Cautionary Tales in Designed Experiments

David S. Salsburg

The correct bibliographic citation for this manual is as follows: Salsburg, David S. 2020. *Cautionary Tales in Designed Experiments*. Cary, NC: SAS Institute Inc.

Cautionary Tales in Designed Experiments

ISBN 978-1-952363-24-5 (Paperback)
ISBN 978-1-952363-23-8 (Web PDF)

SAS Institute Inc., SAS Campus Drive, Cary, NC 27513-2414

September 2020

Contents

Preface

When I was a child, two streets away from our home there was a large area of fallow land with a small stream, hollows where rain water would accumulate, small groups of bushes that tore at one's clothes and, in the spring and summer, a profusion of wild flowers. On a Sunday afternoon, while the grown-ups were doing "important" things, my older brother, Zev, and I would wander through these fields. Sometimes, we would dig up a wildflower and bring it home to plant in our yard. Other times, we would take soil from different parts, which had different colors, and mix them together with creek water to "see what would happen". We thought of ourselves as "experimenting." (Eventually, Zev went on to become a chemist, and I became a statistician, where I dealt with more serious "experiments.")

I did not know it, but we were "experimenting" in much the same way as natural philosophers had been doing since the time of the ancient Greeks. An "experiment" was when you perturb the "natural" order of things to see what might happen. In 1920, years before our "experiments" in that fallow field, something happened to this traditional way of "experimenting." This was the arrival of an irascible genius, named Ronald Aylmer Fisher, at an agricultural research station north of London.

This is a book about the revolution that he created in the nature of experimentation. Fisher's innovation was to require that the experiment start with a mathematical model of the relationships between things that will be used in the experiment and the observed and measured outcome of the experiment. I want to explain the statistical design of experiments to readers who have only minimal mathematical knowledge and skills. This is a difficult task because Fisher was a consummate mathematician. He had been named "senior wrangler" at Cambridge University. This was a title earned by the student with the highest mark on a difficult final math exam. Reading his research papers, one has to be ready for him to suddenly pull something from the arcane subject of number theory or prove some amazing result by invoking multi-dimensional geometry. Aspects of probability theory that Fisher used are now taught in graduate courses of advanced calculus.

Fisher's innovation has remade the world of science. In most fields, you cannot publish a scientific paper describing an experiment if you have not followed Fisher's dicta. In this book, I have tried to explain the statistical theory of experimental design by using examples of studies that have succeeded and studies that have failed. Many of these examples are medical studies because this is where most of my experience has been. I have tried to avoid mathematical notation. Although the use of mathematical symbols and methods clarifies any discussion, there are going to be readers for whom even two lines of algebra are impenetrable. Since I cannot use the precision of mathematical notation, my "explanations" are relatively vague. It is my hope that my examples give such readers an understanding of what is being described.

There is only one place in this book where I could not find a way around the mathematics. To explain Bayes' Theorem in Chapter 10, I had to resort to four lines of algebra. Bayes' Theorem falls out of the mathematical description of probability by invoking a symmetry in the notation. The result is completely unexpected, and I know of no other way to show where it came from.

Some readers will find another problem with this book. I describe experiments run on mice and rats. Some of them expose the animals to painful stimuli. Some of them involve inducing cancer and causing their deaths. We use mice and rats (and other rodents) because their physiologies are similar to ours, and we hope that what happens with a mouse is somewhat predictive to what might happen with a human. But they are also similar to us in other ways. They seem to have emotions. They appear to suffer pain. They are enough like us that we cannot help but feel that some experiments are "cruel." However, it is a fact that some of the experiments that best exemplify Fisher's models are done on living animals, and a great deal of this research has led to beneficial medical treatments for human beings.

So, here it is: Fisher's great innovation explained by starting with the circulation of the blood and the "good" in milk.

Chapter 1: Experimental Design from Harvey to Fisher

1.1 The Circulation of the Blood

In 1628, William Harvey (1578–1657) published the results of his investigations into the nature of blood and its circulation. He showed that blood was pumped by the heart to the rest of the body through large blood vessels (the arteries) and returned to the heart via smaller vessels (the veins). In one of his more spectacular experiments, he placed an evacuated glass column into the artery in the neck of a horse and showed how the column of blood moved up and down with each beat of the animal's heart. His careful observations established the nature of blood very clearly. He postulated the existence of small blood vessels that connected the two systems, although he could not find them.

Figure 1.1: William Harvey

Figure source: https://en.wikipedia.org/wiki/William_Harvey

There was a problem with Harvey’s careful work. The ancient Roman philosopher/physician, Galen of Pergamon (130–210 BCE), had stated that there were four fluids in the body that carried specific diseases. In addition to “blood,” his other humors were “phlegm,” “black bile,” and “yellow bile.” In the Middle Ages, it became the doctrine of the Catholic Church that it was impossible to produce new knowledge beyond what the ancient philosophers had written because the amount of knowledge had been steadily diminishing since the expulsion of Adam and Eve from the Garden of Eden. In fact, the Church had taken the works of philosophers like Galen as part of Church doctrine. Who was this William Harvey that he could challenge the established knowledge and doctrine?

About 100 years before Harvey, the Italian mathematician, Gerolamo Cardano (1501–1576) wrote a book that he called the Ars Magna, in which he described a new method of calculation that had been devised (which we now call algebra). In the introduction to the Ars Magna, Cardano states that he realizes that there is nothing new, but he has been unable to find these ideas in the works of the ancients. He is presenting this material in hopes that someone more knowledgeable would point out where this can be found in the works of the ancients.

No one ever found the methods of algebra in the works of the ancients, but one member of the Church hierarchy did find what was wrong with Harvey’s work. The good bishop noted that Harvey had come to his conclusions through experimentation, and, wrote the bishop, it is well known that Nature abhors experimentation and will purposely do things wrong if one attempts to experiment.

When I tell this story to experimental scientists, they tend to agree with the 17th Century bishop. They tell me about the experiments that went wrong. In a pharmacological experiment, a strange virus swept through the lab, and all the mice died. In a clinical study, the wife of one of the patients told the doctor that her husband had been flushing the medication down the toilet. A carefully laid out agricultural experiment was “knocked for six by some fool of a tractor driver hurrying home to his tea via a short cut across the plot” (Salsburg 2001).

What is an experiment? What is this process that Nature “abhors”?

This is a book about the use of statistical models to design and execute experiments. These models take into account the elements of experimental design that often lead to failing experiments. The nature of statistical experimental design will be developed in the following chapters with a careful examination of one of the first large statistically based experiments, the Lanarkshire Milk Experiment of 1930. There will be side trips looking at other studies (some that failed and some that succeeded). The only mathematics that the reader will need to know is elementary algebra. The more complicated aspects of experimental design will be referred to and described in general without using mathematical notation.

1.2 The Statistical Model of Experimental Design

The genius Ronald Alymer Fisher (1890–1962) was the first person to propose the use of statistical modeling to design an experiment, his classic text appearing in 1935 (Cochran and Cox 1992). Before Fisher, experiments were designed and executed at the whim of the experimenter. Gregor Mendel (1822–1884), for instance, planted rows of beans and peas, carefully examining the frequencies of plants with wrinkled peas, different color leaves, or some other characteristic that he found, which seemed to be influenced by specific aspects of the parent plants. He and his fellow monks counted and sorted, and he kept planting new seedlings based on what he had observed so far. The data that he displayed in his scientific papers are too perfect to be true. When things happen by chance with the probabilities that he proposed, the occurrence of specific inherited characteristics is seldom “perfect.” For instance, suppose a given trait is recessive, needing copies of the gene from both parents, then the probability that the offspring will have that trait is 0.25, but seldom will exactly one fourth of the offspring have the trait. There will be plantings where more than or less than one quarter have the trait. In Mendel’s publications, the counts that he displays are all exactly “correct.”

Was Mendel lying? Not exactly. In the 1860s when he published his work, it was common for scientists to display results “corrected” so that there were no random variations. The designs of reported experiments or sighting (in astronomy for instance) were idiosyncratic to the scientist. Results and the degree to which these results agreed with theory were the important thing, and experiments were seldom described in sufficient detail for someone else to be able to replicate them exactly.

Figure 1.2: R. A. Fisher in 1930

Figure source: https://en.wikipedia.org/wiki/Ronald_Fisher

Fisher was dealing with agricultural experiments, and he realized that the weight of potatoes, for instance, depended more upon the rainfall or the general fertility of the soil than it did upon the experimental fertilizer dressings being tested. Fisher proposed that an experiment could be described with a set of mathematical equations. The general idea was to describe the final measurement—call it Y—as dependent upon the various aspects of the experiment that could influence the outcome, something like Y = an overall effect uninfluenced by the irregularities of weather or soil

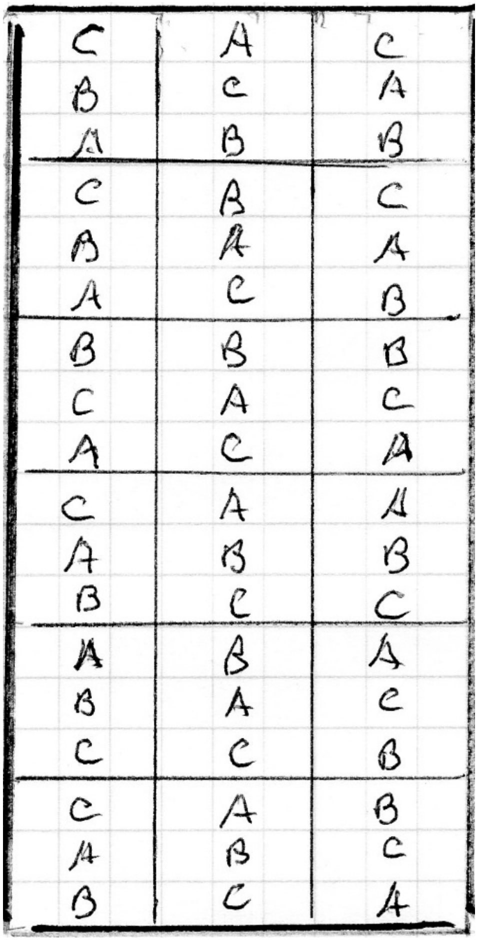
- the effect due to the treatment used in the experiment
- the effect due to the amount of rainfall that year
- the effect due to the general fertility of the land being used
- the effect due to weather conditions other than rainfall
- an additional effect (hopefully small) due to all the things we cannot account for

In addition to this algebraic formula, he proposed formulas based on calculus to describe the final additive effect (which he called the “error”).

When put this way, it is obvious that the experimenter needs to be able to estimate the other effects before she can estimate the treatment effect. Using such mathematical formulas, some possible reasons why experiments fail become clear. For instance, suppose the field is uneven and that one specific part of the field consistently produces a higher weight of potatoes than another specific part. If the experimenter puts one treatment on the first part and another treatment on the second part, then it is impossible to determine how much of the difference in output is due to the differences in treatment and how much is due to the differences in fertility between parts of the field. Fisher called a situation like this “confounding.” The fertility gradient of the field is confounded with the treatment effect.

Figure 1.3 shows a layout of experimental plantings in 18 small plots of land. Three treatments are being compared, and the position of each treatment's plants varies at random from plot to plot. We will examine what it means to assign treatments at random in Chapter 7.

Figure 1.3: Treatments Assigned at Random Within Blocks



C	A	C
B	C	A
A	B	B
C	B	C
B	A	A
A	C	B
B	B	B
C	A	C
A	C	A
C	A	A
A	B	B
B	C	C
A	B	A
B	A	C
C	C	B
C	A	B
A	B	C
B	C	A

In 1938, soon after Fisher published the Design of Experiments, his student William Cochran (1909–1980) spent a year visiting Iowa State University in Ames, Iowa. George Snedecor (1881–1974) had founded the first department of statistics in the United States and written the first undergraduate level textbook dealing with Fisher's methods and insights. (This was not an easy task since Fisher often assumed that his readers were as insightful as he, and parts of his papers were almost impenetrable by ordinary mortals). Cochran returned the next year to take a position on the Iowa State faculty. There, he worked with Gertrude Cox (1908–1978) to produce an undergraduate-level textbook that pulled together all of Fisher's work on the design of experiments and presented the reader with a group of specific designs and their interpretation. Cochran and Cox's *Experimental Design* (1992) became the "Bible" of statistical design of experiments. It has continued to influence experiments and experimental design well into the 21st Century.

Figure 1.4: Gertrude Cox

Figure source: https://en.wikipedia.org/wiki/Gertrude_Mary_Cox

Recall that, in Fisher’s development of experimental design, the first step is to create a mathematical model of the intended experiment. In the chapters that follow, I will examine aspects of experiments that become clear only when the experiment is described in terms of a mathematical model. To start this journey, we will look at one of the first major experiments that used Fisher's insights: The Lanarkshire Milk Experiment of 1930.

1.3 Summary

In the 17th Century, Harvey’s discovery of the circulation of the blood was challenged because “Nature abhors experimentation.” Until the early years of the 20th Century, scientists used idiosyncratic experimental designs, and published results were often ”corrected” to make the results fit exactly to the theory being tested. In the 1920s, R. A. Fisher introduced the concept of statistically based experimental design where the possible outcomes are described by a set of algebraic formulas and the random variation is described through the use of calculus.

References

Cochran, W., and Cox, G. (1992) Experimental Design. 2nd Edition, Wiley, New York.

Davis, R. F., et al. (2017) Bulletin 1177. “Designing Research and Demonstration Tests for Farmers’ Fields.” University of Georgia Extension, http://extension.uga.edu/publications/detail.html?number=B1177.

Fisher, R. A. (1966) The Design of Experiments. 8th edition, Oliver and Boyd, Edinburgh.

Salsburg, D. (2001) The Lady Tasting Tea. Holt, New York, pg. 258.

Chapter 2: Measuring the “Good” in Milk

2.1 Pasteurization and the “Good” in Milk

During the early development of the science of bacteriology in the mid-19th century, many dangerous microbes could be found in milk soon after leaving the cow. Louis Pasteur (1822–1895) found a similar growth of bacteria in the mash used for the production of beer. Pasteur suggested that the mash be heated to just under the boiling point in order to kill off the unwanted microbes.

This process, known as “pasteurization,” was also applied to milk in the first quarter of the 20th century, and it was shown to reduce the incidence of illness among children who were fed pasteurized milk instead of the untreated raw milk. As a result, there was a movement among medical authorities to push for legislation that would require that all milk be pasteurized before being sold to the public.

This, in turn, led to a backlash among people who believed that heating the milk destroyed some or all of the “good” in the milk. Because of this resistance, very few governments passed legislation requiring that milk be pasteurized. It was not until after the Second World War that many of the states in the United States of America passed such legislation. In 1973, the federal government of the United States began requiring that all milk sold in interstate commerce be pasteurized. As of this writing, most western nations require pasteurization of milk and allow for raw milk to be sold only under very strict safeguards to inhibit the bacteria in that milk.

In 1927, when the legislation to require pasteurization of milk was introduced in the British Parliament, there were many who believed that pasteurization removed the “good” in the milk. This opposition was enough to block passage. The Department of Health of Scotland decided to run a study to determine whether this were true.

What is meant by the “good” in milk? Language has at least two purposes. One is to convey information. The other is to convey emotion. These two uses often get mixed up in political campaigns or at football games where the audience applauds or even joins in the emotion-laden chant. Who, for instance, would want to destroy the “good” in the milk?

The investigators at the Scotland Board of Health had to find a way to measure the “good” in the milk or to count something unambiguously that reflects the “good” in the milk. Doctors Gerald

Leighton (1868–1953) and Peter McKinley (1901–1972), who ran this study, decided to look at the growth of children, some of them fed pasteurized milk and some of them fed raw milk. They could measure the growth of children by measuring the gains in weight and height over a fixed period of time. It would have to be a long enough time so that day-to-day fluctuations would not have a measurable effect. They could have chosen to look at the days of illness as a measure and used adults for the experiment, but this would have been a more variable measure since some adults would have had no illness in the period of the study. Similarly, for children, they decided not to use days absent from school due to illness.

One problem with measuring the increase in height and weight is that growth is affected by more than the milk consumed. It is very much a function of the child's economic class. Poor children (especially poor children in the 1920s and 1930s when fewer social services were available) tend to grow less and more slowly than children from more well-to-do homes. They would not want a study in which the children on one type of milk were from different backgrounds than those on the other type of milk. (Fisher would call this confounding the social status with the effect of the milk). A child's gain in weight and height is also very much a function of genetics. They would not want children of short parents to be compared to children of tall parents.

Oh, the things that can go wrong when you experiment on people! It is far easier to experiment on mice, and so, with the reader's permission, I will digress to look at experiments on mice. Although this might disturb some readers, mice are often used on drugs to establish safe dose levels and potential efficacy before exposing humans to them.

2.2 Experiments on Mice

When using laboratory mice, all the units of experimentation have closely similar genetics and are kept in the same environment. The strains of mice used in pharmacological and toxicological experiments are ordered from the breeder, who ships boxes of 24 newly weaned mice of the same sex from carefully bred dams of very similar genetic background.

The mice are gregarious creatures and would not last long if housed singly in cages, so the usual practice is to house 4–6 mice in a single cage. The mice in cages are kept in the same room under controlled climatic conditions. As a result, we have several cages of mice that are as identical as possible.

The only difference between cages is the treatments that we want to compare. Still, something might go wrong. A deadly virus might sweep through the colony. So, one or two cages are left without treatment. These are the "sentinel" mice. If they take sick, the entire experiment is aborted.

Figure 2.1: Caged Mice

Figure source: https://www.hrsa.gov/hansens-disease/research/index.html

In a typical test for cancer treatment, the mice are injected (in the paw usually) with live melanoma cancer cells. All of the mice have the same genetic background and the same conditions of living. The average time of death per cage is a measure of the efficacy of the treatment given the mice in that cage.

Are all the same? Is it only the treatment that differs? Perhaps, but average time to death is not a good measure. True, the conditions and background are as close as we can make them, but, even under these conditions, individual mice differ from one another. Suppose the cage starts with four animals in the cage. The mice are constantly moving around, jockeying for who will be the dominant animal. Then, the disease takes effect, and, one by one, the mice begin to die until there is only one mouse left. There is no competition. It does not have to move around, so it sits quietly. It lives on and on. Depending on differences that we cannot distinguish, some of these solitary remaining mice live longer, much longer than others. This final mouse can increase the average when all the other mice have much shorter lives.

There is another situation where seemingly identical mice differ. Lifetime feeding studies are used to look at the results of exposure to drugs or chemicals over a long period of time. The mice are usually housed four mice in a cage, in vertical racks that hold six cages. When the first such studies were run in the 1960s, they found that the mice in the top cages were dying earlier with a larger number of senile lesions. It looked as if there was something in the air about the higher cages, so some labs introduced laminar flow air conditioning to maintain the same air throughout the room. Still, the mice in the top cages continued to be more sickly.

The solution to this problem came when one of the toxicologists observed the lab tech who was filling the cages from newly arrived boxes of mice. He would open the top cage, reach into the box and pull out one animal, then reach down and pull out another, and so on, until he had four mice in the cage. He would close that cage, open the next cage and reach down for another mouse. What could be wrong with that?

What was wrong is that all 24 mice in the box were not equally healthy—they never are. Reaching into the box, the technician tended to pick up the least lively of the mice, the one that did not scoot around as he reached in. The next mouse was the least active of the rest of the animals. Once the technician had filled all six cages in the column, the box was empty, so the top cage in the next column included the least active mice from the next box…and so on. How does one deal with a problem like this? Fisher had the answer: randomize. Once the cages are filled, the six cages in a column are now randomly re-assigned to different levels.

Thus, even in the “ideal” experiment where all the test animals have the same genetic background and all are subjected to the same environment—even then, differences among the test animals can lead to an unexpected bias in the results. If this is true for lab mice, how much truer must it be for children who are all raised in different environments and have different genetic backgrounds?

Would it even be possible to design a study whose results would not be twisted in some way by the use of different children? Doctors Leighton and McKinley thought they had a way, which we will look at in the next chapter.

2.3 Summary

In the 1920s, attempts were made to require that all milk sold in the United Kingdom be pasteurized to prevent illness and death among children due to bacteria in their milk. This was opposed by those who believed that pasteurization took the “good” out of the milk, and the bill was defeated in the British Parliament. The Scotland Department of Health set up a study to determine whether pasteurized milk was less healthy than raw milk. In planning the study, they had to define what was meant by the “good” in the milk. They decided to look at the growth of children given either raw or pasteurized milk. The study had to take into account the many differences in children that would influence their growth, other than the milk that they drank. Mice provide a seemingly perfect example of experimental units that are of the same genetic background and have the same environment. However, even in these almost identical circumstances, differences in the mice can produce problems in design of the experiments. Two examples are given.

Reference

Salsburg, D. (1986) Statistics for Toxicologists, M. Dekker, New York.

Chapter 3: Designing the Lanarkshire Milk Experiment

3.1 “Man Is Not a Big Mouse”

Bernard Oser (1899–1995) was one of the founders of modern toxicology, and he once published an article entitled, “Man is not a Big Mouse.” Let us consider how experiments involving human beings differ from those involving mice and how this affected the design of the Lanarkshire Milk Experiment.

One advantage of experimenting on mice is that all the animals are genetically similar and all kept in the same environment. In a perfect experiment, we would like to compare two treatments, both being applied to the same subject (humans or mice). In fact, as we will see in Chapter 8, there is a statistical model developed by Donald Rubin (b. 1943) of Harvard for an experiment that postulates two possible outcomes for each experimental unit, one as a result of treatment A, the other as a result of treatment B. But before we look at Rubin’s very sophisticated model, let us look at how Leighton and McKinley approached the problem in planning the Lanarkshire Experiment.

What they really wanted were pairs of children who were as identical as possible so that one member of each pair would get raw milk and the other pasteurized milk. Nature has already given us pairs of children with the same genetic background and raised in the same environment—identical twins. This would mean finding households where they could make sure that the two types of milk were not intermixed in the ice box. (In 1929–30, few households had refrigerators but used ice boxes with blocks of ice delivered once a week.) This would call for a member of the family (usually the mother) to take on a careful, exacting task while continuing her normal duties of making meals, doing the washing, cleaning the house, and taking care of her other children. Besides, how many pairs of identical twins could they find? They knew they had to get a large number to be able to detect a slight difference in treatment effects. The calculation of how many experimental units are needed in a given experiment can be made, but only with mathematical ideas that were not fully developed until 10 years after the Lanarkshire Experiment. But, even without specific calculations, Leighton and McKinley realized they would need far more children than they might find in an identical twins study.

How do you choose children who are not identical twins but who are in something close to the same environment? The county of Lanarkshire in southern Scotland seemed like a good place to start. At that time, the county consisted of only small villages and farms. Choosing children from that county meant that the environment would be very similar for all of them. There would not be the extremes of wealth and poverty found in big cities. Also, the families living in Lanarkshire County had been there for many generations, and almost all were of Celtic background, so the genetics would not differ very much.

As Oser warned us in his famous paper, the children in Lanarkshire may have had similar genetic and environmental backgrounds, but they were not big mice. They would differ, some of them in fairly dramatic ways. How could they account for those differences?

Leighton and McKinley did not have the power of a modern computer to help them, but consider what they could have done with one.

In a paper published in 2017, Jose Zabizarreta and Luke Keele of Columbia University (2017) were trying to analyze several years of data from the National Educational Census of the nation of Chile. In the 1980s, Chile was ruled by a dictatorship, which imposed a number of laws implementing very conservative proposals. One of these laws used the nation's education budget to provide vouchers that students could use in public schools or in private schools. The National Educational Census collected student test scores on standardized tests, along with student names, gender, parent names, and the school attended. Zabizarreta and Keele used the data from 2003 to 2006 to follow students and determine their educational gains as a function of whether they were in public or private schools.

Leighton and McKinley did not have the computers available to Zabizarreta and Keele, who gathered all the data available on each student and used statistical cluster analyses (involving millions of computations) to group together students who were similar with respect to the data recorded about them. Although Zabizarreta and Keele did not have identical twins to compare, they identified small groups of students who were very much like each other but some of whom had been in public schools and some in private schools. (Their conclusions were that there was no indication that children sent to private schools did any better than those in public schools once children from similar backgrounds were matched against each other.)

Leighton and McKinley tried to do something like this with the primitive tools that they had. They paired off schools that had similar student populations. One school in each of a matched pair was provided with bottles, each holding ¾ of a pint of raw milk, to give to the children. The other school in the pair was provided ¾ pints of pasteurized milk. They were afraid to provide both types of milk to a single school because of the complicated logistics that would be involved.

In this way, they were able to mount a complex experiment involving 20,000 children.

3.2 Measuring the "Good" in Milk

Leighton and McKinley decided that the best way to measure the "good" in milk was to look at the gain in weight and height among the children through several months. They would measure each child's height and weight in February and in June. The difference in average height and weight gain would be used to compare children who had been drinking raw milk and those who had been drinking pasteurized milk.

Complications immediately come forward. How do you decide which children would be given which type of milk? How do you guarantee that the children assigned a particular type of milk would get what they were assigned? And, a very skeptical question, how do you know that milk, in either form, plays a role in the increase in weight and height if the children are free to eat other foods, including their normal intake of milk?

You cannot keep children from drinking milk as they normally would. The most you can do is to supplement their diet with additional milk (raw or pasteurized). There is an assumption behind the design of this study—that additional milk will increase the child's growth. This meant that they really had to have another group of children, those who did not receive an extra ration of milk. When you include a group of experimental subjects that receive none of the experimental treatments, you are including "controls." This, of course, makes the experiment even more complicated.

Then, there are the practical problems of running the experiment. Leighton and McKinley could not be at the experimental sites each day and supervise the distribution of the milk. They would have to depend on teachers or someone else in the schools to keep tabs on the children and the milk.

3.3 Adjusting for Differences among the Children

At this point, the purity of the experiment has to be changed to take problems like those noted in the previous section into account. The people living in Lanarkshire were very similar in economic status and genetic background. They would be even more similar if the children given different types of milk attended the same school. The possibility of mix-ups could be reduced if the children in one school received raw milk, while the children of another school were given pasteurized milk—but this would be at the risk of not having comparable subjects.

In the end, Leighton and McKinley decided that all the children in each school would receive either one type of the experimental milk or no extra milk (the controls). They, now, needed to "match" the children in each school. Then, the children given extra milk would be as similar as possible to the control children given no extra milk.

How do you match children within schools? They did not have the sophisticated clustering computer programs of 2017, and there are many variables to describe socio-economic status and genetic background. Do you look at the family income? Do you look at whether the family home has central heating? Do you look at the educational level of the parents? And, what do you do when two children have parents with the same educational achievements but vastly different family income?

We will look at these questions in the next chapter.

3.4 Summary

Designing a study to compare children given different types of milk runs into problems because children differ from each other in many ways that can affect their growth. Identical twins might be useful here if one twin is given raw milk and the other pasteurized milk, but there are not enough twins to create a proper-sized study, and the distribution of each type of milk would depend on a member of the family who would be handing out the milk. Leighton and McKinley decided not to match children within schools but to match schools. Modern, computer-based statistical clustering methods might be able to match children, but Leighton and McKinley did not have that tool.

Reference

Zabizarreta, J. R., and Keele, L. (2017) “Optimal Multilevel Matching in Clustered Observational Studies: A Case Study of the Effectiveness of Private Schools Under a Large-Scale Voucher System”, Journal of the American Statistical Association, 112, Number 518, pp 547-560.

Chapter 4: The Execution of the Lanarkshire Experiment

4.1 Matching the Children

February of 1930 was rapidly approaching. Leighton and McKinley had arranged for the ¾ pints of raw and pasteurized milk deliveries. They identified the schools in Lanarkshire County where the experiment would be run. They had decided to divide the children in a given school into two groups, one to receive an extra ¾ pint of milk per day (raw or pasteurized) and one to receive no extra milk and be used as "controls."

One problem that they had to face was how to make sure that any biases of the teachers would not interfere with the assignment of different types of milk. In a modern study like those sponsored by the National Institutes of Health (NIH) of the federal government, there is a standard procedure for "blinding" the assignment of treatments. In their report, Leighton and McKinley do not describe how they did this, but in a modern NIH-sponsored study, students would be assigned numbers, and the bottles of milk would have been labeled at the dairy with those numbers, randomly shuffled so that neither the teachers nor the students would know what type of milk was being handed out.

(As a sidelight on nomenclature, it is standard practice to refer to this type of procedure as "blinding"—for all except studies sponsored by the National Eye Institute. For such studies, the NEI requires that the procedure be referred to as "masking.")

These were not mice that they were experimenting on. They were children who came from different homes and had different genetic backgrounds. Some of them would fall ill during the school year, affecting their growth. Some would engage in greater physical activity than others. Some would be well-fed at home. Some would go to bed hungry. It says in the Talmud that when an earthly king stamps out coins, they are all alike. When the King of kings stamps out people in the image of Adam, they are all different. How much of an effect could the "good" in milk have on children's growth? Is it so small that it would be swamped by all these other factors? If Leighton and McKinley kept track of all these other differences in the children, could they find some way to "correct" for their influence? Let us look at a simplified version of their problem.

Suppose there were only two aspects of the experiment that affected a child's gain in weight: the milk treatment given and the child's socio-economic status. Suppose, further, that there are only two levels of socio-economic status—"poor" and "rich." One can write a mathematical model where the child's change in weight from February to June is influenced by the treatment (raw, pasteurized, or no milk) and by the socio-economic status (rich or poor).

If the schools where the children were given raw milk had only "rich" children and the schools where the children were given pasteurized milk had only "poor" children, then the differences in weight gain could be due to either the type of milk or the family status. There is no way of determining which. R. A. Fisher, the founder of modern statistical design of experiment theory, called this "confounding."

4.2 Fisher's Model

In Fisher's case, he was looking at 50+ years of agricultural experiments that had been run at the Rathamsted Agricultural Research Station north of London. The typical experiment in the Rathamsted archives would dress an entire field of wheat with an experimental fertilizer, and the output of that field would be compared to the output of the same field in previous years. Or it might compare the output of the treated field that year with the output that same year of a particular field that was always left without treatment. Fisher pointed out that the amount of rainfall each year had a major effect on the wheat output and that comparing the treated field with its previous year's output confounded the treatment with the difference in yearly rainfall. (Recall that by "confounding," Fisher meant that the two factors being confounded always occurred together.) If they compared the treated field with the "control" field in a given year, then the differences in fertility from one field to the other were confounded with treatment effect.

Suppose they were to apply both the experimental treatment and the control treatment to grain in the same field on the same year. The agricultural scientists claimed they could not do this because it was well known that different parts of a given field had different levels of "fertility." In fact, one of the procedures (used before Fisher) was to examine past experiences with a given field. Then they would determine a "fertility gradient," a direction and degree to which the field's fertility diminished. In some previous experiments, they had used the estimated "fertility gradient" to add or subtract from the actual yields of different parts of the field. (In one of his first published papers, Fisher showed that the whole concept of a simple one-directional "fertility gradient" was nonsense.)

Fisher proposed that they break up the field into small plots and use the fertilizers to be tested on different rows of plants in each of those plots. The differences in harvest output would no longer be confounded with yearly rainfall since all the treatments would have the same amount of rain. The fertility differences between different parts of the field would no longer be confounded with treatment since each plot was small enough to have a constant level of fertility for all the treatments within it. But, suppose there is a subtle fertility gradient that runs north to south. And, suppose you use treatment A on the plants on the north side of each plot and treatment B on the south side. Then, the accumulated fertility differences between north and south in all the plots would be confounded with treatment effect.

You do not know what the subtle fertility gradient is (if it exists). Therefore, you cannot counteract the north-south confounding by dividing the plot into east and west, since the gradient might run east to west, or some other pattern that mimics any orderly pattern of treatment that might be tried.

4.3 Randomization

Then came Fisher's genius. Don't apply the treatment in any orderly fashion. Randomly switch treatments from row to row among the different plots. Recall Figure 1.2 from Chapter 1. The different treatments in the University of Georgia example are randomly assigned to each block in different patterns in that figure.

In an intuitive sense, Fisher's randomization would seem to do the trick. There is no way the fertility gradient could shift and twist and turn about to keep up with different random arrangements of treatments. The use of random assignment of treatments has become a fundamental part of most statistically based experimental designs. Randomization is usually posed as a means of counteracting confounding effects, whether they are known about or not. In many large complicated experiments, like drug trials in human patients, attempts are often made to see whether the randomizations actually "worked," to see whether there are any patterns of other effects that might confound the treatment effect.

Fisher, ever the careful mathematician, showed that randomization is more than a clever way to fool the malevolent "Nature" seen by the 17th Century bishop who sought to prove Harvey wrong. Fisher showed that if the treatments are assigned at random, then the probabilities of error associated with the experiment can be calculated using a normal or Gaussian distribution. This is the famous "bell-shaped curve" that is part of elementary first-year statistics courses. Knowing that we can use the normal distribution means that we can calculate the levels of uncertainty associated with the conclusions of the study. All that is necessary is that the treatments be assigned at random.

For 60–70 years after Fisher proposed randomization, statistical analyses had to make do with these approximations even when the number of observations was not very large. This was because any attempt to calculate the exact probabilities was beyond the abilities of one statistician and a desk calculator. The modern computer has changed all of that. There are computer programs now that will calculate what are known as the "permutation probabilities" that result from random assignment of treatment. And, there are purists in the world of statistics who insist that permutation probabilities be used instead of their normal approximations.

In a college course on statistical design of experiments, a great deal of time is spent on the problem of confounding. In these courses, confounding is examined in terms of the sets of equations that define an experimental design. These can get quite complicated, especially when some treatments are tightly linked to possible confounding variables. Experimental designs for specific problems have been derived and published in specialized books for many fields like psychology and medicine.

To avoid the problem of confounding the effects of different types of milk and the effects of many other ways in which the children's growth might be affected, Leighton and McKinley chose schools whose student populations had similar background characteristics. They then assigned raw or pasteurized milk to different schools at random. To reduce confounding when comparing milk to controls, they instructed the teachers to assign students who would receive milk or not at random.

4.4 The Lanarkshire Experiment Begins

In February 1930, the children arrived in the schools that were participating in the study. They were weighed and their heights measured. Information about each child and his or her family was recorded on special forms. An indication was made whether the child would (or would not) receive a daily ¾ pint of milk. Today, in such a study, the information would be recorded on a computer file in a central data office. In 1930, sheets of paper bound into a booklet for each child were filled out.

This recording of data is an important part of any experiment, whether the data are written (often in pencil) in a lab notebook, entered into a computer, or entered by means of a scan. The human interaction at this point is subject to possible error. The lab assistant who enters the data in a lab notebook might have a handwriting difficult to decipher. The scientist who entered the data into a computer file might have used a touch-typing setting for his fingers, but, for a brief period, his right hand slipped over one key and "weight" became "weogjt."

I was once involved with a medical study where the patients were children for whom the dose of drug was calculated based on the child's weight. In one set of case reports, all filled out in the same handwriting, whenever the child had a Spanish surname, the same weight—29 kg—would be written down.

At another time, I was involved in a toxicological study of animal pathology. The pathologists filled out prepared forms by coloring in the ovals next to the specific pathology findings that they saw. These forms (often with spots of dried blood) were fed into a mark-sensing device to be turned into a computer file. Each morning, the mark-sensing device had to be "programmed" by running a specially marked card through it. After many days of usage, the programming card would often become torn. At one point, rather than go to the trouble of marking up a new programming card, the technician pasted the old card together with transparent tape. However, he failed to line the halves up correctly, and for the next several days, all the data entered into the computer were shifted from the appropriate position to the next one in that file.

Hopefully, the report forms used in the Lanarkshire study were filled in correctly, and the paper record was a true representation of what actually happened. We will see in the next chapter what happened after Leighton and McKinley sat down to analyze all that data from 20,000 children.

4.5 Summary

In designing their study, Leighton and McKinley had to find a way of matching children so that children given raw milk would be, as much as possible, like the children given pasteurized milk. However, there are many ways that differences between children can affect their growth. They could differ in family education, in family income, in whether there was adequate food at home or they went to bed hungry.

Fisher's development of experimental design in agricultural experiments led him to consider small plots of land, wherein rows of plants with different treatments would be planted in soil with the same "fertility." To ensure that there were no systematic differences confounded with the application of treatments, Fisher proposed randomly assigning treatment to different rows. Fisher also showed that random assignment provides a mathematical theorem that the random error of the study converges to a normal distribution as the number of plots increases. In the Lanarkshire Milk Experiment, pairs of matched schools were assigned to be either raw or pasteurized milk stations. In each school, children were randomly assigned to be given the extra ¾ pint of milk or no additional milk (the controls).

Chapter 5: The Results of the Lanarkshire Experiment

5.1 Gossett's Criticism

Look at Figure 5.1. It describes average heights of children from the Lanarkshire Milk Experiment. Before we pull it apart to see what it actually tells us, note that the most remarkable aspect is that the average heights of the children who were controls (no extra milk) is consistently greater than the average heights of children who were given extra milk. The lines for those given extra raw milk and those given extra pasteurized milk keep crossing and re-crossing, suggesting no difference between the two forms of milk.

Figure 5.1: Results of the Lanarkshire Experiment

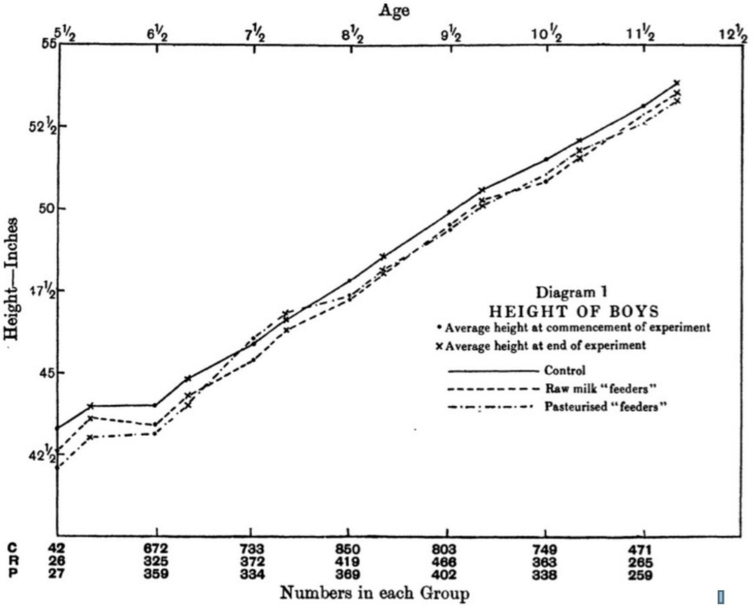


Figure 5.1 is from a paper written by William Sealy Gossett, whose published papers are attributed to "Student" (Biometrika 1931). The Lanarkshire Milk Experiment appeared to have been a failure. There was no consistent indication that children given raw milk fared any better than children given pasteurized milk. However, the study also seemed to show that the children who were given no extra ration of milk did better than those who received either raw or pasteurized milk. Gossett was sent by the Royal Statistical Society to examine the circumstances and the data associated with the Lanarkshire study.

Let us pull Figure 5.1 apart. In this display, the children are grouped by age. Along the top of the figure, the age grouping is defined. Along the bottom of the figure are the numbers of children given no milk, raw milk, and pasteurized milk in each age group. Each segment of the graph shows the average height of that group of children in February (at the start of the study), indicated by a "•". The end of each segment, indicated by an "**x**", shows the average height of that group in June. If we wanted to compare height gain in children given raw milk versus children given pasteurized milk, we would look at the difference, "**x**" minus "•", but Gossett chose this display because it illustrates very clearly the main problem that he found with the study. The children kept as controls were, on the average, heavier and taller than the children given extra milk of either type.

In 1933, Ethel Elderton published a paper in the journal, Annals of Eugenics (1933), in which she examined the data from the Lanarkshire Experiment in great detail and tried to locate groups of children who were very similar in initial height and weight in order to compare the three treatments. The better effect on the control children still held in her careful balancing. There was also some indication that the older girls on raw milk did better than those on pasteurized milk. However, a word of warning. Francis Anscombe (1918–2001), the long-time chairman of the Yale University statistics department, once wrote about what he called "will o' the wisps." If you look at a large amount of data (20,000 children, 27 schools, three treatments) and if you hunt long enough, you are bound to find interesting relationships that are nothing more than random glitches in the data and that have no predictive power for any other set of related data.

(Having mentioned Ethel Elderton (1878–1954), I should digress a little and describe this remarkable person. In 1849, Bedford College opened in London as the first college in the United Kingdom to provide higher education to women. Elderton was one of their most celebrated graduates. Sir Francis Galton had founded the Biometrical Laboratory in London to study the statistical properties of human measurements and health. He needed to engage in extensive and repetitive computations, so he hired women to be his "computers." One of them was Ethel Elderton, whom he hired away from her teaching position. She quickly became one of his most trusted assistants. When Karl Pearson assumed control of the Biometrical Laboratory after Galton retired, he also leaned heavily on Elderton to organize and interpret the extensive calculations that he needed. The pages of the two journals that he controlled, Biometrika and the Annals of Eugenics, contain many articles attributed to E. Elderton. A brief biography and references to more of her achievements can be found in the entry in the References.

5.2 What Went Wrong

What went wrong in Lanarkshire County, Scotland, in the winter and spring of 1930? Gossett traveled among the schools and asked questions.

A number of the teachers admitted that they had purposely assigned the extra milk rations to the children from poorer homes. After all, Lanarkshire consisted of farms and small villages where people had lived generation after generation and knew each other. The teachers were from the same towns and were also aware of which children came to school hungry. The Lanarkshire experiment did not test the difference between raw and pasteurized cow's milk. It tested the milk of human kindness.

Ethel Elderton noted that the children were weighed with their clothes on and speculated that the children from richer homes would have been weighed in February with more and heavier clothes than the poor children. However, that would have skewed the results in a different direction since the children from more wealthy homes would have been wearing heavy clothes in February and lighter clothes in June. Any effect of such confounding would have made the gain in weight for the controls less, suggesting that the difference between controls and treated was even greater.

5.3 Mismanaged Experiments

Studies involving people are often marred by unforeseen prejudices and errors. H. Fairfield Smith had been a student of Fisher's in the early 1930s. The late 1930s found him in British Malaysia, working for a rubber company that had plantations of rubber plants. He related to me the following story.

The rubber trees were tapped, and the sap drained out for a few days until the tap sealed up. Then the tree would be tapped in another spot. The question arose as to whether it was better to tap the tree on the north or on the south side. It was decided to select a stand of trees and tap each one on both the north and the south side simultaneously. The difference in the amount of sap recovered would show if one side or the other had an advantage.

Perhaps, someone suggested, the effects of direction are really based on an east/west comparison. So, it was decided to compare north, south, east and west. However, they did not want to tap a given tree in more than three places, lest the large number of taps harmed the tree.

North, south, east, or west, but only three on any given tree! Fisher's theories of the statistical design of experiments had, in fact, considered just such a situation. Furthermore, H. Fairfield Smith had published several papers on what were called "incomplete block designs." Using the theorems of abstract algebra that are needed to consider complicated designs like this, Smith drew up a plan that would enable the research team to test east versus west and north versus south. He prepared diagrams of tree trunks with the appropriate places of taps indicated, and the rubber tappers were sent out to the trees with bundles of these diagrams.

However, the rubber tappers were, for the most part, illiterate and had no appreciation for the subtle mathematics behind Smith's plans. All they understood was that they needed to put three taps on each tree. The result was a collection of data from trees, some of them tapped three times on the same side, some two on one side and one on another, and so on. Smith's incomplete block design was in shambles.

The data that he was given from the study consisted of drawings of circles with the place of the three taps indicated and with the output of each tap recorded. He drew an east/west line across the circles for each of the trees. If all the taps were on one side of that line, he discarded that data. If some taps were on both sides of the east/west line, he averaged the output of the taps on the south side and of the taps on the north side and took the difference as a single data point. He told me this story to illustrate that, in any sufficiently well-designed study, you can always aggregate data to a point where you have individual answers to the question posed and use the variability among those aggregates to estimate the level of uncertainty.

This is what Ethel Elderton did with the Lanarkshire study data. Her conclusions were that there were no substantial differences in the growth of the children whether they were given raw or pasteurized milk. The slight weight advantage of raw milk for older girls entering puberty could be one of Anscombe's will o' the wisps—so much for the "good" in the milk!

The teachers had been told to choose the children for the extra ration of milk at random. But they were never told how to choose at random. So, how do you choose at random? We will look at that question in the next chapter.

5.4 Summary

The Lanarkshire Milk Experiment produced a strange "finding." The children given no extra milk (the controls) were heavier, taller, and gained more in weight and height on the average than the children given an extra ration of milk. There were no clear differences in growth between the children given raw and those given pasteurized milk. William Sealy Gossett was sent to investigate why the controls (no extra milk) did so much better than the others. He found that the teachers had been told to choose the children at random for extra milk. However, many of them took pity on the children from poorer homes and assigned them to receive the extra ration of milk. Ethel Elderton tried to analyze the study by matching children within groups. An example of another botched study of tapped rubber trees showed how it was possible to recover the essential parts of a study as Elderton did, provided the study was properly designed.

References

Elderton, E. (1933) "The Lanarkshire Milk Experiment" http://onlinelibrary.wiley.com/doi/10.1111/j.1469-1809.1933.tb02098.x/pdf.

Rodkey, E. (2011) "Ethel Elderton" https://www.feministvoices.com/ethel-elderton/.

"Student" (1931) "The Lanarkshire Milk Experiment" Biometrika, Vol. 23, No. 3/4 (Dec., 1931), pp. 398-406. https://www.jstor.org/stable/pdf/2332424.pdf

Chapter 6: How to Randomize Experimental Units

6.1 Random versus Haphazard Treatment Assignment

How does one go about randomizing, anyway? The teachers in Lanarkshire County did not do it correctly. They gave the milk to the poorest children, thereby confounding the difference between milk drinkers and controls with the difference between poorer and richer families. If they were to assign milk at random, how should they have done it?

First of all, random assignment does not mean haphazard assignment. Humans tend to have orderly minds, so if they try to assign treatments at “random” without any structure, the haphazard result is bound to have structure and patterns in it. Let us consider what is meant by “random assignment.”

Suppose we could write down all the possible ways in which different treatments could be assigned to the children. “Random assignment” means that each of these possible ways is equally probable. If there are 1,000 different possible assignments, then we choose one with probability 1÷1,000.

How do you choose a set of treatment assignments that are drawn at random from all possible assignments? In a simple case, this is easy to do. Suppose you have four children and two treatments, A or B. There are six possible assignments:

ABAB
ABBA
AABB
BABA

Initially, to

BAAB

You can make a token for each of the six possible assignments, shake them up, pull one out, and use that assignment. However, if you want to assign a larger number of children to treatment at random, the number of tokens you will need to have for all possible assignments gets extremely large.

Another method that allows for all possible assignments to be equally probable is to start with a table of random numbers. Suppose we know that there are 100 or fewer possible assignment patterns. Suppose we have some means of choosing a two-digit number in such a way that all the numbers from 00 to 99 are equally probable. (This is called a "uniform distribution.") Initially, to get uniformly distributed random variables, R. A. Fisher went to the most recent census of Great Britain and looked at the populations of individual towns. He took the last two digits of each town's population and listed them in a table. It can be shown with a mathematical proof that the least significant numbers in a series of population counts are uniformly distributed.

For instance, here are the 2014 populations of eight towns in the state of Connecticut:

Andover 3,272
East Hartford 51,033
Monroe 19,867
Sherman 3,671
Ansonia 18,959
East Haven 29,044
Montville 19,635
Simsbury 23,975

The last two digits in this set of number are

72
33
67
71
59
44
35
75

If there are any duplicate numbers in the set, one of each duplicate is thrown out. It can be shown that numbers chosen in this way come from a set of uniformly distributed random numbers from 00 to 99

Fisher's table sufficed for much of his early work, but you cannot use the same sample of random numbers over again without destroying the theoretical properties of numbers drawn from that table. In order for the different experiments to be properly "randomized," Fisher needed another and larger table, so he and his student Frank Yates (1902–1994) went to a table of logarithms. They took the 10th through the 19th digits of each logarithm until they had 15,000 digits from 0 to 9, which they put into pairs. (The logarithm of a number, N (denoted log(N)), is the solution to the equation: $10^{\log(N)} = N$.)

Several other tables of random numbers were published for the next 50 years, culminating in a large volume generated by the Rand Corporation (2001). Using the time between emissions of beta rays from a radioactive substance, they constructed one million random digits arranged in groups of five. Martin Gardner (1914–2010), whose column discussing mathematics appeared in Scientific American for many years, called this the “epitome of the 20th Century.” Not only did no other century produce such a book, but no one in a previous century would have ever thought of making such a book.

The Rand tabulation of a million random digits has an introduction, explaining how to use the book. You do not open it “at random.” After all, the first few uses of the book would have broken the binding. Any future attempt to let the book fall open where it might well bring you back to the same region. Here is how one should use the book.

1. Open “at random” and choose a number “at random.” Suppose that number is 47174.
2. The first three digits, 471, tell you what page to go to.
3. The next digit, 7, tells you to start in the 7th row of numbers on that page
4. The next digit, 4, tells you to go four numbers into that row and start there.

Only then are you ready to have a truly random set of numbers.

6.2 Using Random Numbers to Assign Treatments

Now that we have a sequence of random numbers, how do we use that to assign treatments at random? This is best described in an example. Suppose we want to assign three treatments to 21 children. Let the treatments be identified by P (for pasteurized), R (for raw), and C (for controls). We start with 21 random numbers (which I drew out of a table in a widely used textbook by Dixon and Massey (1969)):

36
43
31
84
78
41
13
82
25
69
46
38
04
01
70
73
87
92

47
67
11

Write symbols for treatments next to each number.

36 P
43 P
31 P
84 P
78 P
41 P
13 P
82 R
25 R
69 R
46 R
38 R
04 R
01 R
70 C
73 C
87 C
92 C
47 C
67 C
11 C

Then order the numbers, smallest to largest, along with treatment assignments:

01 R
04 R
11 C
13 P
25 R
31 P
36 P
38 R
41 P
43 P
46 R
47 C
67 C
69 R
70 C
73 C
78 P

82 R
84 P
87 C
92 C

Thus, as the children enter the classroom, the assignments are R R C P R P P R P P R C C R C C P R P C C.

That is how you randomize assignment of treatments, not haphazardly, not with an eye toward the poverty of the child, but strictly following a random sequence of uniformly distributed random numbers.

We now have all the tools needed, the mathematical model, the identification of blocks within which to apply treatments, the avoidance of confounding, and the use of random assignment. We can now go out and run a properly designed experiment. Except that there are times when…but that is the subject of the next chapter.

6.3 Pseudo Random Numbers on the Computer

In actual practice, the Rand Corporation's book of a million digits has been superseded by computer programs called Random Number Generators. Of course, the computer is unable to "create" anything that is truly random. It can only follow the instructions of the programmer, which are determined (not random) in advance. A computer program that can generate uniformly distributed random numbers or engage in any other complicated analysis is called an "algorithm." It produces what are known as "pseudo random numbers." For readers who are interested in such details, section 6.4 of this chapter describes how a pseudo random number generator works.

There are a couple of problems with these pseudo random numbers. Once the sequence of "random" numbers repeats a number that has been seen before, the generator repeats all the numbers that came after it. It cycles. In fact, for many random number generators, the cycle can be quite short, 5–10 numbers. Using mathematical number theory, it is possible to set a lower bound on how many numbers a generator will produce before it cycles. Random number generators that are used in carefully constructed software are usually set to stop before that lower bound is reached.

There is another problem with pseudo random number generators. Mathematician George Marsaglia (1924–2011) plotted successive pairs of pseudo random numbers and discovered that the plots fell into orderly parallel lines. Nobody knows what this means because the pseudo random numbers met all the conditions of uniform random variables. But it was disturbing to have this regularity hidden in the sequence. A "solution" to Marsaglia's problem is what is known as a shuffle random number generator. In a shuffle generator, there are two sequences of pseudo random numbers. One is used to generate the random numbers that will be used. The other generates a number from 1 to 10 at random. When the program is initialized, the main generator creates 10 random numbers, which are stored in slots labeled 1 to 10. When called upon, the program generates a new random number and a number from 1 to 10. The new random number is stored in the slot determined by the second generator, and the number that was in that slot is used. This destroys the regular pattern that Marsaglia found.

6.4 How a Pseudo Random Number Generator Works

A random number generator starts with two large numbers that are relatively prime, that have no factors in common. Prime numbers are numbers that have no proper divisors other than 1. The first eight prime numbers are 3, 5, 7, 11, 13, 17, 19, 23.

The number 21,505 is the product of 5, 11, 17,and 23. The number 5,187 is the product of 3, 7, 13, and 19. Since they share no divisors in common, division of one of them by the other will produce a remainder infinite decimal. In this case, if we divide 21,505 by 5,187, we get 4.14979757… The first five digits of the remainder, 14979 is the first "random number." This first random number, 14,979, is then divided by 5,187, yielding 2.88779641… and our second random number, 88779. In turn, 88,779 is divided by 5,187, and so on. For each random number generator, specific pairs of starting numbers, called "seeds," have been extensively tested to be sure the random numbers that they produce have a uniform distribution and lack any correlation between successive numbers. Useful seeds have been determined for each random number generator.

6.5 Summary

Randomization of students to treatment starts with a table of random numbers, numbers whose order is purely random. Early tables of random numbers used the final digits of populations of English towns. Fisher and Yates took the central digits from pages of calculated logarithms. The Rand Corporation produced a book of one million random numbers based on times between emission of beta rays from a radioactive substance. Modern computers use pseudo random number generators that are dropped before reaching a cycling point.

This chapter shows, from an example, how to use random numbers to establish random assignment of treatments in an experiment.

References

Dixon, F. J., and Massey, W. T., (1969) Introduction to Statistical Analysis, 3rd Edition, McGraw-Hill.

Rand Corporation, (2001) "A Million Random Digits with 100,000 Normal Deviates (reprint) American Book Publishers, New York.

Chapter 7: When Randomization Cannot Be Done

7.1 Observational Studies

William Cochran (1909–1980), co-author with Gertrude Cox of the first textbook on design of experiments, creator of Cochran's C test, Cochran's Q test, and Cochran's Theorem, was chairman of the newly formed Department of Statistics at Johns Hopkins University from 1949 to 1957. The public housing authority of the city of Baltimore was concerned that placing families in public housing projects might lead to breakup of those families. There was anecdotal evidence that putting families in the "sterile" environment of public housing has caused breakups of some of those families. The housing authority developed a questionnaire that measured the degree of family cohesiveness in a given family, and they wanted to test whether public housing made a difference as measured by that instrument.

Figure 7.1: William Cochran

Figure source: https://www.york.ac.uk/depts/maths/histstat/people/cochran.gif

They were about to open a new block of public housing, and they came to the Johns Hopkins Statistics Department to see how they might use this opportunity to run an experiment and determine whether public housing changed the level of family cohesiveness. Cochran took on the case. The solution was fairly obvious. Cochran (along with Gertrude Cox) had written the first textbook that covered Fisher's theories and methods in detail and that provided designs for specific types of studies. All the housing authority had to do was to take the families who had applied for public housing and randomly assign new housing to some and leave the others as controls in their current situation. How many families should be involved in this experiment? That was also easily solved using tables in the text by Cochran and Cox.

There was only one problem. The housing authority had told the families applying for public housing that the new homes would be provided on a first-come-first-served basis. They would lose their credibility if they allowed a computer (and a pseudo random number generator, at that) to decide who would get these coveted spots.

Can you run an experiment where the treated families would be different from the non-treated families because the first ones to apply would surely be different in their attitudes and cohesiveness than those who applied later? Isn't this like the Lanarkshire study where the children from poorer homes got the extra milk?

Suppose, Cochran mused, we could identify other factors that affect family cohesiveness and measure those factors on all families. Suppose we use those other factors to predict the change in family cohesiveness and compare the two groups of families' scores after we subtract the predicted effect of these other factors. How could they do that?

In fact, Fisher had already provided the answer. In one of his papers, Fisher had considered the combined effect of treatment on both the weight of the wheat produced and the weight of the straw. If the treated plants were affected by the fertilizer, then this should result in an increase in both wheat and in straw. Fisher found a way to account for that nuisance accumulation of straw. He called it the "analysis of covariance" and provided the mathematical techniques needed to subtract the influence of the treatment on the production of straw. Cochran applied the analysis of covariance to the "experiment," which did not use random assignment.

In the end, the housing authority study showed that there was a slight increase in the average score of family cohesiveness for the families moved to public housing. This held in spite of the anecdotal evidence that some families broke up after moving to their new homes.

Cochran described this study in a paper that he published in the journal Biometrics (1968). In the paper, he called this an "observational study." The basic idea was that one could compare treatments even when the treatment assignment has not been random, provided you have enough information about the possibly confounding variables to estimate their effect. Cochran's observational study approach was to have an unexpected effect on modern medical research.

7.2 The Development of "Evidence-Based Medicine"

R. A. Fisher's formulation of the statistically based experimental design quickly caught on in a number of fields. In 1925, he published a book entitled, Statistical Methods for Research Workers (1970), which eventually went through 14 editions in English and more than 10 foreign language editions. There are no mathematical proofs in this book, just descriptions of statistical methods that can be used to design and analyze experiments. By the beginning of World War II, statistical methods had come to dominate experiments in sociology, psychology, pharmacology, chemistry, anthropology, and even archeology. In all these fields, one could find a copy of Fisher's Statistical Methods for Research Workers on the shelves of almost everyone engaged in running experiments, and this book is often included in the list of references.

One exception was the field of medical science. If you look into a medical journal from the 1930s or the 1940s, you will find it dominated by case studies. The typical article would describe the patient's symptoms, a differential diagnosis, the treatment given, and the outcome. If not case studies, then a typical article would describe a series of patients presenting with the same symptoms, the author's proposed etiology, and the outcomes of treatments. This was how medicine had advanced for hundreds of years. In the hands of a perceptive physician like Sir William Osler in the late 19th Century, it produced remarkable insights into the nature of illness, as did Osler in his identification of congestive heart failure.

In the hands of less perceptive physicians, however, it produced medical "cures" like powdered mummy to counteract poisons, and bleeding to reduce fever. This was how medicine had been taught since the time of the ancient Roman philosopher Galen.

7.3 The DES Debacle

In the late 1930s, pharmacologists had identified a class of estrogens, female hormones deeply involved in the birth process, and chemists began creating synthetic analogs. One of these was Diethyl-Silbesterol (DES), which had a long half-life, so it remained in the body longer than naturally occurring estrogens. In 1938, it was approved by the Food and Drug Administration of the United States for the treatment of acute acne.

But many obstetricians thought DES might have another use. Some of them began injecting pregnant women who had had a history of miscarriages with high doses of DES. They reported remarkable success, particularly in the Boston area. Women who had suffered two or three or even four prior miscarriages were carrying their babies to full term. In the cases where the treatment failed, the leaders of this new therapy recommended increasing the dose—until the average dose being used was more than seven times the dose recommended for severe acne.

In a 1949 issue of the journal Obstetrical and Gynecological Survey, Dr. O. Watkins Smith of the Free Hospital for Women in Brookline, Mass., reported on a series of pregnant women who had been treated with DES. This became one of the most widely cited papers in that field over the next few years. Dr. Smith reviewed 589 cases, dividing them into 219 where spontaneous abortion was “threatened,” 272 where the DES was used prophylactically, and 98 with premature delivery. He reported that 73% of these women had “live and well babies.” He compared this to his prior experience where over half of such patients would be expected to have miscarriages.

Opposed to the long heritage of case studies that was traditional in medicine, randomized controlled studies based on Fisher’s model were being proposed for medicine in the early 1950s. Austin Bradford Hill (1897–1991) of the Medical Research Council in the United Kingdom and Joseph Berkson (1899–1982) of the Mayo Clinic in the United States were advocating the use of randomized controlled trials instead of case studies to investigate medical claims.

7.4 Deickmann’s Study

In 1951, Dr. William Deickmann (1897–1957) of the University of Chicago School of Medicine began entering pregnant women into just such a trial. Eventually, he had 1,940 women who met his entrance criteria in that trial. Unknown to the patients, half of them had been given DES and half had been given saline injections. Under modern ethical guidelines, he should have asked the women if they would want to be part of a study, explained the nature of the study, and given them the opportunity to opt out. However, at that time, DES injections for women suspected of having difficulty was the standard of care, and Deickmann did not see the need to notify the patients.

When he published the results in 1952, he showed that both the placebo and DES groups had exactly the same incidence of miscarriages. Deickmann’s results were eventually corroborated by controlled studies in the United Kingdom and in Scandinavia. The supporters of DES claimed that Deickmann had not used the “right” type of patient. If he had restricted his study to women who were “prone” to miscarriage, he would have had a different result. However, Deickmann’s entrance criteria were exactly the same as the indications the obstetricians who "discovered" DES had used in deciding which women should be treated.

In spite of these clear findings, DES continued to be used for this purpose well into the 1970s, and it has been estimated that over 10 million women were treated worldwide. However, Deickmann's study was a wake-up call to the editors of medical journals and leading professors at medical schools. Following the publication of his study, editorials began to appear in the medical journals noting the need for controls, that individual case reports or series of patients were inadequate to establish either the safety or efficacy of a given treatment. The effects of the treatment on patients had to be compared to similar patients who had not been so treated in a properly randomized study.

Eventually, this perceived need for controls evolved into today's insistence on evidence-based medicine among the leaders of the medical community. The randomized controlled clinical trial is now considered the gold standard for evidence-based medicine. Most important medical journals ceased publishing case studies and descriptions of series of patients. One journal, Lancet, did continue to publish these but under the heading, "Preliminary Reports."

7.5 The Misuse of Observational Studies

However, it is very difficult to mount a double-blind randomized clinical trial. It is far easier to follow a group of patients given different treatments (but not chosen at random) by practicing physicians. Cochran's paper describing observational studies came as a godsend to many in the medical field. As a result, journals began to publish comparisons of treatments that had not been randomly applied but that were the result of the medical judgment of the attending physician and calling them "observational studies." Seldom, however, do these meet the criteria proposed by Cochran for what he called an observational study.

Since the beginning of the 21st century, exciting new insights have been made about the nature of probability and statistical design, and a new type of design has been—but that is the subject of the next chapter.

7.6 Summary

William Cochran coined the phrase "observational study" to describe a study where it is impossible to randomize individuals to treatment and the analyst has to use baseline data to account for inherent differences among the subjects that might influence the final outcome. He called these "observational studies." The term "observational study" has come to be used in many situations where subjects are not randomized to treatment and no attempt is made to provide the types of correction Cochran proposed. The DES debacle is described as an example of problems that arise in medicine when studies do not have adequate controls.

References

Cochran, W., (1968)“The effectiveness of Adjustment by Subclassification in Removing Bias in Observational Studies,” Biometrics, 24, 295-313.

Fisher, R. A., (1970) Statistical Methods for Research Workers, 15th edition, Oliver and Boyd, Edinburgh.

Chapter 8: Different Designs for Experiments

8.1 Crossover Designs

Let us start with the cow, the source of the milk. In a modern dairy farm, the cow is primarily a machine that makes milk. She is artificially inseminated with carefully selected bull sperm. She is pregnant for around 280 days. A few days after birth, the calf is weaned from its mother and fed a gruel of cooked grain. The mother cow is now ready to be milked two or three times a day. Can we find feed additives that increase amount of milk from the cow?

Figure 8.1: Cow Being Milked

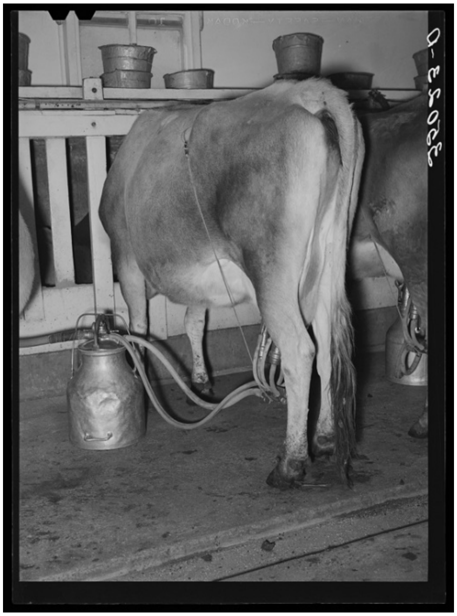


Figure source: https://www.loc.gov/resource/fsa.8b23639/

Suppose we experiment with two different additives to the grain that she is fed. Call them additive A and additive B. The reader should now know the design that we need to use: randomly assign cows to A or B and measure the amount of milk produced in the A-treated cows versus the B-treated cows. However, different cows produce different amounts of milk. In 2016, Gigi, a cow in Wisconsin, produced a record 75,000 pounds of milk in one year (Holstein Association USA 2016).The typical cow produces around 2500 pounds of milk a year (Holstein Association USA 2016). If Gigi were assigned treatment A and ordinary Bessie treatment B, then, regardless of what the other cows produce, the average output for cows treated with A would be much greater than the average output of cows like Bessie.

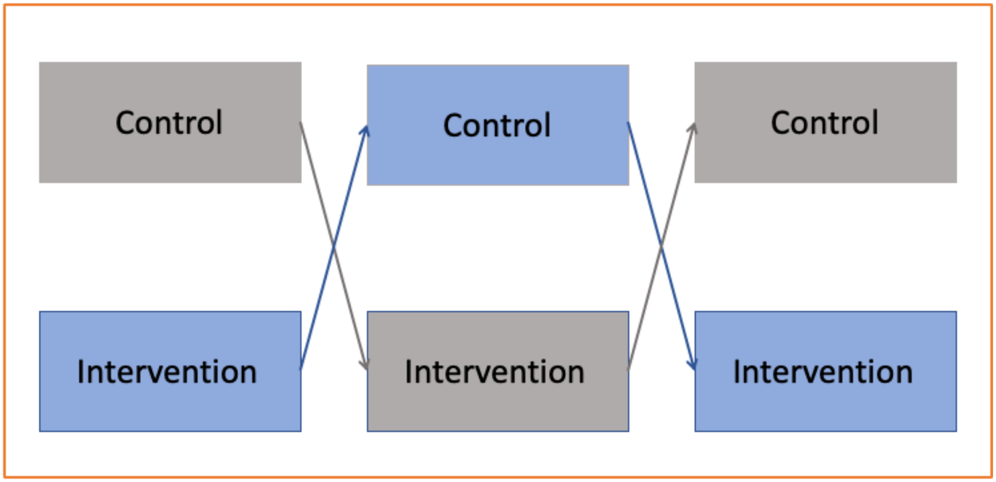
The amount of milk that a cow produces on a given day changes from day to day, but the day-to-day differences for any one cow are much less than the difference in output between one cow and another. In order to reduce the amount of variability (measured by the variance of the error term in Fisher's model from Chapter 1) it makes sense to try both treatments on each cow. In this type of study, cows are randomly assigned to either treatment A or treatment B for the first week. Then they are crossed over to the other treatment for the next week. This is a crossover design. Each cow is her own control.

There is one problem with this crossover design when used to compare the milk output of cows. The amount of milk a cow gives each day steadily diminishes from the time she gives birth. Even champion Gigi produced most of her record amount of milk in the first few months of 2016. At the end of one year, a cow's daily output could have been as much as half her initial daily output. Thus, the treatment effect on the output of milk is confounded with the order in which the animals are treated.

(Recall that to say "treatment is confounded with order" means the following: the algebraic formulas that define the experiment show that whenever treatment changes, so does order. Therefore, it is impossible to disentangle the effects of treatment from the effects of this steadily diminishing output from each cow.)

Theoretically, we could randomly assign treatment orders to a very large group of cows and compare the average output. However, when it is over, someone is bound to point out that we randomized Gigi to go from A to B, and ordinary Bessies went from B to A. How do we know whether the deterioration in output is the same for supercow Gigi as it for the others?

Figure 8.2: Schematic of Cross-back Design



In order to determine whether feed additives can increase the amount of milk, university studies run for the dairy industry use a "cross-back design." Half the animals are randomized to be treated by A, then by B, then by A. The other half goes from B to A to B. See Figure 8.2. Since the output of milk is steadily decreasing for each cow, the amount of milk in the first treatment can be averaged with the amount of milk in the third treatment, and the amount of milk in the middle treatment compared to this average. The BAB cows are treated the same way. Thus, if Gigi happens to be in the ABA group, the relative difference between A and B can be determined, not by Gigi's immense output in general, but by the relative difference between A and B for Gigi.

8.2 Carry-over Effects

In clinical research, there is a strong temptation to use the patient as his or her own control. Different patients respond differently to the same treatment, and human responses to medical treatment often contain a high level of "placebo response," where the act of being "treated" is often enough to bring relief, even when the "treatment" is only a sugar pill. For this reason, it is often useful to "use the patient as his own control," applying both the treatment being tested and placebo to the same patient at different times.

There is a problem with crossover designs when the patient has a chronic disease like diabetes. Symptoms of chronic diseases fluctuate and, often, patients are entered into clinical trials when the disease is at its worst. So the measured symptoms of disease will often have an improving trend over time, regardless of treatment. This is much like comparing the milk output of Gigi and Bessie, but there is an added wrinkle in clinical studies.

Medical treatment seldom provides immediate relief over a short interval of time. Often, extensive time is needed for the treatment to take effect. Drugs given by mouth have to be absorbed, enter the blood stream, do whatever they are designed to do, and then remain in the blood as the drug is slowly metabolized and eliminated from the body. The amount of time the drug is still “working” in the blood is estimated using a statistical model of probability known as an exponential distribution. This particular model implies that small amounts of drug will remain in the body for a very long time. The delay in elimination is measured by the “half-life” of the drug—the amount of time needed for half the drug to be eliminated. Thus, for a drug with a 12-hour half life, the patient’s blood will have half the dose taken after 12 hours, one-fourth of the dose after 24 hours, one-eighth of the dose after 36 hours, and so on.

In a crossover study comparing treatments A and B, this means that the effects of the previous treatment remain (albeit at a reduced level) when the second treatment begins. One way to eliminate the carry-over effect is to have a “wash-out” period between the two treatment periods. The “wash-out” time has to be great enough to make sure that almost all the effect of the first treatment to be over. But chronic illnesses have their own natural ups and downs. If you wait too long before giving the second treatment, the patient’s condition might be drastically different. In that case, you lose all the advantages of having the patient as his own control.

Recall that the statistical design of experiments starts with a set of equations that describe the course of the study in terms of the observed values (indicated in Roman letters) and controlling parameters (indicated in Greek letters) that have to be estimated from the observed data. When involved in a cross-over study, those equations can include an element to describe the diminishing carry-over effect of the first treatment.

One advantage of having this set of equations to describe the study is that we can determine the conditions necessary for us to be able to estimate the values of unobserved parameters like the carry-over effect.

8.3 Latent Response Models

This brings us to the work of Donald Rubin of Harvard. Rubin set up mathematical descriptions of the probabilities associated with a study. This mathematical description can contain symbols for every conceivable aspect of the study, whether they can be observed or not. For instance, if there are two treatments, the model can include the response of each subject to each of the treatments. A particular patient will be given only one of the treatments, but the mathematical function contains a Greek letter for that subject’s latent “response” to the treatment not given. Rubin goes even further, including symbols for the carry-over effect and for missing variables. Once the nature of the experiment is laid out this way, these mathematical functions can be used to predict differences in response and relationships between the observable values.

Figure 8.3: Donald Rubin

Figure source: https://statistics.fas.harvard.edu/people/donald-b-rubin

For instance, we can examine the detailed mathematical model of the probabilities of response to determine which factors can be isolated. We can look at different methods for estimating the values of the parameters in the model. Most important, we can look at the structure of the probability and decide which questions can and cannot be answered using the data from this study.

8.4 Summary

When animals used in experimentation differ greatly from one another, the study can have a crossover design. In such a design, each unit of experimentation is given both treatments that are being compared, one after the other. There are problems with crossover studies. The measurements being made might be steadily changing, regardless of treatment. In such cases, the effects of that trend have to be subtracted off when comparing two treatments. In some crossover studies, there is a carryover effect from the first treatment to the time of the second treatment. Different designs are used to account for these problems. Donald Rubin of Harvard has developed a wide range of designs based on the course of measurement over time.

Reference

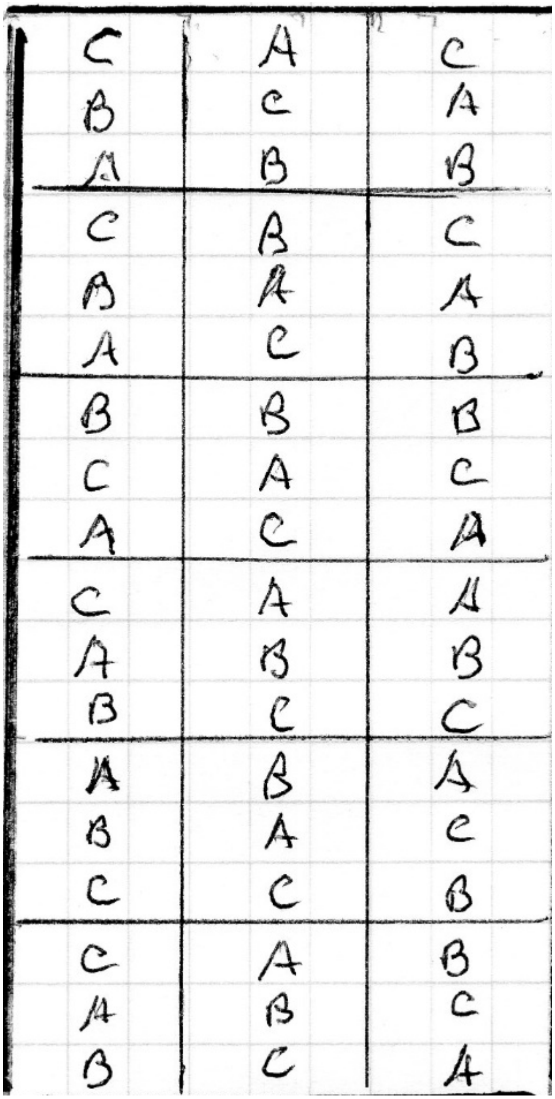
Holstein Association USA, (2016) "Bur-Wall Buckeye Gigi Sets National Milk Production Record." Holstein Association USA. www.holsteinusa.com/news/press_release2016.html.

Chapter 9: Analysis of Variance

9.1 Fisher's Model

Statistical design of experiments began with Fisher's work on agricultural experiments. Recall that Fisher split the experimental field into small plots, each plot sufficiently small so that the variation in fertility across the plot is negligible. Then, within each plot, he planted rows of plants, each row with a different treatment. Finally, the rows used for specific treatments were chosen at random. This schema is displayed in Figure 9.1.

Figure 9.1: Treatments Randomized within Blocks



C	A	C
B	C	A
A	B	B
C	B	C
B	A	A
A	C	B
B	B	B
C	A	C
A	C	A
C	A	A
A	B	B
B	C	C
A	B	A
B	A	C
C	C	B
C	A	B
A	B	C
B	C	A

Let us use Fisher's more general terminology. He called the individual plots "blocks." So his experiment consisted of "blocks" and "treatments." For a given block, the average weight of grain (putting all the treatments together) was a measure of the fertility of that block. For each treatment, the average weight of grain (putting all the values for each treatment in all the blocks together) was a measure of the efficacy of the treatment.

There is a widely used measure of the variability of a group of observations called the "variance." If you ignore the division of the field into blocks and treatments, but just look at all the weights of all the rows in all the blocks, you can estimate the overall variance of the experiment. With a little manipulation of the mathematics involved in these estimations, it turns out that

(overall variance) = (variance of blocks) + (variance of treatments) + (the underlying "error" variance of the measure used)

This last element of the equation Fisher called the "error variance." If the treatments differed in effect, then the ratio

(variance of treatments)÷(error variance)

will be large. Furthermore, that ratio will also tell us how great the differences in effect are among the treatments. Fisher called this procedure, "analysis of variance."

Keep in mind Fisher's basic abstraction:

Blocks + Treatment + Error.

There is nothing in this abstraction about wheat or potatoes in an agricultural field. It can be applied to many other types of experiments. The "blocks" could be schools and the "treatments" different methods of teaching. The "blocks" could be batches of steel and the "treatments" different methods of annealing. The "blocks" could be states of the United States and the "treatments" different minimum wage laws. As we saw with the work of Fairfield Smith, the "blocks" could even be rubber trees and the "treatments" different angles for the taps.

The analysis of variance and experimental study designs involving "blocks" and "treatments" form a model used in almost every scientific field. Computer software for running analyses of variance is available in almost all commercial packages used in scientific work.

9.2 "Errors" to "Residual"

When I first went to work for Pfizer Pharmaceuticals in their clinical research department, I was one of a group of statisticians being hired by pharmaceutical companies because the FDA was now requiring that they prove their new drugs were efficacious, using clinical studies that followed Fisher's pattern. The whole idea of statistical design and analysis was new to medicine at that time, and few in the company's senior management had even seen a formal statistical analysis with analysis of variance tables.

I generated the standard analysis of variance tables for one of the clinical studies. This bothered a member of senior management. Here, in my table, I had indicated a line for "error." What errors? Had we made "errors" in this study? he asked me. I explained that this was the standard table produced by standard computer software. He said, "I still don't like the word 'error'."

I called a statistician I knew, who had been working for a food company for years, and asked her if she had ever run into this problem. She laughed and said that, indeed, her senior management was reluctant to admit “error.” In their analysis of variance tables, they called this final bit the “residual.” And so, in the material we sent to Washington, the analysis of variance tables listed lines for

Clinics + treatment + residual.

In mathematics, we often use ordinary words (like “error”) but give them exact and restricted meanings. This can cause confusion among readers who use these words in a more vague and general sense. For instance, throughout this book, I have used the words “estimation” or “estimator” to describe the process of using observed data in a specific way to calculate a value for an unknown parameter (indicated by a Greek letter in the algebraic formula used to describe the experiment). In ordinary language, these words often carry an overtone of wild guessing. In statistics, “estimators” are always used to reach hard conclusions.

9.3 Extensions of Fisher’s Model

Consider now the amazing flexibility of Fisher’s simple analysis of variance. Suppose, for instance, that the row of wheat seeds given treatment A in one of the plots failed to germinate. What should be done with this missing data? You can go to the set of equations that describe the experiment, insert a dummy symbol (Greek letter) for the missing value, manipulate the mathematics, and come up with an estimate of the missing value. This was done early in the development of statistical design by Frank Yates (1902–1994), and most computer packages will use Yates’ correction to estimate missing values.

Another problem that can arise is when there is not enough “room” in the blocks to apply every one of the treatments being examined. This leads to an “incomplete block design.” This type of design was investigated by Fairfield Smith and was applied to the problem of determining if the direction of the tap had an effect on the amount of rubber tree sap recovered.

The blocks used in a designed experiment need not be identified in advance. As an example, consider a clinical study with a high drop-out rate. The ethics of a study in humans require that the study subjects be free to leave the study whenever they want. In some long-term studies, the patients are brought into a clinic for several hours of measurement at each visit. In some of those studies, patients are put through exercise tests or undergo invasive evaluations. In such studies, some patients drop out. Sometimes the clinician knows the reason for the drop out (such as no improvement in symptoms). Other times, the patient just fails to show up. She might have moved. She might have been hospitalized for a non-related problem.

9.4 Propensity Scores

What can be done in such cases? We have early measures of efficacy on these drop out patients, but seldom enough to evaluate the full course of the experimental treatment. In 1983, Donald Rubin and Paul Rosenbaum of the University of Pennsylvania produced a method of analysis known as “propensity scores”. Here is how propensity scores are being used in tackling the problem of early patient drop-out:

Taking just the material available at baseline, they run a statistical analysis (using a method known as log-odds regression) with which they determine the degree to which baseline variables predict the probability that this patient will drop out, which they called the “propensity score.” Once they have these scores, they can order the propensity scores and divide patients with similar propensity scores into blocks. Like the varying fertility among the blocks in an agricultural experiment, the varying propensities to drop out can be used to block patients into groups.

Propensity scores can be computed for other elements that interfere with the full completion of a clinical study. Rubin and Rosenbaum’s propensity scores have been widely used in modern clinical research. This has been especially true in areas like cardiology, where many minor problems with patient compliance can result from complicated protocols.

Now comes a question that has bedeviled the use of analysis of variance since it was first used on experiments other than Fisher’s agricultural experiments. Consider Fisher’s plots of land, the blocks in which he planted his different experimental seeds. His blocks do not get up and wander all over the field. They are fixed in advance, and all the unexplained randomness can be summed up in the error term. But what about Rosenbaum and Rubin’s propensity scores (predictions of probability of drop out developed out of the data from the study)? You do not know in advance which patients will end up in which blocks. The division of patients into blocks depends on the predictions of the propensity scores. Whether a given patient would drop out is a random event. Thus, the division of patients into propensity groups is the division of patients based on random events. Fisher’s mathematical derivation of analysis of variance assumed that the blocks were fixed and contained no random element in their definition. If the boundaries of the blocks are random events, Fisher’s derivation will not work out. Fortunately for the future of statistical design of experiments, Fisher later recognized this problem and worked out the mathematical derivations needed when the boundaries of the blocks are random events.

Why is it important to know whether the blocks are fixed in advance or arise at random as the study continues? It changes the criteria for deciding if a “significant” difference has been seen between treatments. Which of Fisher’s two derivations is appropriate in analyzing the data? If the blocks were fixed, and you used the random block analysis, then you will miss important differences in treatment effects. If the blocks were random and you used the fixed block analysis, you will flag differences among treatments as “significant” when, in fact, the differences can be attributed to random noise.

It seems obvious when comparing blocks of patients with similar propensity scores or when comparing results in agricultural field studies. But what about studies of hereditary problems? Are families fixed or random blocks? Real-life problems are seldom as obvious as the examples found in textbooks.

9.5 Regression

In addition to developing new studies of artificial fertilizer for Rothamsted, Fisher kept plowing through what he called 50+ years of "muck"—the experimental data derived before Fisher. One field had been left untouched by the different experimental fertilizers and had been used as a control to compare against the treated fields. Careful records had been kept of rainfall, weed infestation, type of seed used, times of planting and reaping, and so on. Was it possible to predict the field's output, using only these measures of weather and other natural conditions during the time the plants were growing?

Francis Galton (1822–1911) had addressed a similar problem in the last decade of the 19th century. He had set up a biometric laboratory in London and invited families to come and be measured. He had hoped to establish the nature and heredity of intelligence by looking at both children and parents. Measurement of intelligence proved to be very difficult, but he was able to measure their heights and weights and examine the effects of heredity on these measures.

Figure 9.2: Sir Francis Galton

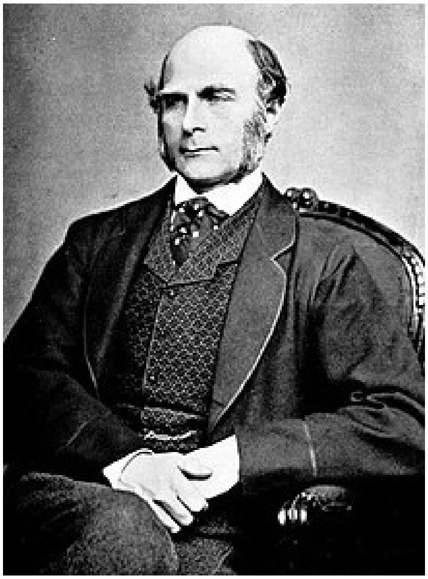


Figure source: https://en.wikipedia.org/wiki/Francis_Galton

It made sense, to Galton at least, that children of tall parents should be tall and that children of short parents should be short. He compared the heights of tall fathers to the heights of their sons and the heights of short fathers to the heights of their sons. He discovered that, on the average, sons of tall parents were shorter than their fathers and sons of short parents were taller than their fathers.

After some thought, Galton realized that this had to be true. Suppose the sons of tall fathers had averaged their fathers' heights and sons of short fathers had averaged their fathers' heights. Then, in each generation there would have to be some sons of tall fathers who were taller than their fathers and some sons of short fathers who were shorter than their fathers. If this happened generation after generation, then the human populations would include some extremely tall men (well over 10 feet?) and some extremely short men (well under 1 foot?).

Galton called this (now obvious) phenomenon “regression to the mean.” He proposed a general biological principle: in any species, there is a theoretical mean configuration toward which all individuals tend.

(At this point, we are using the highly specific meanings of words in mathematics. We distinguish between the “mean” and the “average” in this fashion. The “average” is computed from a group of observations. The average is the sum of all the observed values, divided by the number of observations. The “mean,” on the other hand, is a theoretical parameter of a probability distribution. It is represented by a Greek letter in the mathematical formulas. It is the center of the probability distribution, and, in most situations, it is best estimated by the average of the observed data. Galton’s “regression to the mean” refers to the theoretical center of a distribution.)

Fisher, ever the consummate mathematician, worked out the mathematical relationships that represented regression to the mean. He noted that these formulas for regression could be used in any situation where you have a set of imperfect “predictors” and final outcomes. Galton’s formula could be used to predict a future outcome based on these predictors. Fisher called this mathematical model “regression.” Although it has been used for problems far afield from the inheritance of height, this technique of analysis is still called regression. When embedded in an analysis of variance, Fisher called this analysis of covariance. It was this analysis of covariance that William Cochran used in defining observational studies.

9.6 Uses for the Computer

For most of his professional life, Fisher had to do all his calculations on a hand cranked desk calculator. The algorithms that he invented were designed to be used on a desk calculator. Analysis of covariance, as defined by Fisher, involves tedious calculations that can take hours on such a calculator.

With a modern computer, I can write the mathematics and ignore the difficulties that might arise during the calculations that were difficult or impossible on a desk calculator. The modern computer can now read such complicated mathematical formulas and grind away generating solutions. Russell Wolfinger at SAS Institute is a leading scientist in the use of computers for complicated calculations. He has created programs that are part of the commercial software available from the SAS Institute. Using these programs, modern statisticians have been able to make the computer engage in statistical calculations that were once thought to be intractable.

Among these are problems involving more complicated experiments where the treatments are randomized among more than one type of block, so Fisher’s equation becomes:

Overall variance(Y) = (variance of block type I) + (variance of block type II) +…+(treatment variance) + (error variance).

Wolfinger's programs can deal with situations where some of the block divisions are random and some are fixed and where missing data consists of much more complicated types than envisioned by Yates. These modern computer programs usually start by putting the problem into one big regression equation, examining the relationships among the variables that can affect the output, and then calculating the results for patterns of equations that can be solved for this unique set of data.

These computer programs often require millions of calculations. The computer is not like Fisher sitting at his desk calculator. It does not have a wife and children to go home to. It does not have an arm that gets weary of pulling the lever. It does not worry whether the results make sense. It just grinds on and prints out its results.

The modern scientist running a complicated experiment does not need to know how the mathematics work, but she or he has to be able to interpret the output of the program and needs to understand the nature of the data. When I run an analysis of data, I first look at the data that will be crunched by the computer. I let the numbers stream slowly across the computer screen, looking for anomalies. If these are rats in a toxicological experiment, I know that there is a problem if one of the rats is recorded as weighing several thousand grams. If it is an agricultural study, there might be a section of the field where the plants have stunted growth. If it is an experiment designed to determine whether pasteurization takes the "good" out of the milk, I would want to see whether all the children given extra milk came from the poorer families.

But the subject of cleaning data is a topic for another book than this one.

9.7 Summary

Fisher's first designs used small plots of land (called "blocks") into which he planted rows of grain, each with a different treatment. He showed that the simple formula

Observation = treatment + block + error

carries over into the variability due to different parts of the design, where variability is measured by the variance of the observations

(overall variance) = (variance between treatments) + (variance between

blocks) + (error variance)

and that the differences in output due to different treatments can be tested by the ratio

(variance due to treatments)÷ (error variance).

This method of analysis is called "analysis of variance." Frank Yates used Fisher's formulations to propose a method of dealing with missing data.

In dealing with more complicated situations, Fisher took Francis Galton’s concept of “regression to the mean” and generalized it (under the name “regression”) to take care of the influence of other elements besides treatment. The modern computer has enabled the statistician to deal with much more complicated problems. However, even the most complicated computer program starts with a model descended from Fisher’s analysis of variance structure.

Chapter 10: The Bayesian Heresy

Note to the reader: In all the other chapters, I have been able to explain the basic ideas without resorting to mathematical notation. In this one chapter, I have had to use mathematical notation because Bayes' Theorem falls out of a symmetric relationship in the mathematical notation and makes sense only within the framework of those mathematical formulas.

10.1 What is Probability?

In the 18th century, there was a Swiss family of mathematicians named Bernoulli. There were the brothers, Johann (1667–1748) and Jacob (1654–1765), and Jacob's son Daniel (1706–1787). Although the Bernoulli family was in the spice business, these three all became professors of mathematics at different universities. They were busy measuring and counting. They were followers of Galileo who, about 100 years before, had insisted that knowledge can only be gained by careful measurement. The Bernoullis measured air pressure, the flow rates of water, the weights of different substances, and Daniel decided to look at probability.

The concept of probability had been around for a long time. In the Babylonian Talmud (which records the debates of the rabbis of the 1st and 2nd centuries of the common era), the principle is stated that, if there are two interpretations of the law and if one is as probable as the other, then the more lenient one should be used. Aristotle is recorded as saying, "It is the nature of probability that improbable things will happen." The concept of probability in these ancient discussions referred to something that is not quite certain, and no attempt was made to put a number on it or to compare one probability with another.

The Bernoullis started with games of chance. The probability of getting a 6 with one throw of a six-sided die was

$$1 \div 6 = (\text{no. of favorable outcomes}) \div (\text{no. of possible outcomes}) = 1/6$$

The probability of getting a "6" or a "1" was

$$2 \div 6 = 1 \div 3 = 1/3.$$

Thus, with games of chance, probability was measured on a scale from zero to one. No matter how complicated the game, the basic idea was that probability of an outcome equals

(number of ways to gain that outcome)÷(total number of possibilities).

And, it would always be a number between zero and one.

Following this lead, mathematicians could measure probability whenever there was a well-defined set of possible outcomes. Through the rest of the 18th and most of the 19th centuries, probability calculations were a sidebar in the development of mathematics. To calculate probabilities, mathematicians used tricks in calculating combinatorial events.

By the time Karl Pearson (1857–1936) came on the scene in 1898, the mathematics of probability was a large bag of somewhat related formulas that produced probabilities for specific types of outcomes. To understand Pearson's innovation, consider a living animal as some type of a machine, with muscles and blood all moving about. Anything the animal does starts with discharges of nerve cells, programmed to influence blood flow and muscle movement. We can conceive of the initial nerve discharges as the accumulation of a large number of small changes or events. The normal probability distribution describes random events that originate as a sum of a large number of small random events.

However, Pearson went a little further. We do not measure the discharges of nerve cells. Instead, we measure the final outcome, the purposeful movement of some muscle. Pearson proposed that this final measurement is a distortion of the initial, normally distributed random variable. Its passage through a living animal to our measuring instrument causes the probabilities to be distorted. Pearson assumed that this distortion was smooth and consistent. Using calculus and that one assumption, he derived a class of probability distributions that he called the "skew distributions." He spent much of his professional life after that collecting large amounts of biological data and fitting them to members of his skew distributions.

Figure 10.1: Karl Pearson

Figure source: https://en.wikipedia.org/wiki/Karl_Pearson

Soon after Pearson developed his theory of skew distributions, a group of German physicists (Albert Einstein among them) found that they had to describe the positions and relationships among sub-atomic particles in terms of probabilities, producing methods of calculation known as quantum mechanics.

In the 1920s, John Maynard Keynes (1883–1946) was working on his Ph.D. thesis. In that thesis, he proposed that probability lies at the heart of human activity. People, he claimed, have an innate sense of probability that enables them to anticipate events. You cross a street after observing very few cars because you conclude that the probability of getting hit by a car is low. You do not need to propose specific numbers for these personal probabilities, Keynes noted, you only need to have a feeling for the relative probabilities of different outcomes. You also do not need to know all probabilities. He gives the example of someone looking for a book bound in buckram on a library shelf. In that search, there is no need to know the probabilities that the book's binding is red or green.

In the 1950s, L. J. Savage (1917–1971) picked up on Keynes' ideas and a similar set of ideas proposed by Bruno de Finetti (1906–1985) and built an entire theory of probability based on this idea of personal probability. Savage showed that personal probabilities are just like Bernoulli's probabilities that were based on games—as long as they fulfilled a condition that he called "coherence." If a person believes that the probability of some event A is less than the probability of B and that the probability of B is less than the probability of C, then to be coherent, that person has to believe that the probability of A is less than the probability of C.

10.2 Thomas Bayes and "Inverse Probability"

The Reverend Thomas Bayes (1702–1761) was a dissenting minister of the Anglican Church, which means he did not subscribe to the full body of doctrine espoused by the Church. (Recall that it was an Anglican bishop who, a hundred years earlier, had proclaimed that William Harvey's proof of the circulation of the blood was wrong because it went against established doctrine and because Nature abhors experimentation.)

Figure 10.2: Thomas Bayes

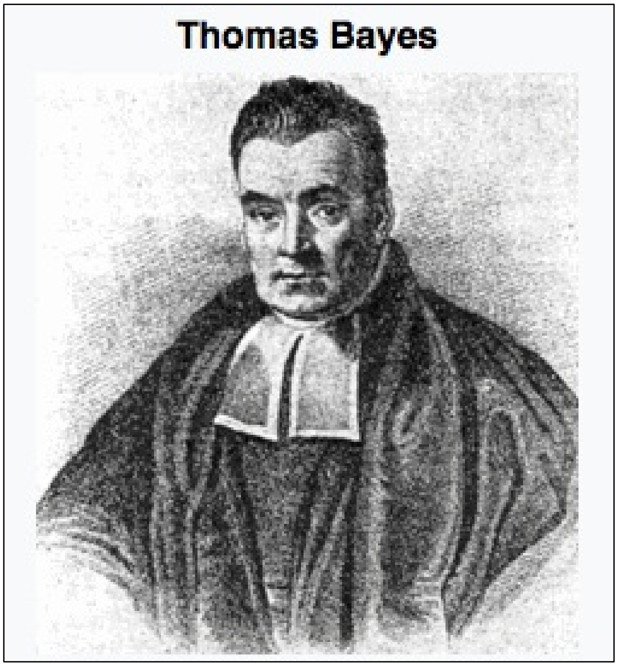


Figure source: https://en.wikipedia.org/wiki/Thomas_Bayes

We know of Bayes in the 21st century, not because of his doctrinal beliefs, but because of a mathematical discovery, which he thought made no sense whatsoever. He was one of the correspondents of the Royal Society in London. The correspondents were natural scientists from all over Europe who sent letters to the Royal Society to be read at their meetings, which described their investigations into chemistry, physics, biology, natural science, or any other aspect of what was then known as "natural philosophy." Most of Bayes' communications have been superseded by later work, but one communication (which he never sent to the Royal Society) has immortalized his name.

To understand Bayes' Theorem, we need to refer to this question of the meaning of probability. As noted earlier in this chapter, the Bernoullis proposed that probability could be measured as a number between zero and one, and they examined probabilities in terms of games of chance. In the 20th century, John Maynard Keynes and L. J. Savage proposed that probability was something that an individual uses to organize life—the concept of personal probability.

However, probabilities get invoked in situations that do not involve games of chance and that are not “gut feelings” of individuals. Quantum physics uses probability calculations to examine the inner nature of atoms and subatomic particles. Statistical design of experiment uses probability calculations to separate the effects of treatments and blocks. The meteorologist on television tells us the probability of rain tomorrow.

In the 1930s, the Russian mathematician Andrey Kolomogorov (1904–1987) proved that probability was a measure on a space of “events.” It is a measure, just like area, that can be computed and compared. To prove a theorem about probability, one only needed to draw a rectangle to represent all possible events associated with the problem at hand. Regions of that rectangle represent classes of sub-events. For instance, in Figure 10.3, the region labeled “C” covers all the ways in which some event, C, can occur. The probability of C is the area of the region C, divided by the area of the entire rectangle. Anticipating Kolomogorov’s proof, John Venn (1834–1923) had produced such diagrams (now called “Venn diagrams”). Venn was a British philosopher interested in the development of symbolic logic.

Figure 10.3: Venn Diagram for Events C and D

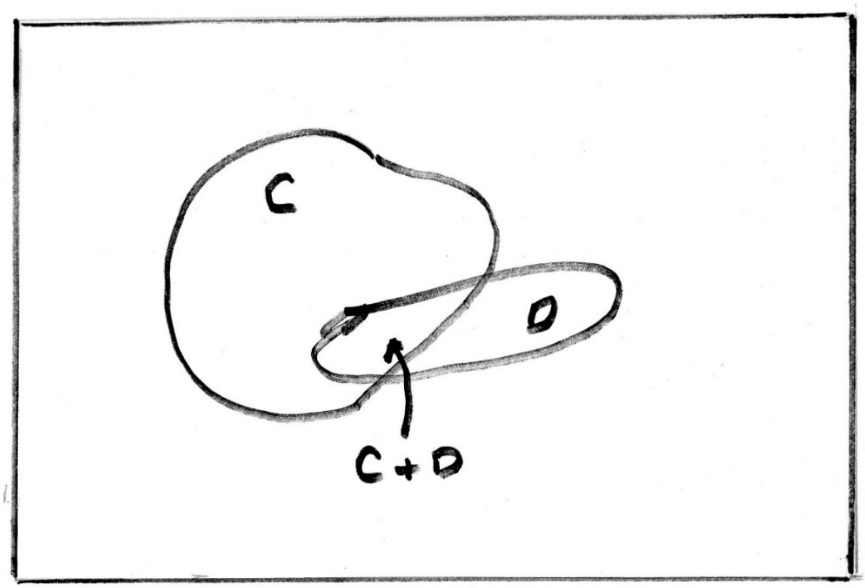


Figure 10.3 shows a Venn diagram for the following situation: We have a quiet wooded area. The event C is that someone will walk through those woods sometime in the next 48 hours. There are many ways in which this can happen. The person might walk in from different entrances and be any of a large number of people living nearby. For this reason, the event C is not a single point, but a region of the set of all possibilities. The event D is that the Toreador Song from the opera Carmen will resound through the woods. Just as with event C, there are a number of ways in which this could happen. It could be whistled or sung aloud by someone walking through the woods, or it could have originated from outside the woods, perhaps from a car radio on a nearby street. Some of these possible events are associated with someone walking through the woods, and those possible events are in the overlap between the regions C and D. Events associated

with the sound of the Toreador Song that originate outside the woods are in the part of region D that does not overlap region C.

The area of region C (which we can write P(C) and read it as “P of C”) is the probability that someone will walk through the woods. The area of region D (which we can write P(D)) is the probability that the Toreador Song will be heard in the woods. The area of the overlap between C and D (which we can write P(C and D) is the probability that someone will walk through the woods and that the Toreador Song will be heard.

If we take the area P(C and D) and divide it by the area P(C), we have the probability that the Toreador Song will be heard when someone walks through the woods. This is called the conditional probability of D, given C. In symbols

$$P(D|C) = P(C \text{ and } D) \div P(C)$$

Some people claim that if the conditional probability, P(C|D), is high, then we can state “D causes C.” But this would get us into the entangled philosophical problem of the meaning of “cause and effect”—a subject that belongs in another book.

To Thomas Bayes, conditional probability meant just that—cause and effect. The conditioning event, C, (someone will walk through the woods in the next 48 hours) comes before the second event D, (the Toreador Song is heard). This made sense to Bayes. It created a measure of the probability for D when C came before.

However, Bayes’ mathematical intuition saw the symmetry that lay in the formula for conditional probability:

$$P(D|C) = P(D \text{ and } C) \div P(C) \text{ means that}$$

$$P(D|C)P(C) = P(D \text{ and } C) \text{ (multiply both sides of the equation by } P(C)\text{).}$$

But just manipulating the symbols shows that, in addition,

$$P(D \text{ and } C) = P(C|D)\, P(D), \text{ or}$$

$$P(C|D) = P(C \text{ and } D) \div P(D).$$

This made no sense to Bayes. The event C (someone walks through the woods) occurred first. It had already happened or not before event D (the Toreador Song is heard). If D is a consequence of C, you cannot have a probability of C, given D. The event that occurred second cannot “cause” the event that came before it. He put these calculations aside and never sent them to the Royal Society. After his death, friends of Bayes discovered these notes and only then were they sent to be read before the Royal Society of London. Thus did Thomas Bayes, the dissenting minister, become famous—not for his finely reasoned dissents from church doctrine, not for his meticulous calculations of minor problems in astronomy, but for his discovery of a formula that he felt was pure nonsense.

$$P(C|D)\ P(D) = P(C \text{ and } D) = P(D|C)\ P(C).$$

For the rest of the 18th century and for much of the 19th century, Bayes' Theorem was treated with disdain by mathematicians and scientists. They called it "inverse probability." If it was used at all, it was as a mathematical trick to get around some difficult problem. Starting in the early 1930s, R. A. Fisher found himself in dispute with another mathematical genius of the 20th century, Jerzy Neyman (1894–1981). Neyman was busy cleaning up some of Fisher's work and proposing carefully reasoned modifications. In his responses, Fisher sometimes accused Neyman of using inverse probability.

10.3 Bayes' Theorem in Practical Use

Since Fisher's time, Bayes' Theorem has proved to be an important element in the statistician's bag of "tricks." Consider the problem of locating a downed aircraft in a mountainous terrain. The searchers have the plane's last known position and its course and speed at the time. The searching aircraft can break the regions of potential crash into small areas, each one capable of being searched in a single pass. Knowledge of the downed plane's position, course, and speed at last contact provides the searchers with probabilities of the crash for each of these small search areas. The obvious thing to do is to search first in the areas of highest probability. Let us suppose that the initial sweep over the most probable sites did not discover the crash. They could go on to less probable areas. However, they know from the terrain and the type of search plan that the probability of finding a crash, if it is in a given area, is less than 100%. In fact, the probabilities of finding a crash site, given that it is there, can be calculated from previous searches for similar areas.

Bayes' Theorem is used to adjust the probabilities that the crash is in a given area, based on these prior probabilities that a crash could not be seen in a given area in a single pass. These adjusted probabilities are then used to plot a new round of area searches.

Fredrick Mosteller (1916–2006) and David Wallace (1928–2017) wrote a classic book on the identification of authors, based on their use of non-contextual words (1964). In most languages, and particularly in English, we link together the words needed to express an idea with words that are not involved in the actual context of the subject but are needed to keep the sentences in good grammar and understandable. These are words like "or," "while," "then," ”of," "to," "and," and "also." The frequencies of the occurrence of specific non-contextual words are unique to a given writer since they are used unconsciously as the writer composes her or his works.

Examining the use of these non-contextual words across many authors, Mosteller and Wallace proposed that we could estimate the rate at which a given author uses each word. For instance, one author might average the use of "also" 15 times in every thousand words. Another author might use it more frequently, averaging 40 times in every thousand words. The average rate at which an author uses "also" is unique to the individual writer, and, if we have enough material written by that person, we can get a good estimate of its value. Since we cannot observe this underlying average but can only estimate it from the data that we have, we call this number a "parameter" (an "almost measurement") that has to be inferred.

But Mosteller and Wallace had some additional information. By examining works of other authors written in different centuries and in different countries, they could see that the parameter describing the average number of times any individual author uses “also” could be thought of as a random variable whose distribution changes from century to century and from country to country. For instance, the word “whilst” is used infrequently by modern American authors, but it is used very often in the United Kingdom today or in 18th-century America.

Here is a place where we can use Bayes’ Theorem to turn the different author-specific estimates into a probability with higher order parameters. If we knew the prior distribution of the frequencies of the use of “also” among authors who were contemporaries, we can detect which author wrote which paper with greater certainty.

Thus, we have

1. Probability of observed data as a formula involving parameters.
2. Prior knowledge that enables us to have a formula for probability of these parameters.
3. Use of observed data to refine the probability distribution of the parameters.

Or, to put it more succinctly,

Prior knowledge ▶ observed data ▶ posterior knowledge.

Since his understanding of probability was based on his understanding of “cause and effect,” Bayes saw his theorem as implying that an event that comes first “causes” an event that comes after with a certain probability, and an event that comes after “causes” an event that came “before” (foolish idea) with another probability. If you think of Bayes’ Theorem as providing a means of improving on prior knowledge using the data available, then it does make sense.

10.4 Bayes’ Theorem in the Design of Experiments

The experimental scientist seldom runs an experiment without having some idea of what the result should be. In 1887, Albert Michelson (1852–1931) was the first person to accurately measure the speed of light. (In fact, his experimental results produced profound problems for physics, which were finally solved by Einstein’s special theory of relativity.) To do so, Michelson set up an experiment where a beam of pure white sunlight was sent on paths of mirrors down two different lengths. His measurement of speed used the relative lengths of the paths that produced rings of interference when the two resultant beams were merged.

Michelson did not begin these experiments without some prior knowledge of approximately what that speed might be. With this prior knowledge, he threw out the results of several runs that clearly produced “wrong” answers. Thus did 18th and 19th century science advance because good scientists like Michelson used their prior knowledge to select specific sets of data and reject others. In the hands of less capable scientists, fields of research like phrenology were cluttered with “findings” that resulted from arbitrary selections of data.

At this writing, Bayesian methods have become respectable in the statistical literature, and computers are busy invoking elaborate mathematical calculations that enable the scientist to incorporate prior knowledge into the interpretation of data. The use of Bayesian methods has not only influenced the interpretation of data, it has also influenced the design of experiments.

In 1995, Kathryn Chaloner of the University of Minnesota and Isabella Verdinelli of the University of Rome published a review of the then current uses of Bayesian techniques in experimental design. Chaloner and Verdinelli looked at a large number of scientific papers where Bayesian techniques had been used to modify experimental designs and found a way to put them into a single unifying concept. This is the way applied mathematics advances. Different approaches to problems are found to be all based on some overall simplifying idea.

In the case of Bayesian experimental design, Chaloner and Verdinelli looked at all these problems from the standpoint of statistical decision theory. In statistical decision theory, the scientist considers a given problem as having a number of choices that can be made. The costs associated with the possible consequences of each choice are listed, along with the best estimate of the probability that a particular consequence will result from that choice. The optimal choice is the one with the lowest average cost over all possible consequences.

Abraham Wald (1902–1950) was the first to propose statistical decision theory as a unifying approach for what appeared to be many different ideas. Once put into the framework of decision theory, the arguments between Fisher and Neyman became greatly clarified. This is what happened with the Chaloner and Verdinelli paper. The basic idea is that the experiment is designed so that prior knowledge about the potential outcomes of different choices can dictate a design with the minimal average "cost." Randomization is still there, but it is restricted so that number of experimental units that are used in specific blocks or treatments depend on prior uncertainty.

Bayesian experimental designs often require vast amounts of computing to reach the design and to analyze the results. Many of the Bayesian algorithms would have been impossible to use in the days of the hand-cranked desk calculator. However, we now have the modern computer, which does not complain if we command it to do millions of calculations. Inverse probability might not have made sense to Thomas Bayes, but it does to the computer.

10.5 Summary

The concept of probability was vague and qualitative until the 17th century when Daniel Bernoulli suggested that probability could be measured on a scale from 0 to 1.0. The first calculated probabilities were based on games of chance. But probability proved useful in many other fields. The 18th and 19th centuries saw the development of complicated probability calculations. Karl Pearson suggested a family of probability distributions in the late 19th century and derived their formulas by assuming that the probability of some biological event can be thought of as originating from a normally distributed probability but is distorted in its passage through a biological event. He called these "skew distributions." Many other systems of related probability distributions have since been proposed.

While dealing with conditional probabilities, Thomas Bayes uncovered a basic symmetry in the idea of conditional probability. Since he saw conditional probability as a form of "cause" and "effect," his newly discovered concept appeared to show that an "effect" could produce its "cause." However, Bayes' Theorem has proven very useful when the experimenter has some prior knowledge and wants to incorporate that into his or her design. In general, Bayes' Theorem allows the experimenter to go beyond the experiment with the concept that experiments are a means of continuing to develop scientific knowledge, so

(Prior knowledge) ► (observed data) ► (posterior knowledge)

References

Chaloner, K, and Verdinelli, I, (1995) "Bayesian Experimental Design; A Review," Statistical Science, 10, #3, 273-304.

Mosteller, F., and Wallace, D., (1964) Inference and Disputed Authorship, The Federalist, Addison-Wesley, Palo Alto, CA.

Chapter 11: The Measurement of Pain

11.1 Measurement in Experiments

Statistical models use mathematics, and mathematics is based on numbers. Thus, any statistical design of an experiment requires that we measure something or count something unambiguously. However, very frequently, the situation that we want to examine in an experiment starts with a vague idea, where it is not obvious how to turn it into a number. As we saw in Chapter 2, pasteurization, claimed its opponents, destroyed the “good” in milk. How can one go about measuring the “good” in milk in order to run an experiment? The important problems in life are usually cluttered with such vague but emotionally loaded phrases. What makes a “good” citizen? How can we measure the effects of anti-cancer drugs? What method of teaching is “best”?

The questions involved in measuring vague, emotionally laden concepts have to be faced before an experiment can be designed. In this chapter, I will examine the problems of measuring vague ideas with a look at the measurement of pain. In measuring human pain, we encounter most of the problems of measurement in experiments.

Pain is a major component of medicine. It is pain that often brings the patient to the doctor. Everyone experiences pain at different times in her or his life. Everyone knows what it is to have pain. But can we compare one person’s pain to the pain experienced by another? Can we determine when pain is reduced but not removed? Let us look initially at the tail of a rat.

11.2 Experimentally Induced Pain

How does one know when an experimental mouse, rat, or hamster is in pain? Before a new medicine designed to relieve pain can be tried out in humans, there have to be successful experiments on animals. Pharmacologists have developed several ways to measure pain in mice or rats. (In keeping with the general principle that all scientific terms should be well-defined without ambiguity, tests like this, which look for a well-defined endpoint and its measurement, are called “assays”.) However, one problem with animal models is that the animal’s discomfort, as measured in a specific assay, might not be predictive of human pain.

One experimental setup that has been successful in identifying drugs that relieve human pain is the rat tail-flick assay. A rat is immobilized, and infrared rays are focused on a spot of the animal’s tail. The measure of the efficacy of a compound is the time it takes for the rat to flick its tail out of the range of the focused heat rays. Another pain assay has the pharmacologist place a mouse on a hot plate. The measure of pain is how long it takes for the mouse to jump off. One problem with the hot plate assay is that about 3% of the mice jump off immediately. Are these mice assay failures to be ignored, or is this a bona fide measure of pain?

In the 1950s, 60s, and 70s, attempts were made to move experimental pain studies from animals to humans. As was done with rats and mice, human volunteers were subjected to pain stimuli, and a measure of pain was derived from how long the volunteer could take the pain before asking it to end. Experimental pain was induced in several different ways. The most widely used procedure was to plunge the volunteer’s hand into ice water, pain measured by the amount of time he (almost all the volunteers in these studies were male) could keep his hand in the ice. Another was to tighten a thumbscrew onto one of his thumbs, tightening it steadily, and pain was measured by the pressure at which the volunteer asked for it to end.

As of this writing, fewer and fewer experimental pain studies in humans are being run. Many critics have raised ethical qualms because the treatment (induced pain) has no medical value and written ethical standards (as embodied in the World Health Association’s much modified Helsinki Declaration) require that any experimental “treatment” given humans has to be of some potential medical benefit. A further reason to drop these studies is that they could not detect a difference between known analgesics like aspirin and placebo. The only type of pain-relieving drugs they could detect were opioids.

11.3 Measuring Pain in Patients

In 1952 and 1953, Henry Beecher (1904–1976) and Louis Lasagna (1923–2003) at Harvard Medical School studied the relief of pain in patients undergoing abdominal surgery. At that time, the most common surgery in the United States was the removal of the gall bladder. This required a surgeon to make a relatively long incision in the stomach of the patients. The recovery from this surgery left patients in considerable pain. Beecher and Lasagna randomly alternated between placebo and a low dose of morphine in responding to a patient’s pain. They did not attempt to measure the pain but used as their endpoint whether the patient stated that the pain had been relieved.

What they discovered was that almost half of the patients found relief from placebo at least once in the course of their treatment. Twenty to thirty percent of the patients had relief almost every time they were given a placebo. With the aid of Frederick Mosteller, who was chairman of the Harvard Statistics Department, they decided to see whether they could identify the type of patient who would respond to the placebo. Among other characteristics of patients that they looked at, they gave patients the Minnesota Multi-Phasic Inventory test, usually referred to as the MMPI. They separated patients who responded to placebo almost all the time and patients who never responded to placebo. They could find no characteristic of patients that predicted whether or not the patient would be a placebo responder. However, the patients who never responded to placebo were peculiar. Their MMPI scores indicated that they were borderline paranoid. They were highly suspicious of medical actions and tended to be loners.

In the 1960s and 1970s, as non-steroidal anti-inflammatory drugs like ibuprofen became available, research began to mature on the measurement of pain. Prominent among these researchers were Abraham Sunshine (1929–2007), Raymond Houde (1926–2016), and Stanley Wallenstein (1921–1996). Eugene Laska of the Nathan Kline Institute for Psychiatric Medicine at New York University has provided much of the statistical backbone to this research. (See Laska et al. 1986.)

Pain was something that ranged from none or mild to severe and seemed to be a candidate for setting up some sort of scale, like measuring pain on a scale from 1 to 10. Various ways of depicting this scale were tried. The patients might be given a 10 mm line with "no pain" at the left end and "unbearable pain" on the right end. The distance from left to right was taken as a measure of pain—except that some patients got mixed up and sometimes graded their pain from left to right and other times from right to left. Furthermore, patients often belonged in one of two classes: One type of patient always remained somewhere in the middle of the line, while the other class, the extremists, jumped from one end of the line to the other.

These researchers tried to overcome the confusion by giving the patients a "pain thermometer," a vertical column of little squares. They tried a "pain speedometer," a curved line with zero on the left end and some number like 100 on the right end. Even with these visual aids, the patients still divided into the thin slicers and the extremists.

Rensis Likert (1903–1981) is known primarily for his development of the psychological aspects of management in his book, New Patterns of Management. However, in 1934 he published a paper on the conversion of ordered categories into a scale of numbers. The problem Likert examined went like this: We can take some subjective feeling (pain?) and produce an ordered set of categories that describe that feeling in an increasing way (no pain, very very little pain, slight pain, moderate pain, uncomfortable pain, severe pain). Suppose we assign a numerical value to each category such as

- 1 = no pain
- 2 = very very little pain
- 3 = slight pain
- 4 = moderate pain
- 5= uncomfortable pain
- 6= severe pain.

Likert asked, can we use these numbers to calculate changes in condition on the average? Likert's answer ran like this: If we want to look at an average, then the numbers that we use to compute the average must measure the same thing in such a way that a change in x units for a patient who started with lower levels of pain is equivalent to a change in x units for a patient who started with an upper level of pain. For instance, we have to be able to equate the patient who has a two-unit change (slight pain (3) to no pain (1)) at one end of the scale to a patient who has a two-unit change (severe pain (6) to moderate pain (4)) at the other end of the scale.

Scales based on ordered categories that have this property are called "Likert Scales." Likert provided no method for determining if a scale has these properties. Since then, investigators who convert an ordered set of feelings into a set of numbers often justify their use of numbers by calling that conversion a "Likert Scale" with no effort made to determine whether it has the appropriate Likert characteristic.

Through the 1970s Sunshine, Houde, and Wallenstein ran a series of studies to identify the best way to turn pain into a numerical scale. They dismissed the use of a large number of categories because of the two types of patients. They eventually decided that patients and attending medical personnel could produce consistent results only if the scale was based on no more than 4 conditions:

- 0 = no pain
- 1 = slight pain
- 2 = moderate pain
- 3 = severe pain

Any attempt to increase the number of categories lead to violations of Likert's condition. In further refinements, they decided that the most consistent results were based on attending nurses' evaluations. They trained nurses, who would put the patient through a sequence of movements and who would evaluate the degree of pain from the patients' responses. They also decided that patients who began the study with "severe" pain represented a different type of patient. Any analysis had to be blocked (recall Fisher's blocks + treatments + error) so that patients with entering severe pain were in one block and the rest of the patients were in another block.

They increased the sensitivity of the process to differences in treatment effects by judging the patient's pain at several points in time and using the sum of the differences from baseline as the measure of efficacy. They called this measure the Sum of Pain Intensity Differences (SPID).

With this design—four-point scale, calculation of SPIDs, and putting patients with baseline severe pain into a separate block—these investigators were able to run experiments that established efficacy and dose-response curves for the non-steroidal anti-inflammatory drugs and modified opioids that were produced over the next 10–15 years.

11.4 Lessons Learned from Pain Scales

What this teaches us is that

1. You might be able to produce an ordered set of categories for a subjective measure, but this does not mean you can convert those categories into a set of numbers.

2. It is necessary to determine whether individuals can produce consistent conclusions when they deal with those numbers.
3. Extensive scales (like using numbers from 1 to 10) can produce severe problems of interpretation.
4. It might be necessary to block on categories that are qualitatively different from the others.

Perhaps the most important lesson to learn from the development of pain scales is that it is very difficult to convert ordered categories of subjective conditions into numbers that lend themselves to statistical calculations.

11.5 Summary

Statistical experimental design and the analysis of results require the use of numbers. When dealing with subjective assessments, it is necessary that the numbers recorded and analyzed fulfill several minimal requirements. Pain is used to illustrate these aspects. The numbers derived to measure pain have to have equivalent, meaning for all patients.

A 10-point pain scale cannot work because some patients stay in the middle of the scale and some jump from one end to another. Attempts to put numbers on categories of pain have to meet the Likert condition that a change of x points from one end of the scale has to be comparable to a change of x points from the other end. Studies done in the 1950s, 60s, and 70s showed that almost everyone responds to placebo at some point. Those who do not are borderline psychotic. The only consistent pain scale is one that identifies only four categories: none, mild, moderate, and severe. Patients with severe pain form a block that is different from patients with lower levels of pain.

References

Laska, E. M., et al., (1986) “The correlation between blood levels of ibuprofen and clinical analgesic response,” Clinical Pharmacology, 40, 1-7.

“Declaration of Helsinki” (2001) Bulletin of the World Health Organization, 79(4), http://www.who.int/bulletin/archives/79%284%29373.pdf.

Chapter 12: When the Experiment Goes "Wrong"

12.1 The MRFIT Study

By 1969, enough epidemiological data had accumulated to provide the medical community with a set of very good predictors of whether a patient would have a heart attack. These factors included age, gender, whether the subject's father had a heart attack, smoking, use of alcohol, lack of regular exercise, obesity, and diets high in animal fats. The first three of these predictors could not be changed, but it seemed reasonable that adjustments in lifestyle that dealt with smoking, alcohol, exercise, obesity, and diet could have a beneficial effect.

However, as most experimental scientists know, what seems reasonable is not always true. That is one reason to run an experiment. Two studies were initiated. One of them dealt with American and Canadian patients and was funded by the National Heart, Lung, and Blood Institute of the National Institutes of Health. The other was sponsored by the World Health Organization and consisted of a group of studies of similar design begun in different European countries.

The basic design of these studies was straightforward. Find a group of men who had a high likelihood of having a heart attack. (They used men because gender was an important factor in predicting whether a heart attack will occur.) Leave some of them to the usual medical practice of the time. Subject the others to intensive education about the preventable factors that appear to "cause" heart attacks. Using the statistical design of experiments, they blocked on clinics and randomly assigned patients to be in the usual care (UC) group or in the special intervention (SI) group.

With the penchant for giving eye-catching names to medical studies, the American study was called the Multiple Risk Factor Intervention Trial or MRFIT and pronounced "Mister Fit." The studies sponsored by the World Health Organization were the WHO MRFIT. Having a catchy name was only the beginning of planning for these studies. Clinical studies have protocols, which are detailed descriptions of how the study will be run, including justifications for the study, the type of patient who will be entered, the nature of the "treatments" to be compared, how often patients will be seen, and how the endpoint will be evaluated. The directors of the clinics that will enter patients into the study review the proposed protocol, make suggestions for changes, and all clinics sign on to the final version.

The American MRFIT study was a multi-clinic study with a single data center where case reports on individual patients were collected, from where the conduct of the study was controlled, and where the final data were to be analyzed. One question that needed to be answered before the finalization of the protocol dealt with the selection of patients to be entered into the study. They needed to start with patients whose risk factors were high. To do this, they created a scoring based on which of the risk factors were present and to what degree. The American study screened over 370,000 potential patients and entered 12,866 of them into the trial. The patients in this study were followed for seven years.

A study this size usually lasts longer than the seven years of follow-up. It takes time, often measured in years, to recruit the almost 13,000 patients. Because of this, the study runs for more than seven calendar years, and when the data are locked and the analysis of data run, there will still be some patients who have not completed the full seven-year follow up.

All the patients in the study were told about lifestyle changes that should reduce the probability of having a heart attack. It would have been unethical not to. The UC patients were told of these factors and then sent back to their primary care physicians who would monitor them during the seven-year span of the study. The SI group were given classes in which proper eating was encouraged. If they smoked, they were assigned to smoking cessation programs. They received weekly telephone calls from a nurse to encourage them to stay with the proper regimen.

Both groups of patients had fewer heart attacks during the seven-year period than might have been expected from epidemiological studies. By the end of the seven years, deaths from heart attacks were 17.9 per 1000 (1.79%) among the SI patients and 19.3 per 1000 (1.93%) among the UC patients. This implied a 7.2% decrease in fatal heart attacks for the SI group. Both groups of patients had fewer heart attacks during the seven-year period than might have been expected from epidemiological studies. By the end of the seven years, deaths from heart attacks were 17.9 per 1000 (1.79%) among the SI patients and 19.3 per 1000 (1.93%) among the UC patients. This implied a mere 7.2% decrease in fatal heart attacks for the SI group.

12.2 What Went "Wrong" with the MRFIT Study?

For many in the medical community, the results of the MRFIT study did not make sense. The patients were engaging in lifestyles that greatly increased their chance of having a heart attack. They were male. Their fathers had had heart attacks. They were smokers. They drank alcohol. They were overweight. Their ordinary meals were filled with animal fats. Many of them had high blood pressure.

The SI (Special Intervention) group were put into smoking cessation programs. They were given weekly calls to encourage them to stay with their regimen. They were given information that included weekly menus for appropriate low-fat meals. Could it be that intensive counseling was a waste of effort and money? True, the UC (Usual Care) patients were told about lifestyle changes that might prevent heart attacks, but there was no concentrated follow-up. They went home and did as they pleased. It is very difficult to cease smoking all by one's self, and this holds for excessive use of alcohol and for changes in diet.

The medical community tore into the MRFIT study. What went wrong? One criticism was that many of the clinics used residents or even medical students to see patients on their clinic visits. This provided excellent training, but can you really expect that such inexperienced personnel would run the study properly?

Another criticism was based on the finding that many of the SI patients who died of heart attacks had high blood pressure at the start of the study. The standard treatment for high blood pressure at the time was to put the patient on relatively high doses of diuretics. Perhaps the standard treatment is at fault. Could the high doses of diuretics be causing heart attacks in some patients? Or was this finding a random glitch in the data that Frank Anscombe called "will o' the wisps," apparent relationships that are purely random noise and have no predictive value?

If the experiment had included hundreds of thousands of subjects, it might have shown a significant, but similarly very slight, difference in death rates between the treated and controls. W. Edwards Deming (1900–1993), who had been instrumental in bringing Fisher's ideas to industrial quality control, once noted that we usually use the data from an experiment to test whether two treatments have the same overall mean effect. But, he wrote, "It is foolish to test whether two means are the same. They are never equal, and with a large enough study, it can be shown that they are not equal." The real question, he proposed, is not whether the difference is greater than zero but, rather, whether the experiment shows that the difference is sufficiently large to make a useful difference in final outcome if the tested treatment were to be adopted.

12.3 The Hawthorne Effect

During the 1920s and into the 1930s, General Electric ran a series of studies at their Hawthorne plant outside of Chicago. The goal was to find ways of reducing the incidence of accidents and increase the plant productivity. They introduced a series of measures designed to prevent accidents and watched the workers as they went about their jobs. The accident prevention measures remained in place whether the workers were watched or not. Whenever the workers were being observed, accident rates went down. When they were no longer being watched, accident rates went back up.

This is the "Hawthorne Effect"—the very act of observing and measuring improves the outcome.

The National Institutes of Health (NIH) of the United States government have been sponsoring clinical studies that examine the effects of different medical measures since the end of the Second World War. If the study is designed to prevent some deleterious event (like a heart attack), then it usually happens that the mere act of putting a patient into a study to be seen at regular intervals causes the incidence of that event to drop—even if the patient is on placebo. There is something about the "hands of the physician." Thus, suppose the study is planned to find out if some procedure prevents an event (like a heart attack) that normally occurs in 5% of this population over the seven-year length of the study. Putting the patients into a controlled clinical trial results in an incidence of less than 2% among the controls.

The MRFIT study seems to have been caught up in the Hawthorne effect.

12.4 Anticipating the Outcome of an Experiment

A good scientist who is about to start an experiment usually has an idea of what the outcome will be. Experiments are sometimes used to refine established measurements or to decide between two well-defined possible outcomes. Signals from two orbiting satellites might be used to plot variations in gravity around a mountain range. But before the data are examined, previous measurements will have provided a general idea of the nature of those variations. Results that differ greatly from what is expected will call into question the assumptions made about the experiment.

Did the outcome of the MRFIT study differ greatly from what had been expected? If one looks upon the MRFIT study as an attempt to validate something that was believed true (intensive counseling can reduce the incidence of death from heart attacks), then it was not a scientific experiment but an exercise in politics. All too often, one can find articles in the medical literature where the introduction describes why such and such a relationship is important and concludes by showing that the relationship holds. While such studies may be useful in advocacy, they can hardly be considered as advancing scientific knowledge.

Can the MRFIT study be used to determine whether intensive counseling would be a useful tool to add to the physician’s armamentarium? If there were unlimited money available, then even the slightest suspicion that this is true would be enough to engage in this practice. However, there is not unlimited money available for the treatment and prevention of disease. The question underlying the MRFIT study was whether money spent on intensive counseling would be better spent elsewhere. How much of a reduction in deaths from heart attacks can be expected from such activity? Many viewed the MRFIT study as a well-done experiment with this conclusive finding: intensive counseling has a minimal effect on the incidence of death from heart attack.

12.5 Cost versus Efficacy

The MRFIT study did not find a decrease in deaths greater than might be expected by random noise alone. A reasonable conclusion is that the MRFIT study showed that the money that might be spent on hectoring patients with lifestyles that predict heart attacks would be better spent on something else. Can this trade-off between outcome and cost be applied to other medical procedures?

Consider a procedure that is engaged in by Emergency Medical Technicians (EMTs) in many states when they are transporting a patient suspected of having a heart attack. The patient is given six baby aspirin tablets to chew and swallow. There is no pharmacological basis for this practice. Aspirin, given at a dose between 75 and 100 mg, reduces the stickiness of blood platelets generated by the bone marrow. The first part of blood clot formation occurs when the platelets stick together to form a web in which white blood cells gather. Blood platelets have a four-day half-life. That is, the platelets are gradually destroyed and absorbed back into the body at a relatively slow rate, so half of the platelets produced on a Monday are still around on Friday. In order for aspirin to be effective in preventing heart attacks, it must be taken steadily for 12 days (three half-lives of platelets). In view of its complete lack of rationale, should this practice be continued? Should it be subject to an experimental study to see if it "works"?

The answer hinges on cost. Unlike special intervention in the MRFIT study, aspirin is very inexpensive, a very small fraction of the cost of other treatments available on the ambulance. Unless the patient is allergic to it, there is very little chance that a single dose of 81 mg (or even 6 baby aspirins, 486 mg) will do any harm. I have gone through this exercise in decision theory with students, and most agree that there is no need for a clinical study and that the practice can continue. The cost is negligible, and it might save lives.

I, then, pose to them the following question: "Philosopher's Stone" (ground up goat gallstones) was a favorite "cure" in Medieval medicine. Suppose the EMTs were offered pills of Philosopher's Stone at no cost. Should the patient suspected of having a heart attack be given such pills? Should the use of Philosopher's Stone be subjected to a proper clinical trial? Most of the students who thought that aspirin should continue to be given without needing a trial objected to Philosopher's Stone—why?

A final note: Throughout these last two chapters, I have used the phrase "heart attack," but this phrase has no clear medical definition. In the medical literature, you will find it replaced by the term "myocardial infraction (MI)," which is defined as the death of the cells in a region of the heart after a blood clot has blocked the flow of blood to that region. For the convenience of readers who might not be familiar with the concept of an MI, I have used the lay term "heart attack" when I meant an MI.

12.6 Summary

The MRFIT study compared the death rates from heart attacks between two groups of men who were of high risk to have a heart attack. The usual care (UC) group were told about their lifestyle practices (smoking, use of alcohol, obesity, high animal fat diets) that increased the risk of a heart attack. The special intervention (SI) group were given intensive follow-up with weekly phone calls, smoking cessation programs, and so on. After seven years, there was no significant difference in the rates of death from heart attacks between the two groups. Was this a failure of the experiment, or did it show that special intervention of this type was not useful? If an experiment seems to show that a very reasonable procedure is a failure, does that mean the experiment went wrong, or does it mean that this "reasonable" procedure is a waste of money?

Reference

Stamler, J. (2008) "The Multiple Risk Factor Intervention Trial (MRFIT)—Importance Then and Now," J. Amer. Med. Assn., 300, p 1343.

Chapter 13: Summing Up

13.1 Ronald Alymer Fisher

This book has been a tour of a subject that has generated untold numbers of scientific papers and books. Courses in graduate schools have been devoted to it. Most importantly, it has remade the face of scientific experimentation.

In 1920, R. A. Fisher arrived at the Rothamsted Experimental Station. He had been hired for one year at a salary of 1000 pounds (equivalent to $71,000 in today's currency) to go over the data that had accumulated in more than 50 years of agricultural experimentation. He was later to call it "raking over the muck heap." He rented a cottage and brought his wife, children, and mother-in-law, left them in the cottage, pulled on his boots, and crossed over a muddy field to see what that year might bring.

A 17th-century bishop had warned us that malevolent Nature is standing by to foul-up any attempt at experimentation, and Nature had had a great time creating the muck of 50+ years' "experimentation" at Rothamsted. Talking to the agricultural scientists, tramping around the experimental fields, and turning through the pages of numbers that had accumulated, Fisher formulated a mathematical structure for experimentation.

He started with the numbers that emerge from an experiment, the weight of the wheat, the ratio of wheat to straw, the number of potatoes. Then, there were the treatments applied to the experimental material. Finally, there were the changing aspects of the material being experimented on, like the annual rainfall, the "fertility gradient," and the infestation of weeds. He put all of that into a small set of algebraic formulas where he had two types of symbols: the things that he could observe and measure or count (denoted in the formulas by Roman letters), and the relationships between the different things that he could observe that have to be inferred from his observations (denoted in the formulas by Greek letters).

Putting the problem into algebraic equations tells us a great deal about the experiment before we even begin. For instance, there is the problem of confounding. The year-to-year variation in the weight of the wheat harvest is affected by both the different fertilizer treatments given in different years but also by the difference in rainfall from year to year. The differences in treatment effect are confounded with the differences in rainfall. In the Lanarkshire Milk Experiment, the increase in a child's weight from February to June was affected by the different type of milk the child drank but also by the child’s socio-economic status. Taking pity on them, the teachers gave the poor children the extra milk, thereby confounding the effect of milk with the socioeconomic status of the child’s family.

Malevolent Mother Nature keeps adding to the mix. Some of the mice who arrive with the same genetic background are, in fact, sicker that the others in the box, and putting the sicker mice in the cages on the top of the racks confounds their initial condition with position in the room. An unexpected infestation of the spores of wheat rust fungi sweeps through a portion of the test field. Slight imperfections in the carbon/iron combination of the steel produce differences in the quality of the annealing. Some of the patients assigned to ordinary care cut back on their destructive lifestyle without being subjected to the intensive coaching given to the “treated” group.

These random glitches, these unexpected differences, were taken care of in Fisher’s modeling in two ways: the nature of these random “errors” was described mathematically through the use of calculus (with more unknown relationships that have to be estimated from the data and are denoted by Greek letters). Then, these random glitches were “tamed” by adding additional randomness to the experiment. Experimental treatments are assigned at random to different units of the experiment (whether these units are fields of wheat or cages of mice or individual patients).

Not haphazardly but AT RANDOM!!

13.2 Computing

All of this requires a great deal of thought and planning, and the data that result have to be analyzed following the complications of the algebraic model and its calculus obbligato. Fisher had a desk calculator called “the Millionaire” because it had enough places on its platen to hold numbers in the millions. It had no electric motor, but it required that Fisher pull a lever whenever he had a calculation set up. It could add and subtract, and, if you rigged it the right way, it could multiply and divide.

Fisher showed in a mathematical proof that the act of randomizing experimental units to treatments provided an approximation to the extremely complicated mathematics that some experimental designs require. When I described these computations, I had to “wave my hands,” because the proof requires advanced calculus, complex analysis, and multi-dimensional algebra, all of which are the subjects of courses in graduate school.

Gossett objected to Fisher’s use of asymptotic theory (which assumes that the number of observations is very large) because he never saw any experiment that involved more than a few hundred units. But if you tried to estimate the values of the Greek letters in the model using just the calculus imposed by the random assignment of treatment and without resorting to Fisher’s approximations, it would involve tens of thousands of calculations, perhaps even millions. Either the gears of his “millionaire” or Fisher’s arm would wear out before coming close to the end.

Then came the computer. Initially, it was a big machine that knew how to add and subtract, which it could do over and over and over… In the last two decades of the 20th century, the computer got smaller and smaller, and sophisticated programs taught it how to multiply, divide, read algebraic notation, and do calculus. Now, in the 21st century, even a laptop computer can handle the standard statistical packages that do all these calculations for you. The scientist needs to know only how to set up the algebraic equations to describe the experiment and which of the different calculus-based choices that she needs in order to describe the randomness.

13.3 The Ubiquitousness of Statistical Designs of Experiment

Statistical design of experiments came first to agricultural experiments, then it was sociology, psychology, physics, chemistry, biology, ecology, and quality control, and (in the 1950s) even medicine. There are still places in science where the initial experiments are probes involving clever methods of measurement, but, in most fields, it is the well-designed randomized experiment that provides the final “proof” of the finding. The terminology often differs from field to field. Atomic physicists look for “six sigma” deviations, structure-activity chemists look for a high percentage of variance accounted for, and medical scientists describe the “specificity” and “sensitivity” of measurements. But all of it starts with statistically based design of experiments.

Afterword

As David Salsburg writes, R. A. Fisher created a revolution in experimentation. It has been nearly 100 years since Fisher introduced his four principles of experiment design—randomization, blocking, replication, and the factorial principle. These principles are as necessary today as they were 100 years ago. Randomization is a vitally important technique that protects against bias due to uncontrolled (but influential) factors that change over the course of an experiment. By removing the variance due to a nuisance factor from the estimate of the error variance, blocking increases the power for identifying the key controlled factors. Replicated runs in an experiment yield an estimate of the error variance that is unbiased even if the fitted model is incorrect. Finally, the factorial principle allows for varying multiple factors over the course of any experiment instead of being limited to the study of only one factor at a time.

It is violations of one or more of these principles that are at the bottom of many of the stories in Salsburg's latest book, *Cautionary Tales in Experimental Design.* This book uses well-researched and interesting examples to trace the statistical design of experiments (DOE) from its origins to recent applications in science and industry.

To solve problems and learn from data, the investigating team needs to carefully consider the data collection process, think critically about what questions they are trying to answer and anticipate what can go wrong. These necessary activities inform the process of coming up with an appropriate design.

In the early days of DOE, tables of experimental plans appeared in textbooks or design catalogs. The problem for the engineer or scientist was to find the catalogued design that most closely matched the information needs and the resource constraints of the process or system being studied. The most useful characteristic of these catalogued (or classical) designs was their orthogonality, which allowed investigators to estimate the effects of the factors with various averages of the observed responses. Such calculations could be done without computers, which, in the early days of DOE did not even exist. The "Achilles Heel" of these classical designs was their lack of flexibility. Their number of experimental runs is fixed, and each run must be performed exactly as specified. The allowed number of runs in a block is also restricted. For example, if the team is trying to block day-to-day variation, the number of runs that can be performed in a day may not match the allowed block size. For industrial applications, another necessary feature is restricted randomization, which occurs when it is logistically required to hold a factor constant for several runs in a row. Classical designs are generally incapable of providing a principled method for accomplishing this in a flexible way.

The reason for all this lack of flexibility in classical designs is the requirement for orthogonality. Orthogonal designs may not exist given a set of information requirements and resource constraints. As a result, engineers and scientists who are limited to these designs must often change their requirements. If they refuse to do so, they may choose to forgo DOE entirely. This is undesirable and no longer necessary.

Due to the ubiquity of high-speed computers, there is no longer a need to compute factor effects by hand. Over the last 30 or so years, computer algorithms have been developed to search the multiple dimensional space of the factors to find a design that simultaneously meets the information requirements given the resource constraints. The resulting designs are called optimal designs. However, it is not their mathematical optimality that matters, it is their flexibility to solve design problems as posed rather than change the problem to suit a prespecified design. Commenting on these computer-aided methods, Stu Hunter, legendary statistician, educator, and co-author of an important DOE textbook, recently stated that “the art of experimental design has changed profoundly” and that if he had to teach DOE now, he would have to teach it in a profoundly different manner.

Machine learning and big data now show promise for faster innovation. However, these tools are dependent on observational data. Therefore, they can establish correlation *but not causation* between a possible input and desired output. By actively interrogating a system or process, DOE can follow up on conclusions from a big data project with a small study that can either establish the validity of these conclusions or prove them false.

While the title of this book is *Cautionary Tales in Designed Experiments*, the beauty of DOE is about learning—from mistakes, from trying new things, from working with others. Given reminders of past mistakes, it is possible to learn to avoid the same errors in the future. The trailblazers in industry who have achieved rapid innovation, solved complex problems, and created significant value have often failed multiple times before a breakthrough.

It is gratifying to be a part of this exciting new era in DOE, further enabling industrial problem solving, faster insights and innovation with less waste.

Bradley Jones
Distinguished Research Fellow, SAS

References

Box, G., Hunter, W. G., and Hunter, J. S., (1978) Statistics for Experimenters: An Introduction to Design, Data Analysis, and Model Building, John Wiley & Sons.

Goos, P., and Jones, B., (2011) Optimal Design of Experiments: A Case Study Approach, Wiley.

Jones, B., and Montgomery, D., (2019) Design Of Experiments: A Modern Approach, Wiley.