

# The Many Faces of WJU

## Using our campus homepage to introduce the concept of a maximum likelihood estimate

Marc A. Brodie  
Associate Professor of Mathematics  
Wheeling Jesuit University

Mathematics professors at liberal arts colleges are frequently called upon to teach introductory courses in probability and statistics; I am no exception. Although many of my students are at best ambivalent about being in MAT105 (the only reason they are there is that “it’s required for my major”), I enjoy teaching it very much. It is a course that is undeniably relevant to the “real world,” and it is not difficult to find interesting examples to illustrate the concepts. One such example, found on many university websites, including ours at WJU, is described in this article.

While my main focus here is on the course-related concepts used in solving the problem posed to the students, my work on this problem led to a new integer sequence: 1, 2, 5, 8, 13, ... which was accepted by the Online Encyclopedia of Integer Sequences and added to its data base as sequence A111097.

When I arrived at WJU in the fall of 2005, I was given a new office computer, which had its internet connection homepage set to the WJU homepage, <http://www.wju.edu>. Each time I accessed the internet, in addition to the standard links to Academics, Admissions, Athletics, etc., I saw a picture of someone connected to our college and a “My Story” link, which lead to a brief vignette of the person pictured. Different visits to the college homepage resulted in a variety of pictures and stories. I was soon wondering, “How many stories are there?” Of course my question could have been answered by contacting IT Services, or perhaps even by exploring the website and finding a list of all the stories, but such approaches are neither interesting nor educational. Furthermore, in actual statistical practice one would not have these options.

The approach I will consider involves collecting data from a simple experiment, then using that data to find a “best estimate” of the number of stories. Specifically, the concept of a maximum likelihood estimator of a parameter is introduced, and the maximum likelihood estimate for our problem is computed. There were many options available for the experiment, but the one I chose was to refresh the webpage repeatedly, making a list of the faces seen, until someone’s face was seen for a second time. The number of different faces appearing was then used to estimate the number of faces on the website. This choice of experiment was ideal for an in-class demonstration because it was straightforward to explain, and because the subsequent work involved no probability ideas or calculations beyond the level of MAT 105.

After introducing my idea to the class, we got on the webpage and refreshed it several times, observing the sequence David Turkaly, James Zamor, Rev. Paul Stark S.J., and then James Zamor again. Although I was surprised to get a repeated face so quickly, it was fortuitous. To

observe such a small number of different faces during the in-class experiment meant that the subsequent calculations would be less cumbersome. The question was then posed to the class, “based on what we just observed, how many faces do you think are on the website?” The first response from a student was, somewhat predictably, “three.” This student was assuming that we had seen every available face before seeing a repeat—certainly a possibility. But perhaps there are more than three faces, and we did not see them all. There could be four faces. There could be 100 faces. Of course, any positive integer greater than or equal to three is a theoretically possible number of faces, but not all such numbers would be reasonable estimates, let alone a best estimate.

Presumably, few people would consider 100 faces to be a reasonable estimate, given that we only saw three faces before a repeat. On the other hand, many people would consider anywhere from three to ten faces realistic possibilities. Good intuition may help narrow our choices, but mathematics is needed to find a best estimate. Do not panic; the following two facts from elementary probability are all that we need:

1. If there are  $n$  possible outcomes in an experiment, and each of the  $n$  outcomes is equally likely, then the probability of any one outcome occurring is  $1/n$ .
2. The probability of a specific sequence of outcomes is the product of the probabilities of each of the individual outcomes, at each stage assuming the previous outcomes have occurred.

We are now ready to address our first formal question: “What is the probability that **if** there are in fact exactly three different faces, we would see them all before we got a repeat?”

In order to proceed with the computations, we must make some (reasonable) assumptions about the way the website works. We assume that each time a face appears it is independent of those faces that preceded it. That is, the fact that we have or have not seen someone’s face already does not affect the chances of that face showing up the next time we refresh the page. We also assume that each time we refresh the page, each face (no matter how many there are) has an equal chance of appearing. Given these assumptions, and assuming there are exactly three different faces, the probability that we would see all three faces before a repeat is:

$$\frac{2}{3} \cdot \frac{1}{3} \cdot \frac{3}{3} = \frac{2}{9} \approx 0.222.$$

(In the above computation,  $\frac{2}{3}$  is the probability that the second face observed differs from the first,  $\frac{1}{3}$  is the probability that the third face observed differs from the first two, and  $\frac{3}{3}$  is the probability that the fourth face observed is the same as one of the first three.)

Thus, **if** there are exactly three faces, **then** there is approximately a 22% chance of seeing three different faces in a row and then a repeat. An event with a 22% chance of occurring is not unusual; it is certainly conceivable that there are, in fact, only three faces.

We may then ask, “What is the probability of observing what we actually observed if there is in fact some other specified number of faces?” For example, if there are exactly four faces, the probability that we would see exactly three different faces followed by a repeat is:

$$\frac{3}{4} \cdot \frac{2}{4} \cdot \frac{3}{4} = \frac{9}{32} \approx 0.281.$$

Similarly, if there are exactly 100 faces, the probability that we would see exactly three different faces followed by a repeat is:

$$\frac{99}{100} \cdot \frac{98}{100} \cdot \frac{3}{100} = \frac{14553}{500000} \approx 0.0291.$$

After working through such computations in class, the students were asked to understand that if there were in fact 100 faces, what we actually observed (a repeat after exactly three different faces) was fairly unusual, with a chance of only about 3% of occurring. We should then agree that 100 faces would not be a good estimate. After some further discussion, we decided that we wanted to find the number of faces,  $k$ , which would result in the largest answer to the question, “If there are exactly  $k$  different faces, what is the probability that we would see exactly three different faces followed by a repeat?” Thus, we want to find the maximum value of the function:

$$P(k) = \frac{k-1}{k} \cdot \frac{k-2}{k} \cdot \frac{3}{k}$$

where  $k$  is an integer greater than or equal to 3. The value of  $k$  which produces this maximum probability is called the **maximum likelihood estimate** for the number of faces. With calculus unavailable as a tool in this course to analyze  $P(k)$ , we used Excel to do computations for many values of  $k$ , and sought out the maximum by inspection. The results were:

| Number of faces, $k$ | Probability of seeing 3 different faces followed by a repeat if there are $k$ faces |
|----------------------|---|
| 3                    | 0.222222  |
| 4                    | 0.281250  |
| <b>5</b>             | <b>0.28800</b>  |
| 6                    | 0.277778  |
| 7                    | 0.262391  |
| 8                    | 0.246094  |
| 9                    | 0.230453  |
| 10                   | 0.216000  |
| ⋮                    | ⋮   |
| 98                   | 0.029682  |
| 99                   | 0.029391  |
| 100                  | 0.029106  |

The largest probability of 0.288 occurs when there are  $k = 5$  different faces. Therefore our maximum likelihood estimate for the number of faces is five.

Are there only five faces on the website? No. Could doing the example again result in a different maximum likelihood estimate? Yes. When I tried the experiment in my office before class for example, I observed eight different faces before seeing a repeated face on the ninth try. The maximum likelihood estimate in this case is the value of  $k$  for which

$$P(k) = \frac{k-1}{k} \cdot \frac{k-2}{k} \cdot \frac{k-3}{k} \cdot \frac{k-4}{k} \cdot \frac{k-5}{k} \cdot \frac{k-6}{k} \cdot \frac{k-7}{k} \cdot \frac{8}{k}$$

is largest:  $k = 33$ , with corresponding probability 0.0965. To reinforce these ideas, students were sent home with an assignment to run their own experiment, note the number of different faces seen, and compute the corresponding maximum likelihood estimate. While not all students were successful (the biggest problem was their inability to use Excel), the smallest number of faces reported seen before a repeat was three, and the largest was eleven. The corresponding maximum likelihood estimates were  $k = 5$  (as we have seen) and  $k = 62$ . Most students who were successful observed six, seven or eight faces. The idea that estimates are subject to such variability is central to understanding inferential statistics, particularly to putting any one estimate into a greater context. Having the students get many different estimates was helpful in driving that point home.

Beyond its usefulness in my MAT105 class, the problem at hand is interesting in its own right. I wrote some simple code in Mathematica (shown below) to compute the maximum likelihood estimate corresponding to the number of faces actually observed in the experiment for 1 through 20 faces.

```
thetable=Table[N[n/k*Product[(k-i)/k,{i,1,n-1}]],{n,1,20},{k,1,300}];
maximums=Map[Max,thetable];
maximumlikelihoodestimates={};
For[i=1,i<=Length[thetable],i++,
  maximumlikelihoodestimates=Append[maximumlikelihoodestimates,
    Position[thetable[[i]],maximums[[i]]]];
maximumlikelihoodestimates
```

The output from this code is:

```
{{{1}},{{2}},{{5}},{{8}},{{13}},{{19}},{{25}},{{33}},{{42}},{{51}},{{62}},
{{74}},{{86}},{{100}},{{115}},{{130}},{{147}},{{165}},{{183}},{{203}}}
```

and the results are summarized in the following table for easier readability:

| Number of observed faces | Maximum likelihood estimate, $k$ |
|--------------------------|----------------------------------|
| 1                        | 1                                |
| 2                        | 2                                |
| 3                        | 5                                |
| 4                        | 8                                |
| 5                        | 13                               |

|    |     |
|----|-----|
| 6  | 19  |
| 7  | 25  |
| 8  | 33  |
| 9  | 42  |
| 10 | 51  |
| 11 | 62  |
| 12 | 74  |
| 13 | 86  |
| 14 | 100 |
| 15 | 115 |
| 16 | 130 |
| 17 | 147 |
| 18 | 165 |
| 19 | 183 |
| 20 | 203 |

It is natural to ask (for a mathematician, anyway) whether or not the resulting integer sequence 1, 2, 5, 8, 13, 19, ... is already known, either from this context or another. As it turned out, this sequence was not recognized by the On Line Encyclopedia of Integer Sequences, which is found at <http://www.research.att.com/~njas/sequences/index.html>. I submitted the sequence and am now the proud “owner” of sequence number A111097.

Postlude: How many faces *are* on the website? It turns out there are 22. One of my students, who was trying to avoid having to do and understand the assignment, did indeed discover a list of all available stories on the webpage. No credit.