

## Probability and Discrete Distributions

### Objectives:

1. To explore relative frequency and the Law of Large Numbers
2. To practice the basic rules of probability
3. To work with the binomial and Poisson distributions in JMP

**Getting Started:** Log onto your machine and start JMP. There are no datasets to download for today.

### Part I. Relative Frequency

Suppose a random experiment is repeated many times, for example, a fair coin is flipped 1000 times. The outcome from a single flip of the coin is either Heads or Tails. If we repeat the experiment, the number of times each outcome is observed, i.e., the number of Heads and the number of Tails, is called the *frequency* of that outcome. The *relative frequency* of an outcome is the proportion of times that outcome is observed. For example, the relative frequency of Heads is the frequency of heads divided by the total number of flips. By the *Law of Large Numbers*, as the number of repetitions of an experiment increases, the *relative frequency* of each outcome approaches the probability of that outcome. In this lab you will use *computer simulation* to find the relative frequency of an outcome and to show how, as the number of repetitions of the experiment increases, the relative frequency stabilizes to a number between 0 and 1.

To illustrate the law of large numbers and the relative frequency definition of probability we will use JMP to simulate the flipping of a coin a number of times. A probability model for this experiment, i.e., flipping a coin multiple times, is called a binomial model. From the JMP Starter window, choose “New Data Table”. We will generate 10 coin flips, one in each row. To get this started, first we need to tell JMP how many coins to flip. To do this, double-click on the header for **Column 1** (at the top of the first column of the data table) to bring up the dialog box (if that doesn’t work, you can also right-click and select **Column Info...**). Near the bottom, where it has a box with 0 after **N Rows**, replace the 0 with 10 and hit “OK”. This will give you a column of dots, which represent missing values. What we’ve done is tell JMP how many random draws we are going to want. Now right-click on the header and choose **Formula**, which will pop up a new dialog box. On the top, just right of center is a menu (titled **Functions (grouped)**) with a scroll bar. Scroll down a bit and click on **Random**, which pops up more choices, and click on **Random Binomial**. It should put “Random Binomial(n, p)” in the big box, and the “n” should be surrounded by a red box, meaning that JMP is asking you to specify how many coins are flipped for each row. Since we’re just putting one coin flip in each row, type the number 1 and hit **Enter**. Now click on the “p” to move the red box to be around it. We want each coin to have probability 0.5 of coming up head, so type in “0.5” and hit **Enter**. Now hit “OK” and JMP will fill in the random flips and return you to the data table. You should now see 0’s (tails) and 1’s (heads) in the second column. In this case it is easy enough to compute the relative frequency by hand. But you can also get it by going back to the JMP Starter window, going to the **Basic** menu, and choosing “Distribution”. Since all entries are 0 or 1, the mean is equal to the relative frequency of 1’s, so you can just scroll down to the mean.

**Question #1** What is the relative frequency of heads?

**Question #2** What proportion of Heads did you expect? Are you surprised at the proportion of Heads you observed?

We want to observe the law of large numbers taking effect by increasing the number of “flips of the coin.” Redo your simulation above separately (create a new data table each time, and remember to tell JMP how many rows you want before you go to the formula window) for (i)  $n = 100$  flips, (ii)  $n = 1000$  flips and (iii)  $n = 10000$  flips. Please fill in the following table:

**Question #3**

<u>Number of Flips</u>	<u>Relative Frequency of Heads</u>
n = 10	
n = 100	
n = 1000	
n = 10000	

**Question #4** The true probability of observing a Head based on this simulation is 0.50. Describe what happens to the relative frequency of the occurrence of a Head as the number of flips increases from 10 to 10000.

↪ **Question #5** If you were to look at someone else’s simulation, would you expect the results for  $n=10$  to be about the same as yours? How about the results for  $n=10000$ ?

---

---

## Part II. Probability

Questions 6-14 regard a sample of 100 college students that was asked what their favorite burger chain is. The following results were obtained:

	Burger King	McDonald's	In-N-Out	Other
Male	10	12	24	9
Female	24	8	7	6

**Question #6** If one person is randomly selected from this sample, what is the probability that they prefer In-N-Out?

**Question #7** If one person is randomly selected, what is the probability that they are female and prefer In-N-Out?

**Question #8** If one person is randomly selected, what is the probability that they are female or prefer In-N-Out?

**Question #9** If one person is randomly selected, what is the probability that they prefer In-N-Out given that they are female?

**Question #10** If a randomly selected person that prefers In-N-Out is selected, what is the probability that they are female?

**Question #11** Do you get the same number for #9 and #10? Why or why not?

**Question #12** Are choice of burger chain and gender independent or dependent? Why?

**Question #13** Are preference for Burger King and for McDonald's mutually exclusive?

↯ **Question #14** Are preference for Burger King and for McDonald's independent?

A newspaper article reported “Finnish men are complaining of sexual harassment in the work place.” A sample of male and female Finnish workers consisted of 48% men. Among all the workers in the sample, 30% reported being sexually harassed in the workplace. Furthermore, 26% of the workers reported being male and being sexually harassed.

**Question #15** Note that the probability a random worker reports being harassed is equal to the probability that they are male and report being harassed plus the probability that they are female and report being harassed. What is the probability that they are female and report being harassed?

**Question #16** What is the probability that they are male and do not report being harassed?

**Question #17** What is the probability that they are male given that they report being harassed?

Now we’ll practice an application of Bayes’ Theorem. Recall that Bayes’ Theorem says:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B|A)P(A) + P(B|notA)P(notA)}$$

The ELISA test was an early test used to screen blood donations for antibodies to HIV. A study (Weiss et al. 1985) found that the conditional probability that a person would test positive given they have HIV was 0.977 and the conditional probability that a person would test negative given they did not have HIV was 0.926. The World Almanac gives an estimate of the probability of a person in North America having HIV of 0.0026.

**Question #18** Suppose a random person is tested and they test positive. What is the conditional probability that this person has HIV given that they test positive?

**Question #19** Are you surprised by your answer? What implications does this have for policies of mandatory testing?

Note that this phenomenon of large and unexpected changes in conditional probabilities is not unusual, particularly when dealing with rare events. What is happening is that the number of false positives is much larger than the number of true positives.

**Question #20** Suppose 10,000 random people are tested. How many of them do you expect to actually have HIV?

**Question #21** Of those with HIV, how many do you expect to test positive?

**Question #22** Of those without HIV, how many do you expect to test positive?

↯ **Question #23** Do your answers to #21 and #22 make sense in light of your answer to #18?

---

---

Of course in practice, people who test positive would be given a follow-up test. Also, more expensive but more accurate tests have subsequently been developed. But the implications on mandatory testing still apply. For rare diseases, it really only makes sense to test in subgroups that are at high risk. Otherwise you get more false positives than true positives.

### Part III. Binomial Random Variables

Recall that the characteristics of a *binomial random variable* are:

1. The experiment consists of a fixed number ( $n$ ) trials.
2. The trials are independent.
3. There are only two possible outcomes on each trial, which for convenience we call “success” and “failure.”
4. The probability of a success ( $p$ ), is the same for each observation.
5. If  $X$  is a binomial random variable then its mean is  $E[X] = \mu_X = np$  and its standard deviation is  $\sigma_X = \sqrt{np(1-p)}$ .

*For the next three questions of the following examples, decide whether  $X$  is a binomial random variable. State why or why not.*

**Question #23** The pool of potential jurors for a murder case contains 100 persons chosen at random from the adult residents of a large city. Each person in the pool is asked whether he or she opposes the death penalty;  $X$  is the number who say “Yes”.

**Question #24** Suppose a jury of 12 people is chosen from the above pool, and this jury hears a case and discusses the verdict;  $X$  is the number who think the defendant is guilty.

**Question #25** A drug is known to be 60% effective in curing a certain disease; that is, the probability is 0.60 that a person with the disease given the drug will be cured. Suppose 30 people with the disease are chosen at random and given the drug. Let  $X$  be the number of people among the 30 who are cured.

**Question #26** Using the information from the previous question, find the expected number of people cured, i.e.,  $E(X)$ , and the standard deviation of  $X$ .

**Question #27** If 25 of the 30 people were cured, would this be unusual?

**Question #28** Use JMP to generate 100 replications of draws from a binomial with  $n = 30$  and  $p = .6$  (create a new data sheet, fill in the first column with 100 rows of dots, then use the formula menu to put in a random binomial with  $n = 30$  and  $p = 0.6$ ). What are the observed mean and standard deviation of your 100 random replications? (Use Basic “Distributions” to get these values from JMP.) Are they close to what you computed in #26?

**Question #29** How many of your 100 replications were 25 or higher? Is this consistent with your answer to #27?

Finally, let's practice with the Poisson distribution. Recall that the Poisson distribution is used to model counts of events that happen independently over time or space. Also recall that the standard deviation of a Poisson is the square root of the mean.

Suppose a pet clinic gets an average of six cats per day to be spayed, and they only have the resources (material supplies and volunteer time) to spay a maximum of eight cats on any given day. Can we find the probability that more cats will show up on a day than they can spay? We could work with the theoretical distribution, but instead, let's estimate it using relative frequencies in JMP. We'll simulate 1000 days of arrivals of cats. Each day will be drawn from a Poisson random variable with mean 6. Create a new data table (from **File** on the JMP Starter window). Double-click on the header for **Column 1** (at the top of the first column of the data table) to bring up the dialog box. Near the bottom, where it has a box with 0 after **N Rows**, replace the 0 with 1000 and hit "OK". Now right-click on the header and choose **Formula**. You should get the familiar formula dialog box. Scroll down to choose **Random** from the **Functions (grouped)** menu, and then choose **Random Poisson**. The red box should now be around  $\lambda$ . Type in 6 and hit **Enter**. Now click on "OK". You should now have 1000 random Poisson draws in your table.

**Question #30** What are the theoretical expected value and standard deviation for the number of cats showing up at the clinic on a given day?

**Question #31** Use JMP to find the mean and standard deviation from your simulated draws. How close are they to the theoretical values?

↪ **Question #32** To get the counts by value, it is easiest to first tell JMP to pretend that the variable is ordinal instead of continuous. In the data table window, in the middle box on the left, there should be a blue triangle to the left of **Column 1** (or whatever you have titled the column). Click on that triangle and change it to **Ordinal**, which should change the icon to grey steps. Now make a new **Basic** "Distribution" analysis, which should give you counts. What fraction of your simulation has values of 9 or higher? What do you estimate the probability is that more cats will arrive than can be handled on a given day?

---

---

**Quit** JMP and please remember to **Log Off**.