Guitar HeroTM is a series of music video games in which players use a guitar-shaped controller to simulate the playing of lead, bass guitar and rhythm guitar across numerous rock music songs. Players match notes that scroll on-screen to colored fret buttons on the controller, strumming the controller in time to the music in order to score points, and keep the virtual audience excited.

Among the many pieces of information the game collects from a player is which notes are successfully “hit”. Those who have played the game before know that different parts of a song tend to be more difficult than other parts. However, one may pose the question as to whether or not the difficult parts of a song actually result in a higher proportion of misses than the other parts (as these difficult areas tend to have far more notes in general to play). We are interested in developing a method to determine if there is evidence that certain portions of songs result in higher proportions of misses than others.

To this end, this project will be split into several components where we will devise a way to answer the question, “Are misses occurring more frequently in parts of songs?” (This is henceforth referred to as the *research question*.) The project steps or components (with tentative time line) are:

1. Develop a method to answer the research question (due 2/8)
2. Data Collection (2/4)
3. Write an R function to compute a value for your methods (due early March)
4. Hypothesis Testing (due late March)
5. Power and Comparison of Methods (due early April)
6. Poster Presentation (4/29)
7. Written Report (due late April)

The first step in this process is to devise a method than can be used to answer this question. Here is some information that might be useful for this step.

1. If misses are at random, then each note would follow a Bernoulli distribution with probability *p* (where *p* is the probability of missing a note). Further, these notes would be i.i.d. A possible notation to use for this is Xi~Bernoulli(p) where Xi = 1 if the ith note of the song is missed, 0 if it is hit.
2. If misses are not at random, then you might expect them to be “grouped” together in certain parts of a song (which may correspond to the more difficult parts of the song). That is, the X’s would not be i.i.d. Bernoulli r.v.’s
3. By treating hits/misses as 0 and 1’s, we can use the data as a numerical value instead of a categorical variable. This will make it easier to devise an estimator.

For this part of the project you will be expected to do the following:

1. Work in a group of 4 people (you and 3 others). I want you to have others to discuss/bounce ideas around with. You must tell me (via email) who is in your group. If you don’t have someone in particular you plan to work with, I will assign you to a group (send me an email to let me know if you are having trouble finding a group to work with, or if you are a group of fewer than 4 people).
2. Devise at least four methods that you believe could be useful in answering the research question. As there are many potential methods to use, you will need to explain to me why you believe yours will work. One thing to keep in mind is that you will eventually need to write functions in R that can compute an estimate from a set of observed data.
3. To aid in your development of your methods and to gather the data we’ll use later on for this project, we will be collecting data on Friday, February 4 2:30 pm – 5*ish*.
4. You should test your methods on the following three “songs” (each with 5 misses, the first two with 20 notes and the last with 17).

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Note (i)** | **1** | **2** | **3** | **4** | **5** | **6** | **7** | **8** | **9** | **10** | **11** | **12** | **13** | **14** | **15** | **16** | **17** | **18** | **19** | **20** |
| **Song A** | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| **Song B** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| **Song C** | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | - | - | - |

Ideally, your methods from part 2 will show that Song B is more grouped than Song A, and should be general enough that they can be applied to different length songs, such as Song C.

1. You will turn in your methods with your explanation of each and your test results on the practice songs. This needs to be typed (and shouldn’t be more than a page or two). If you plan on using MS Word, you may need to use Equation Editor to display your estimators if it contains mathematical notation. However, I encourage you to use LaTeX (a mathematical typesetting language), as it produces far better looking reports.
2. This part of the project is due Tuesday, February 8 in class. I suggest starting early in case you have difficulties developing your estimators.

Some tips that might be useful:

1. In addition to whether or not the note is a hit or miss (0 or 1), you might also want to think about which note (i.e., the value of “*i*” that a miss/hit occurs). For example, Song A can be represented as (1, 7, 9, 12, 17).
2. Remember that since we aren’t clearly interested in a particular *parameter*, we can essentially come up with one that we believe will be able to provide a numerical “answer” to our research question. Then devise a suitable estimator for it. Another way to think about this: devise an estimator and define the parameter to be the population equivalent.
3. Don’t concern yourself with having an unbiased estimator or other concepts like this. These would involve knowing the sampling distribution of our statistic, which may be too complex to answer as of now. Rather, focus on developing an intuitive estimator.
4. If you are having difficulties getting started, you might want to consider applying introductory statistics concepts (such as means, medians, correlation, etc.) that you have learned about in previous courses.
5. Again, there isn’t necessarily an obvious answer to this problem (and many solutions probably exist). My hope is that you will have fun coming up with something as you start thinking like a statistician.

*Example Method (which you can’t use!):  
  
Let Cmax (my method) denote the maximum number of consecutive misses in a song. It would (somewhat) work in the sense that in songs where all the misses are in the same region, there would be higher probability of a longer string of consecutive misses than where the same number of misses would occur at random in a song of the same length.*

*Song A: Cmax = 1*

*Song B: Cmax = 2*

*Song C: Cmax = 4*

*Note: I actually don’t think this is a very good method for this problem, but I just wanted to give you an idea of what could be done. An unofficial goal for the project should be to find a better method than this one (don’t worry about making that comparison just yet).*