

## Chapter 4

# Probability: The Study of Randomness

Probability models are used to describe and analyze real-world phenomena involving randomness. One way to develop an intuition for randomness is to observe random behavior. Computer simulations allow visual penetration into the concept of random variation.

### 4.1 Randomness

#### Simulating Bernoulli Random Variables

A real-world probability can only be estimated through the observation of data. Computer simulations are useful because they help develop insight into the meaning of random variation. Excel is well suited for simulation and provides both a `RAND()` function and a **Random Number Generation** tool for such purpose.

**Example 4.1.** (Exercise 4.7, page 286 in the text.) The basketball player Shaquille O'Neal makes about half of his free throws over an entire season. Use Excel to simulate 100 free throws shot independently by a player who has probability 0.5 of making each shot. The technical term for independent trials with yes/no outcomes is Bernoulli trials. Our outcomes here are hit or miss.

- (a) What percent of the 100 shots did he hit in the simulation?
- (b) Examine the sequence of hits and misses. How long was the longest run of shots made? Of shots missed? (Sequences of random outcomes often show runs longer than our intuition thinks likely.)

**Solution.** Again we will use the `RAND()` function to generate a sequence of 100 free throws and then invoke the **ChartWizard** to dramatically display the results.



- In Step 4 click the button for Data Series in **Rows**. Enter “1” for Use First Row for Category(X) Axis Labels and enter “1” for Use First 1 Column for Legend Text.
- In Step 5 select the radio button **No** for Add a legend?, and label the chart and X axis as shown in Figure 4.1.
- Click Finish.

#### Users of Excel 97/98/2000/2001

- In Step 1 click the **XY (Scatter)** Chart type and the first Chart subtype (upper left on right side).
- In Step 2 on the **Data Range** tab, enter A3:CW4 for the range and check the radio button **Rows** for Series in:.
- In Step 3 on the **Titles** tab, enter the title and labels of the axes, on the **Axes** tab check both Category (X) axis and Category (Y) axis, on the **Gridlines** tab turn off all gridlines, on the **Legend** tab clear the legend, and finally on the **Data Labels** tab select the radio button **None**.
- In Step 4 embed the graph in the current workbook.
- Click Finish.

After the chart appears embedded on your workbook select it for editing and add the text “Hit” and “Miss” on the vertical axis.

Answers to Example 4.1 questions:

- By entering  $= \text{SUM}(B4:CW4)/100$  in an empty cell, we find that the player hit 51% of his shots in the simulation.
- From Figure 4.1, we read that the longest run of hits is 5 and the longest run of misses is 6.

To simulate an additional 100 free throws, press the F9 key as indicated previously or resave the worksheet.

## 4.2 Probability Models

A probability model consists of a list of possible outcomes and a probability for each outcome (or interval of outcomes, in the case of continuous models). The probabilities are determined by the experiment that leads to the occurrence of one or more of the outcomes in the specified list.

Excel provides many distributions that may be constructed in a common fashion with the **Formula Palette** or the **Function Wizard**. The meaning of the required parameters is available online through Excel’s help feature. Because of its prominence, the normal distribution was already discussed in Chapter 1. The

binomial model will be considered in Chapter 5. Here we discuss some other distributions of particular interest in statistics.

### Hypergeometric

$\text{HYPERGEOMDIST}(x, n, M, N, )$  provides probabilities for an experiment in which a simple random sample of size  $n$  is taken from a finite population of  $N$  individuals of which  $M$  are in a so-called “preferred category” called “success” or “1,” while the remaining  $N - M$  are deemed “failure” or “0.” The function returns the probability of  $x$  successes in the sample of size  $n$ .

**Example 4.2.** (Lotto 6/49) Suppose a box contains 49 balls in which one and only one ball is marked with an integer taken from  $\{1, 2, \dots, 49\}$ . The balls are identical otherwise. Suppose that the balls numbered  $\{1, 2, 3, 4, 5, 6\}$  are considered “successes.” If an SRS of six balls is taken at random (without replacement), what is the probability that the sample contains  $k$  successes (for  $k = 0, 1, 2, 3, 4, 5, 6$ )?

	A	B	C
1		<b>Lotto 6/49 Probabilities</b>	
2			
3	k	P(X=k) value	P(K=k) formula
4	0	0.4259650	=HYPERGEOMDIST(A4,6,6,49)
5	1	0.4130195	=HYPERGEOMDIST(A5,6,6,49)
6	2	0.1323780	=HYPERGEOMDIST(A6,6,6,49)
7	3	0.0176504	=HYPERGEOMDIST(A7,6,6,49)
8	4	0.0009686	=HYPERGEOMDIST(A8,6,6,49)
9	5	0.0000184	=HYPERGEOMDIST(A9,6,6,49)
10	6	0.0000001	=HYPERGEOMDIST(A10,6,6,49)

Figure 4.2: Calculating Lotto Probabilities

**Solution.** The answer is provided in Figure 4.2 where column B gives the probabilities and column C the corresponding Excel formulas.

### Student $t$ -Distribution

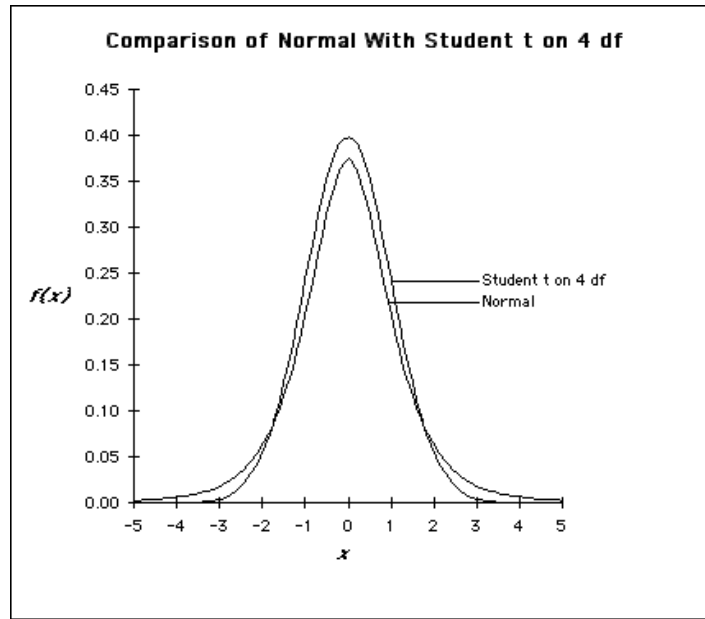
The Student  $t$ -distribution arises as the distribution of the *Studentized* score (similar to a standardized score)

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where  $\bar{x}, s$  are the sample mean and sample standard deviation from a sample of size  $n$  taken from a  $N(\mu, \sigma)$  population. It is determined by a single parameter called the degrees of freedom  $\nu$ . In the above ratio,  $\nu$  takes the value  $n - 1$ .

Excel has an unusual definition of the c.d.f. and inverse c.d.f. for the  $t$ -distribution. The Excel function **TDIST** returns the tail of the distribution, that is, if  $t(\nu)$  is a random variable with a  $t$ -distribution on  $\nu$  d.f., then

$$\text{TDIST}(x, \nu, 1) = P[t(\nu) > x]$$

Figure 4.3: Comparing the Normal and the  $t$  Curves

and

$$\text{TDIST}(x, \nu, 2) = 2P[t(\nu) > x]$$

The argument  $x$  in `TDIST` must be positive. Thus the c.d.f. takes a rather complicated expression using the logical `IF` function:

$$P(t(\nu) \leq x) = \text{IF}(x < 0, \text{TDIST}(\text{ABS}(x), \nu, 1), 1 - \text{TDIST}(x, \nu, 1))$$

where  $\text{ABS}(x) = |x|$ , and similarly

$$P(|t(\nu)| \leq x) = 1 - \text{TDIST}(x, \nu, 2)$$

The inverse function `TINV` is defined by

$$P[t(\nu) > \text{TINV}(\alpha, \nu)] = \frac{\alpha}{2}$$

so  $\text{TINV}(\alpha, \nu)$  is the critical value for a two-sided significance test at level  $\alpha$  of a normal mean (to be discussed in Chapter 7).

**Exercise.** Using `NORMSDIST` and `TDIST`, graph on the same figure and to the same scale the densities of a  $N(0, 1)$  and the Student  $t$ -distribution on 4 d.f. (Figure 4.3).

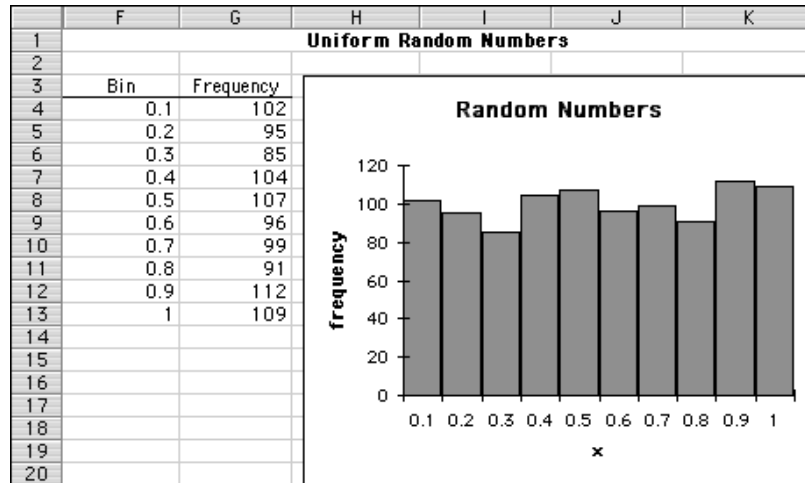


Figure 4.4: Simulating Uniform Random Variables

## Uniform

A uniform random number is one whose values are spread out uniformly across the interval from 0 to 1. Its density curve has height 1 over the interval 0 to 1.

**Example 4.3.** (Based on Example 4.17, page 10 in the text.) Let  $X$  be a uniform random number between 0 and 1. Use Excel to generate 1000 random uniform numbers, and from your simulations estimate the following probabilities and then compare them with the theoretical values.

- $P(0.3 \leq X \leq 0.7)$
- $P(X \leq 0.5)$
- $P(X > 0.8)$

**Solution.** Use `RAND()` to generate 1000 uniform random variables in a column and construct a histogram with bin intervals of width 0.10 beginning at 0 and ending at 1. Figure 4.4 shows the sample output from a workbook where this has been done. The frequencies shown are the number of times the random number generator produced a number  $X$  in the specified interval. The values listed under the heading *Bin* are the right endpoints of the intervals. We count the number of observations in the relevant intervals and divide by 1000 to convert to a probability.

- $P(0.3 \leq X \leq 0.7) = (104 + 107 + 96 + 99)/1000 = 0.406.$
- $P(X \leq 0.5) = (102 + 95 + 85 + 104 + 107)/1000 = 0.493.$
- $P(X > 0.8) = (112 + 109)/1000 = 0.221.$

The theoretical values are 0.400, 0.500, and 0.200, respectively.

### Triangular—Adding Random Numbers

**Example 4.4.** (Based on Exercise 4.54, page 317 in the text.) Generate two random numbers between 0 and 1 and take  $Y$  to be their sum. Clearly the sum  $Y$  can take any number between 0 and 2. It is known that the idealized density curve of  $Y$  is a triangle. Use Excel to generate 1000 pairs of uniform random numbers, add them, and from your simulations estimate the following probabilities and compare them with the theoretical values.

(a)  $P(0 \leq X \leq 0.5)$

(b)  $P(0 \leq X \leq 1.0)$

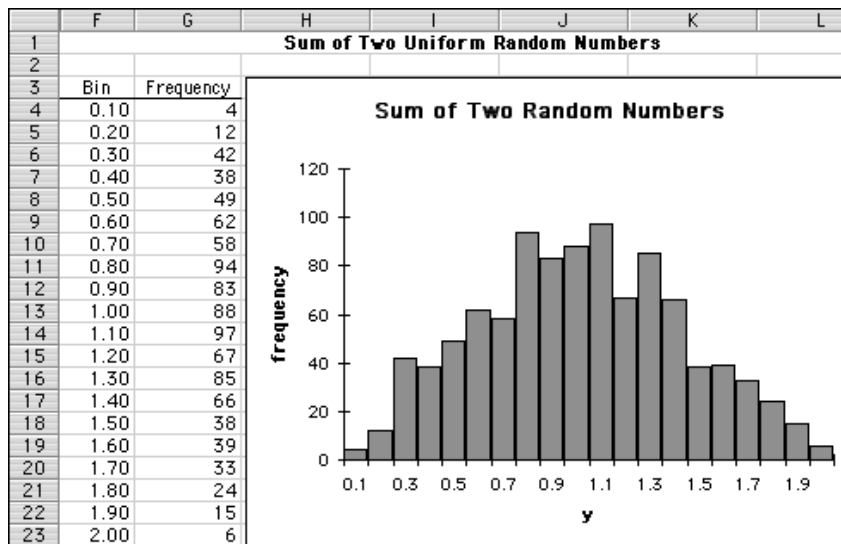


Figure 4.5: Simulating Triangular Random Variables

**Solution.** Again use `RAND()` to generate 1000 pairs of uniform random variables in two columns, add the columns and construct a histogram with bin intervals of width 0.10 beginning at 0 and ending at 2. Figure 4.5 shows the sample output from a workbook where this has been done. The frequencies shown are the number of times the random number generator produced a number in the specified interval.

(a)  $P(0 \leq X \leq 0.5) = (4 + 12 + 42 + 38 + 49)/1000 = 0.145.$

(b)  $P(0 \leq X \leq 1.0) = (4 + 12 + 42 + 38 + 49 + 62 + 58 + 94 + 93 + 88)/1000 = 0.540.$

The corresponding theoretical values are 0.125 and 0.500, respectively.

### 4.3 Random Variables

A random variable is completely prescribed by its probability distribution. This distribution has a long-run frequency interpretation associated with the **Law of Large Numbers**.

Draw independent observations at random from any population with finite mean  $\mu$ . Decide how accurately you would like to estimate  $\mu$ . As the number of observations drawn increases, the mean  $\bar{x}$  of the observed values eventually approaches the mean  $\mu$  of the population as closely as you specified and then stays that close.

Applying this phenomenon to a discrete random variable  $X$ , suppose that

$$P(X = 3) = 0.5$$

In repeated trials, consider the proportion  $\hat{p}$  of times that  $X$  takes the value 3. The random variable  $\hat{p}$  is the sample mean of a sequence of random variables taken from a Bernoulli population with probability of success = 0.5 and whose population mean is therefore also 0.5. The Law of Large Numbers then asserts that in some sense (made precise by the theory of probability)  $\hat{p}$  approaches 0.5 as the number of trials increases, which gives a relative frequency interpretation of probability.

### Simulating Random Variables

#### The Law of Large Numbers Using the RAND() Function

The Excel function `RAND()` picks a number uniformly on the interval (0,1). To generate a uniform random variable on  $(a, b)$  use  $a + \text{RAND()} * (b - a)$ . Using the inverse probability function  $h(a) = \inf\{x : F(x) \geq a\}$ , we can then generate other distributions. Thus `NORMINV(RAND(), Mean, StDev)` returns a random normal with mean given by Mean (either a numerical value or a named reference to a numerical value) and standard deviation by StDev.

By examining the list of functions available (clicking the **Paste Function** button  $f_x$  on the Standard Toolbar), you can determine which distributions Excel can simulate this way and how to describe the required parameters.

**Example 4.5.** Using the `RAND()` function, simulate 1000 independent Bernoulli trials based on tossing a fair coin, calculate the cumulative proportion of heads  $\hat{p}$  after each trial, and construct a graph that demonstrates the law of large numbers in action. Also show on the same graph a horizontal line at the height 0.5

**Solution.** The `RAND()` function produces a number uniformly distributed on the interval (0,1). This can be converted into integers taking the values 0 or 1 with equal probability if this uniform random number is multiplied by 2 and then the integer part is taken. The Excel formula for these operations is `=INT(2*RAND())`.

1. Enter the formula = INT(2\*RAND()) in cell A5 of a new workbook and copy this formula down to cell A1003 by selecting cell A4, then clicking the fill handle and dragging to cell A1003 to generate 1000 tosses of a fair coin (0 representing tails and 1 representing heads).
2. Enter the value “0” in cell B3 followed by the formula = A4+B3 in cell B4. Copy the formula in cell A4 down to cell A1003. Column B tracks the cumulative number of heads.
3. Enter the number 1 in cell C4 and fill to cell C1003 with successive integers {1, 2, . . . , 1000}. This can be achieved efficiently by selecting cell C4 and then choosing **Edit – Fill – Series** from the Menu Bar. Complete the **Series** dialog box with Series in **Columns**, Type **Linear**, and **Step Value 1**, **Stop Value 1000**. Click OK. Column C will label the 1000 tosses.
4. Fill cells D4 to D1003 with the value 0.5. This will represent the horizontal line at height 0.5 on the graph.
5. Enter the formula = B4/C4 in cell E4 and copy to cell E1003.

Figure 4.6 shows part of the workbook with the required formulas.

	A	B	C	D	E
1	<b>Formulas Behind Simulation of 1000 Tosses</b>				
2					
3		0			
4	=INT(2*RAND())	=A4+B3	1	0.5	=B4/C4
5	=INT(2*RAND())	=A5+B4	2	0.5	=B5/C5
6	=INT(2*RAND())	=A6+B5	3	0.5	=B6/C6

Figure 4.6: Simulating 1000 Tosses of a Fair Coin

We next construct a graph displaying the same results. Click the **Chart Wizard** button.

#### Users of Excel 5/95

- In Step 1 enter the data range C4:E1003.
- In Step 2 click the **Line** chart type.
- In Step 3 select **Format 1**.
- In Step 4 click the button for Data Series in **Columns**. Enter “1” for Use First 1 Column for Category(X) axis labels and enter “0” for Use First 0 Column for Legend Text.
- In Step 5 click the radio button **No** for Add a legend? and label the chart and X axis as shown in Figure 4.7. Click Finish.

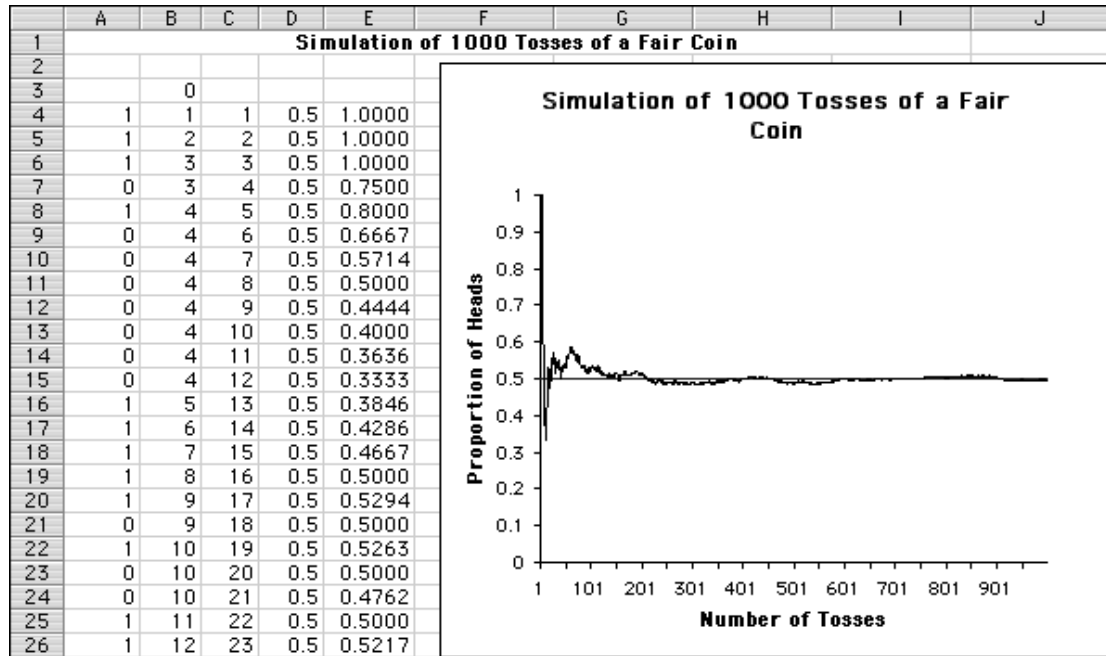


Figure 4.7: Law of Large Numbers

#### Users of Excel 97/98/2000/2001

- In Step 1 click the **Line** chart type and the second Chart sub-type **Line**.
- In Step 2 on the **Data Range** tab enter D4:E1003 for the Data range check the radio button **Series in: Columns**.
- In Step 3 on the **Titles** tab, enter the title and labels of the axes, on the **Axes** tab check radio button **Automatic** for Category (X) axis and check the Value (Y) axis box, on the **Gridlines** tab turn off all gridlines, on the **Legend** tab clear the legend, and finally on the **Data Labels** tab select the radio button **None**.
- In Step 4 embed the graph in the current workbook. Click Finish.

Format the X and Y axes as shown in Figure 4.7, for instance by changing the number of categories between tick marks and reorienting the X axis labels.

Figure 4.7 shows a segment of the completed workbook with the embedded graph. As previously you can reevaluate all functions and the graph will dynamically change. From column A you can see the random sequence of heads and tails generated, while column E exhibits the proportions of heads. These are quite variable at first but then settle down, appearing to approach the value 0.5 (shown by the horizontal line). This behavior is known in statistics as a law of large numbers, commonly referred to as the “law of averages.”

## Simulating Random Variables Using Random Number Generation

In addition to the function `RAND()`, Excel has a **Random Number Generation** tool built into the **Analysis ToolPak** that provides an alternative and more systematic approach to simulation.

The **Random Number Generation** tool creates columns of random numbers, as specified by the user, from any of six probability models (uniform, normal, Bernoulli, binomial, Poisson, discrete) as well as having an option for patterned that creates not random data but rather data according to a specified pattern.

All options are invoked from a common dialog box (as in Figure 4.8 for discrete) following the choice **Tools – Data Analysis – Random Number Generation** from the Menu Bar. Select the distribution of interest using the drop-down arrow and the Parameters sub-box will automatically change, prompting input of parameters.

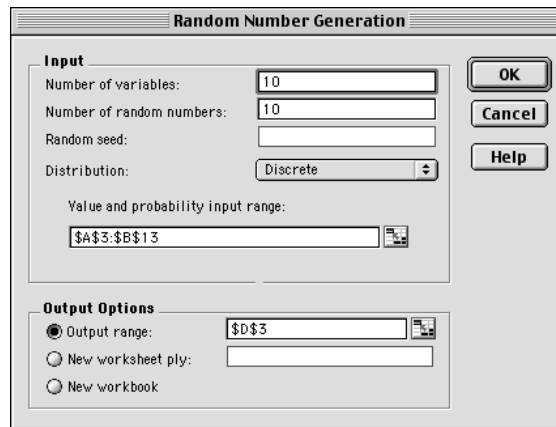


Figure 4.8: Random Number Generation Tool—Discrete

**Number of Variables.** Enter the number of columns of random variables. The default is all columns.

**Number of Random Numbers.** Enter the number of rows (cases) of random variables.

**Distributions.** Use the drop-down arrow to open a list of choices with requested parameters.

**Uniform.** Upper and lower limits

**Normal.**  $\mu, \sigma$

**Bernoulli.**  $p$  = probability of success; Excel unfortunately refers to this as a p Value.

**Binomial.**  $p, n$

**Poisson.**  $\lambda$

**Discrete.** Specify the possible values and their corresponding probabilities. Before using this option enter the values and probabilities in adjacent columns in the workbook.

**Patterned.** This option creates data according to a prescribed pattern of values repeated in specified steps. This is useful if a linear array of data needs to be coded using another variable.

**Example 4.6.** To generate 100 tosses of a pair of fair dice enter  $\{2, 3, \dots, 11\}$  into cells A3:A13 and enter

$$\{1/36, 2/36, \dots, 6/36, 5/36, \dots, 2/36, 1/36\}$$

into cells B3:B12 (Figure 4.9). Excel may interpret the value  $1/36$  as a date Jan 1936. If this happens then format the cells by choosing **Format – Cells** from the Menu Bar and selecting **Number**. Then choose **Tools – Data Analysis – Random Number Generation** from the Menu Bar and complete as in Figure 4.8. The output will appear in cells D1:M10. Since these numbers are random your output will of course be different.

	A	B	C	D	E	F	G	H	I	J	K	L	M
1		<b>Simulation of a Pair of Fair Dice</b>											
2	k	P(X=k)											
3	2	0.0277778		2	8	6	8	4	7	7	8	6	5
4	3	0.0555556		6	9	5	6	8	9	12	9	7	7
5	4	0.0833333		8	5	8	8	12	6	7	8	9	5
6	5	0.1111111		2	9	3	8	5	7	7	6	5	8
7	6	0.1388889		8	6	11	10	6	8	4	8	7	8
8	7	0.1666667		10	6	9	8	2	6	4	7	6	4
9	8	0.1388889		9	3	4	5	10	7	7	6	4	7
10	9	0.1111111		6	6	4	11	9	7	7	2	11	4
11	10	0.0833333		6	9	8	9	5	8	3	12	3	4
12	11	0.0555556		12	8	3	5	8	4	10	7	10	7
13	12	0.0277778											

Figure 4.9: Simulating a Pair of Fair Dice