1 Introduction

1.1 Why do earth scientists need to understand experimental design and statistics?

Earth scientists face special challenges because the things they study – the rock formations, ore bodies, deposits of minerals and fossil species – are often very large, widely dispersed and/or difficult to access. Therefore, it is usually impossible for an earth scientist to study more than a small fraction of any geological phenomenon. For example, imagine trying to measure the length of every brachiopod in the northern hemisphere, the H_2O content of every basalt flow in the USA, the diameter of every volcanic bomb on the island of Hawaii, or the orientation of every single fault plane in an entire formation. You would have to take a sample – a small subset of each – and hope that the results you obtained were representative of the larger group.

Because they are often forced to work with samples, earth scientists need to know how to sample, and they need to know how confident they can be about making generalizations from these samples.

The total number of occurrences of a particular thing (e.g. mineral species, fossil type, rock type) present in a defined area is often called the **population**. But because a researcher usually cannot measure every part of the population (unless they are studying a very restricted location, like the inside of a volcanic caldera), they have to work with a carefully selected **subset** of several **sampling units** that they hope is a **representative sample**, which can be used to infer the characteristics of the population. For example, they might measure the size (usually in terms of diagonal length) of a sample of fifty megalodon teeth from a population of several hundred, or assess the quality of a consignment of several thousand agates by breaking open a randomly chosen sample of twenty. You can also think of the population as the total number of artificial sampling units possible (e.g. all the quadrangles in the United States) and your sample being the subset (e.g. 20 quadrangles) you have chosen to work with as an indication of conditions across the whole country. The concept of a representative subset also applies to experiments where you might take two (or more) samples and expose them to two (or more) different treatments. Here the replicates within each sample are often called **experimental units** to emphasize that they have been artificially manipulated. We will usually refer to replicates as sampling units in this book.

The best way to get a representative sample is usually to choose a proportion of the population at **random** – without bias, with every possible sampling unit having an equal chance of being selected.

Unfortunately it is often very difficult for earth scientists to take a random sample, because they cannot easily access the whole population. For example, it may only be possible to sample rocks that are exposed in outcrops, but these may not be the same as the rest of the formation – the outcrops may only have remained because they have a slightly different composition that makes them more resistant to weathering. A group of rocks sampled at random from float may not represent the variability present in all rocks from that outcrop/formation. Therefore, earth scientists need to know how to take the best possible sample from the part of the population they can access, and be aware of the risk of assuming that the sample is characteristic of the population.

Next, **even a random sample may not be a good representative of the population from which it has been taken**. There are often great differences among sampling units from the same population. This is not restricted to the earth sciences. Think of the people you have seen today – unless you met some identical twins (or triplets etc.), no two would have been the same. But even rock types that seem to be made up of similar-looking minerals show great variability. This leads to several problems.

First, two samples taken at random from the same population may, simply by chance, be very different to each other and not very representative of the population (Figure 1.1).

Therefore, if you take a random sample from each of two similar populations, the samples may be different from each other simply by chance. On the basis of your samples, you might mistakenly conclude that the two populations are very different. You need some way of knowing if the difference between samples is what you would expect by chance, or whether the populations really do seem to be different.

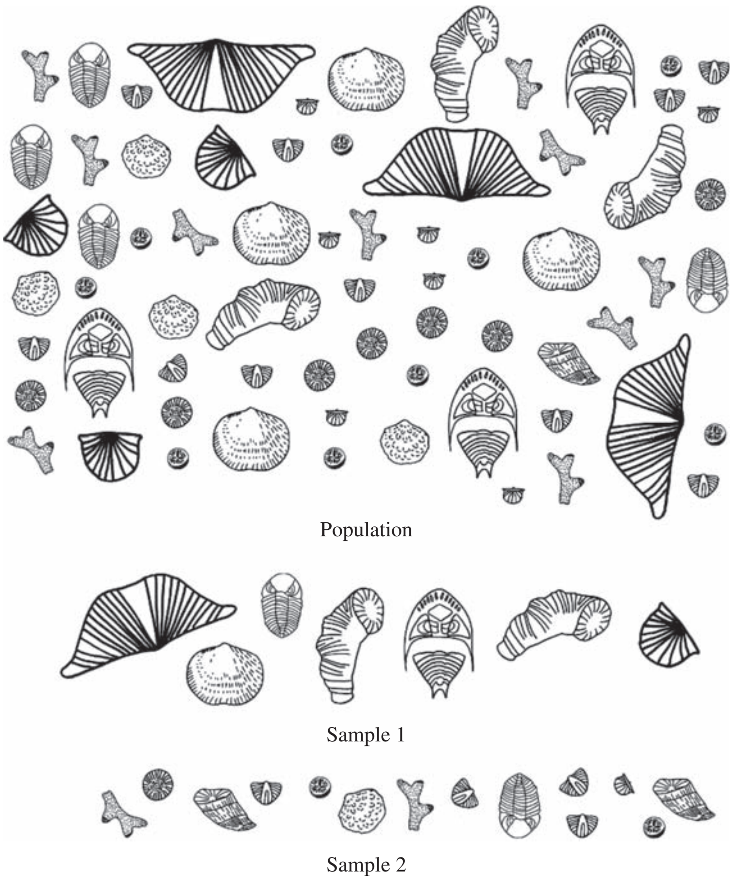

Figure 1.1 Even a random sample may not necessarily be a good representative of the population. Two samples have been taken at random from a Devonian oil field in Ghawar. By chance, sample 1 contains a group of relatively large fossils, while those in sample 2 are relatively small, and the types of fossils in the two samples are also different.

Second, even if two populations are very different, samples from each may be similar simply by chance, and therefore give the misleading impression the populations are also similar (Figure 1.2).

Finally, **variation within samples may make it difficult to interpret any effect of differences in location**. There is often so much variation within a sample (and a population) that differences in location may be difficult to interpret. For example, imagine you are an environmental geologist working to assess a landfill contaminated with lead. The lead content in a sample

Cambridge University Press
978-0-521-74656-4 - Geostatistics Explained: An Introductory Guide for Earth Scientists
Steve McKillup and Melinda Darby Dyar
Excerpt
More information

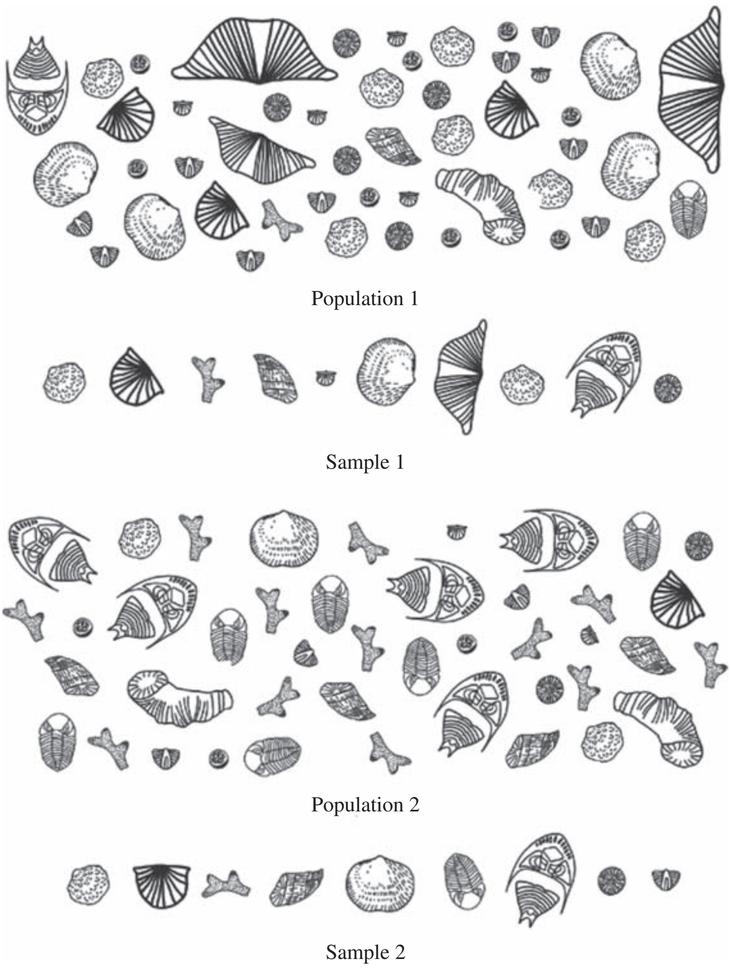

Figure 1.2 Samples selected at random from very different populations may not necessarily be different. Simply by chance the samples from populations 1 and 2 are similar in size and composition.

of ten cores from the oldest part of the landfill is 1000 mg/kg Pb on average, and ranges from 100–9000 mg/kg. In contrast, a sample of ten cores from the youngest part of the landfill contains 2000 mg/kg Pb on average but ranges from 100–7000 mg/kg. Which of these two areas would you consider to be most contaminated?

Variability within samples can also obscure the effect of experimental treatments. For example, opaque brown topaz crystals may change to transparent blue (which people find attractive and pay high prices for) if they are heat-treated. Gamma irradiation also alters the color of topaz. A mineralogist found that 60–80% of brown topaz crystals treated by heating turned various shades of blue. In contrast, when crystals were irradiated and then heated, a few turned bright blue, but others remained quite brown (Figure 1.3). From the extremely variable results for the 12 crystals in Figure 1.3, can you really conclude that irradiation had a significant effect?

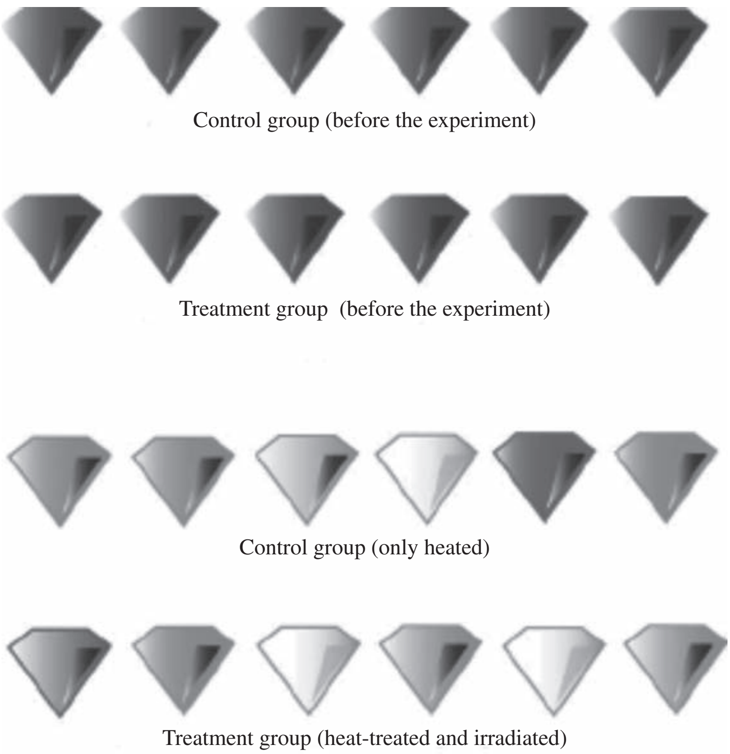

Figure 1.3 Two samples of topaz crystals were taken from the same mine and deliberately matched so that six equally brown individuals were initially present in each group. Those in the treatment group were treated with ^{60}Co radiation followed by heating to 450 °C, while those in the control group were only heated. This caused all crystals to became more translucent and change color to shades of brown, pink and blue. Slightly more of the crystals in the treatment group became translucent gemmy and blue, but this difference is small compared to the variation in color among individuals, which may obscure any effect of treatment.

These sorts of problems are usually unavoidable when you work with samples and mean that a researcher has to take every possible precaution to try and ensure that their samples are likely to be **representative** and thus give a good estimate of conditions in the population. So earth scientists need to know how to sample. They also need a good understanding of experimental design, because a good sampling design will take natural variation into account and also minimize additional unwanted variability introduced by the sampling procedure itself. They also need to take accurate and precise measurements to minimize other sources of error.

Finally, considering the variability within samples described above, the results of an experiment may not be clear-cut. So it is often difficult to make a decision about differences between samples from different populations or different experimental treatments. **Is it the sort of difference you would expect by chance, or are the populations really different? Is the experimental treatment having an effect?** You need something to **help you decide**, and that is what statistical tests do, by calculating the **probability** of a particular difference among samples. Once you have the probability, the decision is up to you. So you need to understand how statistical tests work!

1.2 What is this book designed to do?

A good understanding of experimental design and statistics is important, whether you are a meteorologist, paleontologist, geochemist, seismologist or geographer, so many earth science students are made to take a general introductory statistics course. A lot of these take a detailed mathematical approach that students often find uninspiring. This book is an introduction that does not assume a strong mathematical background. Instead, it develops a conceptual understanding of how statistical tests actually work, using pictorial explanations where possible and a minimum of formulae.

If you have read other texts, or already done an introductory course, you may find that the way this material is presented is unusual, but we have found that non-statisticians find this approach very easy to understand and sometimes even entertaining. If you have a background in statistics you may find some sections a little too explanatory, but at the same time they are likely to make sense. This book most certainly will not teach you everything about the subject areas, but it will help you decide what sort of statistical test to use and what the results mean. It will also help you understand and criticize the sampling and experimental designs of others. Most importantly, it will help you design and analyze your own sampling programs and experiments, understand more complex sampling designs and move on to more advanced statistical courses

2 “Doing science”: hypotheses, experiments and disproof

2.1 Introduction

Before starting on experimental design and statistics, it is important to be familiar with how science is done. This is a summary of a very conventional view of scientific method.

2.2 Basic scientific method

The essential features of the “hypothetico-deductive” view of scientific method (see Popper, 1968) are that a person observes or samples the natural world and uses all the information available to make an intuitive logical guess, called a **hypothesis**, about it or how it functions. The person has no way of knowing if their hypothesis is correct – it may or may not apply. **Predictions** made from the hypothesis are tested, either by further sampling or by doing experiments. If the results are consistent with the predictions then the hypothesis is retained. If they are not, it is rejected, and a new hypothesis formulated (Figure 2.1). The initial hypothesis may come about as a result of observations, sampling and/or reading the scientific literature.

Here is an example. Lead contamination is an enormous environmental problem because in the past many manufacturers discarded wastes containing lead and other heavy metals into pits and landfills. These heavy metals are water soluble so they can leach into aquifers, be transported by groundwater and contaminate water supplies. In the early days, clean-up of these sites involved digging up the contaminated soil and removing it to special disposal facilities where water run-off could be contained and treated. More recently, it has been found that the mineral group apatite has a structure that easily binds to heavy metals, effectively immobilizing them. Luckily, apatite is easy to get because it is readily available in fish and mammal bones, where it is the primary constituent along with collagen.

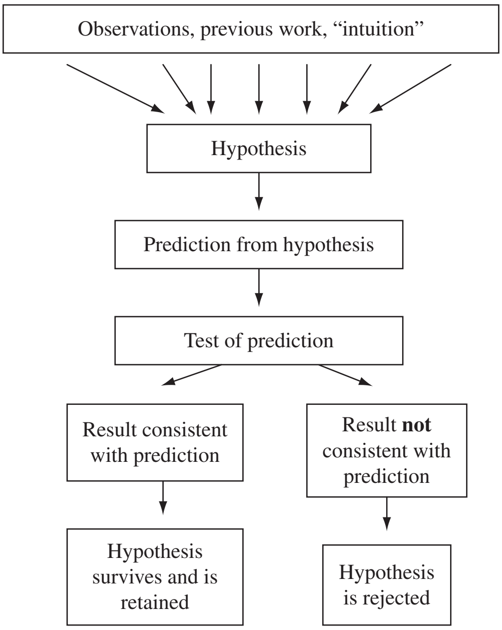

Figure 2.1 The process of hypothesis formulation and testing.

For your first remediation job as an environmental geologist, you decide to contain the lead in a contaminated landfill by mixing the soil with several tons of apatite. Your client balks at the cost, and asks you to demonstrate that it really works. The hypothesis that needs testing is simple: "Apatite will bind lead in contaminated soil."

From this hypothesis it is straightforward to predict, "Lower concentrations of lead will be present in water that has circulated through soils mixed with apatite, compared to soils without apatite."

This prediction can be convincingly tested by doing a simple and inexpensive manipulative field experiment with two treatments: (a) a 90/10 mixture of soil and apatite and (b) a 90/10 mixture of soil and an inert filler (e.g. glass beads) as a control to take into account the dilution that will occur when soil is mixed with anything else.

Because differences in the concentration of lead in the leachate might also result from heterogeneity in lead concentration across the landfill, the treatments need to be **replicated** several times. You could do this by mapping out three locations that are well spaced apart across the landfill. At each you could excavate ~20 cubic meters of soil and divide this into two

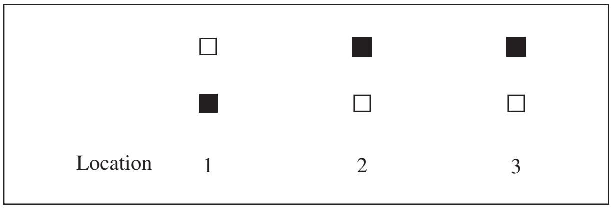

Figure 2.2 Arrangement of a 2×3 grid of treated and untreated areas in a landfill. Black squares indicate areas where the soil was mixed with apatite, and open squares where the soil was mixed with the same volume of glass beads. The treatment and its control are replicated at three locations.

equal-sized heaps (Figure 2.2). One (and here you could toss a coin to decide which) of each pair of heaps could be mixed with apatite, the other mixed with the inert glass beads, and the two heaps isolated and monitored so you could sample the water that drained from them. This arrangement would ensure that replicates of both the treatment and control were dispersed across the landfill, and the coin-tossing is a way of assigning each pair of heaps to the treatment and control at random.

You run the experiment for two weeks. Each day, you sample the water runoff from each of the six heaps, and analyze its lead content. For this manipulative experiment the three locations within each treatment are experimental units (Chapter 1).

From this experiment there are at least four possible outcomes:

(1) Run-off from the apatite-treated soil contains far lower concentrations of lead than run-off from the control. This result is consistent with the hypothesis, which has survived this initial test and can be retained.

(2) Run-off from both the apatite-treated and control soil has high concentrations of dissolved lead. This is not consistent with the hypothesis, which can probably be rejected because it seems that the apatite treatment has no effect.

(3) There is little or no dissolved lead in the run-off from either treatment. It is difficult to know if this has any bearing on the hypothesis – there may be a fault with the experiment (e.g. the 10 m^3 was not enough soil, there was torrential rain during the two weeks, or maybe you did not run the experiment long enough for the rain to percolate through the heaps). The hypothesis is neither rejected nor retained.

(4) Run-off from the apatite-treated soil contains higher concentrations of lead than from the control. This is a most unexpected outcome that is