Rocks in your head data analysis project

BRIAN J. WINKEL
Department of Mathematical Sciences, United States Military Academy, West Point, NY 10996, USA
e-mail: BrianWinkel@usma.edu

(*Received 9 December 1996*)

We present a data collection and analysis project which permits individual creativity and group work, non-linear curve fitting, parameter selection, and comparison between analysed groups. This class activity has been done with and without the software discussed in the paper. The project has been done with calculus students and could be done with any group of students familiar with functions, plotting data and perhaps logarithms.

1. Introduction

We all learn best when we learn actively. In a calculus course we have used a project in which students collect data, analyse the data, and formulate conclusions about different groups of subjects. The project can be done in one or two hours of class time with about a week to collect the data outside of class. Students decide how to collect the data, how to analyse it, and how to compare their analyses for different groups of subjects. In this paper we describe the project.

The requisite mathematics for this project is familiarity with functions, plotting data, and perhaps logarithms. The approach we offer involves the computer algebra system, *Mathematica*. One could use any other such system, a spreadsheet, calculator technology, or, indeed, plotting by hand. The technology is not the issue. The opportunity to challenge the students and give them a window for creativity in their mathematics classroom is the essence of this project.

1.1. *Equipment needed*

- 5–10 small rocks or stones (use stones of the same density, try to get them smooth without jagged edges, and make them in the 50–150 g range);
- 100 g mass from an old science lab scale set (use this as a standard mass or in lieu of the lab piece you can use a rock of 'known' mass as the standard with which you compare the other rocks);
- showbox or small container to hold the rocks and the standard mass;
- pencils, attached by string to the table or box, and forms (described below) which are used to anonymously collect data;
- table on which to set out this equipment (table should be in a public area where students, faculty, and staff can see it, but in a 'protected' area where it would not be subject to 'vandalism');
- 'ballot box' used to collect respondents' forms.

1.2. *The experiment*

The goal of this project is to determine how good people are at distinguishing between different masses. Accordingly, set out some 5–10 rocks in the 50–150 g

range and a standard mass of 100 g. Label the rocks in random order, $1, 2, 3, \ldots, 9$, say. (Do not label them in increasing mass order!) Identify the standard mass and its actual mass for the respondents.

Ask participants to indicate whether or not the sample rocks are of greater mass then the standard mass. Provide a slip of paper in which to ask them some information for later references, e.g. their major, their position (student-undergraduate or graduate), faculty and discipline, administrative, staff, etc. Do not ask for names unless you wish to have a contest for 'best' guesses.

The real information is respondent opinion on whether the mass of the rocks, numbers $1, 2, 3, \ldots, 9$, are more than the standard mass. Provide a form such as the following:

Sample rock number	1	2	3	4	5	6	7	8	9
Check if mass >100 g									

Set the collection table out for whatever time is reasonable to get sample data—one week will be sufficient especially if there is a sign, with words like 'Rock-A-Rama', 'Let's Rock', 'Science Experiment', or 'Which Major is More Perceptive', and some catchy graphics.

Once the data are collected, compute the percentage of respondents who say that a given rock has mass greater than the 100 g standard mass. Plot this percentage against the actual mass of the given rock.

If respondents had perfect knowledge, e.g. a scale, the plot of the returned data would look like the plot in figure 1.

If all or more respondents guessed then the plot of the returned data could look something like the plot in figure 2. A reasonable population with nine sample rocks, e.g., 10, 40, 70, 90, 100, 120, 125, 150, 175 g might yield the plot in figure 3.

1.3. *Statement of the problem and first actions*

Fit a curve (or curves) through the response data, perhaps attempting to differentiate groups with a parameter or feature in the graphs we obtain for different groups. We might even wish to rank the groups based on this criteria, e.g. can geology majors differentiate rock mass better than political science majors?

Students will always, always plot data first, which is a sign of good modelling skill that we work hard to inculcate in them. In our first class period devoted to this

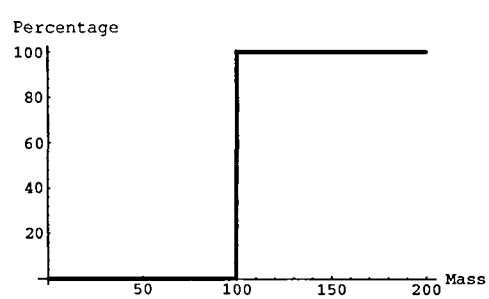

Figure 1. Plot of responses from respondents who had perfect knowledge. Percentage of those who say rock of Mass has more mass than 100 g mass.

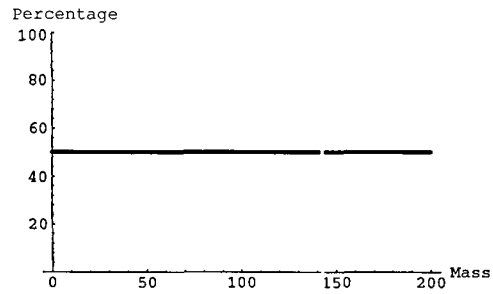

Figure 2. Plot of responses from respondents' guesses. Percentage of those who say rock of Mass has more mass than 100 g mass.

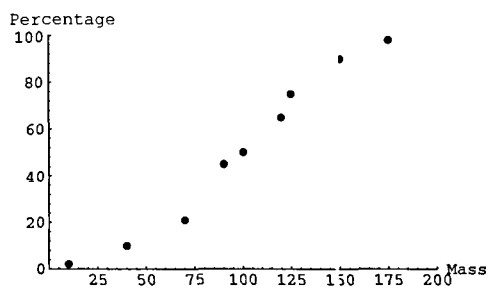

Figure 3. Plot of typical responses from respondents. Percentage of those who say rock of Mass has more mass than 100 g mass.

project, we have the students conjecture what the data response might be. Based on their understanding of the possible variables (mass of rock, percentage who say mass of rock is more than 100 g standard mass) and their response to figure 1 the students are convinced that the data should be S-shaped. Moreover, they quickly convince themselves that the more vertical the central section about the 100 g standard mass is, the better the ability of the respondents to distinguish masses.

Thus the hunt is on for S-shaped curves. A first model students often conjecture is some form of the *arctan*, perhaps, $f(x) = \arctan(a(x-b))$ with constants a and b to be determined by some best fitting criteria. Incidentally, one could have a discussion about such criteria at this point or offer up minimizing the sum of the squares between the observed data and the predicted data from the model for the data points gathered, the so-called least squares approach. We find that students, through minimally guided discussion, arrive at the 'natural' least squares criteria.

Depending on where the class is in terms of topics, this project might be either an introduction or a summary project to a unit on the minimization of a function of several variables, e.g. the parameters a and b in the model function offered above. The function here is the sum of the squares between the observed data and the predicted data from the model for the data points gathered. Or, if you trust your software sufficiently, it might be an introduction to fitting functions to data with little understanding of the underlying minimization algorithm.

Some of the S-shaped curves include

$$f_1(t) = 1/(1 + a\exp(-bt))$$
$$f_2(t) = 1/(1 + at^b)$$
$$f_3(t) = a\exp(b/t)$$
$$f_4(t) = \arctan(a(x - b))$$

It should be noted that students generally go after functions which are asymptotic to 1 as t increases and asymptotic to 0 as t decreases. Some even insist that the function go through the point $(0, 0)$ arguing the fact that it makes no sense to have values of $t < 0$ where t represents mass of sample rock. This insistence is usually dropped in the face of attempting to get a function to do too many things.

Several students will suggest a polynomial (a cubic at least), but others will try to argue them out of this in lieu of the flattening which they deem important at either end of the mass of rock scale, a flattening which a polynomial does not possess. Thus polynomials are usually not assumed. One reason for pushing a polynomial is the fact that in software such as *Mathematica* it is easy to fit a polynomial, say of degree 3, to a data set `data`. One uses the command `p[x_] = Fit[data,{1,x,x^2,x^3},x]` to find a third-degree polynomial $p(x)$ which will fit the data.

1.4. *Timeline and student action issues*

This activity can be done as a long-term project over several weeks from formulation through collection and on to analysis or short term exercise if the collection devices are in place. We have used groups all the time and we permit the groups to go their own way and develop their own analysis within the context of class sharing and discussion so that all benefit from each group's wisdom. The class time devoted to this project is a wonderful time to bring misconceptions and issues out in the open. Indeed, we have found doing such complex problems with students to be a rich time to assess students' capabilities and attitudes [1].

This is a sure starter as students form quick convictions as to what they believe will happen and they like the idea of designing a data collection station and collecting data from their classmates. There is the slowdown when they come to the following: 'We have the data. Now what do we do?' This is where the teacher comes into play, guiding the discussion to bring out as many ideas as possible from the students. Collectively the students have the necessary ideas to do an excellent job. They will need technical guidance, technology assistance, and continual encouragement to complete an analysis and summarize their conclusions.

2. Analysis of the data for selected model

Having selected one of the four models above for example, there are several approaches one can now pursue. Traditionally we often 'linearize' both model and data and then try to fit a straight line through the data, either visually or with some linear regression package, perhaps on a calculator or in some computer package. Certainly *Mathematica* makes this easy with the command `l[x_] = Fit[ldata,{1,x},x]` to fit a linear function $l(x)$ to the linearized data set `ldata`.

We offer a table of the linearized forms of the functions (table 1). Keen and Spain did a nice job of offering up possible function models for fitting to data [2, p. 56].

$y = 1/(1 + a\exp(-bt)$	$\ln\left(\frac{1}{y} - 1\right) = \ln(a) - bt$
$y = 1/(1 + at^b)$	$\ln\left(\frac{1}{y} - 1\right) = \ln(a) + b\ln(t)$
$y = a\exp(b/t)$	$\ln(y) = \ln(a) + b\frac{1}{t}$
$y = \arctan(a(x - b))$	$\tan(y) = at - ab$

Table 1. Linearized models for use by students in fitting data.

Rock mass	10	40	70	90	100	120	125	150	175
Percentage	2	10	21	45	50	65	75	90	98

Table 2.

To apply this linearization approach students have to convert the data from the form $\{(t_i, y_i) | i = 1, 2, \ldots, n\}$ to a linearized data set. If the model used is $f_1(t) = 1/(1 + a\exp(-bt))$ the linearized data set will look like $L = \{(t_i, \ln(1/y_i - 1)) | i = 1, 2, \ldots, n\}$. When we plot this linearized data from L we might well see a linear trend, obtaining $-b$ as its slope and $\ln(a)$ as its vertical axis intercept, thus obtaining a and b easily.

In a course where there is access to *Mathematica* one can form the sum of the squares between the observed data and the predicted data from the model for the data points gathered as a single function of the two parameters a and b in the model. From this one can minimize this sum of squares with *Mathematics*'s `FindMinimum` command.

We consider a typical data set one might get back for a sample of nine rocks set out with a 100 g standard mass for comparison. This data set is plotted in figure 3 and table 2.

In *Mathematica* we can enter the data as:

```
data={{10,.02},{40,.1},{70,.21},{90,.45},{100,.5},{120,.65},
                        {125,.75},{150,.9},{175,.98}};
```

If we had chosen model $f(t) = f_1(t) = 1/(1 + a\exp(-bt))$ we could form the sum

$$h(a,b) = \sum_{i=1}^{9} (f(t_i) - y_i)^2$$
$$= \sum_{i=1}^{9} (1/(1 + a\exp(-bt_i)) - y_i)^2$$

In *Mathematica* we would type:

```
h[a_,b_] = Sum[(f[data[[i]][[1]]] - data[[i]][[2]])^2,
                                    {i, 1, Length[data]}]
```

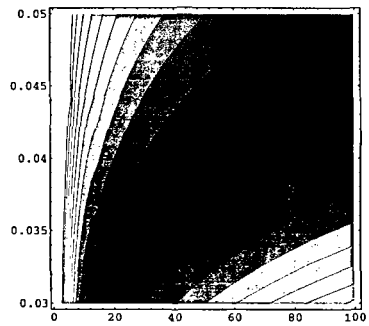

Figure 4. Plot of sum of squares function $h(a, b)$ in the interval $0 < a < 100$ and $0.03 < b < 0.05$. The dark region contains the lowest values of the function and hence a good place to look for the minimum value for $h(a, b)$.

Here `f[data[[i]][[1]]]` is f applied to the first coordinate of the ith data pair, i.e. $f(t_i)$, the predicted value under the model f, while `data[[1]][[2]]` is the second coordinate of the ith data pair, i.e. y_i, the observed value we wish to compare to the predicted value in the least square sense.

We now need to determine the parameters a and b which make $h(a, b)$ minimum. Continuing in *Mathematica* we could get a look at the contour plot of this function to attempt to see where a possible minimum might occur in order to start our root finding algorithm off with a good seed (figure 4).

Finally, we could apply the `FindMinimum` command with reasonable starting points to get values of a and b which minimize $h(a, b)$.

```
sol1 = FindMinimum[h1[a,b],a,40,b,.02]
```

We would obtain the following:

```
0.00601426, a -> 53.9126, b -> 0.0401633
```

where 0.006 014 26 is the actual minimum value of $h(a, b)$, the sum of squares, for values $a = 53.9126$ and $b = 0.040\,163\,3$.

Thus our best fitting function of the form $f(t) = 1/(1 + a\exp(-bt)$ is actually

$$f(t) = \frac{1}{1 + 53.9126\exp(-0.040\,163\,3t)}.$$

We show this best fitting curve ploted over the data for 'eye' confirmation (figure 5).

We could perform the same analysis with this data set for other models and compare our fits, perhaps comparing the sum of the squares as returned by *Mathematica*. An interesting discussion arises when we compare the fits we obtain from the linearized and the original data set fits. They are not the same! We consider the question, 'Which is better?'

3. Comparison between populations

We turn to the problem of a number of different populations for whom we have grouped responses, e.g. geology majors, mathematics majors, and English majors.

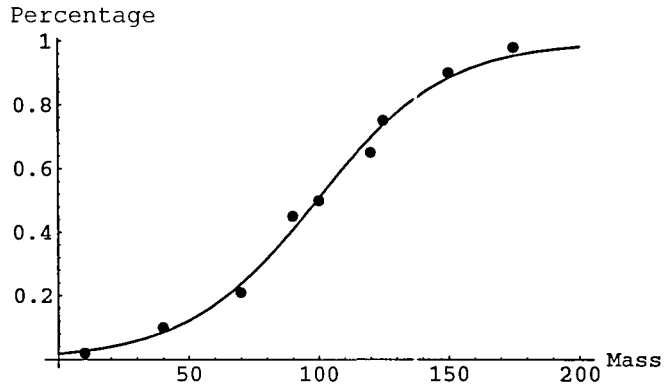

Figure 5. Plot of our best fitting function $f(t) = 1/(1 + 53.9126 \exp(-0.040\,163\,3t))$ for our data set with the data itself.

Students are asked to compare the groups and perhaps determine which group is 'best' at distinguishing masses. Often the parameter which is offered, assuming the same model is fitted, is the corresponding least sum of squares. In seeking other parameters for comparison, one could go back to seeing how far from perfect each group is by simply taking the sum of squares between observed and 'perfect' values (figure 1) and comparing these numbers. Finally, one could use the 'steepness' in the centre of the fitted plot, i.e. find where the first derivative is maximum and compare maximum first derivative values. This idea comes from the fact that the perfect knowledge model (see figure 1) has 'infinite' steepness at the 100 g mass.

4. Conclusion

We have presented a project which permits data collection and analysis, individual creativity and group work, non-linear curve fitting, parameter selection, and comparison between groups. Students' opinions are valued and they are required to back them up with analysis and communication.

Source and acknowledgement

A *Mathematica* notebook and an ASCII version of this problem, with solution and comments is available (under title ROCKRAMA), as part of a larger National Science Foundation project effort, 'Development Site for Complex, Technology-Based Problems in Calculus', NSF Grant DUE-9352849. This notebook, along with other problem sources material developed with support of the grant, is available on the World Wide Web under the address: http://www.rose-hulman.edu/Class/CalculusProbs. Web site preparation was done by Dr Aaron Klebanoff, a colleague at Rose-Hulman Institute of Technology.

References

[1] Winkel, B. J., To appear, Formative Assessment During Complex Problem-Solving, Group Work in Class. *MAA Notes on Assessment Practices in Mathematics.* Washington DC: Mathematical Association of America.

[2] Keen, R. E., and Spain, J. D., 1992, *Computer Simulation in Biology: A BASIC Introduction* (New York: Wiley-Liss).