Chapter 8

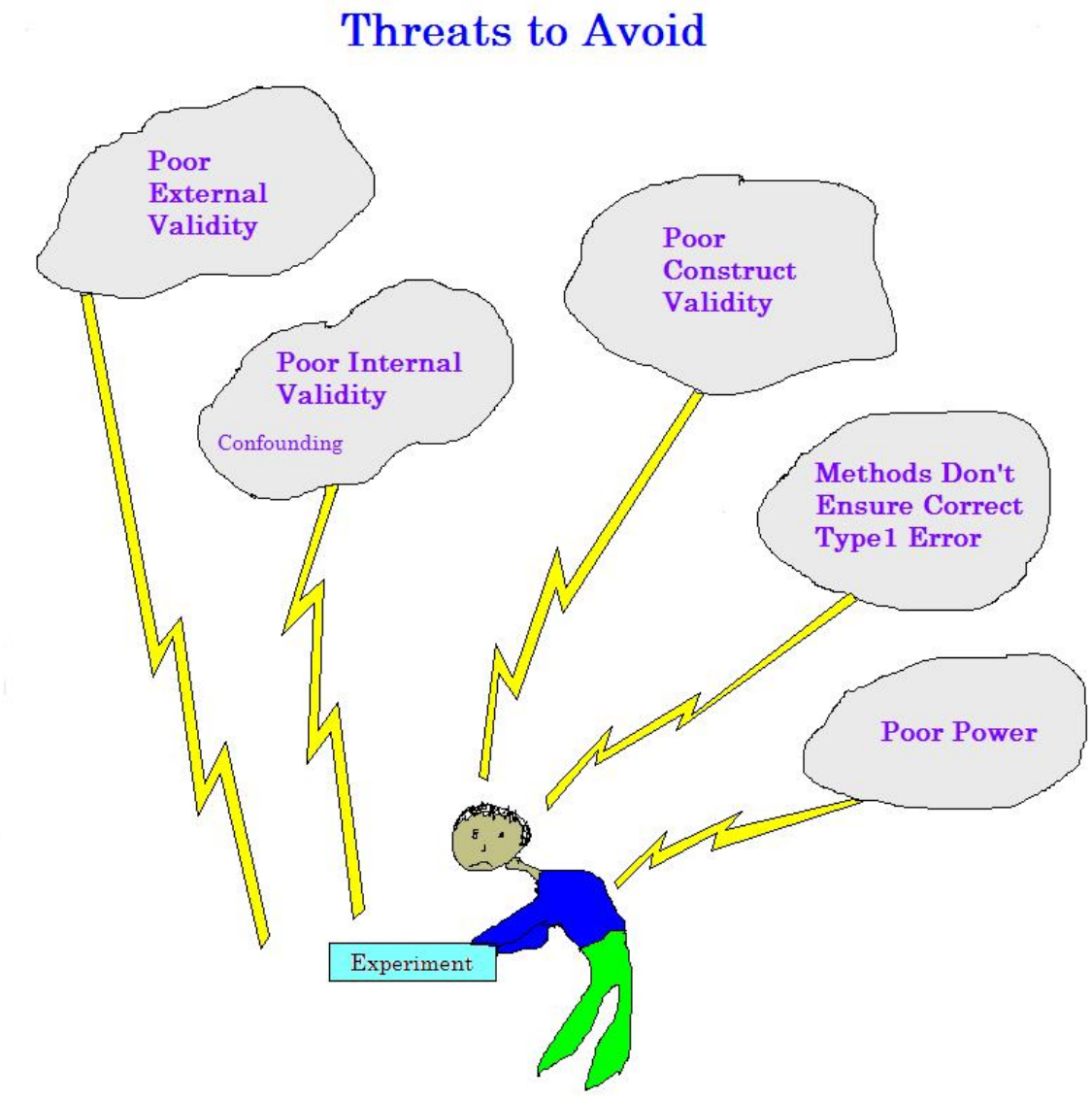
Threats to Your Experiment

Planning to avoid criticism.

One of the main goals of this book is to encourage you to think from the point of view of an experimenter, because other points of view, such as that of a reader of scientific articles or a consumer of scientific ideas, are easy to switch to after the experimenter's point of view is understood, but the reverse is often not true. In other words, to enhance the usability of what you learn, you should pretend that you are a researcher, even if that is not your ultimate goal.

As a researcher, one of the key skills you should be developing is to try, in advance, to think of all of the possible criticisms of your experiment that may arise from the reviewer of an article you write or the reader of an article you publish. This chapter discusses possible complaints about internal validity, external validity, construct validity, Type 1 error, and power.

We are using "threats" to mean things that will reduce the impact of your study results on science, particularly those things that we have some control over.



8.1 Internal validity

In a well-constructed experiment in its simplest form we manipulate variable X and observe the effects on variable Y. For example, outcome Y could be number of people who purchase a particular item in a store over a certain week, and X

could be some characteristics of the display for that item, such as use of pictures of people of different "status" for an in-store advertisement (e.g., a celebrity vs. an unknown model). **Internal validity** is the degree to which we can appropriately conclude that the changes in X *caused* the changes in Y.

The study of causality goes back thousands of years, but there has been a resurgence of interest recently. For our purposes we can define **causality** as the state of nature in which an active change in one variable directly changes the probability distribution of another variable. It does not mean that a particular "treatment" is *always* followed by a particular outcome, but rather that some probability is changed, e.g. a higher outcome is more likely with a particular treatment compared to without. A few ideas about causality are worth thinking about now. First, **association**, which is equivalent to non-zero correlation (see section 3.6.1) in statistical terms, means that we observe that when one variable changes, another one tends to change. We cannot have causation without association, but just finding an association is not enought to justify a claim of causation.

Association does not necessarily imply causation.

If variables X and Y (e.g., the number of televisions (X) in various countries and the infant mortality rate (Y) of those countries) are found to be associated, then there are three basic possibilities. First X could be causing Y (televisions lead to more health awareness, which leads to better prenatal care) or Y could be causing X (high infant mortality leads to attraction of funds from richer countries, which leads to more televisions) or unknown factor Z could be causing both X and Y (higher wealth in a country leads to more televisions and more prenatal care clinics). It is worth memorizing these three cases, because they should always be considered when association is found in an observational study as opposed to a randomized experiment. (It is also possible that X and Y are related in more complicated ways including in large networks of variables with feedback loops.)

Causation ("X causes Y") can be logically claimed if X and Y are associated, and X precedes Y, and no plausible alternative explanations can be found, particularly those of the form "X just happens to vary along with some real cause of changes in Y" (called confounding).

Returning to the advertisement example, one stupid thing to do is to place all of the high status pictures in only the wealthiest neighborhoods or the largest stores,

while the low status pictures are only shown in impoverished neighborhoods or those with smaller stores. In that case a higher average number of items purchased for the stores with high status ads may be either due to the effect of socio-economic status or store size or perceived status of the ad. When more than one thing is different on average between the groups to be compared, the problem is called **confounding** and confounding is a fatal threat to internal validity.

Notice that the definition of confounding mentions "different on average". This is because it is practically impossible to have no differences between the subjects in different groups (beyond the differences in treatment). So our realistic goal is to have no difference on average. For example if we are studying both males and females, we would like the gender ratio to be the same in each treatment group. For the store example, we want the average pre-treatment total sales to be the same in each treatment group. And we want the distance from competitors to be the same, and the socio-economic status (SES) of the neighborhood, and the racial makeup, and the age distribution of the neighborhood, etc., etc. Even worse, we want all of the unmeasured variables, both those that we thought of and those we didn't think of, to be similar in each treatment group.

The sine qua non of internal validity is **random assignment of treatment** to experimental units (different stores in our ad example). Random treatment assignment (also called randomization) is usually the best way to assure that all of the potential confounding variables are equal on average (also called balanced) among the treatment groups. Non-random assignment will usually lead to either consciously or unconsciously unbalanced groups. If one or a few variables, such as gender or SES, are known to be critical factors affecting outcome, a good alternative is **block randomization**, in which randomization among treatments is performed separately for each level of the critical (non-manipulated) explanatory factor. This helps to assure that the level of this explanatory factor is balanced (not confounded) across the levels of the treatment variable.

In current practice randomization is normally done using computerized random number generators. Ideally all subjects are identified before the experiment begins and assigned numbers from 1 to N (the total number of subjects), and then a computer's random number generator is used to assign treatments to the subjects via these numbers. For block randomization this can be done separately for each block. If all subjects cannot be identified before the experiment begins, some way must be devised to assure that each subject has an equal chance of getting each treatment (if equal assignment is desired). One way to do this is as follows. If

there are k levels of treatment, then collect the subjects until k (or 2k or 3k, etc) are available, then use the computer to randomly assign treatments among the available subjects. It is also acceptable to have the computer individually generate a random number from 1 to k for each subject, but it must be assured that the subject and/or researcher cannot re-run the process if they don't like the assignment.

Confounding can occur because we purposefully, but stupidly, design our experiment such that two or more things differ at once, or because we assign treatments non-randomly, or because the randomization "failed". As an example of designed confounding, consider the treatments "drug plus psychotherapy" vs. "placebo" for treating depression. If a difference is found, then we will not know whether the success of the treatment is due to the drug, the psychotherapy or the combination. If no difference is found, then that may be due to the effect of drug canceling out the effect of the psychotherapy. If the drug and the psychotherapy are known to individually help patients with depression and we really do want to study the combination, it would probably better to have a study with the three treatment arms of drug, psychotherapy, and combination (with or without the placebo), so that we could assess the specific important questions of whether drug adds a benefit to psychotherapy and vice versa. As another example, consider a test of the effects of a mixed herbal supplement on memory. Again, a success tells us that something in the mix helps memory, but a follow-up trial is needed to see if all of the components are necessary. And again we have the possibility that one component would cancel another out causing a "no effect" outcome when one component really is helpful. But we must also consider that the mix itself is effective while the individual components are not, so this might be a good experiment.

In terms of non-random assignment of treatment, this should only be done when necessary, and it should be recognized that it strongly, often fatally, harms the internal validity of the experiment. If you assign treatment in some pseudo-random way, e.g. alternating treatment levels, you or the subjects may purposely or inadvertently introduce confounding factors into your experiment.

Finally, it must be stated that although randomization cannot perfectly balance all possible explanatory factors, it is the best way to attempt this, particularly for unmeasured or unimagined factors that might affect the outcome. Although there is always a small chance that important factors are out of balance after random treatment assignment (i.e., failed randomization), the degree of imbalance is generally small, and gets smaller as the sample size gets larger.

In experiments, as opposed to observational studies, the assignment of levels of the explanatory variable to study units is under the control of the experimenter.

Experiments differ from **observational studies** in that in an experiment at least the main explanatory variables of interest are applied to the units of observation (most commonly subjects) *under the control of the experimenter*. Do not be fooled into thinking that just because a lot of careful work has gone into a study, it must therefore be an experiment. In contrast to experiments, in observational studies the subjects choose which treatment they receive. For example, if we perform magnetic resonance imaging (MRI) to study the effects of string instrument playing on the size of Broca's area of the brain, this is an observational study because the natural proclivities of the subjects determine which "treatment" level (control or string player) each subject has. The experimenter did not control this variable. The main advantage of an experiment is that the experimenter can randomly assign treatment, thus removing nearly all of the confounding. In the absence of confounding, a statistically significant change in the outcome provides good evidence for a causal effect of the explanatory variable(s) on the outcome. Many people consider internal validity to be not applicable to observational studies, but I think that in light of the availability of techniques to adjust for some confounding factors in observational studies, it is reasonable to discuss the internal validity of observational studies.

Internal validity is the ability to make causal conclusions. The huge advantage of randomized experiments over observational studies, is that causal conclusions are a natural outcome of the former, but difficult or impossible to justify in the latter.

Observational studies are always open to the possibility that the effects seen are due to confounding factors, and therefore have low internal validity. (As mentioned above, there are a variety of statistical techniques, beyond the scope of this book, which provide methods that attempt to "correct for" some of the confounding in observational studies.) As another example consider the effects of vitamin C on the common cold. A study that compares people who choose to take vitamin C versus those who choose not to will have many confounders and low internal validity. A

study that randomly assigns vitamin C versus a placebo will have good internal validity, and in the presence of a statistically significant difference in the frequency of colds, a causal effect can be claimed.

Note that confounding is a very specific term relating to the presence of a difference in the average level of any explanatory variable across the treatment groups. It should not be used according to its general English meaning of "something confusing".

Blinding (also called masking) is another key factor in internal validity. Blinding indicates that the subjects are prevented from knowing which (level of) treatment they have received. If subjects know which treatment they are receiving and believe that it will affect the outcome, then we may be measuring the effect of the belief rather than the effect of the treatment. In psychology this is called the **Hawthorne effect**. In medicine it is called the **placebo effect**. As an example, in a test of the causal effects of acupuncture on pain relief, subjects may report reduced pain because they believe the acupuncture should be effective. Some researchers have made comparisons between acupuncture with needles placed in the "correct" locations versus similar but "incorrect" locations. When using subjects who are not experienced in acupuncture, this type of experiment has much better internal validity because patient belief is not confounding the effects of the acupuncture treatment. In general, you should attempt to prevent subjects from knowing which treatment they are receiving, if that is possible and ethical, so that you can avoid the placebo effect (prevent confounding of belief in effectiveness of treatment with the treatment itself), and ultimately prevent valid criticisms about the interval validity of your experiment. On the other hand, when blinding is not possible, you must always be open to the possibility that any effects you see are due to the subjects' beliefs about the treatments.

Double blinding refers to blinding the subjects and also assuring that the *experimenter* does not know which treatment the subject is receiving. For example, if the treatment is a pill, a placebo pill can be designed such that neither the subject nor the experimenter knows what treatment has been randomly assigned to each subject. This prevents confounding in the form of difference in treatment application (e.g., the experimenter could subconsciously be more encouraging to subjects in one of the treatment groups) or in assessment (e.g, if there is some subjectivity in assessment, the experimenter might subconsciously give better assessment scores to subjects in one of the treatment groups). Of course, double blinding is not always possible, and when it is not used you should be open to

the possibility that that any effects you see are due to differences in treatment application or assessment by the experimenter.

Triple blinding refers to not letting the person doing the statistical analysis know which treatment labels correspond to which actual treatments. Although rarely used, it is actually a good idea because there are several places in most analyses where there is subjective judgment involved, and a biased analyst may subconsciously make decisions that push the results toward a desired conclusion. The label "triple blinding" is also applied to blinding of the rater of the outcome in addition to the subjects and the experimenters (when the rater is a separate person).

Besides lack of randomization and lack of blinding, omission of a control group is a cause of poor internal validity. A **control group** is a treatment group that represents some appropriate baseline treatment. It is hard to describe exactly what "appropriate baseline treatment" means, and this often requires knowledge of the subject area and good judgment. As an example, consider an experiment designed to test the effects of "memory classes" on short-term memory performance. If we have two treatment groups and are comparing subjects receiving two vs. five classes, and we find a "statistically significant difference", then we only know that adding three classes causes a memory improvement, but not if two is better than none. In some contexts this might not be important, but in others our critics will claim that there are important unanswered causal questions that we foolishly did not attempt to answer. You should always think about using a good control group, although it is not strictly necessary to always use one.

In a nutshell: It is only in blinded, randomized experiments that we can assure that the treatment precedes the outcome, and that there is little chance of confounding which would allow alternative explanations. It is these two conditions, along with statistically significant association, which allow a claim of causality.

8.2 Construct validity

Once we have made careful operational definitions of our variables and classified their types, we still need to think about how useful they will be for testing our hypotheses. **Construct validity** is a characteristic of devised measurements that describes how well the measurement can stand in for the scientific concepts or "constructs" that are the real targets of scientific learning and inference.

Construct validity addresses criticisms like "you have shown that changing X causes a change in measurement Y, but I don't think you can justify the claims you make about the causal relationship between concept W and concept Z", or "Y is a biased and/or unreliable measure of concept Z".

The classic paper on construct validity is *Construct Validity in Psychological Tests* by Lee J. Cronbach and Paul E. Meehl, first published in Psychological Bulletin, 52, 281-302 (1955). Construct validity in that article is discussed in the context of four types of validity. For the first two, it is assumed that there is a "gold standard" against which we can compare the measure of interest. The simple correlation (see section 3.6.1) of a measure with the gold standard for a construct is called either concurrent validity if the gold standard is measured at the same time as the new measure to be tested or predictive validity if the gold standard is measured at some future time. Content validity is a bit ambiguous but basically refers to picking a representative sample of items on a multi-item test. Here we are mainly concerned with construct validity, and Cronbach and Meehl state that it is pertinent whenever the attribute or quality of interest is not "operationally defined". That is, if we define happiness to be the score on our happiness test, then the test is a valid measure of happiness by definition. But if we are referring to a concept without a direct operational definition, we need to consider how well our test stands in for the concept of interest. This is the construct validity. Cronbach and Meehl discuss the theoretical basis of construct validity for psychology, and this should be applicable to other social sciences. They also emphasize that there is no single measure of construct validity, because it is a complex, often judgment-laden set of criteria.

Among other things, to assess contruct validity you should be sure that your measure correlates with other measures for which it should correlate if it is a good measure of the concept of interest. If there is a "gold standard", then your measure should have a high correlation with that test, at least in the kinds of situations where you will be using it. And it should not be correlated with measures of other unrelated concepts.

> It is worth noting that good construct validity doesn't mean much if your measure is not also reliable. A good measure should not depend strongly on who is administering the test (called high inter-rater reliability), and repeat measurements should have a small statistical "variance" (called test-retest reliability).

Most of what you will be learning about construct validity must be left to reading and learning in your specific field, but a few examples are given here. In public health studies, a measure of obesity is often desired. What is needed for a valid definition? First it should be recognized that circular logic applies here: as long as a measure is in some form that we would recognize as relating to obesity (as opposed to, say, smoking), then if it is a good predictor of health outcomes we can conclude that it is a good measure of obesity by definition. The United States Center for Disease Control (CDC) has classifications for obesity based on the Body Mass Index (BMI), which is a formula involving only height and weight. The BMI is a simple substitute that has reasonably good concurrent validity for more technical definitions of body fat such as percent total body fat which can be better estimated by more expensive and time consuming methods such as a buoyancy method. But even total body fat percent may be insufficient because some health outcomes may be better predicted by information about amount of fat at specific locations. Beyond these problems, the CDC assigns labels (underweight, health weight, at risk of overweight, and overweight) to specific ranges of BMI values. But the cutoff values, while partially based on scientific methods are also partly arbitrary. Also these cutoff values and the names and number of categories have changed with time. And surely the "best" cutoff for predicting outcomes will vary depending on the outcome, e.g., heart attack, stroke, teasing at school, or poor self-esteem. So although there is some degree of validity to these categories (e.g., as shown by different levels of disease for people in different categories and correlation

with buoyancy tests) there is also some controversy about the construct validity.

Is the Stanford-Binet "IQ" test a good measure of "intelligence"? Many gallons of ink have gone into discussion of this topic. Low variance for individuals tested multiple times shows that the test has high test-retest validity, and as the test is self-administered and objectively scored there is no issue with inter-rater reliability. There have been numerous studies showing good correlation of IQ with various outcomes that "should" be correlated with intelligence such as future performance on various tests. In addition, "factor analysis" suggests a single underlying factor (called "G" for general intelligence). On the other hand, the test has been severely criticized for cultural and racial bias. And other critics claim there are multiple dimensions to intelligence, not just a single "intelligence" factor. In summation, the IQ test as a measure of the construct "intelligence" is considered by many researchers to have low construct validity.

Construct validity is important because it makes us think carefully whether the measures we use really stand in well for the concepts that label them.

8.3 External validity

External validity is synonymous with **generalizability**. When we perform an ideal experiment, we randomly choose subjects (in addition to randomly assigning treatment) from a population of interest. Examples of populations of interest are all college students, all reproductive aged women, all teenagers with type I diabetes, all 6 month old healthy Sprague-Dawley rats, all workplaces that use Microsoft Word, or all cities in the Northeast with populations over 50,000. If we randomly select our experimental units from the population such that each unit has the same chance (or with special statistical techniques, a fixed but unequal chance) of ending up in our experiment, then we may appropriately claim that our results apply to that population. In many experiments, we do not truly have a random sample of the population of interest. In so-called "convenience samples", e.g., "as many of my classmates as I could attract with an offer of a free slice of pizza", the population these subjects represent may be quite limited.

After you complete your experiment, you will need to write a discussion of your conclusions, and one of the key features of that discussion is your set of claims about external validity. First, you need to consider what population your experimental units truly represent. In the pizza example, your subjects may represent Humanities upperclassmen at top northeastern universities who like free food and don't mind participating in experiments. Next you will want to use your judgment (and powers of persuasion) to consider ever expanding "spheres" of subjects who might be similar to your subjects. For example, you could widen the population to all northeastern students, then to all US students, then to all US young adults, etc. Finally you need to use your background knowledge and judgment to make your best arguments whether or not (or to what degree) you expect your findings to apply to these larger populations. If you cannot justify enlarging your population, then your study is likely to have little impact on scientific knowledge. If you enlarge too much, you may be severely criticized for over-generalization.

Three special forms of non-generalizability (poor external validity) are worth more discussion. First is non-participation. If you randomly select subjects, e.g., through phone records, or college e-mail, then some subjects may decline to participate. You should always consider the very real possibility that the decliners are different in one or more ways from the participators, and thus your results do not really apply to the population of interest.

A second problem is dropout, which is when subject who start a study do not complete it. Dropout can affect both internal and external validity, but the simplest form affecting external validity is when subjects who are too busy or less committed drop out only because of the length or burden of the experiment rather than in some way related to response to treatment. This type of dropout reduces the population to which generalization can be made, and in experiments such as those studying the effects of ongoing behavioral therapy on adjustment to a chronic disease, this can be a critical blow to external validity.

The third special form of non-generalizability relates to the terms efficacy and effectiveness in the medical literature. Here the generalizability refers to the environment and the details of treatment application rather

than the subjects. If a well-designed clinical trial is carried out under high controlled conditions in a tertiary medical center, and finds that drug X cures disease Y with 80% success (i.e., it has high efficacy), then we are still unsure whether we can generalize this to real clinical practice in a doctor's office (i.e, whether the treatment has high effectiveness). Even outside the medical setting, it is important to consider expanding spheres of environmental and treatment application variability.

External validity (generalizability) relates to the breadth of the population we have sampled and how well we can justify extending our results to an even broader population.

8.4 Maintaining Type 1 error

Type 1 error is related to the statistical concept that in the real world of natural variability we cannot be certain about our conclusions from an experiment. A Type 1 error is a claim that a treatment is effective, i.e., we decide to reject the null hypothesis, when that claim is actually false, i.e. the null hypothesis really is true. Obviously in any single real situation, we cannot know whether or not we have made a Type 1 error: if we knew the absolute truth, we would not make the error. Equally obvious after a little thought is the idea that we cannot be making a Type 1 error when we decide to retain the null hypothesis.

As explained in more detail in several other chapters, statistical inference is the process of making appropriately qualified claims in the face of uncertainty. Type 1 error deals with the probabilistic validity of those claims. When we make a statement such as "we reject the hypothesis that the mean outcome is the same for both the placebo and the active treatments with alpha equal to 0.05" we are claiming that the procedure we used to arrive at our conclusion only leads to false positive conclusions 5% of the time *when the truth happens to be that there is no difference in the effect of treatment on outcome*. This is *not at all* the same as the

claim that there is only a 5% chance that any "reject the null hypothesis decision" will be the wrong decision! Another example of a statistical statement is "we are 95% confident that the true difference in mean outcome between the placebo and active treatments is between 6.5 and 8.7 seconds". Again, the exact meaning of this statement is a bit tricky, but understanding that is not critical for the current discussion (but see 6.2.7 for more details).

Due to the inherent uncertainties of nature we can never make definite, unqualified claims from our experiments. The best we can do is set certain limits on how often we will make certain false claims (but see the next section, on power, too). The conventional (but not logically necessary) limit on the rate of false positive results *out of all experiments in which the null hypothesis really is true* is 5%. The terms Type 1 error, false positive rate, and "alpha" (α) are basically synonyms for this limit.

Maintaining Type 1 error means doing all we can to assure that the false positive rate really is set to whatever nominal level (usually 5%) we have chosen. This will be discussed much more fully in future chapters, but it basically involves choosing an appropriate statistical procedure and assuring that the assumptions of our chosen procedure are reasonably met. Part of the latter is verifying that we have chosen an appropriate model for our data (see section 6.2.2).

A special case of not maintaining Type 1 error is "data snooping". E.g., if you perform many different analyses of your data, each with a nominal Type 1 error rate of 5%, and then report just the one(s) with p-values less than 0.05, you are only fooling yourself and others if you think you have appropriately analyzed your experiment. As seen in the Section 13.3, this approach to data analysis results in a much larger chance of making false conclusions.

Using models with broken assumptions and/or data snooping tend to result in an increased chance of making false claims in the presence of ineffective treatments.

8.5 Power

The **power** of an experiment refers to the probability that we will correctly conclude that the treatment caused a change in the outcome. If some particular true non-zero difference in outcomes is caused by the active treatment, and you have low power to detect that difference, you will probably make a Type 2 error (have a "false negative" result) in which you conclude that the treatment was ineffective, when it really was effective. The Type 2 error rate, often called "beta" (β), is the fraction of the time that a conclusion of "no effect" will be made (over repeated similar experiments) when some true non-zero effect is really present. The power is equal to $1 - \beta$.

Before the experiment is performed, you have some control over the power of your experiment, so you should estimate the power for various reasonable effect sizes and, whenever possible, adjust your experiment to achieve reasonable power (e.g., at least 80%). If you perform an experiment with low power, you are just wasting time and money! See Chapter 12 for details on how to calculate and increase the power of an experiment.

The power of a planned experiment is the chance of getting a statistically significant result when a particular real treatment effect exists. Studying sufficient numbers of subjects is the most well known way to assure sufficient power.

In addition to sample size, the main (partially) controllable experimental characteristic that affects power is variability. If you can reduce variability, you can increase power. Therefore it is worthwhile to have a mnemonic device for helping you categorize and think about the **sources of variation**. One reasonable categorization is this:

- Measurement
- Environmental
- Treatment application
- Subject-to-subject

(If you are a New York baseball fan, you can remember the acronym METS.) It is not at all important to "correctly categorize" a particular source of variation. What is important is to be able to generate a list of the sources of variation in your (or someone else's) experiment so that you can think about whether you are able (and willing) to reduce each source of variation in order to improve the power of your experiment.

Measurement variation refers to differences in repeat measurement values when they should be the same. (Sometimes repeat measurements should change, for example the diameter of a balloon with a small hole in it in an experiment of air leakage.) Measurement variability is usually quantified as the standard deviation of many measurements of the same thing. The term **precision** applies here, though technically precision is 1/variance. So a high precision implies a low variance (and thus standard deviation). It is worth knowing that a simple and usually a cheap way to improve measurement precision is to make repeated measurements and take the mean; this mean is less variable than an individual measurement. Another inexpensive way to improve precision, which should almost always be used, is to have good explicit procedures for making the measurement and good training and practice for whoever is making the measurements. Other than possibly increased cost and/or experimenter time, there is no down-side to improving measurement precision, so it is an excellent way to improve power.

Controlling environmental variation is another way to reduce the variability of measurements, and thus increase power. For each experiment you should consider what aspects of the environment (broadly defined) can and should be controlled (fixed or reduced in variation) to reduce variation in the outcome measurement. For example, if we want to look at the effects of a hormone treatment on rat weight gain, controlling the diet, the amount of exercise, and the amount of social interaction (such as fighting) will reduce the variation of the final weight measurements, making any differences in weight gain due to the hormone easier to see. Other examples of environmental sources of variation include temperature, humidity, background noise, lighting conditions, etc. As opposed to reducing measurement variation, there is often a down-side to reducing environmental variation. There is usually a trade-off between reducing environmental variation which increases power but may reduce external validity (see above).

The trade-off between power and external validity also applies to treatment application variation. While some people include this in environmental variation, I think it is worth separating out because otherwise many people forget that it

is something that can be controlled in their experiment. Treatment application variability is differences in the quality or quantity of treatment among subjects assigned to the same (nominal) treatment. A simple example is when one treatment group gets, say 100 mg of a drug. If two drug manufacturers have different production quality such that all of the pills from the first manufacturer have a mean of 100 mg and s.d. of 5 mg, while the second has a mean of 100 mg and s.d. of 20 mg, the increased variability of the second manufacturer will result in decreased power to detect any true differences between the 100 mg dose and any other doses studied. For treatments like "behavioral therapy" decreasing variability is done by standardizing the number of sessions and having good procedures and training. On the other hand there may be a concern that too much control of variation in a treatment like behavioral therapy might make the experiment unrealistic (reduce external validity).

Finally there is subject-to-subject variability. Remember that ideally we choose a population from which we draw our participants for our study (as opposed to using a "convenience sample"). If we choose a broad population like "all Americans" there is a lot of variability in age, gender, height, weight, intelligence, diet, etc. some of which are likely to affect our outcome (or even the difference in outcome between the treatment groups). If we choose to limit our study population for one or several of these traits, we reduce variability in the outcome measurement (for each treatment group) and improve power, but always at the expense of generalizability. As in the case of environmental and treatment application variability, you should make an intelligent, informed decision about trade-offs between power and generalizability in terms of choosing your study population.

For subject-to-subject variation there is a special way to improve power without reducing generalizability. This is the use of a **within-subjects design**, in which each subject receives two or more treatments. This is often an excellent way to improve power, although it is not applicable in all cases. See chapter 14 for more details. Remember that you must change your analysis procedures to ones which do not assume independent errors if you choose a within-subjects design.

Using the language of section 3.6, it is useful to think of all measurements as being conditional on whatever environmental and treatment variables we choose to fix, and marginal over those that we let vary.

Reducing variability improves power. In some circumstances this may be at the expense of decreased generalizability. Reducing measurement error and/or use of within-subjects designs usually improve power without sacrificing generalizability.

The strength of your treatments (actually the difference in true outcomes between treatments) strongly affects power. Be sure that you are not studying very weak treatments, e.g., the effects of one ounce of beer on driving skills, or 1 microgram of vitamin C on catching colds, or one treatment session on depression severity.

Increasing treatment strength increases power.

Another way to improve power without reducing generalizability is to employ **blocking**. Blocking involves using subject matter knowledge to select one or more factors whose effects are not of primary importance, but whose levels define more homogeneous groups called "blocks". In an ANOVA, for example, the block will be an additional factor beyond the primary treatment of interest, and inclusion of the block factor tends to improve power if the blocks are markedly more homogeneous than the whole. If the variability of the outcome (for each treatment group) is smaller than the variability ignoring the factor, then a good blocking factor was chosen. But because a wide variety of subjects with various levels of the blocking variable are all included in the study, generalizability is not sacrificed.

Examples of blocking factors include field in an agricultural experiment, age in many performance studies, and disease severity in medical studies. Blocking usually is performed when it is assumed that there is no differential effect of treatment across the blocks, i.e., no interaction (see Section 10.2). Ignoring an interaction when one is present tends to lead to misleading results, due to an incorrect structural model. Also, if there is an interaction between treatment and blocks, that usually becomes of primary interest.

A natural extension of blocking is some form of more complicated model with multiple **control variables** explicitly included in an appropriate mathematical form in the structural model. Continuous control variables are also called **covariates**.

	Small Stones		Large Stones		Combined	
Treatment A	81/87	**0.93**	192/263	**0.79**	273/350	**0.78**
Treatment B	234/270	**0.87**	55/80	**0.69**	289/350	**0.83**

Table 8.1: Simpson's paradox in medicine

Blocking and use of control variables are good ways to improve power without sacrificing generalizability.

8.6 Missing explanatory variables

Another threat to your experiment is not including important explanatory variables. For example, if the effect of a treatment is to raise the mean outcome in males and lower it in females, then not including gender as an explanatory variable (including its interaction with treatment) will give misleading results. (See chapters 10 and 11 for more on interaction.) In other cases, where there is no interaction, ignoring important explanatory variables decreases power rather than directly causing misleading results.

An extreme case of a missing variable is **Simpson's paradox**. Described by Edward H. Simpson and others, this term describes the situation where the observed effect is in opposite directions for all subjects as a single group (defined based on a variable other than treatment) vs. separately for each group. It only occurs when the fraction of subjects in each group differs markedly between the treatment groups. A nice medical example comes comes from the 1986 article *Comparison of treatment of renal calculi by operative surgery, percutaneous nephrolithotomy, and extracorporeal shock wave lithotripsy* by C. R. Chang, et al. (Br Med J 292 (6524): 879-882) as shown in table 8.1.

The data show the number of successes divided by the number of times the treatment was tried for two treatments for gall stones. The "paradox" is that for "all stones" (combined) Treatment B is the better treatment (has a higher success rate). but if the patients gall stones are classified as either "small" or "large", then Treatment A is better. There is nothing artificial about this example; it is

based on the actual data. And there is really nothing "statistical" going on (in terms of randomness); we are just looking at the definition of "success rate". If stone size is omitted as an explanatory variable, then Treatment B looks to be the better treatment, but for each stone size Treatment A was the better treatment. Which treatment would you choose? If you have small stones or if you have large stones (the only two kinds), you should choose treatment A. Dropping the important explanatory variable gives a misleading ("marginal") effect, when the "conditional" effect is more relevant. Ignoring the confounding (also called lurking) variable "stone size" leads to misinterpretation.

It's worth mentioning that we can go too far in including explanatory variables. This is both in terms of the "multiple comparisons" problem and something called "variance vs.bias trade-off". The former artificially raises our Type 1 error if uncorrected, or lowers our power if corrected. The latter, in this context, can be considered to lower power when too many relatively unimportant explanatory variables are included.

Missing explanatory variables can decrease power and/or cause misleading results.

8.7 Practicality and cost

Many attempts to improve an experiment are limited by cost and practicality. Finding ways to reduce threats to your experiment that are practical and cost-effective is an important part of experimental design. In addition, experimental science is usually guided by the KISS principle, which stands for Keep It Simple, Stupid. Many an experiment has been ruined because it was too complex to be carried out without confusion and mistakes.

8.8 Threat summary

After you have completed and reported your experiment, your critics may complain that some confounding factors may have destroyed the internal validity of your experiment; that your experiment does not really tell us about the real world concepts

of interest because of poor construct validity; that your experimental results are only narrowly applicable to certain subjects or environments or treatment application setting; that your statistical analysis did not appropriately control Type 1 error (if you report "positive" results); or that your experiment did not have enough power (if you report "negative" results). You should consider all of these threats before performing your experiment and make appropriate adjustments as needed. Much of the rest of this book discusses how to deal with, and balance solutions to, these threats.

In a nutshell: If you learn about the various categories of threat to your experiment, you will be in a better position to make choices that balance competing risk, and you will design a better experiment.