Current Trends in Ecology
Vol. 2, 2011

Review

The logic of scientific discovery

Craig Loehle*

National Council for Air and Stream Improvement, Inc., 552 S Washington Street, Suite 224, Naperville, Illinois 60540, USA

ABSTRACT

The widespread view of science as consisting of experiments that test hypotheses obscures what is in reality a multi-stage logical process. Erroneous or vague resolution of issues at any stage can lead to inconclusive or contradictory results. Deriving a prediction from a theory or understanding a datum or experiment in relation to a theory is fraught with difficulties. Thus posing a well-formed scientific question and designing a proper study in relation to this question is difficult in all but the simplest cases. Delineation of objects and system boundaries is subject to definitional ambiguities and requires specification of experimental frames. Proper study execution, data evaluation, and logical inference of the meaning of results all have their own difficulties. In this study, each of these stages of a research project are evaluated to clarify the sources of ambiguities and uncertainties and suggestions are offered for reducing errors and speeding scientific progress.

KEYWORDS: hypothesis testing, scientific method, theory

INTRODUCTION

As presented in introductory science and statistics textbooks, the "Scientific Method" and statistical "Hypothesis Testing" seem relatively simple. However, such apparent simplicity obscures a nuanced process that if not carefully observed and followed can result in spurious, erroneous or conflicting results that do little to advance scientific knowledge. As an anecdotal example, how often have we heard that eggs or coffee or wine are good for us, then bad for us, and then good for us… and this is a relatively simple type of question. Why are the results of so many studies contradictory (e.g., [1]) or later shown to be wrong [2]? We can further note evidence for confirmation bias or lack of deductive rigor in certain fields [3] and the widespread presence of basic errors in published papers [4, 5]. The reason for these deficiencies, I believe, is that the logic of scientific discovery in reality consists of a long chain of assumptions, deductions, and inferences. It is all too easy for an error to occur at one or more steps in this chain and thereby produce an inconclusive or even wrong result. Very often, the simplifying or auxiliary assumptions are implicit and maybe even unconscious, and can likewise bias the result. It is useful to lay out these steps explicitly to perhaps enhance the discovery process.

Science consists of both theories and empirical questions, and both are fraught with difficulties. I use the term "theory" rather than "hypothesis" here to distinguish my point from a statistical "hypothesis" and also from hypotheses which are logical statements describing phenomenon predicted from a theory. The steps of inference surrounding a theory can be complicated and ambiguous (Figure 1). It is possible that more than one theory makes the same prediction (Figure 1a), in which case verifying the prediction does not allow discrimination between theories. For example, Allouche and Kadmon [6] showed that multiple versions of the neutral model of community organization produce

*cloehle@ncasi.org

the same species abundance distribution as predictions. The best theories predict phenomena that are unambiguous, unique, and clearly different from a random phenomenon. Atomic theory is exemplary in this regard. For example, antimatter was specifically predicted before it was discovered. This is both a novel prediction and one that is uniquely testable (no other phenomenon gives the same result). In contrast, a theory may not be at a stage of development where specific predictions can be made (Figure 1b) [7]. For example, before Newton it was understood that the Earth attracted objects, but Newton made the hypothesis more general (all bodies attract each other) and further made detailed predictions of motion possible by his invention of the calculus. In some cases, (Figure 1c) different experiments attempting to test the same theory will lead to different results. The only inference that can be made in this case is that something somewhere is confounded, either by error, experimental device, or sampling regime. Finally, (Figure 1d), data may appear to be in conflict with a theory. However, it is not always clear what should be done about it. Is it a special case? Is the data suspect? Should the theory be discarded or merely modified?

The second stage of scientific enquiry is empirical. We may be testing or refining a prediction from a theory (Figure 1) or asking an empirical question. There are many empirical questions for which no deduction from theory is possible. For example, we might want to know how well a paint withstands salt water, or which fertilizer best enhances corn growth.

For either theoretical or empirical questions, the first link in the chain of logic is only intact if the question is well-posed [8]. For example, the philosophical question, "Why does the universe exist?", is simply unanswerable through scientific inquiry, and is therefore meaningless in an empirical sense. If a theory is not well enough developed or quantitative enough to make specific predictions, any attempt to test it will be ill-posed because it will not be possible to tell if the experimental outcome was or was not consistent with the theory. For empirical questions, it is not meaningful to ask questions with greater precision than our means of measurement, such as wishing to know the exact population of wolves in a forest.

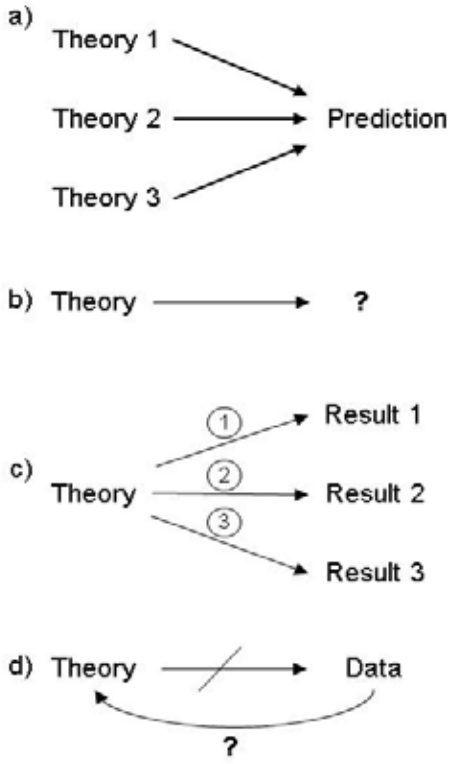

Figure 1. Inference from theories is not necessarily straight forward. a) Multiple theories might make the same prediction, making observation of that outcome not informative. b) It might not be clear what exact outcome is predicted by a theory, either due to theory vagueness or the need for auxiliary assumptions. c) Attempts to test a theory may yield multiple incompatible results due to confounding effects. d) When a fact contradicts a theory, it may not be clear what to do.

The second link in the chain of logic is intact if we conduct the study using a proper statistical approach. Perhaps the most common flaw is to devise an experiment that is only likely to produce a result consistent with a given hypothesis, without excluding other possibilities. Such an experiment is unlikely to refine the theory under test in any meaningful way (see Figure 1a). For example, Loehle [9] showed that studies of the species-abundance distribution relationship tend to only discuss a fit of the model being presented rather than comparing it to other possible models, and that there are inherent difficulties in acquiring a sufficient set of data for such tests. Furthermore, since multiple hypotheses can produce the same abundance distribution [6] a fit to the data per se is not very informative. In other cases, sampling

methods or measurement biases can confound results. It is therefore critical to explicitly lay out the sampling theory and assumptions relative to the question at hand. Are measures of the experimental population (e.g. age, relative health) likely to be normally distributed? Are measurement errors additive? These issues are often dealt with after data are collected, to the consternation of consulting statisticians asked to save the data at the end.

The next link in the chain is intact only if the objects of study are unambiguously defined and relate properly to the theory or empirical question. One of the central concerns of physics and chemistry during their early decades was to develop standard definitions and metrics for objects, substances, and measures. Where pure definitions were not possible, physical standards (e.g., kg standard, meter stick standard) were constructed. In ecology, this is not always done or potentially even possible. Let's say that we have a theory about trees (concerning evolution or life history, for example), and we wish to test it. If we go out to the field, how do we delineate our study population and separate trees from shrubs? Shrubs can be large and trees can be dwarfed in some circumstances. Things that are clearly trees can have multiple trunks and thus be shrub-like. Is bamboo a tree? We can unambiguously identify electricity and magnetism, but what about biodiversity or health or intelligence? We have no universally agreed upon operational definitions for these concepts. Such vague terminology leads to subjective and often undocumented decisions in each study that may produce ambiguous scientific findings, both empirically and in terms of any test of a theory. Ambiguous terminology can lead to debates that go on for decades (e.g., [10]).

The next link in the chain is intact only if what is measured is what was intended. This may seem obvious, but is not so simple in practice. Various metrics may exist for some concept (e.g., diversity [11]), but researchers often do not provide a clear statement as to which one is "really" what they mean by the term (e.g. songbird species richness may or may not reflect "true" biodiversity, which is practically unmeasurable). Studies of ecosystem function often face a boundary problem. It is simple to draw a line on a map around a study area, but verifying that this map area is functionally integrated and characterizable is rarely done and may not even be possible [12].

Divergent metrics pose a unique type of problem. A divergent metric is one that varies with resolution. The classic case is a fractal object. If we ask about the measure of a fractal tree or the surface area of a fractal landscape, this is an ill-posed question because the measure increases without limit (for pure mathematical objects) or without practical limit (for physical objects) as it is examined (measured) in more detail. In this case we can only specify how the measure scales and make scale-specific estimates. For such metrics an empirical estimate does not represent what was intended (e.g., forest fragmentation) unless scale is specified. For example, Butler *et al.* [13] defined forest edge from remote sensing data but noted that coarsening the data to 90 m pixels resulted in almost no edge on their study landscape. The definition of "canopy gap" (e.g., [14]) is similarly scale dependent but is not usually evaluated with respect to scale. Other scale-dependent "objects" in ecology include home range, geographic range, canopy cover, ground cover, forest vs. woodland area, and so on.

It is next critical that effects are not confounded. In the physical sciences, confounding has often been minimized by creating experimental systems that are close to the ideal systems of mathematics. For example, ideal gases or pure substances may be experimentally approximated to evaluate physical properties or study chemical reactions. Acceleration due to gravity can be studied in an experimental vacuum, even if the vacuum is not absolute. Air resistance can then be evaluated by comparing to the vacuum case. Experiments may be shielded to reduce extraneous electromagnetic effects. In ecology this can be very hard to do. In field studies, unobserved processes can easily interfere with those being studied. For example, trees in a forest may appear to be unaffected by a drought because their roots reach the water table, which would not necessarily be observed in the typical study. It might be assumed that the males with the brightest plumage have the best genes, but bright feathers can also result from lack of disease, which may or may not indicate good genes [15]. Fortunately, many of these confounding effects can be statistically

controlled if honest and accurate measures or estimates of possible confounding factors are taken into account.

Another type of confounding occurs when the process of studying a problem interferes with the process being studied. For example, the psychological boost that results from being treated for an illness leads to the well-known placebo effect. The creation of genetically uniform (to reduce variation) white lab rats can produce unusual and extreme responses to toxins or carcinogens and increases their cancer prevalence due to homozygosity. Prolonged observation of primates in the field can change their behavior if the study animals are aware of the observers. The observer can directly interfere with the data when subjective estimates are made of quantities such as vegetative cover (e.g., [16]). When repeated live sampling of animals is conducted, some species become very trap-shy whereas others like the free meal and will seek out the traps.

The next step in maintaining an intact chain of logical inference is to be able to quantify error or deviation (Figure 2). In a well-defined measurement problem, such as weights of deer, the deviation from the mean is a measure suitable for statistical testing. In other cases, the investigator may be unaware that there is error involved. For example, areas of a map may be classified or delineated as belonging to various vegetation types. The delineation is not without error (e.g., [17, 18]), but it is not clear how to account for the error, which is thus often ignored, leading to inconclusive results [19] because the null of random effect is falsely assumed to be due to the treatment.

Once data are collected or an experiment performed, it is critical that the proper statistical analysis is carried out (in contrast to merely asking the right statistical question, discussed above). This involves proper treatment of outliers (not getting rid of them without a good reason, for example). A proper null expectation may need to be formulated. For example, in the absence of the effect being studied (say, competition), how many species of the same life form or same genera would be expected to coexist by chance alone? This may not be an unambiguous question and may itself require study. One must choose between frequentist (e.g., ANOVA) and Bayesian

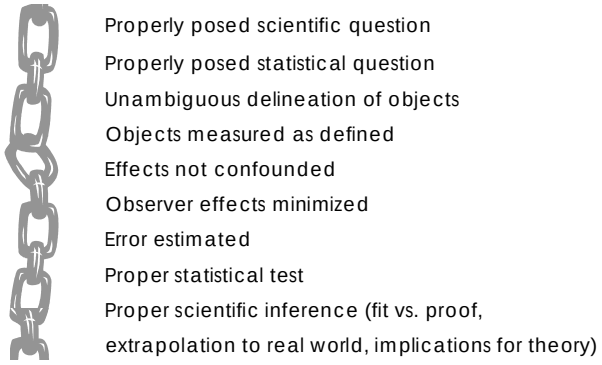

Figure 2. For either testing a theory or asking an empirical question, there is often a sequence of steps in the chain of inference. The study is only as strong as the weakest link.

(or other) approaches to the analysis, data may need to be transformed, proper statistical approaches must be used, assumptions of those approaches will need to be verified (e.g., normally distributed data, heteroskedasticity), along with other considerations (e.g., [20]).

The final link in the chain of reasoning is to make proper scientific inferences based on the results of the study. A study could provide modest support for a theory or for other empirical findings, thus reinforcing current orthodoxy. It could contradict existing theory as well. In the latter case, it is important to be quite clear about how strong the result is. A very weak negative result is not usually sufficient to overturn existing theory, and a very weak positive result can not be considered strong affirmation. Such results might best be qualified as either inconclusive or equivocal. It is also critical to be precise about the implications of the study. Statistical significance does not necessarily equate to biological significance. A very small effect in an ecological field study does not allow predictions to be made.

The difficulty with a long chain of reasoning is that the probability of success goes down as a power law function. For example, in Figure 2, which has 9 sequential reasoning steps, even if you are 95% sure that you have done each step right there is only a 63% chance of executing the sequence without an error that calls the entire result into question. Reviewers and graduate student advisors in fact often encounter such cases, where

a single fatal flaw invalidates an entire study. In the next section, methods are discussed to help reduce the chance of faulty reasoning during scientific discovery in ecology.

Strengthening the chain of logic

It is not my intention to be negative. If we understand the problem, it is possible to overcome the stumbling blocks I have identified. In fact, looking back on the history of science, many of the greatest advances have been responses to the very kinds of difficulties I have been describing. The transition from natural philosophy to modern science was marked by the development of the experimental method which helps reduce subjectivity and confounding. For example, Galileo was successful because instead of asking "why" objects fall to earth, he asked and then quantified "how" they did so. He even took steps to reduce the effects of confounding factors. The double-blind medical study was established as a standard method to overcome the interference of the physician with the response of the patient. Factorial experiments and analysis of variance were introduced to quantify error, clearly separate effects, and reduce subjectivity. Animal model based drug trials that are successful lead to clinical trials to help determine if the results are transferable to humans, reducing the uncertainty associated with non-human trials.

For dealing with issues related to theory (Figure 1) it would be helpful to focus a little attention on foundational issues. In physics, where phenomena can often be studied in isolation and objects (such as atoms) are identical, foundational issues have received significant attention. While not as simple in other fields, such formal treatment of theory is possible. For example, a formal null model of biotic community organization [21] is being mathematically elaborated [6] and deductions developed from it. Various techniques have been proposed for elaborating and clarifying theory (e.g., [7]). It would be helpful if ecology journals were a little more interested in such topics.

For performing empirical studies (Figure 2), particular fields have standard techniques for certain links in the chain, but not for all types of studies or all steps. New topics and new methods of collecting data (e.g., internet surveys, remote sensing) are particularly prone to gaps in the prevention of the failures of logic outlined in Figure 2. Just as surgeons (and patients) benefit from use of a checklist by the surgical team, it could be a useful practice to develop and use a discovery checklist. A first step in this direction is presented next.

Proceeding through the chain of logic (Figure 2), the first and obviously critical issue is the proper posing of a scientific question. There is not a formal method to be applied to this step. However, there are certain things we can check about the question being asked to see if it is well-posed. If it is a trans-scientific question, such as how much social value can be ascribed to a particular species or ecosystem, then clearly it cannot be answered by a scientific study. We should look for unique predictions from a theory (Figure 1a), rather than those also derivable from alternate theories or from null models. In ecology, an important null effect often relates to sampling method (e.g., [22, 23]) or scale. If an empirical question, it should not require more precision to answer than methods can deliver. I showed [24] as an example that while a grazing system might exhibit the characteristics of a cusp catastrophe stability manifold, field detection of the expected properties was unlikely due to constantly changing underlying conditions (e.g., rainfall, animal stocking). We can also ask whether predictions were properly deduced from the theory. These questions help us evaluate whether the scientific question can, in principle, be answered.

Translating a scientific question into a statistical one is not straightforward in ecology [25]. If a theory can make specific predictions of relationships or distributions rather than just more/less predictions, this makes for a more robust test. If the theory says (or we want to simply ask if) an intermediate level of disturbance enhances plant diversity (e.g., [26]) it is necessary to specify spatial and time scales, disturbance types and how they will be quantified, and how diversity will be sampled and quantified (at a minimum).

How do we ensure that the objects of study are unambiguously defined? At one level, if a question or theory invokes vague terms such as competition, health, or biodiversity, then there is an up-front ambiguity that makes it unclear

whether any actual measurements properly represent such terms. Age-adjusted excess mortality or body mass index would be more specific terms than health, for example, for a wildlife study. An example of an operational definition of a term is presented by Godsoe [10] who suggests how the species niche concept might be quantified. A characteristic of proper terminology is that units can be specified and the measurements related to a specific experimental, measurement, or sampling methodology, such as pan-evaporation in agrometeorology. The importance of standard metrics and measurements, such as weather instruments and heights/times for taking observations, cannot be over-emphasized.

The problem of measured objects that are not what was intended or defined is a little trickier. If there is a protocol for a measurement, such as the appropriate location and structure of a weather station plus time of day for observations, then compliance with this protocol can verify that the data are appropriate. If data values depend on the details of sampling (e.g., [22, 23, 27, 28, 29, 30]) then results are likely to be inconsistent between studies and care is needed to justify and perhaps standardize data collection or experimental methods.

The prevention of confounding is difficult and becomes more so in field studies or as time/space scales increase. It is helpful to be alert to past discussions of potentially confounding factors and to not ignore them. It can also be helpful to get outside comments on a study plan specifically on this issue before commencing the research.

The reduction of observer effects is so critical that certain studies such as drug trials are built around double-blind safeguards. Observer interference can come into play also in the selection of encountered data. For example, when multiple data sets exist for species abundance, the choice of which data to include or exclude (butterfly data? bird data? trees?) opens the chance for observer subjectivity to influence the outcome of the study. The same can occur when a literature review is written or a meta-analysis conducted and certain results or points of view are simply ignored. In these cases, journal reviewers should insist on completeness and documentation of data selection methods, though of course it is better if the scientist himself is alert to these issues.

The issue of error is one that is easy to neglect. Without being careful about map classification error, it is not possible to properly evaluate landscape changes over time [19, 31]. Computer models are subject to multiple types of errors, including parameter error, model structure errors, errors resulting from spatial discretization, numerical errors, and others. Error propagation methods are available for some but not all of these error types, but modelers can be reluctant to admit how wide the error bounds on their models actually are, and thus often do not show error estimates or confidence limits on model output. As noted by Berthouex and Brown [32], “A guiding principle of statistics is that any quantitative result should be reported with an estimate of its error.” Any time the output or result is simply a number with no way to estimate error, it is time to re-examine the analysis. This is good advice for reviewers as well.

Statistics has become an enormous field. To make it even more confusing, there are competing paradigms within statistics. In spite of this, there are things that can be done to prevent serious errors. Use of statistical packages is helpful, but it must remembered that the package knows nothing about potential outliers, sampling bias, and special circumstances, or whether you are doing the right analysis. Involvement of a statistician can be helpful if one can afford it. Reviews of the data analysis by colleagues can also help reduce errors of this type.

Finally, how do we more often make more valid scientific inferences based on the results of a study? In many studies, the results are modest. Some percent of variation is explained. The results agree with some past studies and disagree with others, for various reasons. Only rarely does a study by itself prove or disprove a theory, provide an adequate basis for managing an endangered species, or identify the ideal diet. At this stage what is needed is clear logic and equally clear language to state exactly what the results do and do not mean. Over-stating the significance of a study does not help science progress.

CONCLUSIONS

The discovery process is not simple. It is subject to the difficulty that human logic and powers of deduction are imperfect. Furthermore, nature does

not always agree that things which we think are logical and necessary actually are so. The categories we like to believe are obvious (ecosystem, diversity, life form) may not be discrete or even definable. There is no formula for the general case. Some simple types of scientific studies have been more or less formalized (e.g., agronomy experiments) but even here new techniques open up new opportunities to make mistakes. Because the entire chain of logic must be intact, it is not enough to do part of a study right. In this essay I have suggested some techniques for reducing error. The general idea is to ask oneself questions. Did I miss any confounding factors? Why are there outliers? Did I influence the outcome in any way? The second general method is to check and recheck. Redo the derivation and check the data for artifacts. On rereading, are your conclusions too grandiose? Finally, getting comments from colleagues is invaluable because they are not as in love with your ideas as you are. Perfection is not possible, but maybe the worst mistakes can be avoided.

For further reading on the discovery process, see Loehle, C. 2009, Becoming a Successful Scientist, Cambridge U. Press.

ACKNOWLEDGEMENTS

No outside funding was obtained for this study. Thanks to Sleep, D. and Verschuyl, J. for helpful comments.

REFERENCES

1. Whittaker, R. J. 2010, Ecology, 91, 2522.
2. Freedman, D. 2010, Wrong. Little Brown and Company, New York.
3. Fanelli, D. 2010, PLoS ONE 5:e10068, doi:10.1371/journal.pone.0010068.
4. Altman, D. G. 2010, JAMA, 287, 2765.
5. Altman, D. G., Goodman, S. N., and Schroter, S. 2010, JAMA, 287, 2817.
6. Allouche, O. and Kadmon, R. 2009, Ecol. Lett., 12, 1287.
7. Loehle, C. 1988, Oikos, 51, 97.
8. Loehle, C. 2011, Ecol. Compl., 8, 60.
9. Loehle, C. 2006, Ecology, 87, 2221.
10. Godsoe, W. 2010, Oikos, 119, 53.
11. Barrantes, G. and Sandoval, L. 2009, Revista de Biología Tropical, 57, 451.
12. Loehle, C. and Pechmann, J. H. K. 1988, Amer. Nat., 132, 884.
13. Butler, B. J., Swenson, J. J., and Alig, R. J. 2004, For. Ecol. Manage., 189, 363.
14. Busing, R. T. and White, P. S. 1997, Oikos, 78, 562.
15. Loehle, C. 1997, Ecol. Mod., 103, 231.
16. Bergstedt, J., Westerberg, L., and Milberg, P. 2009, Plant Ecol., 204, 271.
17. Castilla, G., Larkin, K., Linke, J., and Hay, G. J. 2009, Landscape Ecol., 24, 15.
18. Langford, W. T., Gerge, S. E., Dietterich, T. G., and Cohen, W. 2006, Ecosystems, 9, 474.
19. Linke, J., McDermid, G. J., Pape, A. D., McLane, A. J., Laskin, D. N., Hall-Beyer, M., and Franklin, S. E. 2009, Landscape Ecol., 24, 157.
20. Prairie, Y. T. and Bird, D. F. 1989, Oecologia, 81, 285.
21. Hubbell, S. P. 2001, The Unified Neutral Theory of Biodiversity and Biogeography, Princeton University Press, Princeton, NJ.
22. Dengler, J. 2008, Geobot, 43, 269.
23. Dengler, J. 2009, Ecol. Indic., doi:10.1016/j.ecolind.2009.02.002.
24. Loehle, C. 1985, Ecol. Mod., 27, 285.
25. Sleep, D. J. H., Drever, M. C., and Nudds, T. D. 2007, J. Wildl. Manage., 71, 2120.
26. Connell, J. H. 1978, Science, 199, 1302.
27. Burke, A. 2007, Afr. J. Ecol., 46, 788.
28. Dengler, J., Löbel, S., and Dolnik, C. 2009, J. Veg. Sci., 20, 754.
29. Heino, J. 2008, Boreal Environ. Res., 13, 359.
30. Rocchini, D. 2005, Spatial Sci., 50, 25.
31. Pancer-Koteja, E., Szwagrzyk, J., and Guzik, M. 2009, Plant Ecol., 205, 139.
32. Berthouex, P. M. and Brown, L. C. 1991, Statistics for Environmental Engineers, CRC Press LLC, Boca Raton, FL, USA.