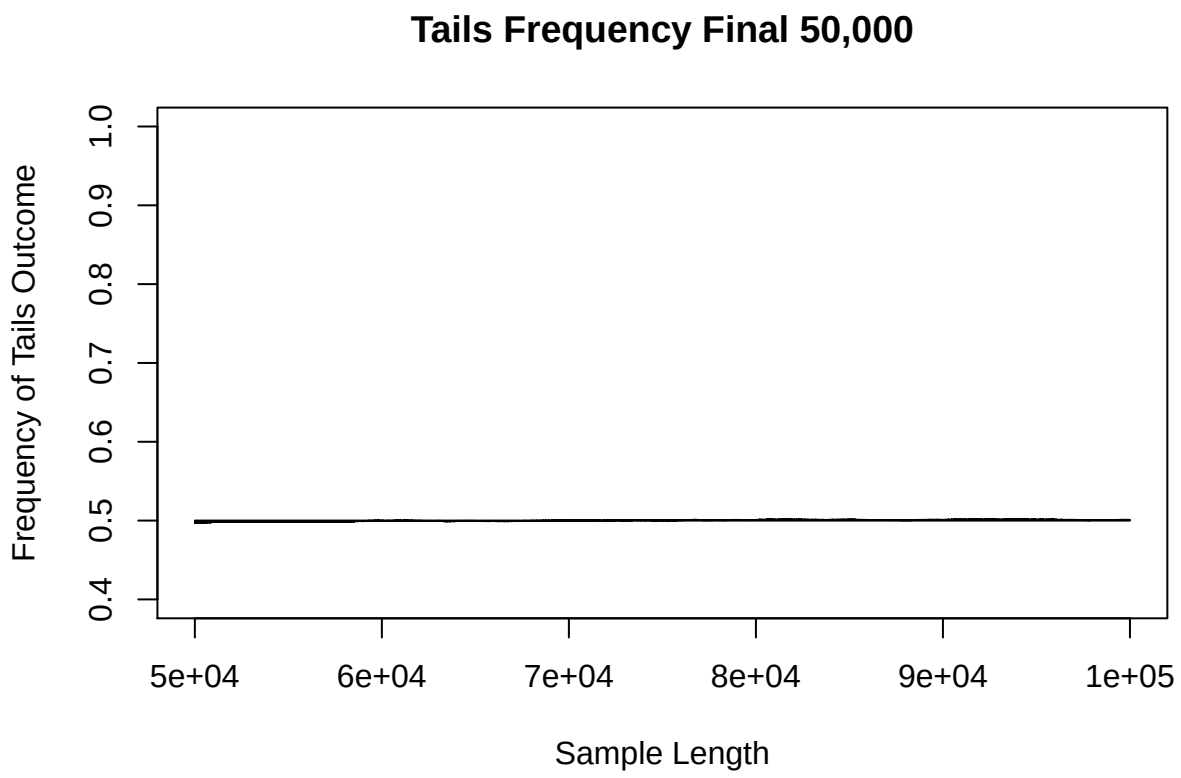
```r
(min(freq.tails[20:1000]))
```

```
## [1] 0.4090909
```

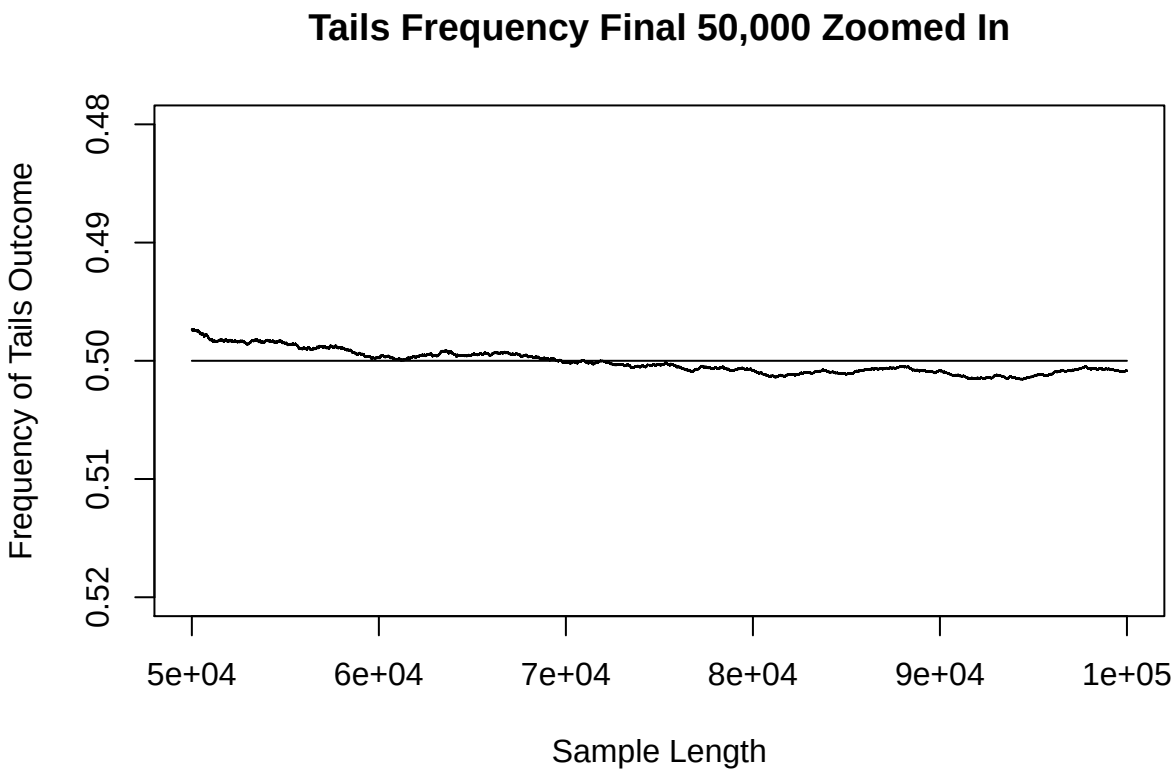
All of which simply demonstrates that our random experiment is unpredictable in the short term. While it may be surprising that the trajectory has such an extended tendency toward tails (lying above 0.5 for most of our experiment), as the trajectory becomes longer and longer (n becomes larger and larger), we can see the frequency of flips with the outcome Tails moving closer and closer to 0.5, the theoretical probability of getting Tails.

Let's take a closer look at the final 5000 entries:

```r
# Create a line graph honing in on the next 50000 entries in the tails frequency vector
plot(50000:100000, freq.tails[50000:100000], ylim = c(0.4, 1), type = "l",
     ylab = "Frequency of Tails Outcome", xlab = "Sample Length",
     main = "Tails Frequency Final 50,000")
lines(c(50000, 100000), c(0.5, 0.5))
```




```r
# And Zoomed In
plot(50000:100000, freq.tails[50000:100000], ylim = c(0.52, 0.48), type = "l",
     ylab = "Frequency of Tails Outcome", xlab = "Sample Length",
     main = "Tails Frequency Final 50,000 Zoomed In")
lines(c(50000, 100000), c(0.5, 0.5))
```



When we change the scale, we can see that our experiment still varies and is varying both above and below the 0.5 line. But at the original scale the frequency is nearly indistinguishable from a straight line at the 0.5 frequency mark. At the 100,000th flip, the tail frequency is 0.50084.

2. Check Your Intuition About Random Walks

2.1 One Trajectory

```
# Seed for reproducibility
set.seed(12345)

# Increase the number of flips
num.flips <- 1000000

# Create a random sample of num.flips coin flips to simulate a gambling game, where
#   Heads loses $1 and Tails pays $1
flips.wealth <- (sample(0:1, num.flips, replace = T) - 0.5) * 2
(table(flips.wealth))
```

```
## flips.wealth
##     -1      1
## 499933 500067
```

a. Find at least one alternative way of simulating variable Flips (in my code, flips.wealth).

```
# This is a transformation of ~ Binom(1, 0.5), so we can also produce the
#   sample with rbinom()
set.seed(12345)
binom.wealth <- rbinom(num.flips, 1, 0.5)
flips.wealth <- (binom.wealth - 0.5) * 2
(table(flips.wealth))
```

```
## flips.wealth
##     -1      1
## 499933 500067
```

```
# We can also simply hard code our options
set.seed(12345)
flips.wealth <- sample(c(-1, 1), size = num.flips, replace = T)
(table(flips.wealth))
```

```
## flips.wealth
##     -1      1
## 499933 500067
```

b. Check your intution by answering questions before calculation:

How much do you expect the trajectory of wealth to deviate from zero?

Intuitively we can expect the trajectory of wealth to deviate from zero as much as $2-$3, as we imagine that the coin could perhaps immediately land on Tails 2-3 times.

What is the probability that this intuitive guess is underestimating, and that instead the coin will land on Tails 4 or 5 times in a row in its initial flips?

??? XXX ???

How long do you expect it to stay on one side above or below zero?

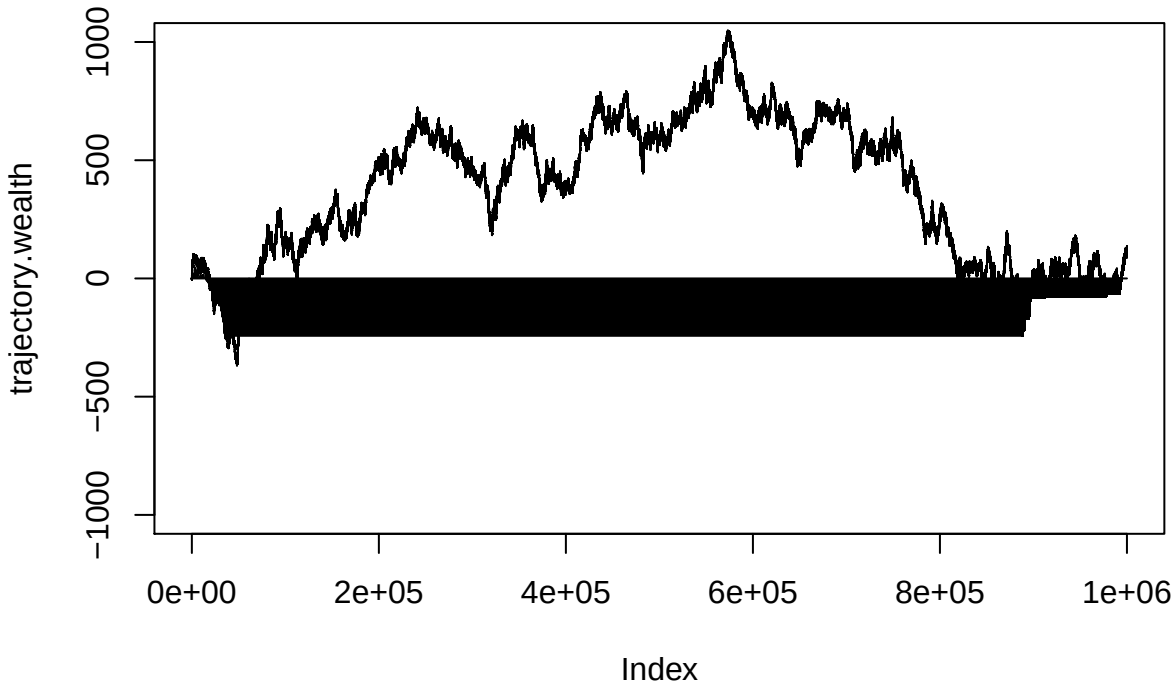
Intuitively have a sense that the trajectory will erraticaly jump above and below zero. Since remaining above or below the line would represent consecutive flips of Heads or Tails, we estimate that, similar to our above guess, that the wealth of our hypothetical gambler would not often remain consecutively negative or positive longer than 3-4 flips.

c. How do the observations match our prior expectations?

We run the experiment:

```
# Create our random walk, a vector of length num.flips containing the cumulative sum of the flips,
#   incrementing by 1 for each outcome of Tails, and -1 for each outcome of Tails.
#   This is a one-dimensional random walk, moving +1 or -1 with equal probability.
trajectory.wealth <- cumsum(flips.wealth)

# Create a line graph of our wealth trajectory. In this case rather than plotting the frequency
#   of Tails we are plotting the position, in terms of wealth, of our gambler, as she
#   'walks along randomly' in either the +1 or -1 direction.
```

```
plot(trajectory.wealth, ylim = c(-1000, 1000), type = "l")
lines(c(0, num.flips), c(0, 0))
```

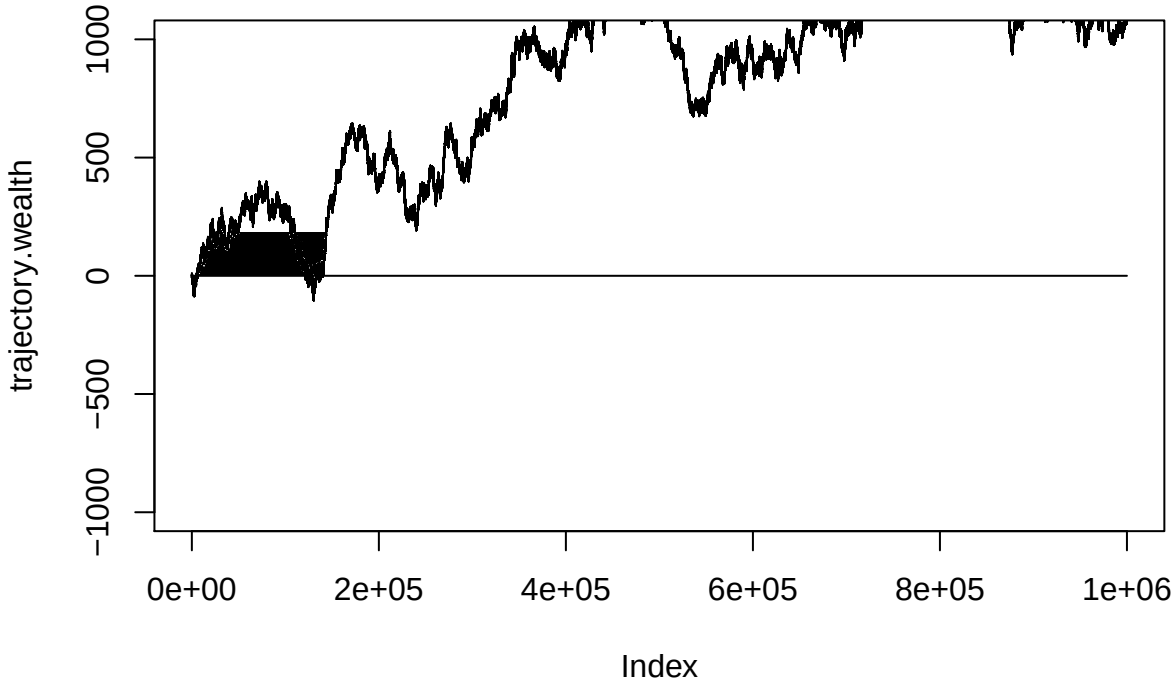


And our expectations are wildly inaccurate!

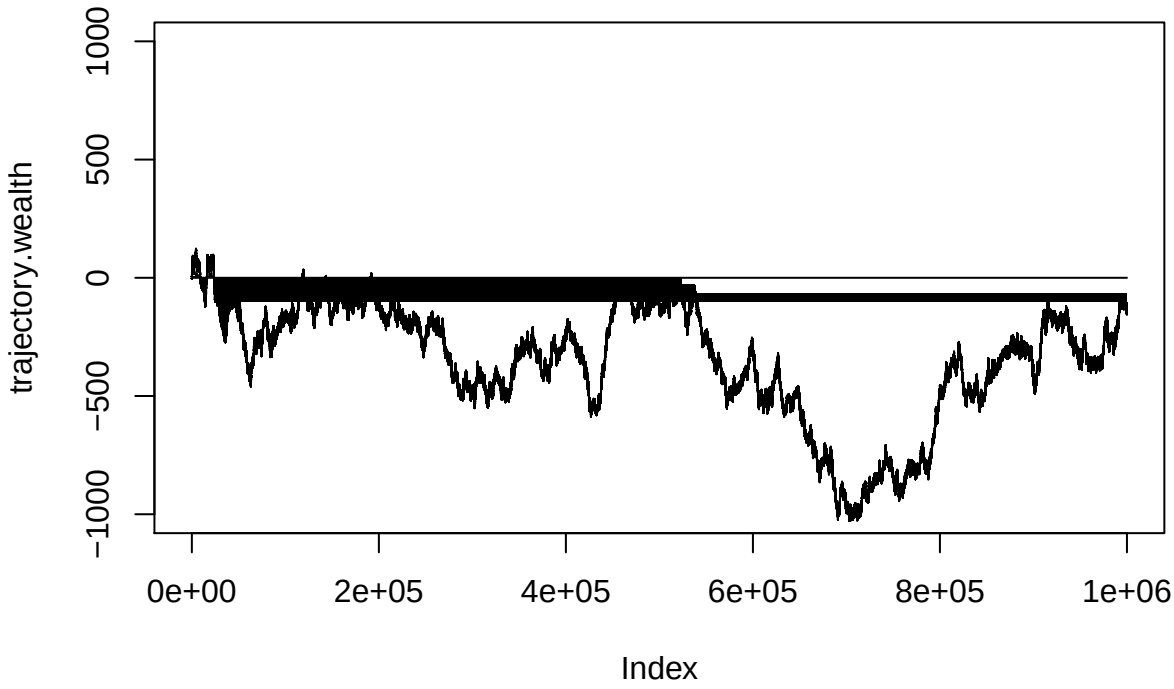
While we did consider the probability the coin would land on 4-5 consecutive heads or tails, and that probability is small, what we did not consider is the probability that the experiment would take a general upward (or downward) trend over time, despite these constant fluctuations.

Running the experiment with alternate seeds:

```
set.seed(54321)
flips.wealth <- sample(c(-1, 1), size = num.flips, replace = T)
trajectory.wealth <- cumsum(flips.wealth)
plot(trajectory.wealth, ylim = c(-1000, 1000), type = "l")
lines(c(0, num.flips), c(0, 0))
```




```
set.seed(33333)
flips.wealth <- sample(c(-1, 1), size = num.flips, replace = T)
trajectory.wealth <- cumsum(flips.wealth)
plot(trajectory.wealth, ylim = c(-1000, 1000), type = "l")
lines(c(0, num.flips), c(0, 0))
```



Taking a closer look at the data, we can see that

??? XXX ???

2.2 Multiple Trajectories

In our next experiment we look at the probability of our trajectory ending a certain distance from zero.

a. What do you expect the probabilities of the following events to be?

If N_h ***is the number of "Heads" and*** N_t ***is the number of "Tails" in 500 coin flips, then what is:***

$P(|N_h - N_t| < 5)$?

??? XXX ???

$P(|N_h - N_t| > 5)$?

??? XXX ???

b. Estimate the Probabilities

Convert the flips.wealth sample of 1,000,000 coin flips into a matrix of 2000 random walk samples, each 500 long:

```r
# First convert the sample into a matrix with 500 columns
matrix.wealth <- matrix(flips.wealth, ncol = 500)
# Then apply cumsum over each row and transform, turning our sample into 2000 random walks
trajectories.matrix <- t(apply(matrix.wealth, MARGIN = 1, cumsum))

# Observe the dimensions of our matrix
dim(trajectories.matrix)
```

```
## [1] 2000  500
```

```r
# What proportion of the
```

How many times out of 2,000 runs:

Do trajectories end less than 5 points away from zero (5 is 1% of 500 tosses)?

??? XXX ???

Do trajectories end more than 25 points away from zero (25 is 5% of 500 tosses)?

??? XXX ???

Time On One Side

a. How long do you expect trajectory of random walk to spend on one side from zero, below or above?

??? XXX ???

b. Interpret the results. Was your intuition correct?

Use the