EXPERIMENTS: RANDOM SAMPLES FROM CONCEPTUAL POPULATIONS

S. R. Searle

Biometrics Unit, Cornell University, Ithaca, New York

BU-729-M January 1981

Abstract

Some elementary features of designed experiments are described and illustrated.

Paper BU-729-M in the Biometrics Unit.

EXPERIMENTS: RANDOM SAMPLES FROM CONCEPTUAL POPULATIONS

S. R. Searle
Biometrics Unit, Cornell University, Ithaca, New York

1. Election Polls

Statistics is concerned with gathering (and interpreting) information about populations gained from samples. Often, the populations are very large. For example, the pre-election polls last November were intended to give information about voter intentions for the whole country. That's a large population. Yet the information used was gathered from only some 1500-2000 would-be voters.

How were these 1500 or so people chosen? Did the pollster go to Wall Street to get his sample? Or did he take a sample of 1500 people from Plains, Georgia? He did neither of these things. The sample of people used as a representative sample of the whole U.S.A. was selected as a random sample of the population; maybe each person was not selected at random from the population of the whole country, but from one of several carefully designed sub-populations so that the total sample is a collection of random samplings.

Why this emphasis on selecting at random? Because the pollster wants his sample information to be representative of, or a reliable indication of, the information in the whole population. He does not want the information in the sample to be unduly affected by what he knows, or thinks he knows, about the political persuasion of certain groups of people – those who work on Wall Street, or live in Plains, Georgia, for example. He does not want them to be the only people in the sample, nor to be purposefully disproportionately represented therein.

2. Experiments with New Drugs

2.1. A conceptual population

The idea of sampling a population also applies in conducting experiments, but with a few changes in outlook. Consider testing the efficacy of a new drug. The major distinction between this situation and that of the pollster is that whereas the pollster knows that his population exists, what it is and where it is, in the drug-testing situation there is no population of users of the new drug from whom a sample can be selected. Because it is a new drug and has not yet been approved for the consumer market, those millions of users constituting a population of users do not exist in the way that millions of voters exist.

2.2. Sampling the conceptual population

So an experiment has to be done. We have to try out the drug on a limited number of people. And in doing this, we keep in mind that from this group of people we hope to draw conclusions about the new drug that apply to everyone – or at least to everyone who might have genuine need for the drug. In this way we do have a population, a conceptual population, the population of future users. And in performing an experiment we think of the people in that experiment as being a random sample, or a collection of random samples, from the conceptual population of users. And the experiment must be designed accordingly.

In planning the experiment we must take care not to select people just from a group of people whom we know, or think we know, have special characteristics that would make their reaction to the drug different from the reaction of the "average" person. True, the whole conceptual population of users may consist of several such groups of people (for example, people at different ages), and we may want to randomly sample each of these groups to make up our experiment.

As an illustration, the new drug might be a new progesterone formulation for the treatment of dysmenorrhea. Then we would not want the experiment to consist solely of women in their 70's, any more than the political pollster wanted just people from Plains. The experiment may well consist of a sample of women from each of several age groups, say in their teens, 20's, 30's and 40's, or maybe even finer age groupings. And in each case the women selected for the experiment would be a random sample of women of their age groups.

2.3. Random sampling

Why must we insist on this "selecting at random"? For basically the same reason that we do not want just 70-year olds. Women in their seventies do not suffer very much from dysmenorrhea; we know this, and so we take care not to base our experiment on them and them alone. Women of other ages may have characteristics that affect their reaction to the drug, too, even though we may not know about them. So, to avoid any possibility of these unknown characteristics disproportionately affecting the information we get from our sample, we must make sure that the sample is chosen randomly.

Suppose, for some reason, that the new drug is being administered in pill form with the instruction "to be taken in the morning with your first cup of hot coffee or hot tea". Suppose, too, that it is being used in a public health clinic, to which bus loads of patients come throughout the day, and the protocol says "use the new drug on the patients who come between 11:00 and 11:15 a.m.". Finally, suppose that the drug becomes ineffective if dissolved in milk. What would be the effect on a day's data if the 11 a.m. bus load of patients was, quite unbeknownst to us, from a Mormon church?

The principles of randomization are precisely for lessening, or averaging out, the effect of unknown characteristics like this, which occur with every individual of a population.

2.4. Comparison with a control

The object of our experiment is to assess the efficacy of a new drug. This might suggest administering just the new drug and measuring its effect. But good experimental procedure dictates that a control drug, or even a placebo, should also be administered as part of the experiment. In situations where the effect of the control drug is well known (maybe a drug that has been on the market for many years), one may wonder why it needs to be used in an experiment concerning a new drug. The reason is again basically the same as the reason for randomizing — to take into account the unknown vagaries, in this case those connected with running the experiment: for example, the ability of the nurses giving the injections, the quality and sterility of the needles, and the laboratory standards and practices where the subsequent blood work and lab analyses are done. All these things and more can affect the information obtained from the sample. So, in order to get some measure of them other than as part part and parcel of the response to the new drug, we use a standard drug, the response to which, for this experiment, also includes the effects of those vagaries of running the experiment. Then the effect of the new drug is measured as the difference between the responses to the new and standard drugs:

$$\begin{aligned}\begin{pmatrix}\text{Response}\\ \text{to new drug}\end{pmatrix} - \begin{pmatrix}\text{Response}\\ \text{to control}\end{pmatrix} &= (\text{effect of new drug} + \text{effects of experimental conditions})\\ &\quad - (\text{effect of control} + \text{effects of experimental conditions})\\ &= (\text{effect of new drug}) - (\text{effect of control}) \quad .\end{aligned}$$

Of course, the moment we have two different drugs to administer the question arises: for each person in the experiment how do we decide which drug to give them? Again we must rely on randomization. Within each class of people constituting the experiment (the teen-agers, the 20-year olds, 30-year olds, and so on), one of the two drugs will be allocated to each person at random. This should be done in such a way that neither the doctor administering the drug, nor the person receiving it, will know which drug is being used. Again, this randomization is used as insurance against possible effects on the reported response that might arise either from the doctor deciding what drug to give each particular patient, and/or from a patient's knowing which drug has been administered. This is, of course, the well-known double-blind technique that is frequently used in drug trials.

3. Unsuccessful Experiments

Rather than extolling the virtues of randomization in purely general terms let us look at some experiments, some real, some imaginary, where the lack of randomization shrouded the intended purpose of the experiment – i.e., nullified its value. Not many such experiments get into the statistical literature – for the very reason that ultimately they are seen to be poorly designed and so never get reported. It is therefore not easy to find written descriptions of ill-fated experiments. But there are some.

3.1. Vaccinating children

For the Salk vaccination trials against poliomyelitis in 1954, parental consent was needed before a child could be vaccinated. Data were collected from two trials. One was a double-blind trial of 400,000 children for whom consent had been given, and the other involved 350,000 children for whom 225,000 had parental consent and they were vaccinated, and 125,000 had no parental consent and they were not vaccinated. The results (Freedman *et al.*, 1978, p. 6) are shown in Table 1.

Table 1. Incidence of Polio: Number of Cases per 10^5 People

	Double Blind	Vaccinate all Consenters
Vaccinated	($\frac{1}{2}$ of consenters): 28	All consenters : 25
Not vaccinated	($\frac{1}{2}$ of consenters): 71	All non-consenters: 44
Reduction in incidence	1 - (28/71) = 61%	1 - (25/44) = 43%

The double-blind data give an unbiased estimate of the reduction in the incidence of polio; but the other data do not. The reason is that the results from the other data are affected by two factors: (i) high-income parents consent to their children being vaccinated more readily than do low-income parents, and (ii) children of high-income parents are more vulnerable to polio than are those of low-income parents. These two facts lead

(a) **in** the double-blind experiment, to an over-estimate of both incidence rates but they are both over-estimated by the same factor, and so their ratio provides a good estimate of the effectiveness of the vaccine, and

(b) in the vaccinate-all-consenters data, to over-estimation of the incidence among the vaccinated, and under-estimation among the non-vaccinated – and hence the effectiveness is under-estimated.

3.2. The value of a control treatment

Freedman *et al*. (1978, p. 8) summarize data from 51 different studies on the *portacaval shunt* operation to redirect the flow of blood in acute cases of cirrhosis of the liver. Some of the studies were done without controls, and of those done with controls some used randomization and some did not. Enthusiasm for the operation was as shown in Table 2.

Table 2. Summary of 51 Studies on the Success of an Operation

Study	No. of Studies	Measure of Enthusiasm		
		High	Moderate	None
No controls	32	75%	22%	3%
Controls				
No randomization	15	67%	20%	13%
Randomization	4	–	75%	25%

The contrast between the randomized and non-randomized situations is startling. One possible explanation is that when there is no randomization patients often get assigned "to treatment or control according to clinical judgment [and] there is a natural tendency to [give the treatment to] only the patients who are in good shape".

3.3. Testing sun-tan lotion

The efficacy of a new sun-tan lotion was once tested in London and New York using a random sample of 900 people in each city. For each person, carefully placed circles of lotion were applied to the back of the hands, the new lotion on the left hand and the old on the right. This was done in early April, and in mid-June light-meter readings were taken of the lotion-covered and of the un-lotioned areas of the hands. The average difference between these two readings came to be as indicated in Table 3. With the estimated standard error of an observation being 6, each of these differences is significantly different from zero. In experiment 1 the new lotion is superior to the old in New York, but vice versa in London. Why? Probably because the experiment was done without regard for the fact that cars in New York are driven on a different side of the road from London. Hence in New York the left hand is

Table 3. Average Difference for 900 People, between Light Meter Readings of Protected and Unprotected Skin on the Back of the Hand Using New and Old Sun-tan Lotion

Experiment	New York		London	
	Left	Right	Left	Right
1. Not randomized				
900 in each city	New 10	Old 4	New 6	Old 8
New minus old	6		-2	
2. Randomized				
450 in each city	New 10	Old 4	New 6	Old 8
450 in each city	Old 8	New 6	Old 4	New 10
Mean difference (new – old)	2		2	

exposed to more sun than the right and vice versa in London. This effect is accounted for in experiment 2 where, in each city, the new lotion was put on the left hand for a random half of the sample, and on the right for the other half.

3.4. Rabbit ears

Some twenty years ago a pharmaceutical company tested the healing powers of a new antibiotic by scratching the ears of a number of rabbits and then treating the left ears with the new antibiotic and the right ears with a placebo. The left ears healed about three days more quickly than the right ears. The medical staff were delighted: the antibiotic looked like a winner. But the statistician wasn't happy with the unrandomized experiment, so he repeated it with placebo on both ears of another set of rabbits. Lo and behold, the left ears still healed three days more quickly than the right – a fact (albeit unbelievable, it seems to me) that was then well known with certain strains of rabbits.

3.5. Subjective assignment

Cox (1958) discusses an experiment in England in 1930, where some schoolchildren got free milk each day and others did not. Average increases in weights might have been something like those of Table 4.

Table 4. Average Weights (lb) of School Children Receiving, and not Receiving, Free Milk at School

Month	No Milk	Milk
February	$82\frac{1}{16}$	81
June	$83\frac{5}{8}$	$82\frac{1}{2}$
Difference	$1\frac{9}{16}$	$1\frac{1}{2}$

The surprising feature of these results is that those not receiving milk appeared to gain more weight than those receiving it.

On reviewing the experiment it was found that in each school, teachers had assigned proportionately more under-nourished and poorly dressed children to receive milk than to not receive it. Furthermore, all children were weighed in their clothes, and when differences in clothing weights were taken into account, then the children receiving milk were indeed found to have gained more weight than those not receiving it (see Table 5). True randomized assignment alleviates difficulties of this kind.

Table 5. Average Body Weight and Weight of Clothing
Corresponding to Table 4

Month	No Milk			Milk		
	Body	Clothing	Total	Body	Clothing	Total
February	81	$1\frac{1}{16}$	$82\frac{1}{16}$	80	1	81
June	$82\frac{7}{8}$	$\frac{3}{4}$	$83\frac{5}{8}$	82	$\frac{1}{2}$	$82\frac{1}{2}$
Difference in body weight	$1\frac{7}{8}$			2		

3.6. Experiments with caged animals

Maybe some of the preceding examples of unsuccessful experiments seem so obviously unsuccessful that one might wonder what application the principles illustrated therein have for laboratory experiments involving cages of animals such as mice, rats and rabbits. For a well-run, antiseptic laboratory with modern equipment, the temptation is to say that randomization is not necessary – that all cages and racks of cages are the same.

Nothing could be further from the truth. Remember, randomization of animals to cages and treatments thereto is desired, not only to level out the effects that we know or think we know about but, much more importantly, to also take into account effects we do not know about, cannot measure, or tend to ignore. As Finney (1960, p. 11) says, "If an experimenter finally insists that his experiment cannot or shall not be randomized, the ultimate responsibility is his; the statistician can analyze it as though it were randomized, but part or all of that analysis will depend for its validity upon personal judgement instead of statistical theory."

Even as someone who has never worked in a laboratory, a mere desk-hugging statistician can think of several things which might affect animals differently, depending upon the precise location of their cages.

1. First, some animals, rats for example, are nocturnal in habit. Rats in cages near the bottom of the racks might therefore have more darkness than those at the top — especially if lights get left on all night.

2. Windows and doors can cause draughts of air to circulate through a room in some unknown fashion, and temperature changes accompanying them might affect animals in some cages more than others.

3. Air conditioners and heaters can malfunction, and create temperature differentials of unknown effect.

4. Night watchmen may sometimes cause and/or rectify such malfunctions, without the experimenter ever knowing about it.

5. Suppose all cages receiving the same treatment are put on the same tier of a rack, or in the same column of a rack. Suppose also that one treatment does indeed begin to have deleterious effects on the animals. If those cages are all close together the animal debilitation will be easily noticed by a lab technician who, if kind hearted, might give those animals preferential treatment. But if their cages had been allocated randomly throughout the laboratory this might be less likely to occur.

An experimenter must ensure "not only that his experiment shall satisfy himself but also that it shall convince his brother scientists; he will be wise to remember that 'It is not merely of some importance but is of fundamental importance that justice should not only be done but should manifestly and undoubtedly be seen to be done'". Thus says Finney (1960, p. 9), quoting from a Lord Justice of England, in 1924.

4. Analysis of Experiments

Much is written on the statistical analysis of data gathered from experiments. Relevant to randomization in designing an experiment, there is one vitally important feature of analysis: the analysis of experimental data depends precisely on how the experiment was carried out, and particularly on what randomization has been used. If, by some strange quirk of fate three different experiments all gave the same set of data insofar as the numbers and layout of them was concerned, there would nevertheless be three different analyses.

Consider the following example of observations on people of four different age groups, using three different drugs all intended for the same purpose. For each drug there are two observations at each of the age groups, a total of $4 \times 3 \times 2 = 24$ observations.

Table 6. Data

Drug	Age Group			
	1	2	3	4
I	10 14	16 22	12 18	9 19
II	23 25	17 21	24 32	18 24
III	13 17	8 12	16 12	7 19

Suppose these data have arisen from one of the following three experiments:

Experiment 1 The two observations on a drug in an age group are from two different people (one observation per person), selected at random from people of that age group – a total of $2 \times 12 = 24$ people.

Experiment 2 For each age group, two people were selected randomly and they then used all three drugs in turn (no carry-over effects), a total of 2 x 4 = 8 people.

Experiment 3 The two observations on a drug in an age group are from the same person selected from the age group, the first observation in each case being made 30 seconds after taking the drug and the second one being taken one hour later, a total of 1 x 12 = 12 people.

The question of interest in each experiment is

Is there a difference between age groups?

Confining attention to the analysis of variance of these data, the question revolves around the F-statistic for testing the hypothesis that age group effects are equal. This comes down to the question "what is the denominator mean square for this F-statistic?". It is different for each of the three experiments. The analyses of variance are shown in Table 7. They illustrate why computers will never put statisticians out of a job!

References

Cox, D. R. (1958). Planning of Experiments, Wiley, N.Y.

Finney, D. J. (1960). An Introduction to the Theory of Experimental Design, The University of Chicago Press.

Freedman, D., Pisani, R. and Purves, R. (1978). Statistics, Norton, N.Y.

Table 7. Analysis of Data in Table 6

Source	d.f.	S.S.	M.S.	F-statistic for H = Age groups equal
Experiment 1				
Mean	1	6936		
Drugs	2	448		
Age groups	3	36	12	
Interaction	6	136		$F_{3,12} = 12/20\frac{5}{6} = .576$
Residual	12	250	$20\frac{5}{6}$	
Total	24	7806		
Experiment 2				
Mean	1	6936		
Drugs	2	448		
Age groups	3	36	12	
Interaction	6	136		$F_{3,4} = 12/49\frac{1}{6} = .244$
People	4	$196\frac{2}{3}$	$49\frac{1}{6}$	
Residual	8	$53\frac{1}{3}$		
Total	24	7806		
Experiment 3				
Mean	1	6936		
Drugs (D)	2	448		
Age groups (A)	3	36	12	
D X A	6	136	$22\frac{2}{3}$	$F_{3,6} = 12/22\frac{2}{3} = .529$
Times (T)	1	$160\frac{1}{6}$		
D X T	2	$6\frac{1}{3}$		
A X T	3	$36\frac{1}{2}$		
D X A X T	6	47		
Total	24	7806		