Category Theory and Statistics Working Group

Paul Marriott

July 25, 2020

1 Comments on 'What is a statistical model'

This informal set of notes I wrote to help me understand the notation that's being used in the McCullagh (2002) paper. I'm just going to write down some very simple, basically classroom, examples and see what the notation being used in his Section 3 would corresponds to. And just to be clear, examples here are stylized because they are classroom examples, but they are not the strawman example of the type being used in the paper. There's nothing here which should be surprising to a statistician and so he wouldn't have included them in the paper as everyone in his audience we have these examples in their minds. Having said that, classroom examples are not fully representative of real statistical practice, and it will be interesting to explore those differences later.

In Section 3.1 he sets up the notation naming objects which are very familiar to all undergraduate statisticians, and provide the framework for the linear model. Let's keep everything very concrete and at a level that might be taught in an introductory statistics course.

We start with a question about the real world and see if an empirical investigation can help us learn, at least something about the answer. The question of interest does *not* come from the statistician, but from a client or collaborator. Our job is to pass the result of the analysis back.

Example 1.1. *Suppose we were interested in the relationship between mathematics students' mark in a calculus course, x, and their final mark in their first statistics course y. There are different possible questions that you could ask about this relationship.*

1. *You might be interested in the predictive power of x on the mark that the corresponding student gets for the statistics course. [Often prediction is in term of $E(Y|x)$, but we need a random structure to define this.]*
2. *You might be interested in testing a hypothesis. An old professor of statistics might have claimed that you don't need calculus to study statistics and you're trying to test, through an empirical experiment what evidence says about that claim.*

3. *You might be interested in quantifying the strength, and type, of the relationship between x and y, assuming that there is one.*

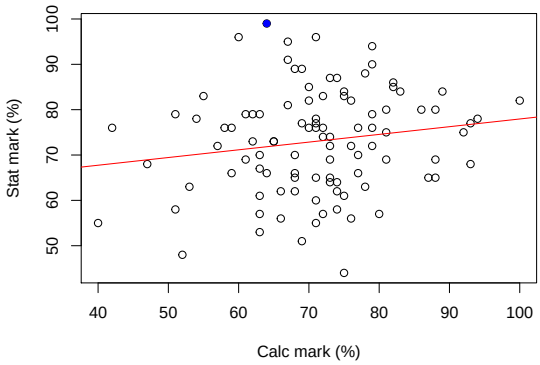

Figure 1: Classroom dataset

In an introductory class you might start by plotting some data, see Fig. 1, and computationally fitting some sort of linear regression model before trying to explain how the data has relevance to the questions above above.

We always have to ask where the data came from - I regard this as part of the design. Of course in teaching we typically simulate to get the figure that makes the pedagogical point we want.

Here we might have told the students that the marks came from $N = 100$ (fixed by design) students randomly and independent selected from a well defined population ($\mathcal{P}_{\mathbf{study}}$), say all students from U Waterloo who have complete both courses. However there are many choices here depending often on the question of interest.

The parametric model we might introduce at this stage is simple linear regression

$$Y|\{x, \alpha, \beta\} \sim N\left(\alpha + \beta x, \sigma^2\right) \tag{1}$$

and the fitted line, $\widehat{\alpha} + \widehat{\beta}x$, is shown as the line in Fig. 1.

In McCullagh (2002, §3.1) he starts defining notation which we quickly summarise it here. The *set of statistical units* is denoted by $\mathcal{U}$, the set of things that we might have measured or controlled on $u \in \mathcal{U}$ is called the covariate space, Ω, the particular variable of interest is called the response, y, chosen in light of the underlying question. This is measured on a selected scale denoted by $\mathcal{V}$. Further, the sample space $\mathcal{S}$ contains all possible values of y. The design $x : \mathcal{U} \to \Omega$ maps units to covariates (this really matters for those covariates which are controlled by the experimenter). The statistical model is a

choice of function $P : \Theta \to \mathcal{P}(\mathcal{S})$, where P(S) is the set of all probability distributions on S. This choice is also influenced by the objective of the study.

In Example 1.1 we would then have:

1. $\mathcal{U}$ could be a set of students, or a set of student numbers and their associated academic records. **Question: This only has a finite set structure in category theory?**
2. Ω would be the set { (student number, calculus mark) } as a pair. **Question: category theory structure?**
3. The y associated with a student is the set { (student number, statistics mark) } as a pair. **Question: category theory structure?**
4. The design, x, is simple here since no covariates that are controlled by the experimenter – its an observational study. The design is just the process of looking up student records, but I would include information on the way that the data arrived to us. **Question: category theory structure?**
5. $\mathcal{S}$ would be $[0, 100]$, or more precisely $\{0, 1, \cdots, 100\}$, as percentages. **Question: category theory structure?**
6. $P : \Theta \to \mathcal{P}(\mathcal{S})$ is defined by by Equation (1), and as noted by McCullagh does indeed depend on x.

Example 1.2. *This example is the classroom stylised version of a* causal analysis *and its an experimental study. The question of interest is to investigate if a new treatment for a disease, T, is better than the current best, CB, in terms of one year outcome. The outcome is treated here as binary: recovered, (R), or not recovered (NR).*

A sample of patients with disease, of size N (fixed by design), was recruited from a hospital. They were randomly allocated to received, T or CB. The experiment ran for a year and the final counts (i.e. the data for analysis) has the structure shown in Table 1

	Recovered	Not Recovered	Counts
New Treatment	N_{11}	N_{12}	N_{1+}
Current Best	N_{21}	N_{22}	N_{2+}
Counts	N_{+1}	N_{+2}	$N = N_{++}$

Table 1: Results of RCT.

Data in contingency tables such as Table 1 are analysed by Generalised Linear Models (GLM). These were championed by McCullagh and are the main workhorse of applied statistics. In this case a simple multinomial regression would be used, say with the `glm( )` *function in R.*

In Example 1.2 we would then have (with the same questions as above)

1. $\mathcal{U}$ could be a set of patients selected for the study
2. Ω would be the set { (anonymised i.d., Treatment Type) } as a pair

3. y is the set { (anonymised i.d., Outcome) } as a pair.
4. x, the design maps randomly selected units to treatment type.
5. $\mathcal{S}$ would be the binary set {R, NR}. **Note a more refined study could have an ordered set of outcomes, better to worse. Would this have a category theory structure?**
6. $P : \Theta \to \mathcal{P}(\mathcal{S})$ is selected to be a GLM, depend on x.

In his §3.2 McCullagh looks at the *inferential universe*. We can explore what this might mean in our Examples 1.1 and 1.2. In the classroom situation we would first explain that we are not interest just in the units in the sample but in some much larger *population*. Here are some brief comments:

1. The inferential universe should be part of the consideration of the question of interest. It is part of the first things we would be thinking about at the start of the analysis.
2. It often is not completely well-defined, or has a nominal definition
3. In Examples 1.1 Question 1 is really about the behaviour of future cohorts of maths students at the same University
4. In Examples 1.1 Question 2 might be more general and part of a consideration of the relationship between maths and statistics. This single study can only be part of a much larger study.
5. In Examples 1.1 Question 3 could be the same as 2.
6. In Examples 1.2 the data might have been collected from a hospital in Toronto, but we might really be interested in how well the treatment does for all possible patients with the disease. This single study can only be part of a much larger study
7. **Key question: How does the category theory concept of a *natural extension* play a role here?**