

Elements of Experimental Design

Dick DeVeaux

Williams College

I Why Experimental Design?

What is (are?) Data?

Data are everywhere. And everyone wants to use them. As sensible as this seems, the process of turning data into information is not always simple. In fact, sometimes there is no information at all. Or, even worse, we can be lead to the wrong answer. A recent medical study was undertaken to determine which risk factors are most important for predicting heart attacks in adult males. Having mined a large database, the researchers found a strong negative correlation between men's height and the incidence of heart attacks. The alarm went out that short men face a much greater risk for hheart attacks than taller men. Now, it's certainly true that the correlation between shortness and incidence of heart attacks for males is very high. But, why? It turned out that the researchers had neglected to consider one other piece of information that just might be relevant - namely, their ages. Men over 70 tend to be shorter than men under 30, sothe correlation analysis was unable to distinguish the effect of height from the effect of age. And the researchers mistakenly pointed to height as the culprit.

Humans are very good at explaining patterns in data. Probablytoo good. It is easy to imagine some of the explanations for the connection between height and heart attacks -- short men have shorter blood vessels and therefore higher chance of clotting, short men are under more stress to prove themselves, *etc. etc.* The problem is that *many* effects are plausible. We can reject some explanations on first principles, or when the conclusions are obviously silly. But data analysis in the real world is not always so kind. Imagine a marketing study finding a similar connection between short men and sales of a new product. After the data analysis, the recommendation might be to develop a co-marketing relationship with an elevator shoe manufacturer, when, in fact, a hearing-aid vendor would have been the better partner!

The problem of turning large amounts of data into useful information is an important and, at times, difficult task. Sometimes there is extremely valuable information hidden in data, but one has to be extremely careful inassigning *causes*. Often, two variables are related because they are both caused by another variable, a so-called *lurking variable*, an example of which we saw with men's heights and heart attacks. In that case, age

was the lurking variable, responsible for both the variation in men's heights and the increased risk of heart attacks.

Why not analyze data?

Rather than try to interpret happenstance data, we 'll discuss and analyze data from designed experiments instead. Here, the conditions during the experiment are designed to be the same, or at least dealt with in such a way as to make fair comparisons between treatments. Unlike historical data, the input variables in a designed experiment will be *actively manipulated* by the experimenter rather than be allowed to vary on their own. This manipulation distinguishes designed experiments from observational studies.

This seemingly simple distinction is not immediately grasped by everyone. In my experimental design classes, I always have the students conduct their own experiments at the end of the semester. Many years ago, one student wanted to study the relationship between demographic factors such as parental income and parental education on the performance by high school students on SAT tests. While this is a perfectly reasonable question, can you imagine the *experiment* that one would have to perform? How can the input variables be *manipulated*? I was having difficulty getting this student to see the difference between an analysis of an observational study and a designed experiment. I kept asking what the experiment was, what factors was he going to manipulate. The answer I got repeatedly was that the demographic variables would change, thus enabling him to measure the effect of the variables on performance. It wasn't until I told him that he would have to manipulate the parent's incomes himself that he finally got the idea!

When the factors are not actively manipulated, it's sometimes better just to ignore the historical data. Why? Here's a story to illustrate. Several years ago, I received a call from a large, well known, chemical company just before Christmas. The engineer told me that they had been manufacturing a polymer at her plant for nearly 30 years. For the first ten years or so, because they owned the patent, life was relatively easy as they enjoyed an environment with very little competition. During the next ten years or so, even after the patent expired, other companies had to play catch up, so they still maintained the lion's share of the market. During these years, a vast folklore developed on how best to produce the polymer. For this polymer, the molecular weight, or viscosity determined its quality. The job of the control engineers was to keep the process running and to keep the viscosity within certain limits. If it fell outside the limits, the batch was not sold. But all was not lost. If the viscosity was too high or too low, the batch wasn't scrapped, but was

Introduction

kept in storage until another blend of compensating viscosity (its “evil twin”) was produced by accident. At a production rate of 12,000 pounds an hour, with only 30% (!) of the product meeting spec, you can start to imagine the storage problems. Even worse, a large customer discovered that the polymer was being blended in this way and started looking for another vendor. The incentive to make even small improvements on this “production opportunity” was great. The engineer told me that for every percentage point in spec increase translated into about \$1,000,000 per year in increased profit.

My job, according to the engineer, was “simply” to look at the vast amounts of data that had been collected from many different batch runs, and to discover a model for the production process. From this model, the company would then be able to pick the levels of the input variables needed to optimize the process and target the viscosity. She assured me that there was no dearth of data --- I could have access to as much as I wanted. Not only that, but there were as many inputs measured as I could want: nearly 6000 of them!!

After more discussion, though, it turned out, that the vast majority of these variables were either the same input measured at different time points, or variables that had been constructed as ratios, differences, or other transformations of a set of common inputs. After a half hour of fairly intense negotiation, we managed to pare the list down to about 23 essential inputs, collected at various time intervals. Unfortunately, there was nothing we could do about the fact that the viscosity could be measured only at 4 hour intervals since it involved a time consuming lab measurement and measuring it more often wasn't economically feasible.

After hearing all of this, I told the engineer that I wasn't willing to even look at the data. A long silence followed, after which she asked if I understood that the company was willing to pay me to perform the analysis. I assured her that I, but that I would hold my ground and refuse to analyze the historical data. Why? Because, from past experience I suspected that in such a process, analyzing historical data would lead to many spurious correlations, much like the short men and heart attacks, and very little, if any, usable knowledge about controlling viscosity. However, both because she was very insistent, and to prove a point, I finally agreed to spend a limited amount of time analyzing some data. To make the knowledge discovery process as simple as possible, we focused on two very different production runs -- one that produced very good polymer, and one that produced an extremely bad batch, for each of two different plants.

The controllable input variables used in the model included things like flow rates, pressures, temperatures, levels, and amperages. In addition, there were five other variables having to do with properties of the raw input materials; important, but uncontrollable. Armed with every tool in the data mining toolkit, I went at the data with a vengeance. I gave them my best shot; I lagged them, transformed them, checked for outliers, high influence points: the works.

(We'll discuss all of this later in the book).

Two days later I called the engineer back, asking whether she wanted the good or the bad news first. She chose the good, and I was able to tell her that for each time period, I could explain nearly 80% of the variability in viscosity with my best model (giving an R^2 value of .80). She was thrilled, unable to imagine the bad news, which was that the models for the four time periods were completely different, containing completely different input variables. Not only that, but each model was incapable of being used effectively for any other time period other than the one on which it had been based. In other words, each model could explain its own past reasonably well, but was useless for predicting the future.

What happened?

The above scenario is an example of what is sometimes referred to as PARC analysis, a wonderful acronym that stands for Practical Accumulated Record Comparisons. These analyses are typically performed to answer questions after the data have been collected. They're certainly practical, since, as my engineer pointed out, the data are there anyway, with no further effort needed. But, as we saw, such analyses often amount to little. This has inspired another explanation of the same acronym: Planning After the Research is Completed.¹

It is tempting to use historical data to make comparisons, and to draw conclusions from them. The problem is that there is no guarantee that conditions are comparable across the different periods. In our polymer example, the plants are exposed to the open air where ambient temperature can range from well below zero in the winter to the mid 90's in the summer. But ambient temperature was not one of the input variables that I got to consider.

¹ The term PARC analysis is due to Stu Hunter

Introduction

What we wound up with from our analysis of the four different production runs was four different models for the same process with almost completely different explanations (input variables). Most likely, we've developed four different models for the noise, rather than the signal. This is worse than useless. Worse, because someone might act based on one of these models and believe that by changing the inputs to certain settings, the output will behave accordingly. To test the models, we used each model to predict the other three time periods. The results were *worse* than I would have gotten just by guessing the historical mean viscosity for the entire run.

But how could we have four different models each explaining the same phenomenon, each with a credible R^2 value? Among other problems:

- The conditions were different
- Other important variables were not measured -- for example, ambient temperature
- The predictors are modeling noise
- Two inputs were varying similarly, and the model chose to keep only one

The trouble is, we don't know which of these might be the case. Usually we analyze data from one time period and don't have the benefit of comparing models across different times to see that they're all equally worthless. To summarize the dangers of this type of analysis, the usual value of PARC analyses can be found by considering what the acronym PARC spells backwards!

After she calmed down, I suggested that perhaps we wanted to start a process of experimental design, investigating systematically which factors affect viscosity. For the next six months we planned, carried out and analyzed a series of experimental designs on the production plant. Because this was an operating plant, and not a pilot plant or simulation, our changes were small and certainly not radical. But, with enough observations, we were able to distinguish the signal from the noise. After six months, we increased the average in spec percentage from 30% to 45%. Certainly not perfect, but the increased \$15,000,000/year profit was appreciated.

The philosophy and implementation of experimental design are crucial in process improvement.

For readers familiar with the work of W. E. Deming, this is consistent with the cycle known by the acronym PDCA, for Plan, Do, Check and Act. This cycle (known as the Deming or Shewhart cycle) serves as a model for process improvement. The techniques taught in this book form an integral part of all four components in the cycle. We hope to make this clear throughout the remainder of the book.

So what *is* this Book About?

In this book we'll layout both the basics of experimental design and the associated techniques for analyzing data from these experiments. We'll be practically motivated, pointing out what to watch out for when using data from experiments to make real world decisions. Sometimes the real world can get complicated and the idealized world outlined in text books can seem far away. When that happens, we'll be careful to point out what can go wrong and if the solution is beyond the scope of this book, we'll point you in the right direction for further reading.

Calculating and Computing

There's no denying that Statistics does involve some calculations. These days, almost all statistics calculations and graphs are made with computers. Throughout the book we will discuss ways to use computers to perform the calculations – and we will concentrate on interpreting the results rather than on calculating them.