Chapter 14. Experimental design and statistics

Note: I have provided clickable links to all software (www.geoff-hart.com/books/journals/software.html) and Web pages (www.geoff-hart.com/books/journals/web-links.html) in this book, plus updates and error corrections (www.geoff-hart.com/books/journals/2014-errata.html).

Ask anyone what scientists do and the odds are good they'll tell you that scientists collect numbers. But before you can collect numbers, you need to design an experiment that will generate reliable and useful numbers; that's the part of science referred to as "experimental design". It might seem odd to be describing this aspect of science in a book on writing for journals, since one might reasonably assume that the data has already been collected if you're reading this book. However, a large proportion of the problems my authors encounter during the peer review process result from following inadequate experimental designs that produce data of insufficiently high quality or that provide data that cannot answer their research questions, for reasons that I will discuss in the rest of this chapter.

Part of the problem is that experimental design does not seem to be taught as a specific course at many universities, though it is usually addressed to some extent within well-designed statistics courses. Students seem to be required to learn this skill on their own, whether by finding a good textbook on the subject or absorbing the necessary skills subconsciously while reading journal manuscripts. For this reason, a brief discussion of experimental design seems necessary to fill in what seem to be common gaps in the education of many researchers.

A second problem is that it is no longer possible to present experimental data without performing at least some basic statistical analysis of that data, and these tests must be appropriate for the data you are analyzing and the question you are hoping to answer. For an experimental design to be effective, it must both generate reliable data and generate data that you can analyze statistically to inform your readers whether your results are likely to be meaningful. A great many interesting and potentially important studies are undermined by an inadequate experimental design that makes it difficult to extract statistically significant results from the mass of data. Although basic statistics is

part of the curriculum in most fields of research, this course seems to be poorly integrated with the key courses in a student's major area of study. Again, students seem to be required to learn how to integrate this knowledge with their main expertise without much help. Although that help is provided to some extent in graduate school, it seems secondary to the more urgent tasks of mastering advanced topics in a study area and finding time to perform research and write a thesis.

Since most textbooks on experimental design and statistics are considerably larger than this entire book, it's clearly not possible for me to cover either subject in any depth. Instead, I will provide general guidance based on the most frequent problems that my many authors have encountered during the peer review process. There are undoubtedly other problems that the journal's reviewers are not catching, but short of performing a large and rigorously designed survey of the review process (something that is beyond my ability), it's not possible for me to tell you what those problems are. Thus, in this chapter, I'll focus on the most common problems in the hope that by helping you to avoid them, your papers will have a smoother review process.

Note: Although I've studied statistics, both formally during my education and informally during many years of editing journal manuscripts and needing to understand what my authors were describing, I'm not a statistician. Although I'm confident that my advice in this book is generally reliable, the best advice I can provide is that you ask for expert help from a statistician at your university or research institute. If that advice contradicts what I have written here, follow the expert's advice, not mine.

Experimental design

In modern science, most research is designed to test a *hypothesis*, which represents your informed belief (based on a rigorous review of our knowledge of some subject) about some aspect of your experimental system. Experimental design is the art of determining how to test that hypothesis. It begins with a careful consideration of the questions you are trying to answer. It concludes with the development of a data collection plan that will provide the data that you will use to answer those questions. There are three main types of experimental design that are worth distinguishing. The first is a primarily observational de-

sign, in which your goal is to accurately describe some variable so that you can determine whether its value differs between two or more conditions. The second involves manipulating a study system and measuring the results so that you can accurately describe the effects of the manipulation. The third type involves modeling or mathematical simulation. To meet the goals of this chapter, it is not necessary for me to distinguish among these designs, and I will use the term "treatment" to represent any specific set of experimental conditions, regardless of which of the three types of experiment I am discussing. However, when you are planning an actual research project, the different characteristics of these three types can greatly affect both the design of your study and the results produced by that design, and you should consult a statistician to learn how to account for these effects.

Choice of variables

For each question, it will be necessary to measure the values of one or more specific variables so that you can analyze these values and determine their meaning. There are two main types of variables to consider:

- *Independent* (*explanatory*) variables drive changes in other variables. They may be variables that you will control, such as reaction temperature, or factors such as time that are only partially under your control. In both cases, they define the treatments or experimental conditions that you will use. These variables are often referred to as "factors", and when you design an experiment, you must consider how many factors are important and how many levels of each factor you should test. For example, chemical reactions are strongly affected by the temperature and by the reagent concentrations, and in a chemistry experiment, these factors are independent variables whose effects you should explore.
- *Dependent* (*response*) variables are the ones that change in response to any changes you make in the independent variables. In a chemical reaction, these variables might be the amount of product that results from a given combination of reagents, under a given set of reaction conditions, and the rate at which the product is produced.

Choice of methods

Having chosen your variables, you must now decide how to measure their values. In any field of science, there are both proven traditional methods and promising new methods that you should consider.

Thus, any proposed design should begin with a careful consideration of how previous researchers have designed their experiments to measure your chosen variables. Older methods that have been used for years, and perhaps even for decades, are a good starting point because history has shown that they work. An advantage of reading the research literature on these methods is that you will discover the problems that other researchers have encountered when they used these methods and any special tricks of the trade they have learned that make the methods more effective. Most importantly, the recent literature will reveal situations in which certain methods should *not* be used, whether because they will be ineffective or because they will produce results that are "confounded" (i.e., the results that might be explained by factors other than the ones you studied).

Note: Choosing appropriate research methods requires an understanding of the precision and accuracy that you must achieve to answer your study questions. On the one hand, there is no point performing highly precise measurements when less precise measurements will be adequate. On the other hand, it may be difficult to achieve the desired level of precision or accuracy because of the high cost or long time required to obtain those measurements; in this case, you will need to redesign your experiment to work within your time and money constraints.

Most methodological innovations have been developed for good reasons, and if you don't understand those reasons, you cannot account for their effects on your research. Understanding the origins of a method and the assumptions it depends on is the only way to be sure that a given method is an appropriate way to answer your research question. If you cannot determine those assumptions, ask your more experienced colleagues for an explanation; if they cannot provide one, you'll have to look to other experts for help. Every method depends on one or more assumptions, and if those assumptions are not valid for your experimental system, then there is a high risk that the method will produce invalid results. A great many mistakes have been made over the years by authors who didn't adequately understand the method they chose and its relevance for the system they were studying.

However, even if the assumptions that underlie a given method remain valid, it may still be possible to improve on the method. You should never abandon a method simply because it's old; many old methods continue to be useful, albeit with minor revisions. But it's likely that much new knowledge has been obtained since an older method was originally developed, thereby challenging the underlying assumptions and requiring modifications of the method to account for the newer, more complete understanding. Some older methods, though potentially powerful and able to produce high-quality results, were rarely used because of their difficulty or the need for expensive or rare technology; instead, researchers used less powerful but more practical methods that became the standard approach in a field. More powerful methods may have recently become feasible based on new technology that solves the problems with the old method and that makes it a practical research tool.

Error-proofing your research

Once you've identified the most important factors and the most appropriate ways to measure them, the next step is to think carefully about how your research could fail. Early in your career, this knowledge comes from reading enough of the literature to understand how your colleagues work, which is usually based on emulating successful studies published in the literature. Later in your career, that knowledge comes from years of personal experience as a researcher or with a particular experimental system. The researchers who have gone before you have made a great many mistakes during their research, and learning how an experiment has gone wrong for others can help you avoid the same fate. (You'll discover enough of your own mistakes; there's no need to repeat the errors of others.) This advice is particularly true if you're early in your career and have not yet fully mastered your field and its research techniques.

One major source of error, particularly in the biological sciences, involves factors that you are not investigating but that can nonetheless affect the factors that you are investigating. For example, physiological and psychological differences between men and women can affect the outcome of a medical trial, so men and women must be treated as different groups in your design and analysis. Similarly, variations in solution temperature and in the purity of the reagents can affect the outcomes of a chemical study that focuses on the effects of using dif-

ferent reagents or reagent combinations. Your review of the literature will reveal all or most of the factors that can affect the outcome of your experiment, and one of your first steps must be to determine which of these factors you should control, how you can control them, and the implications of failing to control other factors. In some cases, you can account for these factors. In the medical trial, you may be able to include separate male and female groups, or you may need to focus on only one of those groups and leave the other group for subsequent studies. In the chemical study, you may be able to control the solution temperature precisely, and leave the effects of other temperatures (e.g., different reaction kinetics for different reagents) for subsequent studies.

It's not possible to eliminate all possible mistakes or sources of failure, particularly if you're working with living organisms (which are famously difficult to control), but a deep understanding of the most common problems in your field of research will let you incorporate techniques to avoid these problems. For example, in many chemical or physical analyses, the measurement apparatus must be calibrated repeatedly during the course of measurements using a "blank" that contains none of the substance you're seeking and a laboratory standard with a known concentration of that substance. This calibration may be performed using a laboratory-standard chemical solution provided by a lab supply company (or created carefully in your own lab), a cylinder of gas with a known concentration at a given pressure, a radioisotope with a known activity, and so on. Calibration before, during, and after your measurements lets you detect any developing problems that would render your data meaningless so that you can correct the problem before it affects the rest of your data. For example, if a non-zero value is obtained for the blank, that suggests the possibility of contamination, whereas an incorrect value for the laboratory standard indicates the need to recalibrate the instrument. Other forms of calibration are available for most measurement devices in most fields of research.

Error-proofing your research also requires a careful consideration of each step in your method in search of things you might do wrong even if the method itself is robust. A thorough literature review will help because it will alert you to the potential problems that other researchers have encountered and solved rather than relying only on your own experience, which is more limited than that of the research community as a whole and is likely to be biased in different ways. One option is to create a list of symptoms that will tell you when something

has gone wrong, so that if you see one of these symptoms, you can stop your analysis and correct the problem before you risk analyzing the remaining samples. This kind of thought process will reveal additional ways to reduce the risk of error or outright failure.

In many cases, and particularly in the most interesting research—research that is pushing into unknown territory—consensus may not exist about the best methodology or about how to error-proof your use of that methodology. In that case, it may be necessary to perform a series of exploratory studies that will let you understand your experimental system thoroughly before you begin your main experiments. Such preliminary research sometimes seems to be a waste of time, but it's the only way to ensure that you understand your study system sufficiently well to begin collecting more interesting data. This step also provides a good reality check that confirms your ability to analyze the data you will subsequently collect and provides an idea of the amount of variation in your study system (i.e., the sample size you will require to obtain statistically significant results).

Note: For surveys and questionnaires, always test and revise your initial set of questions before using them to collect data. The goal is to confirm that they are as clear as you think they are, and that potential respondents will answer the questions you think you are asking. Using a subset of your eventual survey population is most effective, since the results will reflect the real population's understanding of your questions. Where this is not possible, a good editor can help you review the questions to ensure that they are clear.

There are several additional possibilities to keep in mind as a means of reducing the risk of errors. *Blocking* is the creation of groups composed of similar treatment units so that you can compare those units both within and between the blocks. For example, you might compare three drug treatments with a control (e.g., with no treatment or with the standard treatment) in each block, and each block might represent a different subset of a larger population. Similarly, you might compare two fertilizer treatments with a control (e.g., no fertilizer or the standard fertilizer level) for the same crop in several regions. A blocked design would also let you compare the effects of the same treatment between blocks to gain insights into the amount of variation for each

treatment. If the differences between a given pair of treatment units are consistent among the blocks, you can have more confidence that this difference is real and consistent; if the differences vary among the blocks or contradict each other, you may have discovered that some unexpected factor (e.g., a factor that leads to differences in the conditions between the blocks) explains the variation. This hypothesis is stronger if the same treatment produces significantly different results in different blocks.

Factorial experiments are more complicated designs that let you test all combinations of the levels of the independent variables simultaneously instead of testing each independent variable by itself. Although this is a powerful technique for increasing the efficiency of a design, it necessarily involves a large number of samples, so it may be prohibitively expensive or time-consuming. A fractional or partial factorial experiment may accomplish much of the same goal while minimizing the size of the experiment. The Wikipedia article on factorial experiments (http://en.wikipedia.org/wiki/Factorial_experiment) provides a good introduction to this topic, but if you don't already understand this complicated topic, you should consult a statistician for assistance.

When you choose among the various potential variables to measure, keep in mind the possibility of orthogonality. Two independent variables are orthogonal if they are not correlated—that is, if neither one affects the other. For example, temperature and absolute humidity (the mass of water per unit volume of air) are not orthogonal because the amount of water vapor the air can hold increases with increasing temperature. Two non-orthogonal variables tend to have a high and significant covariance, so calculating the covariance is a way to detect orthogonality. In contrast, an orthogonal design lets you predict the dependent variable using two or more independent variables either separately or in combination, thereby providing a way to estimate the relative importance of or contribution by each variable. Because interactions among factors often reveal important phenomena, a lack of orthogonality should not be considered a failure of your experiment; on the contrary, it may reveal an important relationship that you must account for in future research.

It is also important to look for spurious relationships. Never forget the famous advice that "correlation does not imply causality", particularly if you cannot propose a causal explanation for the correlation. Sometimes variable A appears to control variable B, when in fact vari-

able C controls A, and A cannot control B in the absence of C. In that case, C is the truly important variable. Thoroughly understanding your study system will help you to understand whether it is necessary to study only A, or whether you should also study C.

A final but very important way to confirm that you understand the system you are studying is by obtaining independent lines of evidence. This is referred to as *triangulation*, a term that originates from the use of trigonometry to identify the position of the third point of a triangle when you can determine the directions of that point from two known positions and the distance between them. In the context of experimental design, triangulation means that you look for ways to provide confirmation by measuring the same factor from different perspectives. If both lines of evidence point to the same conclusion, you can have more confidence the conclusion is correct. Conversely, if the evidence points to different conclusions, more research will be required to learn the cause of this discrepancy.

Choosing a standard of comparison

Having chosen your variables and a method for reliably and accurately measuring their values, the next step is to choose a standard for comparison. That standard is usually a "control" for which you can obtain a known result, but sometimes it is a standard practice that requires improvement. The most basic control is the unperturbed experimental system, since a common research goal is to perturb that system and observe the consequences. For example, a known biochemical reaction observed under a standard set of conditions (e.g., temperature, pressure, reagent concentrations) should produce a predictable result against which all other combinations of reaction conditions can be compared. In genetics, levels of expression of so-called "housekeeping" genes that are active in basic cellular functions provide that control; the genes that encode actin and ubiquitin are common examples, though other genes may be more appropriate in specific circumstances.

Combining measurements of such a known situation with measurements of a poorly understood situation that is your real study goal provides two forms of control: First, if the absolute values of the control measurements lie outside the expected normal range, then you know there may be a problem with your experimental conditions that must be solved, or perhaps you have discovered something new that can explain that unusual result. Second, if the results for the control

fall within the expected range of values, they provide a way to normalize your results. "Normalization" (also called "standardization") relies on the assumption that in the absence of perturbation, a standard process or an indicator such as a housekeeping gene will function at 100% of its normal efficiency, and dividing all results by that base level will give a true measure of the change relative to the standard conditions. In some cases, the standard of comparison is the result produced by an existing method (e.g., the current form of medical treatment), and the goal is to see whether a new treatment produces a result that (i) differs from the absence of any treatment (the control) and (ii) provides a better result than the current standard.

Note: Although "normalization" has a range of specific meanings in statistics, non-statisticians commonly use the word to refer to the comparison of a value with a standard of comparison, in the form of a ratio. This differs from "transformation", which involves mathematically adjusting values so that the dataset more closely resembles a normal distribution. See Osborne (2002) and Maciejewski (2011) for details.

Eliminating bias

The next step is to eliminate as many forms of bias as you can. Because calibration greatly reduces the risk of equipment-based errors that would introduce a consistent bias, human bias becomes the most common problem—and the most pernicious form of bias because it is often subtle and difficult to detect.

When you will be studying multiple instances of your study system, whether those instances are individuals within a population, sites within a region, ore bodies within a geological province, or depths within a body of water, randomization is a good approach to eliminating human bias. Randomization does not mean that you carelessly throw your test subjects into different groups or pick samples with your eyes closed. Rather, it means that you distribute your subjects among groups or choose your sample locations without conscious or unconscious bias. Many popular randomization techniques, such as drawing playing cards from a deck or using the randomization function provided by many computer programming languages, are actually closer to pseudo-random. That is, they may appear to be random for small

sample sizes, but in statistical studies with large numbers of iterations (e.g., Monte Carlo simulations, computer models of the environment that run over long time periods or with short time steps), patterns in the randomization algorithm may become apparent that bias your results—potentially seriously. If you need to use true randomization that will be suitable for the latter type of experiments, explain your needs to a statistician and ask for their advice.

Bias can arise from other causes. The first form of bias arises from temporal effects. Some effects are delayed and cannot be measured for some time after you apply a stimulus or disruption of the experimental system. For example, it may take some time for applied heat to fully penetrate an object or to raise an entire volume of liquid to a consistent temperature, and it may take time for a plant to change its physiology in response to changes in gene expression. Examining the system before that time has passed can lead to an incorrect understanding of the effects of a treatment. Effects can also change over time, as in the case of plants that respond differently to a treatment during their vegetative growth and reproductive stages or that respond differently if the daylength, light intensity, or other factors change during the year. For these reasons, you must always carefully determine the initial state of the study system and decide both how long will be required before you can observe the effect of a treatment, and whether that initial state is suitable for the factor or factors that you are studying. If you choose the wrong time or an unsuitable starting condition, this can bias your results.

An important form of bias arises when the subject's knowledge of the treatment can potentially influence the outcome, as in the case of studies with living organisms and particularly studies with humans. In such a case, you must use *blinding* to ensure that the subjects cannot learn which treatment they are receiving. If your knowledge of the treatment could also affect the subject's perception of the treatment or your interpretation of the results, then you should also use blinding for yourself (a so-called *double-blind* experiment) to ensure that you do not know the nature of the treatment. In this case, someone who is not directly involved in applying the treatment or collecting and analyzing the data should assign subjects to each treatment for you. Although this is essential in fields such as medicine and psychology, it is rarely, if ever, necessary in the physical sciences, as Smith and Pell (2003) point out quite clearly.

Randomization is important when you do not know enough about your study system to predict whether it is subdivided into distinct groups. Sometimes these groups are clear and obvious, as in the case of vegetation studies that distinguish between grassland and forest sites. In other cases, it may be necessary to perform some preliminary research to determine whether subtler groups exist. When you know of or can detect such groups, it is better to divide your sample among these groups, a process referred to as *stratification*. This approach reduces variation in your data that results from differences between the groups; for example, studying human populations for the incidence of a beard would provide an average incidence of 50% if you combined men and women into a single population, whereas performing separate analyses of these groups would provide an incidence close to 0% for women and close to 100% for men. In such a case, you should divide your sample between these groups and then sample randomly within each group. Correct stratification requires you to identify background factors (ones that you will not directly manipulate) that could affect the dependent variables, the independent variables, or both.

Replicate your results

Replication is a fundamental technique in science, because it increases your confidence that what you are observing is real. Indeed, an interesting result obtained by one researcher may not be considered real until similar results have been obtained by other researchers using the same techniques. Replication is particularly important for work with living organisms, which have much higher variability than non-living systems. Replication is also important within a single experiment, since it provides more confidence that you have obtained a representative sample of the population that you're studying instead of accidentally selecting the one or few individuals who will respond in a specific way. Replication comes in two forms:

- Increasing your sample size to increase the likelihood that you have obtained a representative sample.
- Increasing the number of replicates (groups of measurements, each with a similar sample size) to determine whether the results are highly predictable (i.e., are similar for each replicate) or may be less predictable (i.e., vary widely among the replicates).

One of the most common experimental design failures that I've see during the past 25 years results from an insufficiently large sample size

and a lack of sufficient replication. The most common reasons for this problem are:

- **Financial constraints:** Research budgets are fixed, and the more expensive the measurements, the fewer you can afford to obtain.
- **Time constraints:** The longer it takes to obtain each measurement, the fewer measurements you can obtain within a given period.
- **Labor constraints:** The fewer people (including graduate students and undergraduates) who are available to help you obtain your measurements, the fewer measurements you can obtain within a given time period.
- **A combination of financial, time, and labor constraints:** When you have a limited budget, time, and pool of assistants, this combination severely constrains the number of measurements you can obtain. For example, research in remote and hazardous locations (e.g., the surface of Mars) permits only a limited sample size.

Although these constraints cannot be ignored, they are sometimes less serious than a recurring problem in the papers that I've edited: the researcher did not attempt to predict the sample size they would require to obtain statistically significant results. Although it is never possible to be absolutely certain that you'll obtain significant results, or non-significant results that you can be confident are really not significant, you can greatly increase the likelihood of success by using available knowledge to estimate an appropriate sample size. There are several possible approaches, each of which has a different goal (e.g., calculating a reliable mean versus testing a hypothesis) or which relies on different statistical assumptions (e.g., that the population resembles a normal rather than a skewed distribution):

- The simplest approach, which is therefore least likely to be effective for your specific conditions, is to review the literature and learn what sample sizes previous researchers have used successfully. The more similar their study system is to yours, the more likely this approach will be productive.
- A more sophisticated approach involves reviewing the literature on your subject to learn the range of variation (typically the variance, standard deviation, or coefficient of variation) that has been reported for a given variable. Where a subject has been studied in depth, you may be able to select studies that are similar to your study system and use their estimates of variation. Based on that variation

and the type of statistical test that you plan to use (e.g., Student's *t*, the *F* statistic in analysis of variance), you can consult published tables of probability to determine the required sample size.

- It is also possible to calculate the sample size based on the required power of the test (e.g., a *P*% likelihood of being able to detect a difference of *D* units between two treatments).

Wikipedia provides a good discussion of some of the factors to keep in mind when calculating a sample size (http://en.wikipedia.org/wiki/Sample_size_determination), but for difficult or demanding situations, it would be wise to consult a more authoritative reference or a statistician who has expertise in experimental design.

Note that increasing your sample size will not always produce better results. Even if you ignore certain issues such as the statistical characteristics of the study population, there comes a point of diminishing returns when increasing the sample size increases the cost, the time requirements, or the labor requirements to complete the research. On that basis, a smaller sample size may be more efficient while still providing acceptable power to obtain statistically useful results.

In all experimental designs, it is wise to remember the risks involved in taking shortcuts (choosing a smaller sample size) to reduce the cost or the time and labor requirements of a study. Compare the cost of a more demanding design with the cost of spending large amounts of time and money but failing to achieve publishable results, whether those results are statistically significant differences between treatments or a defensible lack of such differences. When statistical calculations suggest the need for a larger sample size than you can afford, some degree of compromise will be necessary; often, this means that you must reduce the scope of your project to investigate a smaller subset of the overall problem. The larger the difference you expect between two treatments and the lower the variation within each treatment, the smaller the sample size you can afford to use.

Test your design to confirm that it produces data you can analyze

When you believe that you have created a sound experimental design, test that your design produces data you can actually analyze. To test your design, create an "artificial" dataset based on your expectation of the range of values that you will measure. You can sometimes obtain this data from published research (e.g., by extracting data from

a scatterplot); in other cases, you can generate a random population of data using statistical software (e.g., using a Monte Carlo simulation). In the latter case, you must carefully consider what assumptions the software makes about the nature of the statistical distribution. For example, a randomization procedure that is based on a normal distribution will not provide a valid test for a population with a non-normal distribution of values. Use the same equations or statistical tests that you will use to test your real data to test the artificial data. If you detect any problems, re-examine your design to see whether you can eliminate them or whether you will need different tests or measurement procedures.

One common problem is to base your design on the assumption that the data will be normally distributed (i.e., will follow a normal distribution). Never accept this assumption uncritically. If the statistical software you use for your data analysis does not automatically test data for normality before applying a statistical test that requires normally distributed data, you must perform that test yourself before you use that statistical test. Data that seems likely to follow a given type of statistical distribution may require an experimental design based on that distribution. Although it is common to transform data (often using a logarithmic transformation) until it meets the criterion of normality, thereby allowing the use of standard statistical procedures, this approach is suboptimal because it can introduce certain statistical problems (Osborne 2002).

Design your study to provide publishable results

If possible, try to design your study to produce publishable results, even if it fails to support your primary hypothesis. Choosing a sufficiently large sample size, with adequate replication, and using error-reduction techniques such as the use of blank controls and laboratory samples with a known value of some property means that even if you do not find statistically significant differences between two treatments, you can at least claim with some confidence that it is possible that no difference exists. Traditionally, it has been difficult to publish such negative results, since reviewers tend to assume that something was wrong with your experimental conditions or with your analytical procedure. You'll have an easier time persuading journal reviewers to accept your negative results if you provide evidence that neither assumption is correct.

Dirty secret: To increase the likelihood of obtaining publishable results is to limit your ambitions. If a larger study can be broken into several smaller experiments that each tell a complete story, it may be possible to obtain publishable papers from one or more experiments even if the other experiments fail.

If it's not possible to publish all of your results as a single long research paper, some subset of your results may be publishable as a shorter paper, such as a letter, research note, or short communication. Although these types of paper are less prestigious than a full paper, it is better to publish them than to obtain no published result from your study. Where it is possible to design your experiment to produce a series of small publishable papers, incorporate this possibility in your design. If the overall study produces a consistent story based on assembling the smaller stories from each experiment, that's the best outcome. If that is not possible, then at least you will receive some publication credits from your hard work.

Obtain a reality check

The final step before you begin using an experimental design to collect data is to ask at least one colleague to review your design rigorously in search of flaws, including flawed assumptions. Fixing these flaws before you begin collecting data can save thousands of dollars of your research budget and tens or even hundreds of hours of research time. Although your ability to critique your own designs will improve as you gain experience with experimental design, and although this review is less necessary when you are using a design that your previous research has proven to be effective, someone who was not involved in the design process will always have a perspective you lack. They may have knowledge you lack about better methods or problems with the methods you have been using, and can therefore suggests ways to improve the power of your design or reduce the risk of failure.

When you form a hypothesis, it is usually based on your educated belief about how a system is likely to function and how that functioning will respond to experimental manipulation. It is therefore appropriate to focus on tests of your belief. But you must never forget that your belief (and thus, the hypothesis that is based on it) may be wrong, or that something very interesting may be happening that is entirely unrelated to your hypothesis. Focusing too narrowly on your hypoth-

esis can lead to what is called *confirmation bias* (noticing only the data that supports your preconceptions), and this can blind you to negative results that contradict your hypothesis. It can also blind you to potentially interesting phenomena that should be studied in more depth. Thus, the best experimental designs both provide support for identifying contradictory evidence and encourage you to take a step back and look at the larger system in which your study system exists. Some of the most interesting research results are obtained by serendipity rather than careful design.

Wikipedia provides a good summary article on experimental design that includes an extensive bibliography (http://en.wikipedia.org/wiki/Design_of_experiments). Your colleagues will also be a good source of recommendations for textbooks, software, and other resources. Wikipedia also provides a useful glossary of the terminology of experimental design (http://en.wikipedia.org/wiki/Glossary_of_experimental_design) that will help you decipher the literature on this subject.

Statistics

Modern science relies so heavily on statistics that it is almost impossible to publish a study that lacks a statistical analysis. As I noted in the previous section, it's important to confirm that your experimental design supports statistical analysis—and ideally, that it supports *powerful* statistical analysis, since some statistical tests provide weaker evidence than others and may be unpersuasive to a journal's reviewers. If you lack sufficient expertise to choose an appropriate test, speak with a statistician, and revise your experimental design (if necessary) to permit the use of that test. In the rest of this chapter, I'll discuss some aspects of statistics that confuse many researchers and that have caused problems for many of my authors.

Note: Lang and Altman (2013) provide a useful discussion of how to report statistics in journal papers, along with many literature citations that provide additional details.

A few words about significance

The first thing to understand about statistical significance is that a *significant* result is not the same thing as a *true* result. Significance means only one thing: that we can be reasonably confident that a result did

not arise purely from chance. However, although a high level of significance is reassuring, differences in significance levels are far less meaningful than most people believe because of several important but sometimes forgotten points:

- All thresholds for significance are completely arbitrary. The standard significance level of $p < 0.05$ means an error rate of 1 in 20, which is unacceptably high in many contexts.
- The same significance level has very different meanings for a study with a sample size of 10 and a study with a sample size of 1000. Any given p level is far more meaningful in the second case.
- A result that is statistically significant may have no practical significance, whereas a non-significant difference may be so large that it has high practical significance that demands a closer look at the study system.
- Where the range of values in an observed statistical distribution is large, the extreme values (those that fall outside the 95% confidence interval, which are often called "outliers") may be too large to ignore, particularly in large populations that are expected to contain one or more outliers.
- A difference between two treatments that is statistically significant may be a mathematical artifact rather than a real difference if the measurement resolution is sufficiently coarse. For example, if the maximum spatial resolution of a satellite photo is 30 m, a statistically significant 1-m difference in positioning would not be meaningful; that difference is less than the resolution of the image.
- Gelman and Stern (2006) note another underappreciated problem: that a small change in the underlying data can lead to a disproportionally large change in the significance level.

Some authors I've worked with also make the mistake of focusing so narrowly on the statistical results that they don't stop to think about the larger context for the data. Examples of some common errors provide insights into why you should think as carefully about the meaning of the data as you do about its statistical significance:

- **The "base rate" fallacy (failing to account for the probability of a given result in a population):** Consider the case of an identical frequency of computer defects (1 in 100 units require repairs) for two brands: Brand A, which sells 1000 units per month, and Brand B, which sells 100 units per month. Even though the failure rate is identical, the expected number of repairs per month will

be 10 units for Brand A and 1 units for Brand B. Without considering the base rate (sales per month), it is easy to reach the erroneous conclusion that A requires 10 times as many repairs as B, and is therefore an unreliable brand.

- **Extrapolation based on a regression:** A linear or multiple regression can provide an extremely high goodness of fit and strong statistical significance, but without understanding the range of possible values for each parameter in the regression, we cannot tell how far we can safely extrapolate using the regression equation. For example, many natural phenomena exhibit predictable behavior only within a narrow range of conditions; the regression becomes meaningless beyond those conditions, or may require separate regression analysis (e.g., piecewise regression) for two or more subsets of the overall range of conditions. If your goal is to extrapolate beyond a certain range of values, you must adopt an experimental design that will let you identify the limits of a regression.
- **Choosing the wrong form of regression:** Researchers commonly assume that a linear regression is the best form for a given set of data, and fail to test this hypothesis by comparing alternative forms of equation to see whether they provide a better fit to the data. As a result, many processes that are described using a simple but statistically significant linear regression should instead be described using nonlinear regression or (as noted in the previous point) a segmented (piecewise) regression that produces different curves for different ranges of conditions.
- **Correlation may or may not be related to causality:** The fact that two variables are significantly correlated may not be meaningful if you cannot propose an explanation of that correlation based on some underlying physical or biological mechanism. If such a mechanism exists, then the correlation is more likely to provide a measure of the strength of the mechanism.
- **Simpson's paradox (https://en.wikipedia.org/wiki/Simpson's_paradox):** When you perform regression analysis using combined data for two or more groups with different characteristics, you may reach a conclusion that differs dramatically from the conclusion you would reach if you had analyzed the two groups separately. This example illustrates the importance of efforts to detect significantly different groups that would lead to a need for stratified sampling and separate analyses for each group.

The lesson of these examples is that you must think carefully about the meaning of your data and its underlying characteristics *before* you consider whether your results are statistically significant. Reinhart (2013) provides an informative discussion of various common errors related to statistics-based reasoning.

Reporting significance levels

How do you report significant differences? In the text, the most common approach is to state that one value was significantly higher or lower than another value, then add the test name and p level in brackets. However, you should never say only that two results were significantly different, since that does not tell the reader the direction of the difference; always use more precise wording such as "significantly heavier" or "significantly negatively correlated" to clarify the nature of a difference or correlation. To avoid the need to provide this information each time you describe a difference, some journals will let you define the significance level for each test only once, in the Methods section, using wording similar to the following: "Unless otherwise noted, differences were statistically significant at $p < 0.05$." You can then report the p level only for important exceptions to this rule. Note that except for unusually precise statistical comparisons, it is almost never necessary to report a more rigorous criterion than $p < 0.001$.

In figures and tables, the best approach depends on the nature of the statistical test that you used and the specific comparisons that you tested:

- If you are comparing only one value at a time with some reference value, such as the value in the control, you can often label the values that differ significantly with asterisks, then add the following description in the figure caption: "Significance of differences compared with the control [or other named reference]: +, $p < 0.10$; *, $p < 0.05$; **, $p < 0.01$; ***, $p < 0.001$." Delete any of these definitions for significance levels that did not occur in your analysis.

  Note: These four symbols for significance levels are used by most journals, and represent an informal standard that you should not change.

- For multiple comparisons, where you are comparing values both with a reference value and with each other, use the following simple

wording: "Values in a column [or "in a row"] labeled with the same letter do not differ significantly." For more complicated comparisons, you may need a more complex wording such as the following: "Values in a column [or "in a row"] labeled with the same capital letters do not differ significantly among treatments. Values in a column [or "in a row"] labeled with the same lower-case letters differ significantly among times for the same treatment."

A note if English is your second language: Capital letters are A, B, C, ... Z; lower-case letters are a, b, c, ... z.

Although more complex comparisons are possible, the resulting notation may become too complex for this kind of description to be effective. In that case, consider dividing the figure or table into two or more parts, each of which focuses on efficiently presenting only one subset of the comparisons. To avoid duplication of information in the printed version of the journal, it may be necessary to present the less important comparisons as Online Supplemental Material. (See Chapter 18 for details.)

A final note about significance: To avoid confusion, you should only use the word "significant" in numerical comparisons when you are referring to statistical significance. In all other contexts, you should use words such as *important* or *meaningful* to describe the relevance of a result, or words such as *greatly*, *markedly*, *substantially*, *dramatically*, or *clearly* to describe the magnitude of a difference.

Use the right test statistic

As I noted earlier, every statistical test depends on certain assumptions, including assumptions about the underlying distribution of the data. Before you use any statistical test, learn what assumptions it requires, and test to confirm whether those assumptions are valid. This seems like an obvious point, but I have worked with so many young scientists (and some who were not so young) who don't test their data for normality before applying a test that is only valid for a normal distribution that I feel it necessary to remind you of this point. Part of the problem is that many researchers assume that their statistical software will automatically examine the distribution of the data before it allows the use of a statistical test. Often, that is not the case, and you must remember to perform this test yourself. If the test's requirements are not

met, using the test will potentially produce meaningless results. To inform readers that you have performed this test and remind young researchers to follow your example, it is worth explicitly stating that you confirmed that the test was appropriate (e.g., that you tested that the data follows the distribution required by the test).

When your data does not follow the distribution required by a test (commonly, a normal distribution), researchers commonly apply various mathematical transformations to the data until the transformed data follow the required distribution. Two common transformations are based on power functions (e.g., $x' = x^p$, where x' is the transformed value, x is the original value, and p is the power, ranging from -1 to +1) or logarithmic functions (most commonly, log = $\log_{10}$ and ln = $\log_e$). Note that just because you have applied a commonly used transformation, this does not mean that you have produced a normal distribution; you should always confirm that you achieved this result. If the distribution is still not normal, it is tempting to apply an additional transformation, but each new transformation progressively distorts the data you are using for your analysis. Instead, it may be wiser to use a non-parametric statistical test. Many parametric tests that require a normal distribution have non-parametric equivalents; these include the Kruskal-Wallis test instead of one-way analysis of variance, and Spearman's rank correlation instead of Pearson's correlation.

Tip: Even though "log" is assumed, by convention, to mean $\log_{10}$, it is clearer if your write this as $\log_{10}$. So many of the authors I have worked with did not recall this definition that I am convinced that being explicit is safer than assuming that readers will understand the correct meaning. The cost of this explicitness—adding only two characters (10)—is acceptable.

One problem with transformations is that even when there are few outliers (often defined as values that lie more than three standard deviations from the mean), the transformation may compress the data into a small area of the graph, making it difficult to distinguish patterns within the resulting tightly clustered data. In some cases, it may be helpful or even necessary to present both an overall graph that includes all of your data, and an enlarged version of the key parts of the graph that contain areas of interest. Maciejewski (2011) provides a use-

ful discussion of choosing a transformation that is optimal both for the characteristics of the data and for visualization of that data in a graph.

The set of parameters that relate to what are referred to as "measures of central tendency", and which describe the position of a distribution's center, must also be carefully considered. By default, most researchers calculate the mean value, and in doing so, forget that the mean works best for a data with a symmetrical distribution. For strongly skewed distributions, the median (the point where half of the population has a larger value and half has a smaller value) or the mode (the value or range of values with the highest frequency) represent better choices. In some cases, the range of values or the interpercentile range may be more appropriate. Similarly, the standard deviation (SD) provides an estimate of the size of the variation of the distribution around the mean. In contrast, the standard error (more correctly, the standard error of the mean, SEM) represents the precision of estimates of the mean; that is, it is the standard distribution of the sampling error when you use a sample mean to estimate a population mean, and is particularly important in regression analysis. It should not be used in place of the SD just because it appears to indicate a lower magnitude of variation.

Tip: Whenever you provide values in the form A±B, always specify whether B represents the SD or the SEM.

As I noted in the previous section, correlations are often used to represent the strength of the relationship between two variables. However, many authors are confused about the different types of correlation. Pearson's correlation coefficient is not the same parameter as the goodness of fit in least-squares linear regression. How to capitalize these two parameter names varies, but lower-case r is most often used for Pearson's correlation coefficient, whereas capitalized R^2 (not r or R without an exponent) is the regression "goodness of fit" or "coefficient of determination". Pearson's r provides an indication of how closely a relationship between a dependent variable and the independent variable follows a straight line, whereas R^2 represents the proportion of the variation in the dependent variable that can be explained by the dependent variable. The two are clearly related quantities, but because r typically equals the square root of R^2, they cannot be used as synonyms. Because I find considerable confusion about this difference, it's

important to clearly state the parameter name that you are reporting. For example, write "the coefficient of determination (R^2)". Once you have defined your terminology, the reader will understand the meaning.

Tip: For any value that expresses the strength of the agreement between two variables, always report the p value to communicate whether the relationship is statistically significant. For very large values (e.g., values > 0.9), readers will assume significance, but for small or intermediate values, the significance of the result is unclear if you do not state it explicitly.

The final point I'll discuss relates to the benefits and drawbacks of presenting "normalized" versions of your data, which are the values expressed relative to some reference point. Although normalized values are sometimes expressed by subtracting a baseline reference value, the most common approach is multiplicative: the normalized value is expressed as some multiple (ratio) of the reference value. This is particularly important for variables with different units of measurement or from different categories (e.g., mass versus volume), since the normalized values then reflect the proportional change. This calculation is usually done by setting the value for one treatment (usually the mean value for the control, but sometimes a different reference value) to 1.0 or 100%; all other values are then divided by the reference value used in this calculation. In essence, this is no different from calculating the ratios of one value to another.

This approach offers the advantage of providing an intuitive, easily understood explanation of the relative values of the two numbers. However, it suffers from the flaw I described earlier in this chapter as the "baseline fallacy": by concealing the baseline value, it also conceals the meaning of the normalized value. For example, if the mean value for the control is 1.0 units and the value for a treatment is 2.0 units, the normalized value is 200% of the reference value; the same result is obtained if the control has a value of 100 units and the treatment has a value of 200 units, even though the difference (200 − 100 = 100 units) is 100 times the former difference (2 − 1 = 1). This can lead to a problem similar to one that I discussed earlier in the context of significance: a failure to consider the magnitude of the actual difference. The solution is obvious once you know it: provide both the relative val-

ue and the actual value, and carefully distinguish between their different meanings.

Key points to learn

- An important reminder: I am not a professional statistician. Although my advice in this chapter is sound for most typical situations, more complex situations will require different solutions. When in doubt, obtain the advice of an expert statistician when you plan your own research.
- Experimental design seems to be poorly taught at most universities, and there seems to be little guidance on how to integrate statistics with the core subjects in a given field of research. As a result, the authors I've worked with make many mistakes in their experimental design and the associated statistical analysis. In this chapter, I provide some advice based on 25 years of observing these mistakes in the hope that you will be able to avoid the same mistakes.
- The process of experimental design can be summarized as follows: start by identifying the variables that you will control or measure, and methods and instruments suitable for measuring their values. Conclude by looking for ways to error-proof your measurements, and include those ways in your design. Choose an appropriate standard against which to compare your results.
- Look carefully for potential sources of bias in your measurements, and take precautions to eliminate or minimize the bias.
- Don't guess at the replication and sample size that will be necessary to increase the likelihood of obtaining statistically significant results. There are statistical techniques to estimate this sample size, and you should determine which technique is best for your study's goals.
- Always test your design using artificially generated data to ensure that you can successfully analyze the data. Be careful not to generate test data with characteristics (e.g., the distribution) that differ greatly from those of your actual data.
- Design your study to ensure that at least one of the experiments is likely to succeed, thereby producing publishable results.
- Obtain a reality check from your colleagues to ensure that you have not forgotten anything important or made any incorrect assumptions in your experimental design.
- Statistical significance is a complex concept, and there are many misunderstandings of its true meaning. Watch for several logical er-

rors when you describe significance, and clearly distinguish between *statistical* and *practical* significance.

- Before using a statistical test, confirm that your data conforms with the requirements or assumptions of that test. Explicitly confirm this instead of assuming that a widely used test is appropriate for your data; statistical software often does not test to confirm that a given test is appropriate for your data. Although it may be appropriate to transform your data (e.g., to provide a normal distribution), this also has consequences that may be unacceptable for some datasets.
- Normalization of data makes relative trends clear, but conceals the meaning of the actual values. It should therefore be used carefully, and the results should always be interpreted in terms of both the relative value and the actual value.