Weining Bao weiningb@usc.edu 8287723653

Project 1 Coin Tossing

A) Simulate tossing a coin 50 times and record
- the number of heads and
- the length of the longest run of heads
B) Simulate repeatedly tossing a coin (until the event specified occurs)
- the number of tosses until the first head, second head, third head, fourth head
- the number of tosses needed until there are 2, 3, and 4 heads in sequence
Make 100 runs (with different seeds) to find the distribution of the items recorded (random variables).

Deliverables for this assignment include (documented) source code for your program, the data from your experiments and a discussion of the results. The presentation should be: concise, attractive (use graphics when appropriate), understandable.

Analytic Theories

Underlying Sets in Probability Models

Set of all possible outcomes of some random / uncertain / non-deterministic experiment / trial is called the sample space, S. For example, for the one time coin flip in the simulation, S = {H,T}, where H means head and T tail. An event E is a set of outcomes. For example for the coin flip simulation, E might be $\varnothing$, {H}, {T} or {H,T}. So $E \subset S$.

Assignment of Probability to Different Events

Relative frequency of events' occurrence can be assigned as their probability. For example, the simulation runs experiment N (100) times and record the number of times / frequency of occurrence, f_A, of the event of interest (event A) occurs, 50 tosses give 25 heads for example.

Then say: $P_A = \lim_{N\to\infty}\{\frac{f_A}{N}\}$. This approach is used in the simulation to compare the results from analytic calculation covered below.

Another way to assign is using probability as a measure to different events from one sample space. Thus a map P is defined from event E to [0, 1] that satisfies
1, $0 \le P(E) \le 1; P(S) = 1(certain); P(\varnothing) = 0(impossible)$;

2, For all sets of events that are mutually exclusive, i.e. $E_i \cap E_j = \varnothing$, $P\left(\bigcup_i E_i\right) = \sum_i P(E_i)$.

In many cases with a finite discrete valued state space (with N outcomes), we may consider the outcomes (or simple events) to be equally likely; then above implies that the probability of any particular outcome (simple event) is 1/N. For the coin toss, we have P({H}) = P({T}) = 0.5.

However, if we have a biased coin we could estimate the probability of H or T by using the relative frequency of occurrence in a long sequence of trials mentioned above which will lead to the right intuitive interpretation.

Random Variable as Mapping between Different Events and Values

A random variable (RV) is a mapping from the set of outcomes to numbers which can be either discrete or continuous, e.g. a random variable X can be defined with the outcomes of one time coin flip in the simulation. X = 1 if the toss gives head or X = 0 if tail.

Then we can define the Cumulative Distribution Function (CDF) for a RV, $F_X(x) = P(X \le x)$. For a discrete RV, we can also define the Probability Mass Function (PMF), $f_X(x) = P(X = x)$. For the X defined with the outcomes of one time coin flip,

$$F_X(x) = \begin{cases} 0, x < 0 \\ 1/2, 0 \le x < 1 \\ 1, x \ge 1 \end{cases}, f_X(x) = \begin{cases} 1/2, x = 0,1 \\ 0, elsewhere \end{cases}$$

Expectation and Variance of a Random Variable

Then expectation / mean and variance of a RV can be defined:

$$E[X] = \sum_i x_i f_X(x_i), VAR(X) = E[(X - E[X])^2] = E[X^2] - (E[X])^2 ;$$

The symbol μ_X is commonly used for E[X], also known as the mean. The standard deviation is the square root of the variance. The symbol σ_X is commonly used for the standard deviation of the RV X. So for the Random Variable associated with one time coin flip, the expectation E[X] = (1/2) * (1 + 0) = 1/2, VAR(X) = E[X^2] - (E[X])^2 = 1/2 *(1 + 0) - 1/4 = 1/4.

Simple Experiments in Coin Tossing Simulation

Question 1. If a coin is flipped 50 times, how many times does a head occur?
Question 2. If a coin is flipped 50 times. What is the length of the longest run of heads?
Question 3. A coin is flipped until k heads occur in sequence (for k = 1,2,3…). How many coin tosses are needed?
Question 4. The number of tosses until the first head, second head, third head, fourth head?

To answer question 1, we can use a random variable X to represent the number of heads. There are 2^50 permutations possibilities as outcome. Among them, there are $\begin{pmatrix} 50 \\ X \end{pmatrix}$ outcomes give X heads. The symbol means picking X items from 50 items. Here X tosses are picked to give heads while others give tails. So the PMF $f_X(x) = \begin{pmatrix} 50 \\ x \end{pmatrix} / 2^{50}, 0 \le x \le 50$. This is a binomial distribution. However we can use continuous Gaussian Distribution to estimate it with Central Limit Theorem.

Central Limit Theorem

$$\text{As } n \to \infty: \quad \frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\sigma} \quad \to \quad N(0,1)$$

Equivalently we can say, as $n \to \infty$:

$$\frac{\sum_{i=1}^{n} X_i}{n} \to N(\mu, \sigma^2 / n)$$

So if we treat the number of heads in 50 tosses as the sum of 50 RV X from one time coin toss, the mean of head number is 50 * E[X] = 25 and the variance is 50 * VAR(X) = 12.5, thus standard deviation $\sigma = \sqrt{12.5} \approx 3.54$.

Let us think about question 4 now. Still use a random variable X to represent the number of tosses before the first head. It takes X tosses to get first head only when first X - 1 tosses all gave tails. So $f_X(x) = \frac{1}{2^x}, x = 1,2,3....$ The expectation is 2. As we can see $\sum_{i=1,2...} f_X(i) = 1$ satisfies the requirement of probability. Similarly, random variable Y to represent the number of tosses before the second head has PMF $f_Y(y) = \frac{y-1}{2^y}, y = 2,3....$ The expectation is 4.

Matlab Implementation

```
run = 100; % number of runs
toss = 50; % number of tosses per run
heads = zeros(run, 1); % number of heads per run
longests = zeros(run, 1); % length of the longest run of heads per run
H1 = zeros(run, 1); % number of tosses until first head per run
H2 = zeros(run, 1); % number of tosses until second head per run
H3 = zeros(run, 1); % number of tosses until third head per run
H4 = zeros(run, 1); % number of tosses until fourth head per run
S2 = zeros(run, 1); % number of tosses till two heads in sequence per run
S3 = zeros(run, 1); % number of tosses till three heads in sequence per run
S4 = zeros(run, 1); % number of tosses till four heads in sequence per run

for i = 1:run % run 100 times
    head = 0; % the number of heads per run
    longest = 0; % the longest sequence of heads per run
    serial = 0; % current length of head sequence this run
```

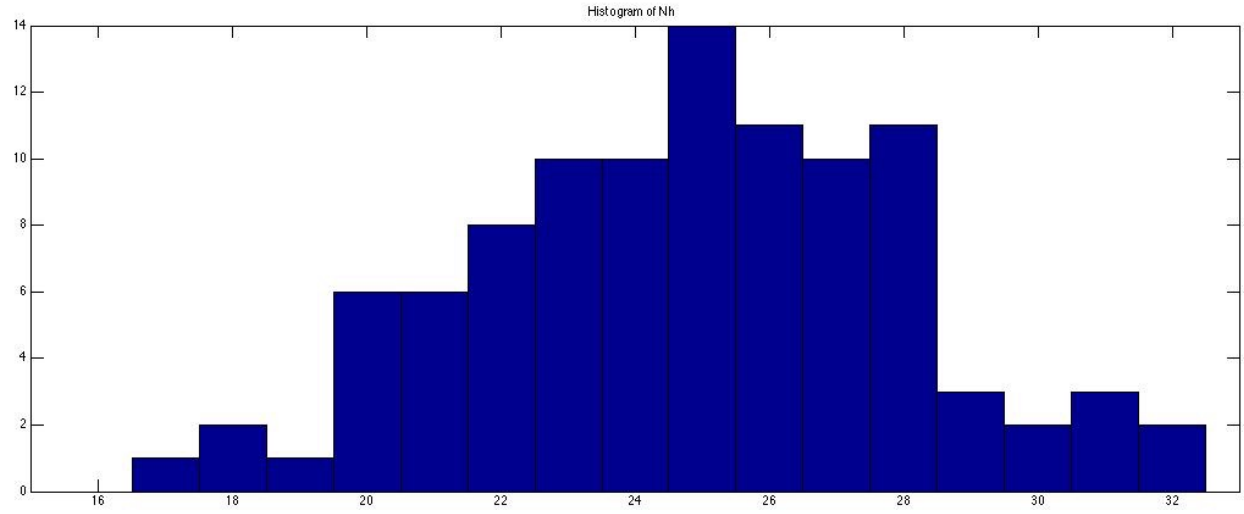
```
    for j = 1:toss % toss 50 times per run
        c = round(rand()); % random number 1 or 0, 1 as head, 0 as tail
        if (c == 1)
            head = head + 1;
            serial = serial + 1;
            if (serial == 2 && S2(i) == 0) % give two heads in sequence
                S2(i) = j;
            end
            if (serial == 3 && S3(i) == 0) % give three heads in sequence
                S3(i) = j;
            end
            if (serial == 4 && S4(i) == 0) % give four heads in sequence
                S4(i) = j;
            end
        else
            if (serial > longest) % update the longest sequence of heads
                longest = serial;
            end
            serial = 0; % start new sequence
        end
        if (head == 1) % give one head cumulatively
            H1(i) = j;
        end
        if (head == 2) % give two heads cumulatively
            H2(i) = j;
        end
        if (head == 3) % give three heads cumulatively
            H3(i) = j;
        end
        if (head == 4) % give four heads cumulatively
            H4(i) = j;
        end
    end
    if (serial > longest) % update the longest head sequence if at end
        longest = serial;
    end
    heads(i) = head;
    longests(i) = longest;
end

% plot the histogram of all data in 100 runs

figure;
hist(heads, 1:50);
title('Histogram of Nh');
figure;
hist(longests, 1:50);
title('Histogram of Lh');
figure;
```

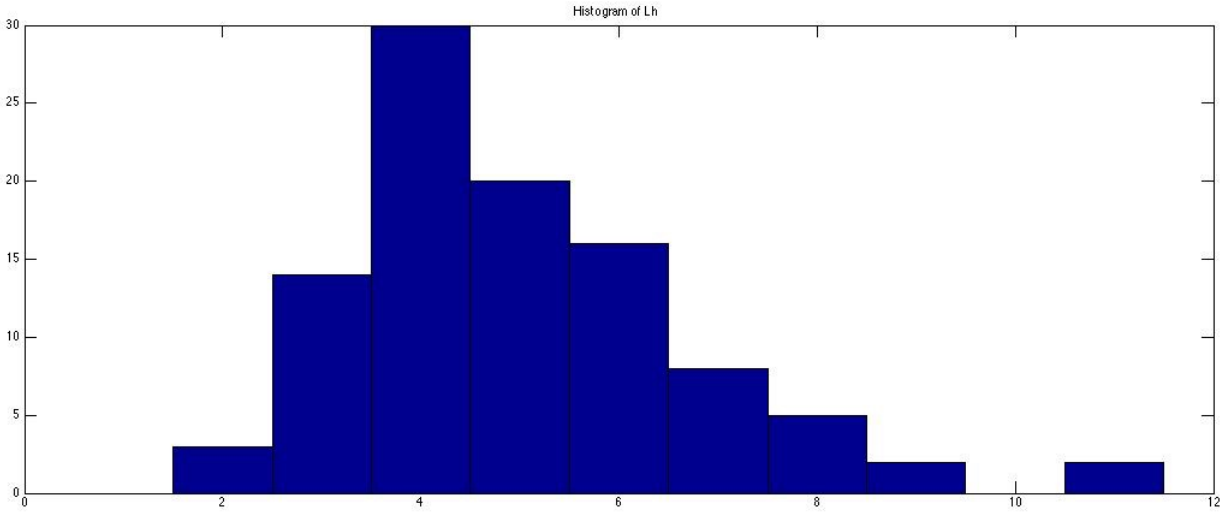
```
hist(H1, 1:50);
title('Histogram of H1');
figure;
hist(H2, 1:50);
title('Histogram of H2');
figure;
hist(H3, 1:50);
title('Histogram of H3');
figure;
hist(H4, 1:50);
title('Histogram of H4');
figure;
hist(S2, 1:50);
title('Histogram of S2');
figure;
hist(S3, 1:50);
title('Histogram of S3');
figure;
hist(S4, 1:50);
title('Histogram of S4');
```

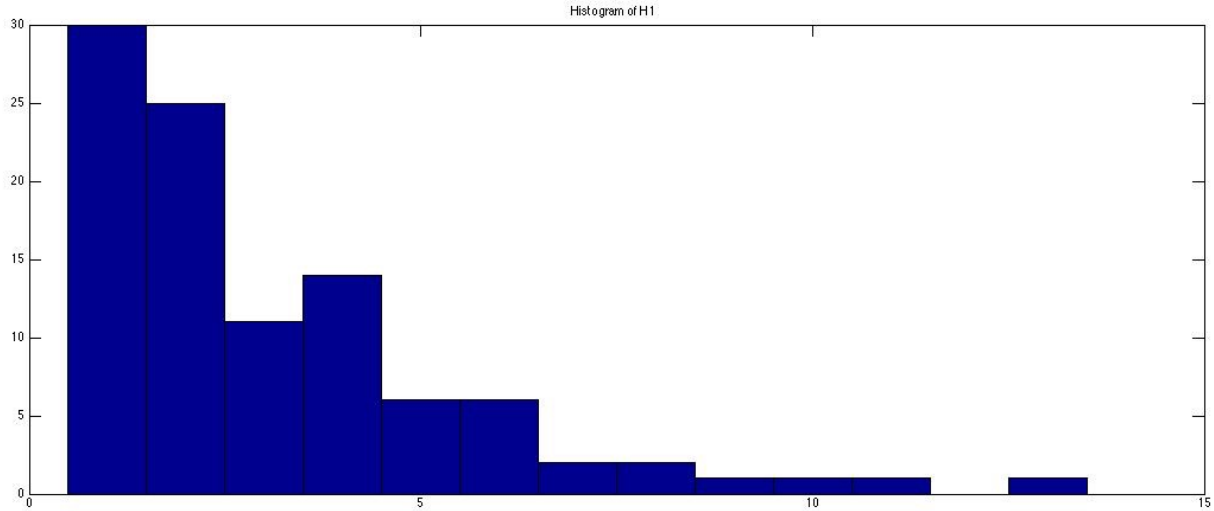
The implementation basically follows the instructions for each run and measure the data of different questions. Number of tosses till N heads in sequence per run is the total number of flips done so far, including the tosses used for N - 1 heads in sequence.

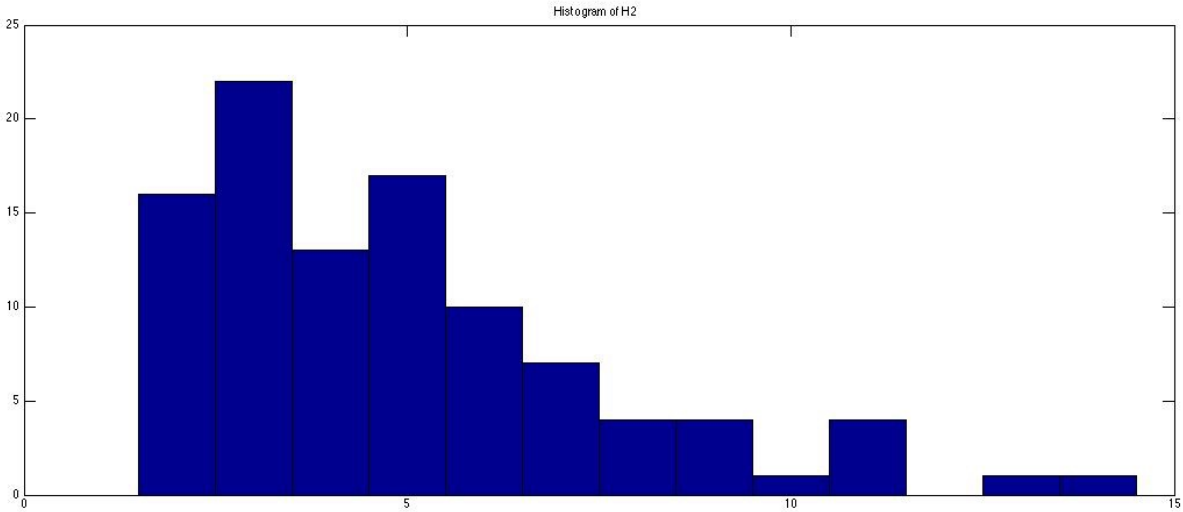
Result

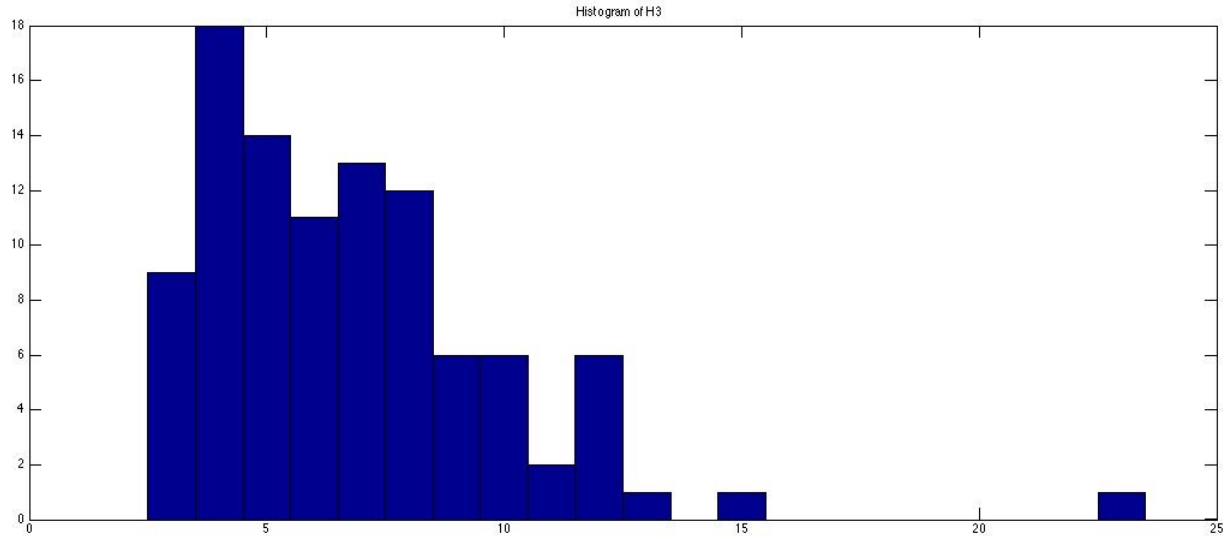
All data are presented in form of histogram:

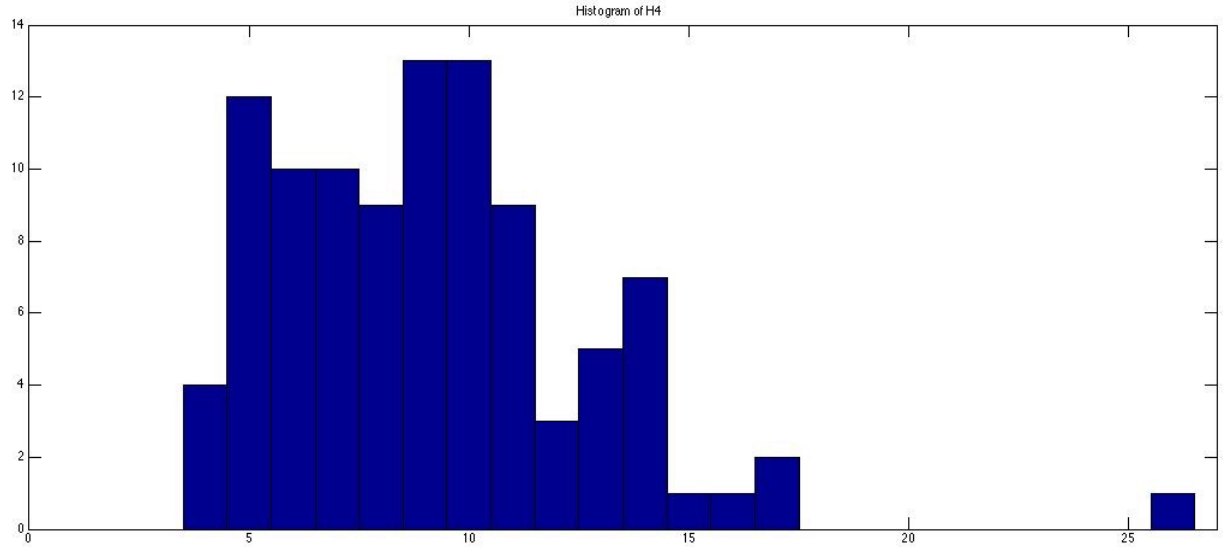


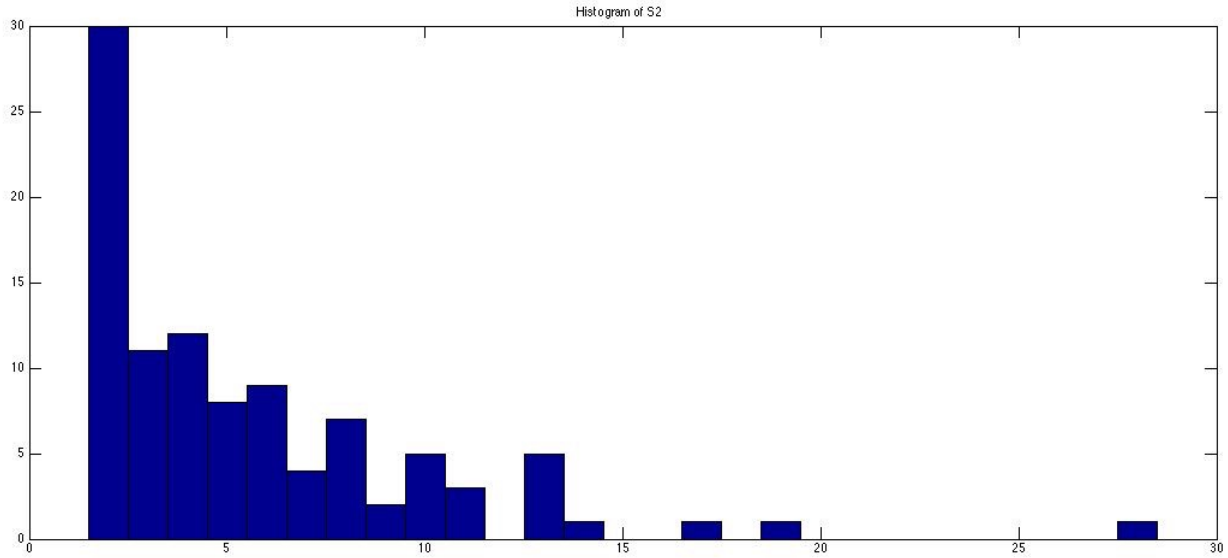

Histogram of Lh
30
25
20
15
10
5
0
0
2
4
6
8
10
12

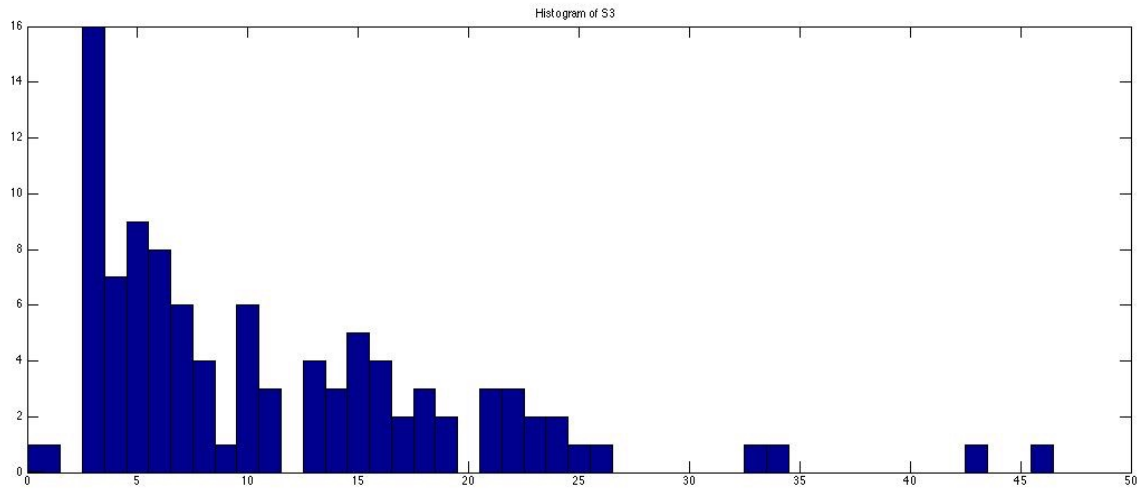

Histogram of H1
30
25
20
15
10
5
0
0
5
10
15

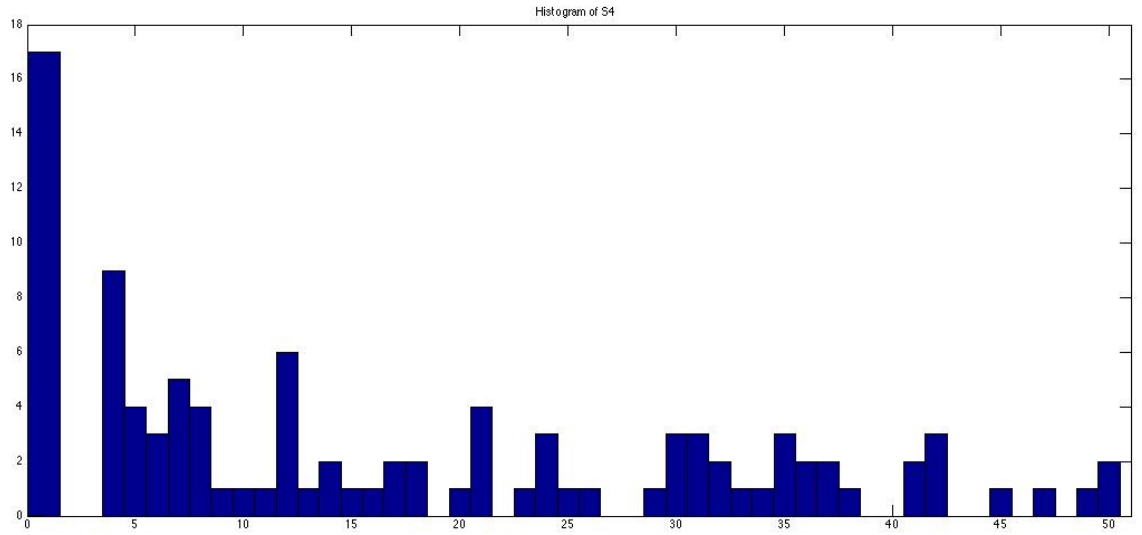

Histogram of H2
25
20
15
10
5
0
0
5
10
15


Histogram of H3
0
2
4
6
8
10
12
14
16
18
0
5
10
15
20
25


Histogram of H4
0
2
4
6
8
10
12
14
0
5
10
15
20
25


Histogram of S2
0
5
10
15
20
25
30
0
5
10
15
20
25
30





The numbers of heads for 100 runs are: 25 27 29 24 24 27 28 20 22 28 25 22 28 23 25 23 23 28 34 29 29 25 25 23 26 28 28 24 26 27 27 20 25 24 22 23 24 23 25 28 25 30 29 27 20 25 24 29 28 24 19 22 24 26 22 29 22 26 25 25 22 26 27 26 27 23 28 27 26 29 22 31 27 23 21 21 27 24 20 28 25 32 25 28 29 22 24 26 25 22 24 25 25 21 21 24 24 24 27 32. The mean is 25.27 and the standard deviation is 2.93. This is close to the analytic theory of 25 and $\sqrt{12.5} \approx 3.54$.

The numbers of tosses till one head for 100 runs are: 2 1 2 3 5 2 3 7 3 1 5 7 1 1 4 4 4 1 1 3 1 4 2 2 4 2 3 4 1 2 3 2 2 1 2 3 4 4 2 2 2 1 2 6 2 1 7 3 1 3 3 1 1 3 4 3 4 3 7 3 2 3 1 6 3 1 2 3 1 1 4 4 4 2 3 3 1 3 5 1 2 3 1 2 3 8 2 4 5 5 3 2 1 3 8 2 1 3 3 4. The mean is 2.88 and the standard deviation is 1.68. This is close to the analytic theory of 2 as mean.

The numbers of tosses till two head for 100 runs are: 3 4 3 7 7 3 4 8 4 2 8 8 2 2 6 5 5 3 2 4 2 5 5 3 5 4 4 5 6 4 4 3 3 4 8 5 8 6 3 3 3 2 3 8 3 2 8 4 2 5 5 2 2 5 5 4 6 4 10 4 4 4 2 9 6 3 3 4 2 4 5 6 5 4 4 5 3 4 11 2 3 4 2 3 5 10 5 5 6 8 4 3 2 6 10 4 2 5 5 5. The mean is 4.54 and the standard deviation is 2.09. This is close to the analytic theory of 4 as mean.