Math 326 Project 1 Part A – Due Tuesday Feb 2 Spring 2010

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 pieces 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 first step in this process will be to devise an “estimator” 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 1s 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 with a partner. I want you to have someone to discuss/bounce ideas around with. You must tell me (via email) who your partner is, if you don’t have someone in particular you plan to work with, I will randomly assign you to another person in class. If you haven’t contacted me by class on Thursday, January 21, I will assign you a partner.
2. Devise an estimator that you believe is 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 a function in R that can compute an estimate from a set of observed data.
3. You should test your estimator on the following three “songs” (each has 20 notes with 5 misses)

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 |
| **Song C** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |

Ideally, your estimate from part 2 will show that (c) is the most “grouped” and that (a) isn’t (with (b) potentially in between the two).

1. You will turn in your idea with your explanation of it and your test results on the practice songs. This needs to be typed (and shouldn’t be more than a page). If you plan on using MS Word, you may need to use Equation Editor to display your estimator if it contains mathematical notation. I might be willing to give a small bonus if you use LaTeX (which is a mathematical typesetting language).
2. This part of the project is due Tuesday, February 2 in class. I suggest starting early in case you have difficulties developing your estimator.

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).
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.
3. Don’t concern yourself with the 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.
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 your introductory statistics course.
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 Estimator (which you can’t use!):*

*Let Cmax (my estimator) 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 estimator for this problem, but I just wanted to give you an idea of what could be done.*

Math 326 Project 1 Part B – Due Tuesday March 2 Spring 2010

Write an R function from your estimator(s). There are two parts to this.

1. I would like you to start this on your own. This may seem like a daunting task, but you will have several weeks to work through it.
   1. General Programming Guidelines:
      1. Understand the problem
      2. Work out a general idea how to solve it (I typically map out my program on paper first)
      3. Translate your general idea into a detailed implementation by:
         1. Writing out the whole program in a small number (1-5) steps
         2. Expand each step into a small number of steps
         3. Keep going until you have a program
   2. Note that in the projects folder on the T drive, I have placed a file containing the R code for the Cmax estimator I proposed in Part A. You might find it helpful when writing your function.

1. Any time after February 11, your group can schedule an appointment with me to hash out any problems and get your program fully functioning. Before doing so, I will expect that you have made an honest attempt of trying to work on your program first. You may also email me R code to receive help with errors and such.

After your program is functioning, there will be several small pieces remaining of the project. You will evaluate your estimator under several situations (which I will provide) and then apply your estimator by conducting a hypothesis test using the real data we collected while playing in the QRC. I anticipate that this project will be finished sometime in mid-to-late March (after Spring Break).

Further, at any stage of the project before its final due date, you will be allowed to redo any part. If at some point you think of a better estimator, you can change it. If you notice a better (more efficient) way to program it…you can change it. This means that I will hold off grading the project until the final version needs to be in and you will just receive credit for completeness at each stage. However, I reserve the right to dock points if I feel that you have not made an honest effort on a current stage by its deadline.

Math 326 Project 1 Part C – Due Tuesday March 23 beginning of class Spring 2010

Use the R function for your estimator from part B to conduct a hypothesis test for Songs A – C from the Part A of the project as well as each of the (real) songs in located in the Songs subdirectory in the Projects folder. You will probably need to use the bootstrap and should model your code after cmax\_boot.R

The following are the hypotheses you are testing.

H0: Notes are missed completely at random

HA: Notes are not missed completely at random.

(Note that some of you have developed a test that does involve a statement about a parameter. If you are doing this, try to write your hypotheses involving that parameter.)

What to turn in: At class on March 23, bring a list of your p-values for each song so I may collect the results.

Some helpful R code to read in the data . (Do something similar for the other songs.)

wolf=scan(‘T:/Ramler/Math326/Projects/SongsHungry\_Like\_The\_Wolf\_JP.txt’)

Part D of the project will be assigned shortly. Following that you will need to put it all together in a nice project report (more details will follow).

Additionally, those interested in doing a “Guitar Hero” poster at the Festival of Science should contact me by March 30.

Math 326 Project 1 Part D – Due Thursday April 1 beginning of class Spring 2010

Use the R function that calculates a p-value for your estimator to determine the (approximate) power under the several scenarios described below. We will use Monte Carlo simulation on 100 replicates of 11 different scenarios to approximate the Type II error (and hence power). You can model your R code after the file Cmax\_power.R

You should turn in a copy of the table provided to me by class on Thursday. Additionally, you should keep a copy for your records (to use for the final report). Further, be sure to understand why using Monte Carlo methods such as this can be used to approximate power of a hypothesis test for a given alternative hypothesis scenario.

Scenario 1: Negative Pairwise Correlations

i) Short Song

200 notes per song

Each song was generated using a probability model with:

Overall (i.e. marginal) Probability of missing a note is p=0.3

Correlation between note *i* and *i+1* for *i=1, …,199* (i.e. any 2 consecutive notes) is **= -0.3

ii) Long Song

Identical to the Short Song but with 600 notes per song

Scenario 2: Positive Pairwise Correlations

iii) Short Song

200 notes per song

Each song was generated using a probability model with:

Overall (i.e. marginal) Probability of missing a note is p = 0.2

Correlation between note *i* and *i+1* for *i=1, …,199* (i.e. any 2 consecutive notes) is **= 0.3

iv) Long Song again has 600 notes

Scenario 3: Autoregressive model of order 1 – denoted AR(1)

v) Short Song type 1

200 notes per song

Each song was generated using a probability model with:

Overall (i.e. marginal) Probability of missing a note is p = 0.1

Correlation between note *i* and *j* for *i,j=1, …, 200* (i.e. any 2 notes) is **= 0.5|i-j| (eg. Correlation between

notes 1 and 2 is 0.5, between notes 1 and 3 is 0.52=0.25 and so on)

vi) Long song type 1 has 600 notes

vii) Short song type 2

200 notes per song

Each song was generated using a probability model with:

Overall (i.e. marginal) Probability of missing a note is p = 0.1

Correlation between note *i* and *j* for *i,j=1, …, 200* (i.e. any 2 notes) is **= 0.3|i-j| (eg. Correlation between

notes 1 and 2 is 0.3, between notes 1 and 3 is 0.32=0.09 and so on)

viii) Long song type 2 has 600 notes

ix) Short song type 3

200 notes per song

Each song was generated using a probability model with:

Overall (i.e. marginal) Probability of missing a note is p = 0.2

Correlation between note *i* and *j* for *i,j=1, …, 200* (i.e. any 2 notes) is **= 0.5|i-j| (eg. Correlation between

notes 1 and 2 is 0.5, between notes 1 and 3 is 0.52 = 0.25 and so on)

x) Long song type 3 has 600 notes

xi) Songs of 200 random notes with p = 0.1.

Math 326 Project 1 Part D – Due Thursday April 1 beginning of class Spring 2010

Keep this copy for your records.

|  |  |  |  |
| --- | --- | --- | --- |
| Type I Error Rate 🡪  Scenario | =0.01 | =0.05 | =0.10 |
| i) Neg. Pair Corr, n=200 |  |  |  |
| ii) Neg. Pair Corr, n=600 |  |  |  |
| iii) Pos. Pair Corr, n=200 |  |  |  |
| iv) Pos. Pair Corr, n=600 |  |  |  |
| v) AR(1), p=0.1, **=0.5,  n=200 |  |  |  |
| vi) AR(1), p=0.1, **=0.5,  n=600 |  |  |  |
| vii) AR(1), p=0.1, **=0.3,  n=200 |  |  |  |
| viii) AR(1), p=0.1, **=0.3,  n=600 |  |  |  |
| ix) AR(1), p=0.2, **=0.5,  n=200 |  |  |  |
| x) AR(1), p=0.2, **=0.5,  n=600 |  |  |  |
| xi) Random notes, p=0.1 |  |  |  |

Math 326 Project Report Due Thursday, April 29, 2010

The very last part of this project consists of submitting a technical report concerning all aspects of the project. Since most of the statistical work is already done, you now need to write your findings and disseminate your results to the world (or at least me). This report must contain the follow.

Title (and please don’t use the words Math 326, Course, and Project in it. Pretend that you have done this for something outside of class)

1. Introduction:

Give a brief description of the problem and the goals in solving it. You should also introduce the dataset(s) that we are using. You should also provide a brief outline of the remainder of the report in the final sentence of the introduction.

1. Methodology:

Here is where you discuss the development of your estimator and how to use it to conduct the hypothesis test we are interested in. This is probably the most important part of the project, so don’t skimp on the details. You may use Songs A, B and C as examples to illustrate your procedure in this section.

1. Power Study:

This section should contain a description of the power study from part d of the project, report the results (in a table) and describe your findings.

1. Application:

In this section, you will apply your methodology to the actual songs. Be sure to provide some type of description of what your results say about the hypotheses and the song(s).

1. Discussion:

This is essentially the conclusion to your report. Summarize the major findings/aspects of your method, describe scenarios where your method works well and discuss any shortcomings to your method. You may also describe any further ideas you have or other scenarios (which may not just include Guitar Hero) where you believe your methodology could be useful.

Appendix:

Please submit the R code used to

1. Calculate your estimate
2. Obtain a p-value
3. Compute the approximate powers

A few notes:

1. Your project will NOT be graded on how well your methodology performed for the actual songs and scenarios used in the power study. There are many different types of scenarios possible; it could be that your method works well for something we didn’t specifically investigate.
2. Instead, you will be graded on how well you communicate your ideas to me (this does include using correct spelling and grammar). It is safe to assume that the intended audience would be someone who has had Math Stats, but would not be familiar with the specifics of what you did. My general recommendation would be, if in doubt, assume someone doesn’t know what you are talking about and thoroughly describe it.
3. You must use section headings. I want to clearly be able to see where each section starts.
4. There is no minimum or maximum page limit. Long manuscripts will not automatically be considered better than shorter reports. If you feel comfortable that you have included everything that is necessary, then whatever length it is will be fine. However, that said, please try to keep it below 10 pages ☺.
5. A small bonus (up to 5% extra credit on the project) will be given to those who use Latex to write their report. Latex is a great typesetting system for scientific reports, and I would consider this report to be of that nature. If you need help with Latex, please feel free to ask.
6. Remember to make full use of what you wrote in part A of the project as I suspect that much of Section 2 will be very closely related to it.
7. Thanks for your willingness to do this project. I personally found it quite fun and I hope you had fun doing it (and of course found that it had high educational value!). If you have any suggestions for how to improve this project, or other projects that could be based on video games, please let me know so I can continue to use this for future classes.