1

Introduction: Science Hypotheses and Science Philosophy

Thomas C. Chamberlin (1843–1928) was trained as a geologist but had a keen interest in and impact on science philosophy. His 1890 paper in *Science* advocated the use of "multiple working hypotheses" and is central to the information-theoretic approaches. He was the director of the Walker Museum at the University of Chicago, president of the American Association for the Advancement of Science, and the founder and editor of the *Journal of Geology*. Chamberlin was the president of the University of Wisconsin at the time the paper was prepared. The paper was republished in *Science* in 1965 and is still very worthwhile reading as much of science turned, unfortunately, to testing null hypotheses by the early part of the twentieth century.

1.1 Some Science Background

Science is about discovering new things, about better understanding processes and systems, and generally furthering our knowledge. Deep in science philosophy is the notion of hypotheses and mathematical models to represent these hypotheses. It is partially the quantification of hypotheses that provides the illusive concept of *rigor* in science. Science is partially an adversarial process;

hypotheses battle for primacy aided by observations, data, and models. Science is one of the few human endeavors that is truly progressive. Progress in science is defined as approaching an increased understanding of truth – science evolves in a sense.

Philosophy of science is concerned with the justification of scientific practices. For instance, it may be obvious that an experimenter would want to avoid confounding; however, it may be far less obvious why randomization or parsimony is often so critical in empirical science. Establishing causation and making proper inductive inferences are the domains of a scientist. These are deep issues where science philosophy has played an important role. A good scientist should try to further understanding of philosophy during her or his lifetime of work.

The scientific method tries to formalize, and make efficient, the everyday process of "finding things out." Good science is strategic. Science is fundamentally about understanding, not so much about decisions (however, there are many solid approaches to making "scientific decisions" but that is another subject). Aldo Leopold (1933:231) stated, "We are not trying to render a judgment, rather to qualify our minds to comprehend the meaning of evidence." We will see that evidence can be formally quantified – this is the science of the matter. However, in application, we then often want to qualify such evidence to aid comprehension. Such qualifications are value judgments, are not unique, and can be contentious.

Science is not so much about what is known (although we do speak of the "body of scientific knowledge"), as it is the process of finding out about new things. Science makes progress by providing evidence that good hypotheses are poor so that they can be replaced by even better hypotheses. Science never stops; it is always looking for more.

Ideally, perhaps scientists should be disinterested and unbiased observers. I suspect that human nature prevents this ideal in most of us; instead, we should admit that we often have some "leaning" on many subjects. This leaning reflects, partially, our interest in the subject in the first place. This predisposition can be accounted for as hypotheses are evaluated objectively, relative to one another. For example, when deriving a small set of hypotheses to be evaluated with data and models, one investigator may have a favorite hypothesis, while a colleague may favor another. This can lead to a spirit of competition for ideas and new hypotheses that can be healthy in learning. This is where good data and a sound approach to evaluating the relative strength of evidence for the set of hypotheses become fundamentally important.

Evidence is defined by the American College Dictionary as "grounds for belief" and "something that makes evident." Proof is evidence so complete and convincing as to put a conclusion beyond reasonable doubt. Strict proof may be rare in life sciences. Faith is belief without evidence.

Substantial elements of personal judgment enter in scientific research, especially in the choice of topics of study and in deeper issues of interpretation. To some extent, however, the goal of scientific methods is to minimize that personal element and subjectivity. Often, we will see that the science of a

matter consists of various pieces of quantitative evidence: things like ranking of hypotheses derived *a priori*, the probability of hypothesis *j*, estimates of parameters, and a measure of their precision. Perhaps it is useful to think that science stops there. Then, value judgments can be offered to qualify the result and therefore aid in its interpretation. Such interpretations can be offered by anyone and these may be fairly similar across individuals or may vary quite substantially. The value judgments by the investigator might be of special interest; this is why a Ph.D. level of education becomes important in scientific studies. In the end, the qualification of the quantitative evidence (the science result) involves value judgment that may vary by individual. Goodman and Royall (1988:1573–1574) note:

> *…the use of evidential measures forces us to bring scientific judgment to data analysis, and shows us the difference between what the data are telling us and what we are telling ourselves.*

1.2 Multiple Working Hypotheses

Thomas Chamberlin wrote several papers over a century ago calling for scientists to adopt what he called "multiple working hypotheses." Francis Bacon advocated a similar science strategy 400 years earlier. Their proposal is a sterling blueprint for an effective science strategy but the approach has been underused during the past century.

Under Chamberlin's strategy, one carefully derives several *plausible* science hypotheses (H_i) that become the entire focus of the investigation:

$$\{\boldsymbol{H}_1, \boldsymbol{H}_2, \ldots, \boldsymbol{H}_R\}, \text{where } R \geq 2.$$

These hypotheses are to be well thought out and derived prior to studying the specific data and ideally prior to data collection. In his time, I believe Chamberlin was thinking that R was in the 2–4 range (i.e., small). Forming a small set of plausible hypotheses is where science enters the issue and is the most important step. Research investigators need the ability to think hard about plausible explanations (hypotheses) for a system of interest. Our present science culture places too little emphasis on the derivation of multiple working hypotheses. Many scientific hypotheses seem shallow and uninteresting and most cases are a single science hypothesis to be contrasted with a "null" hypothesis. Such practice cannot be considered twenty-first century science.

Once the *a priori* set of hypotheses has been carefully defined, then one can begin to ask about their relative empirical support. Royall (1997) asks: Given the data,

> *…how do we quantify the strength of evidence for one explanation over the alternatives?*

Edwards (1972) states:

> *Our problem is to assess the relative merits of rival hypotheses in the light of observational or experimental data that bear upon them.*

Chamberlin (1890:758) posed the question,

> *…what is the measure of probability on the one side or the other…?*

Stated another way,

> *What is the empirical evidence for hypothesis j relative to the others in the set?*

These are different ways to ask the *fundamental methodological question* in empirical science. Until fairly recently, science had no general methodological approach to providing answers to these questions. Certainly, null hypothesis testing is quite distant from these serious questions. Hypotheses not in the set remain out of consideration (but more on this later). Finally, one must always consider the possibility that none of the hypotheses have any substantial merit. In such cases, more experience and thinking are required.

Chamberlin said little about models and associated quantification (modeling was the subject of contributions in the twentieth century) and he said even less about *how* the various working hypotheses might be evaluated (what we now often term "strength of evidence"). Given his education in geology, it is possible he was thinking of questions where the answer was effectively deterministic and where there was little uncertainty concerning the evidence. Effective ways are needed to provide relative *evidence* for members in this set of science hypotheses. Such ways are the focus of this text.

Chamberlin believed that the derivation of a "family of hypotheses" had special merit and by its very nature promoted thoroughness. He felt the value of working hypotheses was in its suggestiveness of lines of inquiry that might otherwise have been overlooked. This approach leads to certain "habits of mind" – special ways of thinking carefully about new problems ("thinking outside the box"). Collaboration with peers can often lead to interesting insights concerning alternatives.

1.3 Bovine TB Transmission in Ferrets

Caley and Hone (2002) present a nice example of hypothesis generation concerning the force of infection of bovine tuberculosis (*Mycobacterium bovis*) in feral ferrets (*Mustela furo*) in New Zealand. Caley (personal communication) provided a synthesis as to how they approached the science issue. Their analysis is quite comprehensive; I will only highlight some simple aspects to illustrate their approach to deriving multiple working hypotheses (I encourage readers to study their paper for deeper issues). Caley and Hone (2002)

derived 12 alternative hypotheses concerning disease transmission; this took place over several months and they made a major effort to get an exhaustive set of hypotheses. Caley was closest to the issue and his beliefs were centered around a hypothesis of a dietary infection hazard (H_4 below). Hone was less close to the polarized political debate and the authors viewed this as a benefit as he facilitated a more open perspective to developing plausible alternatives.

They examined the ecological and epidemiological literature as an aid in the derivation of alternative hypotheses, but this examination was not restricted to either ferrets or bovine tuberculosis. They had both science colleagues as well as natural resource managers to debate the merits of various alternative hypotheses. Over time, the tentative hypothesis set narrowed and expanded as a result of a deliberate attempt to "think hard." The first five hypotheses (given below) became somewhat "obvious" and the seven remaining hypotheses arose from recognizing that the first five were not mutually exclusive. They eventually described about 20 hypotheses using logical combinations of these five. The framework for these hypotheses and the analysis to follow includes gender and site as factors in each case.

The most difficult issues involved decisions about the more complex hypotheses and the potential lack of uniqueness of some combinations of the five base hypotheses. Eventually, they decided on 12 hypotheses. To keep this example manageable, it will suffice to focus attention on their five base hypotheses below:

H_1 Transmission occurs from mother to offspring during suckling until the age of weaning, which occurs at 1.5–2.0 months of age

H_2 Transmission occurs during mating and fighting activities associated with it, from the age of 10 months when the breeding season starts

H_3 Transmission occurs during routine social activities from the age of independence, estimated to be about 2–3 months, such as sharing dens simultaneously

H_4 Transmission occurs during scavenging/killing tuberculosis carrion/prey from the age of weaning (1.5–2.0 months of age)

H_5 Transmission occurs from birth because of environmental contamination

Note that each hypothesis asks about *How* and *When*; these are often better science questions than merely *What*, as this tends to be merely descriptive. The question is not "is there an effect" rather there is interest in the size of the effect and this is measured by estimates of model parameters.

If these five hypotheses could be ranked (simple ranking is a form of evidence), based on the data, many people would realize how much more relevant this would be compared to an array of classic *P*-values. I am unsure what the null hypothesis might be: transmission is random, but that seems unlikely.

Most of us have difficulty with complex issues and some forms of quantification. Chamberlin warned that it was easier and seemingly more pleasing to think in terms of simple interpretations than to recognize and evaluate the

multiple factors that may often be operating. He provided an example where he felt people like to be told that the Great Lakes basins in the United States were scooped out by glaciers, than to be taught that three or more factors working successively or simultaneously were responsible and to then try to partition the relative importance of these factors. This is an important insight, while realizing that effective theory often requires some idealization and simplification.

Scientists should think long and hard about the *a priori* hypotheses to include in the set for study and evaluation. This critical step can often take months of thinking and rethinking the issues (*a la* Caley and Hone). Oliver (1991) said it well,

> *When you come across some observation that does not fit the standard explanation, let your mind wonder to see whether some radically different interpretation might do a better job. Perhaps you will think of something that will fit both the new data and the old data and thereby supplant the standard explanation. Toy with different perspectives. Look for the unusual. Try consciously to innovate. Train yourself to imagine new schemes and innovative ways to fit the pieces together. Seek the joy of discovery. Always test your new thoughts against the facts, of course, in rigorous, cold-blooded, unemotional scientific manner. But play the great game of the visionary and the innovator as well.*

Ken Burnham (personal communication) advises,

> *Ideally, one should have a firm justification for including certain hypotheses in the set and, conversely, have an equally firm justification for excluding other hypotheses from the set.*

The definition of the hypotheses in the set is perhaps the most important part of the investigation. This set defines the science at the moment. Statistical science is most successful when full attention is given to problem formulation and hypothesizing creative, plausible alternatives. This is often the case where "two heads are better than one" and where there is a competition for ideas and alternatives. Ball et al. (2005) provide a nice example where 15 hypotheses were developed *a priori* concerning predictions about vegetation and substrate affinities for Palm Springs ground squirrel (*Spermophilus tereticaudus chlorus*).

1.4 Approaches to Scientific Investigations

Much of both science and statistics is about *inductive inferences*. This is a formal process whereby a conclusion about a sample is extrapolated to the population from which the sample was drawn. The data come from the sample only; the remaining members of the population are not observed.

Inductive inference can also be thought of as a conclusion from the past about the future as in forecasting or prediction. Inference is an act or a process.

For such inferences to be valid, in principle, assumptions must be met; e.g., some type of probabilistic sampling of the well-defined population. Furthermore, there are results from logic stating that there is always uncertainty in making inductive inferences. This uncertainty leads to the need to carefully quantify the uncertainty of such inferences (e.g., variances, covariances, standard errors, various types of confidence intervals) and worry about possible biases.

Inference in many of the life sciences is challenging because of the inherent *variation* in living systems. In addition, there are often multiple causal factors and, thus, the need for replication, controls, and worry about confounding.

1.4.1 Experimental Studies

Experiments are the Holy Grail of science because they allow inferences about causation. In science, the word experiment implies treatment vs. control groups, where experimental units are randomly assigned to these groups, and there is deliberate replication. Anything less than these three conditions should not properly be called an experiment. In cases where random assignment has not been done (but treatment and control groups are defined and replication is in place), they are often called "quasiexperiments" and there is a large literature on this important case. Studies without a control group are likely to yield disappointing results as the effect size cannot be estimated (there are exceptions), while studies without replication yield results that are usually tenuous at best.

The main purpose of an experiment is to estimate the size of the effect on a response variable of interest *caused* by the treatment. This is primarily an estimation problem: One wants to have an estimate of the effect size and its standard error (or some other appropriate measure of precision). The experimenter wants to know something about the effect of the treatment on some response variable: Is the effect trivial, small, medium, large, or extra large? This has little to do with *testing* the null hypothesis that the treatment had exactly zero effect. In these cases, I find relatively little use for extensive tables of sums of squares, mean squares, F statistics, various degrees of freedom, and the ever-present *P*-values, followed by a decision to "reject" or "fail to reject" the null hypothesis.

Even in strict experiments, the null is almost never particularly plausible; it is the size of the effect caused by the treatment that is of scientific interest. For example, "I fed bison calves in the treatment group a special dietary supplement and, at that dose level, it caused them to gain an average of six pounds per week over that of the calves in the control group fed the same volume or weight." This science finding might be followed by an evaluation of costs of the supplement and other factors thought to be relevant. If ordered treatments are part of the experimental design, then the "effect size" becomes the nature of the (causal) functional response.

A great deal of excellent information exists concerning the design and conduct of experiments and the analysis of experimental data. It is often useful to think of experiments as "design based inference" as the *design* stipulates a (single) model and the causal inference stems from the experimental design. For example, a randomized complete block design implies (only) a two-way ANOVA model and subsequent analysis. This is an area of statistical science that is quite mature – hundreds of books extol its virtues; it is usually well taught, and a huge variety of computer software exists for this. If the scientific situation allows, experiments are highly regarded and recommended – they represent a philosophical "gold standard" of scientific endeavor because they address causation, not merely association or correlation.

There are close links between experimental design and sampling design (see Snedecor and Cochran 1989). While the objectives clearly differ, the estimators can be viewed within a common statistical framework.

1.4.2 Descriptive Studies

The harsh reality, in many cases, is that a strict experiment simply cannot be done. A quick survey of journals such as *Ecology, Journal of Animal Ecology*, or *Journal of Conservation Biology* will reveal many papers that are not about experimental results. In some cases, an experiment could have been done but the investigators did not realize this and then the results are nearly always compromised. In most cases, however, there are a host of valid reasons why a strict experiment is not feasible. Ethical concerns often prevent strict experimentation in human medical research. In many cases, investigators turn to descriptive work. Descriptive work certainly has its place in science but such inferences are (or should be thought of as) more shallow and tentative.

Related to what I call descriptive studies is the notion of "exploratory data analysis" and much has been written about this approach. I believe too much emphasis has been placed on descriptive work. I also believe that some types of exploratory data analysis are a relatively poor way to make rapid progress in the empirical sciences. It is too easy to mistakenly think that results from *post hoc* analyses (data dredging) will lead to interesting new hypotheses, when often such results have high probabilities of being, in fact, spurious. The best ways to obtain interesting alternative hypotheses is to think, read, study, attend scientific meetings, and communicate with both colleagues and rivals.

1.4.3 Confirmatory Studies

Another alternative, confirmatory investigation, lies in between strict experiments that may provide evidence of causation and the descriptive studies that often provide only "what?" Confirmatory investigations begin by hypothesizing alternatives prior to data analysis and, ideally, even before data collection. When data are analyzed and "results" appear, these are confirming prior hypotheses. This is a level above descriptive studies where findings come

almost by surprise in many cases. The investigator thinks, "Wow, who would have thought that nest success of species X was related to estimated cloud cover on the day and time an observer found the nest." On the contrary, under a confirmatory format, the investigator notes something like, "Oh, we have thought for some time that nest success is influenced by concealment at the time of nest initiation, now we have some (confirmatory) evidence of this." Better still, "now we ask some hypothetical questions about why and when concealment is important."

Clearly, there is a close link between the confirmatory approach and Chamberlin's ideas of multiple working hypotheses. Platt's (1964) well-known paper on "strong inference" is closely related and addresses the issue of strategy in science. One still lacks a notion of strict causation but the strength of evidence is nearly always above that for descriptive studies. The only price to be paid to achieve a confirmatory result is *a priori* thinking. We must ask why more confirmatory research is not being done. A little hard thinking before data collection and analysis provides a much stronger inference. This approach sets up a basis for formal evidence. Putting in place the *a priori* hypotheses is just good science procedure and it yields a superior result. Of course, some *post hoc* analysis can be done, and I promote this, but these inferences must be treated with appropriate caution. The reader should be informed which results were from the confirmatory process and which came from *post hoc* analyses of the same data. Confirmatory investigations are the domain of model based inference and the primary focus of this text.

Fundamental differences between strict experiments and other types of studies can be further understood in terms of residual variation (e.g., the $\in_i$ in regression). In experimental data, the residuals are the component of variation that is considered to be random, where the model is defined by the experimental design, and, in this sense, is "known." In contrast, nonexperimental (observational) data must also (i.e., in addition to) treat the residuals as containing the effect of as yet unknown confounding variables on the estimated response variable, (somehow) given the model. Here, the model is not known and must be estimated using some model selection procedure. Strict experimentation is quite distinct from other approaches to science questions.

I find that many investigators have a fear during data analysis that they will miss an effect that is real and perhaps important. They worry that the data are wanting to tell something, but that they will miss this finding by "not looking hard enough." This is a valid fear and it may be common, in complex settings, that some second-order effects are missed, even though they might be in the data. Perhaps an interaction is missed or a nonlinearity is left unnoticed. Investigators try to minimize missing effects by examining "all possible models" or using some multivariate software – there are huge inferential risks associated with these seemingly logical approaches.

The associated risk is finding a spurious effect. That is, the analysis picks up some effect that is particular to the data set that is not part of the process of interest. In a sense, noise is being detected and modeled as if it were part

of the process. One has no way to know, based on a given data set of limited size and scope, if a particular effect is spurious or real. This is the risk that people tend to forget and misunderstand. They want to be sure they "did not miss anything" and while doing so they, instead, find results that are spurious. Many things can be done to lessen this nasty issue, but too many investigators continue to forge ahead, unaware of the risks involved. Spurious results arise with high probability when one has little subject matter theory, measured many variables (e.g., more than the sample size in extreme cases), had small sample size (e.g., 20 or 50), and many models (hundreds, thousands, or even millions of models). It takes some experience and maturity to really begin to understand these issues. Model selection theory can help pinpoint these problems but by then it may be too late to salvage the study.

1.5 Science Hypothesis Set Evolves

Expanding on the ideas of Chamberlin and Platt, we want the set of hypotheses to evolve over time. That is, a team of researchers might start with a set of five hypotheses and find, after a careful empirical evaluation, that two of these were implausible to the point they could be dropped from further consideration. Thus, the set is reduced to only three hypotheses that survived the evaluation. These three might then be further refined and elaborated upon and perhaps one new hypothesis introduced. Hence, a set of four hypotheses are now available for evaluation with new data. If the data set is fairly small, one must be careful not to discard more complex hypotheses, particularly if a much superior data set is expected for the next evaluation. Science can progress very rapidly as the evolving set of hypotheses are constantly challenged with new data (information) and careful evaluation.

The hypothesis set is made to evolve over months, years, and decades; the goal is to keep careful focus on the hypotheses that remain plausible and in the set. It is these hypotheses where improved understanding is sought. The strategy is to constantly make this set move along as knowledge is broadened and further understanding is gained. *A priori* reasoning and hard thinking are both critical and difficult. Scientists should not fail to acknowledge that there may be more than one process that would yield a particular outcome (Platt 1964; Pigliucci 2002a). A real focus needs to be placed on the addition of very new and different hypotheses as the next set is defined. Ideally, the model set would be built to consider those outcomes using experiments or observational studies to separate the alternative hypotheses.

The hypothesis set might ideally evolve as national or international teams vie for understanding and knowledge. Peer pressure and national pride might help drive progress on some interesting problems by showing that one or more hypotheses are implausible, relative to the others in the set. By then other teams are already formulating new hypotheses, perhaps suggested by the prior

results. Pressure to evolve the set might be within a laboratory, university, or an agency, where fast learning is important. Of course, one's own personal scientific progress can be based on the notion of wanting the set to evolve so that rapid understanding may be achieved.

One must not be too eager to rule a particular hypothesis "implausible"; if there are seemingly valid reasons to retain it, then its retention may be appropriate. The decision to retain it must be based on the quantitative evidence and ways to obtain such evidence are given in the following chapters. Sometimes a complex hypothesis is rejected largely because the data set is small and there is little information in the data. In this case, the hypothesis should probably be retained if the next data set is thought to be larger and more informative.

1.6 Null Hypothesis Testing

Soon after Chamberlin's science strategy was published in 1890 and widely accepted, investigators found it easy to derive a single science hypothesis (H_a) and then contrast it with a "null" hypothesis (H_0),

Null hypothesis H_0 vs. Science hypothesis H_a.

This approach was prompted by emerging developments for randomization and strict experiments in statistics in the early 1900s. "Student," Fisher, Neyman, Pearson, Wald, and many other pioneers in statistical theory developed methods for "testing" null hypotheses and this became the dominant analysis paradigm for perhaps seven to nine decades. The approach advocated by Fisher differed substantially from that of Neyman and Pearson and this led to a heated and protracted debate. Now, these approaches are combined in a fashion that would have greatly displeased the combatants involved. Such testing has come under increasingly harsh criticism since at least 1938, particularly when used for the analysis of data from observational studies. It now seems clear that this standard "testing" approach is of limited value as new approaches have several important advantages. Still, I find statistics departments around the world teaching primarily null hypothesis testing methods developed in the early 1900s. Worse yet perhaps is the continued focus on the myriad approaches to multiple comparison tests. This focus on relatively poor methods seems increasingly senseless and few people seem to have any idea why good statistics departments cannot do better in teaching current theory and application (e.g., likelihood, information-theoretic, and objective Bayesian approaches).

It is important to realize that null hypothesis testing was *not* what Chamberlin wanted or advocated. We so often conclude, essentially, "We rejected the null hypothesis that was uninteresting or implausible in the first place, $P < 0.05$." Chamberlin wanted an *array* of *plausible* hypotheses

derived and subjected to careful evaluation. We often fail to fault the trivial null hypotheses so often published in scientific journals. In most cases, the null hypothesis is hardly plausible and this makes the study vacuous from the outset. Chamberlin noted, "The vitality of the study quickly disappears when the object sought is a mere collection of dead, unmeaning facts." For example,

H_0: the population size of species X is the same in urban and rural areas,
H_0: species diversity does not change through geologic time,
H_0: the population correlation between variables X and Y is exactly 0, and
H_0: bears do not go in the woods.

Surely, these null hypotheses are false on simple *a priori* grounds – data collection and analysis are hardly needed in cases such as these. A hundred thousand such null hypothesis tests have appeared in the journal *Ecology* in the recent past (over a 20-year period, Anderson et al. 2000). C. R. Rao (2004), the famous Indian statistician, recently said it well, "… in current practice of testing a null hypothesis, we are asking the wrong question and getting a confusing answer."

We must encourage and reward hard thinking. There must be a premium placed on thinking, innovation, synthesis, and creativity; the computer will be the last to know! We need to ask more about *how*, *when*, and *why*, which are more interesting and potentially important, instead of such a focus on *what*, which is often only descriptive.

1.7 Evidence and Inferences

Scientific evidence lies in a triangular plane surrounded by philosophy, statistics, and subject matter science. Statistical inference is the foundation of modern scientific thinking. Many people are unaware of the extent that fundamental scientific thought processes have been influenced by philosophers. This may be especially true in life sciences.

Evidence is something less than proof! Evidence provides a foundation leading to understanding and this ideally leads to useful theory. Theory is a body of knowledge and understanding that has stood the test of considerable time and effort to disprove it. A theory makes predictions and has been found to be generally useful. Theory might eventually be accepted as a law.

Chamberlin (1890) asked, "What is the measure of probability on one side or the other." Methodology to allow such probabilities took nearly 100 years to develop! The field of statistics became sidetracked and much of the twentieth century was occupied with "testing" null hypotheses and resulting P-values that were often used as if they represented a strength of evidence. I believe that P-values reject or fail to reject dichotomies, and the often trivial null hypotheses that they represent are being replaced with formal methods to quantify and allow comprehension of the evidence for members of a set of alternative science hypotheses.

Examination of the differing probabilities shows a stark difference in meaning. A traditional null hypothesis test is based on a summarization of the data into an appropriate test statistic T. The associated P-value is the probability of T being as large or larger, *given* the null hypothesis is true. The P-value is a so-called "tail probability" and has been criticized as assigning probability to data not collected. Shortening the definition slightly, the P-value is

Prob(observed data or more extreme | null hypothesis).

Conditioning on the null hypothesis might seem odd as it is uncommon to believe in the null. If most null hypotheses seem implausible on *a priori* grounds, why condition on such a notion? I think one should condition on the data – that is why data are collected. Data serve as the arbitrator, the jury, the judge; conditioning on the null hypothesis is not intuitive. The relevant probabilities deal directly with the individual science hypotheses, *given* the data,

$$\textbf{Prob}(H_j \mid \textbf{data}), \quad \text{for } j = 1, 2, \ldots, R.$$

1.8 Hardening of Portland Cement

I will use a well-known problem involving cement hardening as an example. This is a relatively simple problem that can be made to illustrate several important issues and I will use it in several chapters to follow. Woods et al. (1932:635–649) published the results of a small study of the hardening of Portland cement; Daniel and Wood (1971) and Burnham and Anderson (2002) provide further details on these data for the interested reader. Interest was in the calories of heat evolved per gram of cement after 180 days; this relates to hardening and was their response variable, denoted here as y. The objective of the study was twofold: (1) identify the important variables related to the response variable and (2) use these to predict the response variable. Four predictor (or explanatory) variables were of interest:

x_1 = % calcium aluminate ($3CaO \cdot Al_2O_3$)
x_2 = % tricalcium silicate ($3CaO \cdot SiO_2$)
x_3 = % tetracalcium alumino ferrite ($4CaO \cdot Al_2O_3 \cdot Fe_2O_3$)
x_4 = % dicalcium silicate ($2CaO \cdot SiO_2$)

In arriving at a set of plausible hypotheses, several lines of reasoning could be followed. The first observation might be that cement is a mixture of ingredients, thus hypotheses involving only a single variable might be deemed implausible. A more shrewd observation might be that x_2 and x_4 have very similar chemical make up,

x_2 tricalcium silicate $(3CaO \cdot SiO_2)$,
x_4 dicalcium silicate $(2CaO \cdot SiO_2)$.

Also, the chemical composition of variables x_1 and x_3 also seem somewhat similar,

x_1 calcium aluminate $(3CaO \cdot Al_2O_3)$,
x_3 tetracalcium alumino ferrite $(4CaO \cdot Al_2O_3 \cdot Fe_2O_3)$.

This line of thinking might avoid hypotheses involving the pair of variables x_2 and x_4 and the pair x_1 and x_3 as they are so similar. We can further check these issues when we have data available (i.e., look at the sampling correlations between these pairs of variables). The investigator might consider the importance of an interaction such as $x_1 * x_2$ and add these as hypotheses. A skeptic (scientists should be skeptics) might ponder the notion that none of the four predictor variables had any merit; if this seems plausible, then it should be included in the hypothesis set. This would be the intercept only model (a null model of sorts) with two parameters: the mean β_0 and the residual variance σ^2. Here this seems unlikely if we assume people in the 1930s had some basic notion of what they were doing in regard to cement making. Still, if this is deemed plausible, it should go in the set; if not, then it should be excluded. We will leave it just for this example (ordinarily I would deem this implausible in this case and omit it). In summary, we might have the following five hypotheses in the set (defined *a priori*; i.e., before data analysis):

H_1 No variables
H_2 x_1 and x_2
H_3 x_1 and x_2 and $x_1 * x_2$
H_4 x_3 and x_4
H_5 x_3 and x_4 and $x_3 * x_4$

Data and simple models of these hypotheses will be given in Chap. 2.

1.9 What Does Science Try to Provide?

There are several common goals in scientific inquiry. Science is very broad and actually defies a simple but adequate definition. Science is a process leading to discovery, understanding, and solutions of well-defined questions about effects; some strategies are better than others. Some parts of science might be classified as:

- Evaluating the strength of evidence for alternative science hypotheses, *a la* Chamberlin
- Prediction of an outcome, given data, a model, estimates of model parameters, and specific values of the predictor variables

- Determination of model structure (e.g., concave or convex in a simple setting)
- Selection of "important" variables from a larger set (variable selection in regression or discriminate function analysis)
- Pattern recognition or smoothing (parsimony)

All of these classes are examples of model based inference. I will address these matters in the material to follow.

There are some distinctions between science and technology that are sometimes worth noting. Estimates of effect size or predictions might often be best classed as technology. Science might best be thought of as discovery in an exciting sense; for example, providing a strength of evidence for fundamental alternatives. Classification of science vs. technology is rarely distinct and there are wide areas of overlap; still the distinction is often useful to keep in mind.

1.10 Remarks

Developing interesting hypotheses is an art but people can become adept at this with dedication and practice. Ford (2000:Chaps. 4 and 13) and Gotelli and Ellison (2004:Chap. 4) provide relevant reading. Cox (1990, 1995) reviews the relationships between hypothesizing and modeling. Krebs (2000) offers a relevant and easy-to-read review of hypotheses and models. Peirce (1955), Moore and Parker (1986), Abelson (1995), Pigliucci (2002b), and Cohen and Medley (2005) give valuable perspectives on hard thinking, statistical reasoning, and science principles. Careful reading of journal papers in one's field can be enlightening as you can begin to understand how others thought about their science. However, only a minority of papers are exemplary in this important regard. Beyond reading one must think broadly about alternatives, draw from conversations with colleagues, attend conferences, and use new technologies (e.g., the Internet and e-mail) to forge science ideas.

Mead (1988) and Manly (1992) provide methods for the design and analysis of experimental data. Cook and Campbell (1979) and Shadish et al. (2002) deal with quasiexperiments, both design and analysis, in readable books. Observational studies are well covered by Rosenbaum (2002). There are many dozens of good books on experimentation and applied statistics. I will let the readers make their own choices; however, I find that Resetarites and Bernardo (1998) and Williams et al. (2002) make many deeper issues clear with examples.

It might seem surprising but during Chamberlin's time there was an effort to minimize theorizing; this dead end was an attempt to reform "ruling theory" that was in place then. Chamberlin believed his strategy "promotes thoroughness and suggests line of inquiry that might otherwise be overlooked." Additional insights on Chamberlin's method are found in Elliott and Brook (2007).

Many science problems arise where the objective is a structural issue (see Blanckenhorn et al. 2004 for an evolutionary issue that focuses on linearity vs. nonlinearity).

Rao's (2004) short comment is full of interesting insights and Krebs (2000) provides a readable treatment concerning hypothesizing. Forsche's (1963) one-page paper in *Science* is delightful reading. More troublesome is O'Connor's (2000) paper on lack of progress in ecology compared to other areas of biology (also see the interesting follow-up by Swihart et al. 2002). However, see Mauer (1999) for evidence of substantial progress.

Kendall and Gould (2002) respond to the criticism that statistics departments often provide poor course material for people in the life sciences. Their excellent point is that biologists often arrive for statistics courses with a poor science background! Biology students arrive without a grounding in experimentation, causation, confounding, and other subjects, making statistical concepts seem out of context (because students often lack that context). Biology students should have a better grounding in the history of science in general and in their field in particular. Science philosophy is equally fundamental in university courses.

Anderson et al. (2001a) provide an overview of the issue of spurious effects and how to minimize this risk. Hobbs and Hilborn (2006) examine alternatives to null hypothesis testing in a readable paper aimed at ecologists. Their Fig. 1 is interesting to compare with the results shown by Anderson et al. (2000).

O'Connor (2000) states: "Moreover, many of the critical breakthroughs in molecular biology have come from experiments that discriminate between alternative outcomes. Critically, ecology seems to have substituted the statement of what will be observed as the hypothesis."

Freedman (1983) used MC simulation methods and stepwise regression to illustrate the difficulties faced when one has (1) little or no theory, (2) a large number of "models" (as there is little theory leading to *a priori* hypotheses, and (3) small sample size. In such cases, inference is very risky and a plethora of spurious results are found (also see Flack and Chang 1987, Freedman 1983, and Rencher and Pun 1980). Freedman demonstrated this phenomenon by a large matrix of uncorrelated random numbers and stepwise regression. While there were clearly no relationships underlying the data, numerous "significant findings" resulted. This has become known as Freedman's paradox. Unfortunately, many studies in the life sciences are done under these conditions and find their way into the published literature.

The paper by Goodman and Royall (1988) contains many philosophical insights and are well worth careful reading. Recently, Keppie (2006:244) provides fresh perspectives, including mention of the "…temptation to advocate value beyond evidence." Hilborn and Mangle (1997) and Kuhn (1970) contain valuable insights.

There are many good books on science philosophy; I enjoy Horner and Westacott (2000), and Ford's (2000) book is a standard one. Papers by Platt (1964) and Popper (1972) are highly recommended. Taper and Lele (2004) summarize several philosophies of interest. Platt (1964) is quoted by Pigliucci

(2002:92) saying, "Scientists become method oriented rather than problem oriented. Stop doing experiments for a while and think."

1.11 Exercises

The following exercises are provided to strengthen the understanding of some of the conceptual issues in this introductory chapter. These questions can be addressed individually but I find it more effective and fun to tackle these issues in small groups of people. This assumes everyone in the group has read the paper and is ready to think hard about the issues.

1. Obtain a good scientific journal in your subdiscipline of interest and look for an article where the investigators used a confirmatory approach or model based inference and begin to carefully critique it. Consider the following questions:

   a. Were only two alternatives hypothesized? Or, were there more alternatives hypothesized?
   b. In any case, do you consider these to be plausible?
   c. Did the investigators justify their set of alternative science hypotheses or was there a rush to models?
   d. Was there a clear statement of the *a priori* hypotheses in the set?
   e. Did the authors entertain all possible models? What problems might this cause (advanced question)?
   f. Did the authors clearly present the body of evidence to the reader?
   g. Did they then offer some qualification (value judgment) of the evidence to aid in interpretation?
   h. What alternative hypotheses would you have added? Before data collection and analysis? After data analysis (*post hoc*)?
   i. As an associate editor, what would be your basis for rejection of this paper (had this been a submitted manuscript)?
   j. As a reviewer, what constructive advice would you have given the authors of the submitted manuscript?
   k. What other issues might be considered in your critique?

2. Using the same scientific journal as above, find a paper in your subdiscipline of interest that used null hypothesis testing as the basis for the results. Consider the following questions:

   a. Isolate the null hypotheses tested (these are often not stated explicitly). Can you offer your judgment as to the plausibility of these null hypotheses? That is, before data analysis or reading the *Results* section of the paper, were the null hypotheses to be tested plausible? Interesting?
   b. Were estimates of effect size given with a measure of their precision? If not, can you explain this omission?

c. Define exactly what is meant by *P*-value. Did the author use this definition correctly or (possibly inadvertently) redefine it in an *ad hoc* manner? How are data involved in *P*-value? Is *P*-value an estimate?
d. Which hypothesis is being formally "tested"? Is it H_0 or H_a? Why?
e. Is *P*-value a measure of strength of evidence?
f. Do the authors provide qualitative statements concerning "significance"? Do they differentiate statistical vs. biological significance?
g. Was a causal result implied or claimed? Do you feel this was justified? How? Why?
h. In what sense were the units (e.g., experimental, observational, sample) taken from a well-defined sample from a population studied? This issue relates to the proper scope of inference. A very large number of similar questions could be asked here. What other issues might you consider?

3. Using a journal of choice in your field of interest, find a paper that provides the results of an experiment. Read it carefully and consider the following questions:

a. Were there clearly defined treatment and control groups?
b. How much replication was used? Was the sample size given?
c. Were the experimental units randomly assigned to treatment vs. control groups? How, specifically?
d. Did the authors imply a causal result? Does this seem justified to you?
e. Did the authors provide an estimate of the size of the effect caused by the treatment? And some measure of its precision?
f. How could the experiment been better (list 3–4 ways)?
g. Assuming that the answer to at least one of the questions a, b, or c was negative, what are the consequences? Elaborate. What was lost? What could have been done differently? Is the result still of interest, in your opinion?

4. Cohen (1966, 1967, 1968) used a combination of imagination and modeling leading to several science hypotheses about optimizing reproduction of a desert plant species in randomly varying environments. This set of papers triggered perhaps dozens of field experiments to gain further insights into this issue. This set of papers would make a great brown bag discussion for those interested in developing sophisticated alternative hypotheses.