Topics: Equally Likely Probability
Activity: Random CDs

Content: This activity explores the meaning of probability and equally likely outcomes through physical and computer simulations.

Prerequisites: For the computer simulation section, limited familiarity with the software Minitab is useful. Students will be shown how to create a Minitab macro in this lab.

Materials:
Four index cards and one sheet of scratch paper per student.

Goals:
- To develop intuition for probability as long-term relative frequency
- To learn or reinforce some elementary counting rules
- To emphasize consideration of when the "equal likeliness" assumption applies and when it does not
- To gain an introduction to the concepts of sample space, random variable, and expected value

Situation:
Suppose that you have four music CDs in your CD player. The corresponding four CD jewel boxes are on the table. In this lesson, we want to explore what would happen if the CDs were returned to the jewel boxes completely at random. We begin by using *simulation* to investigate what will happen in the long run, if you repeat this experiment over and over again.

First, Make Some Conjectures:
Given the four CDs are randomly returned to four jewel boxes,
- How many CDs do you think, on average, will be returned to the **correct** jewel box? _____
- What do you think the probability is of returning **exactly one** CD to the correct jewel box? _____
- What do you think the probability is of returning **more than one** CD to the correct jewel box? _____

Part 1: Simulation Analysis:

Set-up: As a class, decide the name of four music groups whose initials are all distinct. Each student has four index cards which represent the Compact Disks for each music group. Write one music group's name on each index card. A sheet of paper will represent the CD jewel boxes. Divide the sheet of paper into four areas and write one music group's name on each area of the paper.

(a) Shuffle the four index cards well, and then deal them out randomly with one index card going to each area of the sheet. Turn over the cards to reveal which CDs were randomly assigned to which jewel box. Record how many CDs were placed with the correct jewel box.

number of matches = _____

(b) Repeat the random "dealing" of CDs a total of five times and record the number of matches that occur in each case. (Shuffle the index cards well in between "deals.")

Repetition Number	**1**	**2**	**3**	**4**	**5**
Number of Matches					

(c) Combine your results on the number of matches with the rest of the class, obtaining a tally of how often each result occurred. Record the counts and proportions in the table below.

Number of Matches	**0**	**1**	**2**	**3**	**4**	**Total**
Count						
Proportion						**1.00**

(d) In what proportion of these simulated cases did **at least one** CD get matched with the correct jewel box? _____

The ***probability*** of a random event is the long-run proportion (or relative frequency) of times the event would occur if the random process were repeated over and over under identical conditions. One can *approximate* a probability by *simulating* the process a large number of times. Simulation leads to an ***empirical*** estimate of the probability.

Part 2. (Exact) Enumeration Analysis:

In situations where the outcomes of a random process are ***equally likely*** to occur (e.g. tossing a fair coin), exact probabilities can be calculated by listing all of the possible outcomes and determining the proportion of these outcomes which correspond to the event of interest. The listing of all possible outcomes is called the ***sample space***.

The **sample space** for the "random CDs" consists of all possible ways to distribute the four CDs to the four jewel boxes. Let ***wxyz*** mean that the CD ***w*** went to the first jewel box, CD ***x*** to the second jewel box, CD ***y*** to the third jewel box, and CD ***z*** to the fourth jewel box. For example, **1243** would mean that the first two CDs were randomly

assigned to the correct jewel boxes, but the third and fourth CDs were switched with their jewel boxes.

(e) Below is the beginning of a list of the possibilities for the "random CD" process. Fill in the remaining possibilities, using this same notation. [Try to be systematic about how you list these outcomes so that you don't miss any. One sensible approach is to list in a second row the outcomes for which jewel box 1 gets CD 2, and then in the third row list the cases where jewel box 1 gets CD 3, and so on.]

Sample Space:

1234 1243 1324 1342 1423 1432

(f) How many possible outcomes are there in this sample space? That is, in how many different ways can the four CDs be returned to their jewel boxes? _____

You could have determined the number of possible outcomes without having to list them first. The first jewel box could have any of the 4 CDs put in it. Then there are three CDs to choose from for the second jewel box. The third jewel box gets one of the two remaining CDs and then the last CD goes to the fourth jewel box. Since the number of possibilities at one stage of this process does not depend on the outcome of earlier stages, the total number of possibilities is the product $4 \cdot 3 \cdot 2 \cdot 1 = 24$. This is also known as $4!$, read as "4 factorial."

(g) For each of the above outcomes in your sample space, indicate how many jewel boxes get the correct CD.

(h) In how many outcomes is the number of "matches" equal to exactly:

4: _____ 3: _____ 2: _____ 1: _____ 0: _____

(i) Calculate the (exact) probabilities by dividing your answers to (g) by your answer to (e). Comment on how closely the exact probabilities correspond to the empirical estimates from the simulation above.

The "number of matches" is an example of a ***random variable***, which is a function assigning a numerical output to each outcome in a sample space. Here, each of the 24 outcomes has a corresponding numerical value for "number of matches." This is a *discrete* random variable in that it can assume only a finite, or countably infinite, number of values. The ***probability distribution*** of a discrete random variable is given by its set of possible values and their associated probabilities.

(j) Are the possible values of the "number of matches" random variable equally likely? Explain.

(k) For your class simulation results, calculate the average (mean) number of matches per repetition of the process.

The long-run average value achieved by a numerical random process is called the ***expected value*** of the random variable. To calculate this expected value from the (exact) probability distribution, multiply each outcome of the random variable by its probability, and then add these up over all of the possible outcomes.

(l) Calculate the *expected* number of matches from the (exact) probability distribution, and compare that to the average number of matches from the simulated data.

(m) What is the probability that the number of matches equals this expected value exactly? Based on this probability, do you literally expect to find this number of matches in one realization of the process? Explain.

Computer Simulation to Investigate Sample Size Effects:

Now that you have determined the (exact) probability distribution of this random variable, it might be instructive to simulate its behavior in order to examine how closely the results of a given sample match (or do not match) the long-run probabilities. You can use a computer package such as Minitab to simulate the random assignment of CDs to jewel boxes much more quickly.

(n) Set up Minitab to conduct this simulation by naming column `c1` "**jewel box**" entering the numbers 1, 2, 3, and 4 into column `c1`. The numbers in `c1` represent the 4 jewel boxes.
- Label column `c2`, **"CD."**
- Next, randomly assign the 4 CDs to a jewel box. Sample all four numbers in `c1` and put them in `c2` to represent the CDs randomly assigned to each jewel box. Be certain the MTB prompt is
  `MTB> sample 4 c1 c2`
- Name column `c3` "**match? (1=yes, 0=no)**"
- Set up `c3` as an indicator variable (1 if yes, 0 if no) of whether the number in `c1` matches its corresponding number in `c2`:
  `MTB> let c3=(c1=c2)`
- Now count how many matches there are in this sample:
  `MTB> sum c3`

(o) Now write a macro for conducting this simulation many more times much more efficiently by creating a text file (using Notepad) containing the following commands:

```
sample 4 c1 c2
let c3=(c1=c2)
let c4(k1)=sum(c3)
let k1=k1+1
```

Save the text file as "matching.mtb." [Be sure to make its extension .mtb; remember to use quotes around the file name when you are saving it.] In this macro, `k1` will be a counter that will keep track of how many times you have simulated the process. Before running the macro, you need to initialize this counter in Minitab:

`MTB> let k1=1`

(p) Test the macro that you have written by executing it ten times (select File> Other Files> Run an Exec…), tell it to execute 10 times, click on "Select file" and choose the "matching.mtb" file that you just created.

(q) Assuming that the macro works, tally and display the number of matches in your ten simulated samples. Also calculate the mean number of matches:

```
MTB> tally c4
MTB> hist c4
MTB> mean c4
```

Record the simulated proportions in the first row of the table, along with the sample mean:

# of matches	0	1	2	3	4	mean
10 repetitions						
100 repetitions						
1000 repetitions						
2000 repetitions						
Theoretical	.375	.333333	.25	0	.041667	1.00

(r) Repeat this simulation for sample sizes of 100, then 1000, then 10,000 (by altering the number of times that you execute the macro). Remember to reset the counter (k1=1) each time and to erase the earlier results before you run the macro again:

```
MTB> let k1=1
MTB> erase c2-c4
```

Record the simulated proportions in the appropriate rows of the table above. Also record the sample means in the last column.

(s) Comment on how close the simulated proportions come to the theoretical probabilities as the sample size increases. What does this reveal about the interpretation of probability as a long-term relative frequency?

(t) Are the sample mean number of matches all somewhat close to the expected value that you calculated in (k)? Do they tend to get closer as the sample size increases? Explain what this reveals about the interpretation of expected value as a long-term average.