Appendix A

Some notes on experiments and statistics

A.1 Various Types of Experimentation

Science proceeds by selecting research questions, devising hypotheses (i.e. possible answers to these questions), and then experimentally seeking evidence to support or refute the hypotheses. What research questions are interesting, and the general class of methods appropriate for answering them, are defined by a scientific paradigm. If you read scientific papers or follow popular science you could be forgiven for supposing that hypotheses are obvious — the popular view is that scientific conclusions are "natural laws" and carry some necessary rightness, the truth is that they are tentative conclusions not yet shown to be wrong. Some have endured for hundreds of years without contradiction; this doesn't prove their truth, but we can be confident that they work in most situations we are likely to encounter.

But where do hypotheses come from? If they aren't immediately obvious once the research question is posed (they aren't), how do scientists find them? The answer is that they guess, make them up, use their imaginations, mess around for a while to see what's going on, or just have accidents (many of the most profound scientific discoveries have been the result of careful work investigating a surprising accident). The scientific process or method deals with weeding out the incorrect hypotheses — we are free to find hypotheses where we can.

In the light of this experiments fall into two broad categories: exploratory studies, where you are not certain what you are looking for (a kind of disciplined messing-about); and confirmatory studies, where you are trying to answer a carefully formulated experimental question relevant to your research question. Exploratory studies generate data which raises interesting questions, suggest relationships

[0] This chapter was almost entirely taken from notes by John Hallam

between aspects of the system being investigated, and allow important parameters to be quantified. The results of exploratory studies are preliminary models of what causes what, or what influences what, and more precise questions to ask to narrow further the range of possibilities; they are indications of what aspects of the behaviour of a system are interesting and of where to look next.

Confirmatory experiments follow from exploration. A precise and specific question is formulated based on the results of exploratory work and the hypotheses they suggest; an experiment is designed to answer that question; and serious thought is given to whether the answer provided by the experiment is unique (there may be more than one explanation for a set of results; the best-designed experiments produce results consistent with just one explanation).

Another useful way of categorising experiments, independent of the exploratory-confirmatory division, is as observation or manipulation experiments. In the latter case, the experimental structure looks at the influence some factor has on some output, and proceeds by altering the factor in a controlled way and measuring the resulting changes of the output. The factor is manipulated by the experimenter.

Observation experiments are done when the factor cannot be manipulated, for example when the subjects are people and it is impossible or unethical to manipulate them in the necessary way (e.g. one cannot make people have sisters to see whether the number of sisters one has affects one's performance at hang gliding). In these experiments a population is divided into groups using the factor(s) of interest so that each group's members have the same combination of factors (other criteria for selecting groups may also apply, see below), and we look for differences in the output between the groups. These differences, we hope, are the result of the groups' differing constitution.

A.2 What to Do with Data

Experimental results are data — often numerical representations of the behaviour of the system we are trying to investigate. What can we do with such data? What form can the data take? How can we analyse it? These are the questions we consider in this section.

It is worth stressing that data is only a representation of reality and, as such, it is coloured by our presuppositions about the nature of reality. Because of this, we must be careful how we deal with data. Consider the example below (from Tukey), where the following sequence of numerical data was obtained by experiment:

4 7 9 3 4 11 12 1304 10 15 12 13 17 ...

What can we do with this? One possibility is to look for trends (maybe by curve-fitting); it is fairly obvious that the early data are around 3–9 and the later data

12–17, say. But what do we do about 1304? If we believe the data should have a smooth trend, we could argue that 1304 is an anomaly, a spurious datum to be ignored. If we have no such prior belief, we must account for it some other way. The point is that the data *just is*; what we think it means depends on what we believe about the reality it represents and the encoding process by which it was obtained.

A.2.1 Types of Data

Data comes in three qualitatively distinct types, the distinction being to do with how data items inter-relate. The three are *categorial, ordinal* and *numerical.*

Categorial data falls into a set of classes. For instance, we may record data about the weather by classing each day as 'bright', 'cloudy', 'wet' etc. The datum is the class.

Ordinal data can be put in order or ranked. For example, days could be classified as 'cold', 'warm' or 'hot', with ranking in that order (so 'cold' < 'warm' < 'hot').

Numerical data comprises numbers, for instance recording the actual temperature each day would generate this kind of data. Numerical data is probably the most common type, but is also the most abused. The problem is that when interpreting numbers we need to know what kinds of comparison are valid, where the origin of the scale is, and so on.

A.2.2 Tools for Analysing Data

In addition to the distinction between types of data, we can also say that data normally comes in sets. A single experiment may involve running a variety of tests, or repeating a certain test a number of times, and the resulting collection of data needs to be analysed. What we do then will depend on whether the experiment is exploratory or confirmatory.

In general, visualisation techniques are used for exploratory data. These techniques try to make the patterns within a set of data apparent to the human analyst, by displaying visually (or aurally, or in other ways) the relationships between different data variables. There are various tools for doing this, of which a useful one is MATLAB, a matrix manipulation system with excellent graphical display abilities. Apparent interrelationships (effects) can be confirmed using simple statistical techniques.

For confirmatory experiments, statistical testing allows us to determine the precise extent to which a particular effect we anticipated is present in the data from our experiment. Visualisation plays a much less significant role here, though we may need to produce pictures or other summaries of our data for reporting our results.

A.2.3 Looking for Effects

Suppose we have a collection of data from an exploratory experiment and want to examine it to see whether any effects are apparent. What can we do? Before we look at this question, let's introduce some jargon:

- A *population* is the set of all instances of items of interest. For example, the set of all A. I. students, the set of all runs of a program with a certain parameter setting, the set of all runs of that same program for any parameter settlings, and so on. Typically, an interesting population will be too large for us to study it exhaustively.

- A *sample* is a subset of the population, chosen by some means, small enough to work with, from which we hope to draw conclusions about the population as a whole. For instance, we could regard the set of all A. I. students as a sample of the population of Edinburgh University Science students, who are themselves a sample of the set of Edinburgh University students; or we could take 100 runs of a program with a given set of parameters as a sample of all possible runs of that program with those parameters. This latter example makes sense only if we cannot predict exactly what output a single run of the program will give (more of this below).

  The idea of a sample raises certain difficulties. If we wish to extrapolate conclusions obtained from the sample to the population of which it is a sample, the clearly the sample must be representative. It is unlikely that students taking A. I. are representative of students at Edinburgh University — there is no reason why they should be 'typical', and we could not accept such a sample as representative without further evidence. On the other hand, a *random* sample of, say, 20 AI1 students could be considered representative of the whole class.

  The two key questions that arise are: How is the sample to be selected? and How large must the sample be? The general answer to the former is that selection must ensure a representative sample, for the purposes of the experiment at hand (it need not necessarily be representative in other ways, provided that those other ways are unrelated to the effects being investigated, and random samples are not necessarily representative), and that the larger the sample the more work is involved but the more secure the conclusions will be.

- A *statistic* is a numerical encoding of some property of a population or sample, which we hope is characteristic of that population or sample, and which we can use instead of the sample for reasoning about the properties of the sample. For instance, the average age of students taking AI1 is such a statistic: it summarizes certain properties of the population of such students. *Statistics* is the subject that studies the properties of such encodings.

To return to the main question, what can we do with data from our exploratory experiments? There are various ways to visualise such data to make relationships apparent. Here are a few:

Histograms record how many data fall into each of a number of classes. For instance, we might record the temperature at noon each day for a year, and then count how many days there were whose recorded temperature was between 16 and 17° Celsius, how many between 17 and 18, how many between 18 and 19, and so on. A plot of the number of days in each category (vertical axis) against the various categories (horizontal axis) is a histogram, and shows the *distribution* of the data, that is, how likely a randomly chosen value from the set is to be in each category.

One sign in a histogram of something going on is for there to be multiple peaks. This means that there are two or more clusters of similar data, and it is possible that some cause other than chance determines in which cluster a particular datum will lie. Splitting the whole set of data into the clusters associated with the peaks then allows you to investigate whether members of the clusters differ from each other in consistent ways.

For example, the temperatures histogram might show a peak at around 25° and another at 16° with a trough in between. On investigation, you find that all the days in the former cluster are noted as 'bright' while most of those in the latter cluster are 'cloudy'. You can now infer, tentatively, that bright days are hotter than cloudy ones for some reason.

Scatter Plots allow you to look at data in more than one dimension at once. For example, suppose you collect data from a large group of people: for each person, you record age, height and weight. This gives you a large set of triples of numbers. Consider plotting these numbers as points in space — draw three axes at right angles to each other, and consider the triples as specifying the coordinates of a point with respect to those axes[1]. This procedure scatters points, one for each datum in the set, over the space spanned by the axes.

If there is no relationship between the individual measurements, the points ought to be scattered randomly (therefore approximately evenly) throughout the space. On the other hand, if there is an effect, the points will be clustered more densely in parts of the space — in the example above points will tend to cluster along a line sloping upward away from the origin with a sharpish bend in it, since age, height and weight are related loosely proportionately for

[1] If you aren't happy with three dimensions, imagine drawing a graph with age as horizontal axis and height as vertical axis, and marking in those points given by the age-height numbers in each triple.

young people (the older a person, the taller and heavier they are), and then height and weight become roughly independent of age for older people. If you had also noted the gender of each person and were to colour the points in the scatter plot red for male and green for female, say, you would also notice that the relationships between the three variables differ slightly between the sexes. Scatter plots, with or without colour, make these kind of relationships visually obvious.

Summary Statistics express a property of the data set in a single number or set of a few numbers. There are many such statistics. The most common are the mean, mode, median, variance and standard deviation. Interquartile range is also used. For a fuller discussion of these and a few others, consult Cohen's book (section 2.3.2, pages 23–27). Here we just note the following:

- The *mean* $\bar{x}$ of a set of N items of data is obtained by adding up the items and dividing by N. It represents the centre of mass of the distribution of the data, and is the value you might expect to get if you generate another datum with no other information about it[2].
- The *variance* of a set of data, written σ^2, is computed as follows. Calculate the mean of the data set. For each data item, work out its deviation from the mean. Sum the squares (since we don't care which way the datum deviates) of all these deviations, and divide by N or $N-1$ to get the variance. It measures the spread of the data away from the mean: wide distributions, with members likely to be far from the mean, have high variance, while tight distributions where all are close to the mean have low variance.

  The only subtlety here is whether to divide by N or $N-1$. The simple answer is that if the data set is the *whole population*, divide by N; otherwise $N-1$.

  The *standard deviation* of the data is the square root of the variance, i.e. σ.
- The *standard error* is given by $\frac{\sigma}{\sqrt{N}}$ for N data with standard deviation σ. It measures the amount of variation we may expect in the mean of the data.

  This is quite a subtle idea. If an individual experiment gives a result x, say, and repeating the experiment gives a somewhat different result because

[2] but note that the mean value may not be a possible outcome — e.g. consider the mean value of throws of an unbiased six-sided die (which is 3.5). More correctly, the outcome of an individual experiment can be thought of as a 'known value' plus some 'random error' — using the mean as the 'known value' makes the size of the random variation least (in the sense of having least variance).

of randomness or other variation in conditions, the mean indicates what result we expect on average and the variance tells us how close actual results will be to that average one. However, if we only have a sample of the population, we don't know the true average, only that of the sample, *and the value of the sample average depends on the particular sample we have chosen.* In other words, the experiment of doing N tests and computing the mean result $\bar{x}$ of them will also give a different answer each time we do it, depending on exactly which N tests we do out of the whole population. We need to know how much variation to expect in the sample mean.

Intuitively, the mean is the average of a collection of data. If the variations in the data are independent of each other, we might expect them to cancel out to some extent in the averaging, like a tug-of-war team pulling in different directions instead of together. This is exactly what happens, and the standard error tells us how much 'cancelling-out' takes place: the variance of the mean is $\frac{1}{N}$ of the variance of the individual data, and the standard error (or standard deviation of the mean) is thus $\frac{1}{\sqrt{N}}$ times that of the individual data.

Incidentally, the fact that not all the variation cancels out in the averaging process that yields the mean is the reason for division by $N-1$ to calculate the variance of a sample. Deviations are measured from the mean of the sample, and that contains variation from the true mean which is *not* independent of the variations in the data themselves. The deviations are all a bit smaller than they ought to be, because of this, and the effect is that $N-1$ units of variation rather than N (one per datum) are included in the sum of square deviations, the missing one being cancelled by the variation in the mean.

A.2.4 Statistical Measures of Independence

Given histograms, scatter plots and summary statistics, how can we look for effects in our data. In general, we look for suspicious data: we assume that nothing is going on, that no effects are present, that our data are independent of the factors that might influence them, and we search for evidence that we are wrong. Note that the whole of statistical testing is based on this approach: assume no effect and look for evidence to the contrary.

What signs are there of independence or otherwise in our data? Consider scatter plots first. Independent data results in points that spread uniformly over the plot or that line up parallel to the axes (so that one variable effectively remains constant as the other changes widely). However, suppose a scatter plot looks suspicious — the data seem to line up in an interesting way, but the pattern is not clearly visible because of the variation in the data. One option is to use a statistical measure, linear correlation, to assess how likely the data are to be interrelated.

Linear correlation measures how well the data fit the model of a straight line relationship. It is computed thus:

(1) Compute the means of the x and y data from the scatter plot separately.

(2) For each point in the scatter plot (pair of data) calculate the deviation of each datum from its mean and multiply, that is, compute $(x - \bar{x})(y - \bar{y})$.

(3) Sum these products for all the data pairs and divide by $N - 1$ for N data.

(4) Work out the standard deviation of x and y separately, and divide the sum from step 3 by the product of these standard deviations.

The result of this procedure is a number called Pearson's Correlation Coefficient, that lies between ±1, and measures how well the data fit the straight line model it assumes. Intuitively, if the data fit a straight line rising to the right, an x larger than its mean will tend to be associated with a y larger than its mean, while an x less than its mean is paired with a y smaller than its mean; thus the deviation products tend to be positive, and the resulting correlation will be bigger than zero. Dividing by the standard deviations removes any dependence on the size of the data themselves.

A value of $+1$ represents perfect proportional increase between x and y while -1 represents perfect proportional decrease of y for increase of x. Rarely do we get values as obvious as this; but given a little more computation or a book of statistical tables we can also find the chance that independent random data would generate a correlation value at least as large as our data — if this chance is very small we can be correspondingly sure that our data really is correlated and infer, tentatively, that the data are related.

A final comment on correlation: just because two variables are correlated doesn't mean that one causes the other. For instance, time spent watching television and incidence of lung cancer are correlated, but neither causes the other: both are caused by economic factors providing people with leisure time and money to buy cigarettes! Statistical dependence is not the same thing as causal dependence.

Correlation works for scatter plots, but maybe our data is not of that type. For example, in class we collected a set of data as follows. Each person present at a certain lecture chose a number between 1 and 10. The number of people selecting each was recorded for three categories of people: AI/CS degree students, AI/Psychology degree students, and Other AI students. The results are reproduced below. Something seems to be going on here, given the preponderance of people choosing 7. But maybe this is just luck. We can also see that each of the three groups of people singled out in this experiment are reacting similarly, which reassures us. But can we be more precise? And what if the data had been less obvious — is there some procedure that will highlight things for us to consider?

	1	2	3	4	5	6	7	8	9	10	Total
AI/CS	1	3	2	3	2	3	10	2	2	1	29
AI/PS	1	0	2	1	2	1	4	1	1	1	14
AI/Other	2	0	1	1	0	1	6	3	3	1	18
Class	4	3	5	5	4	5	20	6	6	3	61

The answer to these questions is ‘yes’: a simple statistical test will tell us how likely it is that these data could arise by chance if no effect were present. This test is called the χ^2 test, and it is one way of measuring the similarity of distributions of data.

We proceed as follows. First, as usual, *suppose there is no effect present*, and our data are independent. Let’s just consider the bottom line of the table for the moment (labelled ‘Class’). If the number chosen were genuinely random, we would expect around $\frac{1}{10}$ of the class to choose each number. We can make a table of expected frequencies, and actual frequencies, with which each number was selected. The difference between the expected and actual frequencies is an indication of how much variation is present. We square each difference, divide it by the expected frequency, and sum across the table to obtain a single number summarizing this variation. The number is called χ^2.

	1	2	3	4	5	6	7	8	9	10	Total
Expected	6.1	6.1	6.1	6.1	6.1	6.1	6.1	6.1	6.1	6.1	61
Class	4	3	5	5	4	5	20	6	6	3	61

Now we need to work out how many degrees of freedom there are in our data. Recall the discussion about standard deviations above: the factor $N-1$ is used because $N-1$ units of variation get into the sum, rather than N — one is cancelled out because the mean is computed from the data. Similar things go on here. For a table such as the one above, with a single row of actual data, the number of degrees of freedom is one less than the number of classes, i.e. 9.

This is because the total of the row is fixed (to 61) by the size of the sample. (We are interested in knowing what other data we might have obtained from the experiment, in a sample of the same size as the one we actually have. The answer to this question tells us how likely the actual sample of data we have is.) Suppose we generate another set of data by randomly varying the individual entries in the row, and suggest that this data could have been obtained from an equivalent sample: clearly this is only possible if the row sums to 61, as the original data does. We are not free to alter the data as we like: our alterations must respect the constraint that the sample size stays constant. Thus we can only change 9 of the data freely, the last datum being constrained to equal 61 less the sum of the other nine.

For the data above, the value of χ^2 works out to be 36.87 and we can look this up in statistical tables to discover that the chance of obtaining the actual frequencies from an experiment with true frequencies equal to those expected is around 0.00004 — 1 chance in 25000. It is thus very unlikely that the frequencies are the way they are by accident, and we can conclude that something is going on. (Of course, we don't know what, only that the class prefers 7 to other numbers for some reason. A more involved exploratory study would be needed to suggest reasons for this preference.)

We could also ask whether there is any real difference in preferences between the three categories of people identified in the experiment. To answer this, we need to know how many people in each category might have been expected to choose each number. We can work as follows. Consider the AI/CS group: there are 29 of them out of 61 so they represent 47.5% of the class. Of the class as a whole, 20 out of 61 chose number 7, i.e. 32.8%. If the choice is independent of category, we would expect the proportion of AI/CS students choosing 7 to match that of the whole class who chose 7. In other words, of the 20 people who chose 7, 47.5% of them ought to be AI/CS students. This gives an expected frequency of 20 times 47.5% or 9.50 students. In fact 10 chose 7 from this group. We do this calculation for each box in the table, 30 of them in all, to get the set of expected frequencies for the experiment. Then we calculate χ^2 as before, for the whole table. The answer is 11.1, in fact.

How many degrees of freedom are there now? In this case, the alternative samples we have in mind must match the real sample in size *and structure*, so they will have three groups of ten class frequencies, the sizes of the groups will match the sizes of those in our sample, the total number choosing each class will match, and the total sample size will match. How many of the data could we vary freely, given these conditions?

For each row, we have one constraint in that the row total is fixed; so in each row 9 numbers are potentially free. However, each column total is fixed too, and if we vary the first two rows, those variations fix the value of the third row. In each column we have only 2 free numbers, not 3, because of the fixed total. Thus, for this table, with 3 rows of data in 10 columns, we have $2 \times 9 = 18$ degrees of freedom. In general, there are $(r-1)(c-1)$ degrees of freedom for a table with r rows and c columns (both greater than 1).

The statistical tables for χ^2 show that there is a chance of 0.89436 of obtaining a value of 11.1 or more with 18 degrees of freedom. In other words data like ours (or more extreme compared to the expected frequencies) will be obtained from a sample of the size and composition of the one we have here in 89.4% of experiments. We have no justification for concluding that there is any difference in preference between the three groups.

To summarize, the χ^2 test proceeds as follows:

(1) Assume that the data are independent.

(2) Calculate the expected frequencies of each kind of result for a sample of the same size and composition as the one you have, given the independence assumption. For a data table with multiple rows and columns, work out the proportion of the whole sample each complete row represents; then for each column the expected frequencies are computed by multiplying the column total by each of the row proportions, as we did above. This assumes that each column splits up by row the way the whole sample does, i.e. that row and column splitting are independent of each other.

(3) Calculate the square deviation between corresponding actual and expected frequencies, divide each by the expected frequency, and sum over the whole table. Don't include the totals at the ends of the rows and bottom of columns in this calculation.

(4) Work out the degrees of freedom as described above.

(5) Consult the tables of χ^2 distribution probabilities to find the chance that your data could have been generated by accident given that the assumption of independence is true.

A.3 Experimental Controls

As mentioned above, exploratory experimental work looks for suspicious data in order to discover potential relationships between causes and effects which warrant further investigation. Visualisation techniques allow us to spot patterns in mounds of numerical data (not the most perspicuous form of data for humans — could you see patterns in the telephone book, for instance?), while some statistical techniques can be used to reassure us that patterns are likely to be real, or to extract patterns when the data is less clear than the examples above. Once exploratory work is done, we move into confirmatory experimentation, what we often think of as 'proper science'.

In this regime, we are interested in formulating a precise experimental question or hypothesis (suggested by our imaginative analysis of the results of our explorations) and testing whether the evidence supports our hypothesis or not. We can think of hypotheses in terms of factors and behaviours: all hypotheses have the form "factor X affects behaviour Y". For example, we might assert that living near a power cable (the factor) increases the likelihood of your suffering from certain cancers (the behaviour); or that setting the rate of mutation too high in a genetic algorithm (the factor) results in slow convergence or poor solutions being found (the behaviour).

Having chosen an hypothesis, we need to design an experiment to test it (normally to disprove our hypothesis, since a positive result could be caused by something we haven't thought of, but a single negative result disproves the hypothesis). This means finding a way to answer the question "are measurements of X and Y related?" Notice the tacit change of subject here: we don't investigate the factors and behaviours directly, but rather consider measurements of them. These measurements, our data, represent the factors and behaviours we wish to study, and clearly we must choose them with care so that they do represent what we wish to attend to.

Given this basic strategy, there are two ways we can proceed, depending on what we may do with factor X. The experiments that result are the observation or manipulation experiments mentioned above.

Observation experiments are necessary when we cannot directly manipulate factor X for some reason. In our examples above, it would be unethical to make people live under power lines to see whether they suffer from cancer: the factor is not manipulable in practice. There are also cases where the factor of interest is not manipulable in principle.

In this situation, what we do is group our experimental subjects based on the measurement of factor X. Thus we might make two groups, one of people living close to power lines and one of those living far from power lines, and look at how the incidence of cancers vary between the groups.

Manipulation experiments are used when the factor of interest is directly manipulable. The genetic algorithm example is such a case: we can run the program with different values of the mutation rate parameter and see what happens[3], that is, we examine whether there is any relationship (e.g. correlation) between the values of the measurements of the factor and those of the behaviour.

In either case, though, we have a problem. How do we know that the effects we see (variations in the measured behaviour) are due only to our changes in the factor of interest. There may be other factors that influence the behaviour we are interested in, and they may contaminate our experiments. We need to consider this during experimental design: a well-designed experiment allows us just one explanation for the effects we see in the data it produces, while a poor design may allow many. When you look at data, therefore, and consider people's conclusions based on it, you need always to ask what else (apart from what they suggest) might account for the effects described. Consider the following quotation: do you believe its conclusions?

> ALMONDS It may sound pretty nutty, but even though almonds are very high in fat ... they may be good for your heart! A major study of 26,000 members of the Seventh Day Adventist Church in the United

[3] Strictly speaking, this is not obvious —- for example, it may take so long or be so costly to do this that our sponsors will not allow us to do just what manipulations we choose, and we are forced to use an observation experiment instead even with a computer investigation such as this one.

> States showed that those who ate almonds, peanuts and walnuts at least six times a week had an average lifespan of seven years longer than the general population, and a substantially lower rate of heart attack. *(p. 77, The Food Medicine Bible, by Earl Mindell and Carol Colman, 1994.)*

The problem is that, with the data as presented here, we cannot draw the conclusion that almonds are good for you! What other possibilities are there? Well, peanuts or walnuts might be good for you instead; or it might require a combination of nuts to achieve beneficial results; or maybe you need to become a Seventh Day Adventist in order to live longer; or maybe something else is going on.

To resolve these issues we would need to do more experiments (or do this one more carefully). We would need an experiment to demonstrate that it was the almonds which accounted for the healthier people and not the other nuts; and we would need an experiment to demonstrate that the comparison between Seventh day Adventists and the general population was a fair one — for perhaps Seventh Day Adventists are atypical people in some way related to health and heart disease.

Such experiments are called *control experiments*, or just *controls*, and their purpose is to eliminate alternative explanations of the data obtained from an experiment. They are vitally important: many an interesting experiment has been rendered useless by poor controls.

Let's consider this further. Apart from the factor we are interested in, there will typically be other factors that may affect the behaviour we are investigating. If we call the particular factor we wish to study the *independent variable* (the thing we can vary as we choose) and the behaviour of interest the *dependent variable* (because it depends on the factor(s)), then the other factors are *extraneous variables*, things that vary without our particularly wanting them to. Variations in extraneous variables cause disturbances in the dependent variable; control experiments also try to eliminate these disturbances by controlling the extraneous variables in some way.

Three straightforward ways of controlling for extraneous variation are

- Make the extraneous variable an independent one, that is, include it in the experiment. For manipulation experiments, this means varying the value of the extraneous variable together with that of the independent variable we are interested in. Effectively, we investigate two variables rather than one. Obviously, this is only possible if we can actually control the factor concerned and if there are not too many such extraneous variables to include (because the number of combinations of values to investigate multiply together, the amount of work involved in adding several variables to an experimental design can be prohibitive).

- Partition the test cases for the experiment in such as way that the extraneous variable effects should cancel out. For example, to investigate the effect of gender on measured intelligence we might collect a large number of pairs of people — one male and one female — such that each pair were as closely matched as possible in age, socio-economic class, domestic situation, training, etc. and argue that the differences between the members of each pair must then be solely due to gender. This can be done for more complex cases too, and is frequently necessary in human experimentation. Note that this is essentially an observation experimental procedure: we cannot or chose not to manipulate the extraneous variables, and try to minimize their effects by matching.

- Choose a random sample of the population of individuals with each of the values of the independent variable, and compare the behaviours of these samples. For instance, run 100 randomly different runs of a genetic algorithm for each chosen value of mutation rate. The effects of other, extraneous, variables should appear as random variation in the dependent variable whereas the effects of the independent variable will not be random, and a statistical test can distinguish them.

  In this case, we must be careful that the samples really are random with respect to the extraneous variables. If there is some non-randomness present, because of a cause-effect relationship we don't know about, the effects of extraneous variables may add up or compound instead of cancelling out. This means that we have to be very careful in selecting random samples.

Other questions to consider concern the process of measurement. How do we choose the set of tests that vary the factor X of interest and how do we make measurements of the behaviour Y we are studying. In Cohen's book (section 3.1.2), the MYCIN expert system is discussed as a case study illustrating these issues. For a full discussion, consult that section. Here are a few points.

- Often, when testing a behaviour, we make up a set of test problems on which to assess performance of our system. The system performs very well and we are pleased, or very badly and we are distressed. But should we be? What do these results tell us?

  The answer is that it depends what we are comparing our system against. We need to control for the possibility that the problems do not represent a fair test of our system, that they are either so hard that no comparator system could do well on them or so easy that any comparator system could do well.

  The authors of MYCIN wanted to show (amongst other things) that the expert system could perform as well as human experts. This was achieved by generating a set of test problems for MYCIN and for human experts. To control for the possibilities above, human novices were also included in this comparator set. Now if the novices and experts performed equally well on the test set, one might conclude that the problems were too easy (if the novices could do them) or too hard (if neither novices nor experts could). A test set that split the novices from the experts could be considered a fair test for the program.

- Given our test set, what do we measure? Since we are looking for systematic variations in the behaviour, we must use a measurement procedure that doesn't introduce any such variation. For MYCIN, the responses it produced to the test problems were checked by human experts. However, humans might be prejudiced for or against the expert system (depending on their beliefs) so they were not told which solutions were generated by MYCIN and which were generated by the comparator set of humans. In this way their possible biasses were controlled for by *blinding* them to the information which might bias their response. The MYCIN trial was a *single blind* trial, since only the judges were unaware whether a solution was human or machine generated. *Double blind* trials are also widely used (for instance in drug testing or parapsychology) when knowledge available to the subject or even to the experimenter might cause a systematic variation in the measured effects.

To summarise this important section: experiments are indirect, working with *measurements* of factors and behaviours of interest rather than the factors and behaviours themselves; a well-designed experiment answers its question unambiguously, and this is achieved by careful use of controls; controls are implemented by sampling or partitioning the experimental subjects, by careful choice of test cases and evaluation procedure so that the measurements of both factor and behaviour are unaffected by extraneous variables and are not biassed by the measurement procedures.

A.3.1 Statistical Tests for Confirmatory Experiments

One common situation with confirmatory experiments is that we wish to demonstrate statistically that some significant effect is occurring, by showing that the results obtained from two sets of tests are different. We discuss this case here.

Suppose, for example, that we generate a sample of 100 runs of a genetic algorithm with mutation turned off and a second sample of 100 runs of the algorithm with mutation at 2%, say. All other parameters are the same for both samples. We record the fitness of the best individual found in each run, giving us 100 measurements in each sample. Our hypothesis is that mutation contributes something to the genetic algorithm's search, so turning it off will have a bad effect on the performance of the system and we believe this will be reflected in the fitness of the best individual found during a run.

We could answer this question "Yes, mutation does help the genetic algorithm" if we could show that the fitness of the best individual, all other things being equal, was greater for runs with mutation on. However, the best fitness achieved is dependent on the other parameter settings (which we'll ignore for the moment) and on the random effects of selection and mutation which are built into the genetic algorithm.

A more precise question we could ask is as follows: "Given the two sets of 100 data obtained by our experimentation, what is the chance that they represent two different samples of the same population?" The population here is the set of all possible runs of the genetic algorithm with the parameter set we have chosen. If it is likely that they do, then we cannot claim any significant difference in performance; but if they appear to come from different populations (i.e. it is very unlikely they come from the same population) then some factor must account for the difference and mutation is the only candidate.

So, do the samples seem to have been drawn from the same population? If they have been, they should have (for example) the same mean. However, the random variation of the results will ensure that two samples, even of the same population, do not yield exactly the same mean. All is not lost, though: the standard error tells us how much variation we might expect in the mean of a sample, given the variance of the sample and the size of the sample. Furthermore, there is a theorem (the Central Limit Theorem) which tells us that whatever the distribution of data from which the sample is drawn, the mean of a sufficiently large sample has a standard distribution called the normal distribution.

Now, for an experiment whose results have the normal distribution, it is known that 95% of all such experiments will produce an answer within 1.96 standard deviations of the mean, and only 5% will produce an answer that differs more than that from the mean[4]. The experiment of running N tests and taking their mean is this kind of experiment — its result has a normal distribution, for large enough N — and the standard deviation of the mean is the standard error of the sample. Therefore, the actual mean of the population of data from which the sample is drawn lies within 1.96 standard errors of the sample mean for 95% of the possible samples we could make.

We can proceed as follows: for each sample of 100 runs, we calculate the mean and the standard error, and from these we compute the range of values in which the true mean of each sample must lie assuming that our actual samples are among the 95% of well-behaved ones. There is a 1 in 20 chance we will be wrong, and the true mean will lie outside the interval we have computed, for each sample so there is a roughly 1 in 20 chance that at least one of them will be bad, and around a 90% chance that both will be good. If both are good, the two samples can only have come from the same population if the ranges we computed overlap, since a population can have only one mean and it must lie in both ranges. (The ranges are called 95% confidence

[4] In fact, we know the proportion of data that lie at more than a certain deviation from the mean for any possible deviation. Tables of these probabilities are available, or MATLAB (or other statistical tools) can compute them.

intervals, as we are 95% certain the true mean lies inside them.)

The test we do, then, is this. If the intervals overlap, the samples could come from the same population and we cannot confirm our hypothesis. On the other hand, if the intervals do not overlap, the samples are of different populations and our hypothesis is supported. In this latter case, there is a 5% chance we are wrong in drawing this conclusion, as this is the chance that at least one sample is badly-behaved in the way discussed above, i.e. the true mean of the population lies in only one of the two ranges.

In the standard jargon of the subject, we construct a null hypothesis "No effect is present, so both samples are from the same population" and we work out how likely it is that we obtain the data we actually have if the null hypothesis is true. The less likely that is, the more confident we can be about our alternative hypothesis, that (in this example) mutation rate causes changes in search efficacy in genetic algorithms.

The test described above is a version of the Student t-test — the usual form of this extends the reasoning above to calculate how big the confidence intervals can be before they overlap, given the data. The wider the confidence interval, the less likely it is that the true mean should lie outside it, so the less likely it is that a sample is badly-behaved as described above. Properties of the distribution of means of samples of size N allow us to relate the width of the confidence interval to the chance that the sample is bad (i.e. that the true mean lies outside the confidence interval) for that interval. We can work out the actual chance that both samples could have come from a single population by expanding the confidence intervals until they just touch, keeping the probability of badness the same for each interval. The probability of badness when the intervals just touch is the chance that the samples come from the same population.

A.3.2 Student's t-test in more detail

This test compares two sets of data, of size N_i with mean $\overline{X}_i$ and standard deviation s_i, for $i = 1, 2$. Compute:

$$s^2_{pooled} = \frac{(N_1 - 1)s_1^2 + (N_2 - 1)s_2^2}{N_1 + N_2 - 2}$$

Then compute:

$$s_{\text{diff}} = \sqrt{s^2_{pooled}(\frac{1}{N_1} + \frac{1}{N_2})}$$

The Student's t statistic is then calculated from:

$$t = \frac{(\overline{X}_1 - \overline{X}_2) - (\mu_1 - \mu_2)}{s_{diff}}$$

To find the "level of significance" (that is, the chance that the two populations have different means rather than being separate samples from the same population) it is customary to consult a statistical table which, given the t value and the combined degrees of freedom $N_1 + N_2 - 2$, shows a percentage probability. The C program in `~peter/gagp/t-test.c` is a complete replacement for such a table. Just compile it:

```
% gcc -o t-test t-test.c -lm
```

and then, given a t value such as 3.02 and a combined degrees of freedom of 47 (for example because your two sample populations had sizes 25 and 24, so that $47 = 25 + 24 - 2$), you would use it as follows:

```
% ./t-test -t 3.02 -d 47
Probability that true means differ = 0.995922
```

The underlying maths assume that the populations concerned are normally distributed and have equal variances, but the test tolerates a cosiderable departure from this abstract ideal pretty well. It would still be unwise to work with sample sizes that are too small, eg less than about 10 each, or too vastly different from each other in size.

A word of caution is in order. Suppose you had seven sets of data, and it turned out that Student's t test showed that there was a 95% chance that each set had a different mean from any other. There are 21 pairs of sets, so the chance that there are seven different means involved is $0.95^{21} = 0.34056$. Remember, the usually-invoked significance level of 95% is far from meaning "beyond all reasonable doubt". Be very cautious about multi-way applications of Student's t. If you are interested in comparing the means of a number of different sets of data, a more customary approach is to conduct an 'analysis of variance'. The idea here is to estmate variance in two different ways. One is to look at the mean of each set, and work out the variance of these means. The other is to work out the variance of each set, and take the mean of them. If and only if these two quantities are very similar is it credible that all the sets are drawn from the same (essentially normal) distribution.

The t test was developed by William S. Gosset (1876-1937) who worked for Arthur Guinness and Son, ultimately becoming head brewer for them in London. He published almost all his work under the pseudonym 'Student' in order to protect the details of the firm's quality control procedures from competitors.