CHAPTER 16: IS SCIENCE LOGICAL?

16

An earlier chapter revealed that all models are false. This chapter reveals another blemish on the face of science -- how we decide the fate of models is arbitrary.

SECTION 1

Introduction

Once the data have been gathered according to the ideal data template, a challenging phase of scientific inquiry is faced next: what do the data mean? That is, what models can be rejected? The fact that data have been gathered to ensure accuracy does not guarantee that they will be particularly useful for the goals of the study. They need to be sufficiently accurate, but they also need to address the models at hand. Even assuming that the data DO address the models at hand, how do we decide when to abandon one model and move on to a new one? A surprising feature of the scientific method is that this aspect is arbitrary -- not everyone uses the same criteria, and the criteria of one person change from situation to situation. Thus, two objective scientists can evaluate the same data yet come away supporting different models.

The Language of Evaluation

No one can prove that a model is correct, but we nonetheless want to use good models and avoid bad ones. Yet there are many different degrees of accepting/rejecting a model. A modest terminology surrounds the evaluation of models. The most extreme evaluations are

refute: the data are not compatible with a model and force us to reject it

support: the data are not only compatible with a model but refute many of the alternatives and lead us to think that it is possibly useful.

A model cannot be supported unless the data would (had they turned out different in certain ways) have refuted the model. That is, "support" means that the model could have failed the test but didn't. Refuting a model is an absolute classification -- there is no returning to reconsider a refuted model (for those data). Supporting a model, however, is a reversible designation -- additional data may ultimately refute it.

A lesser degree of compatibility between data and a model is

consistent: the data don't refute the model

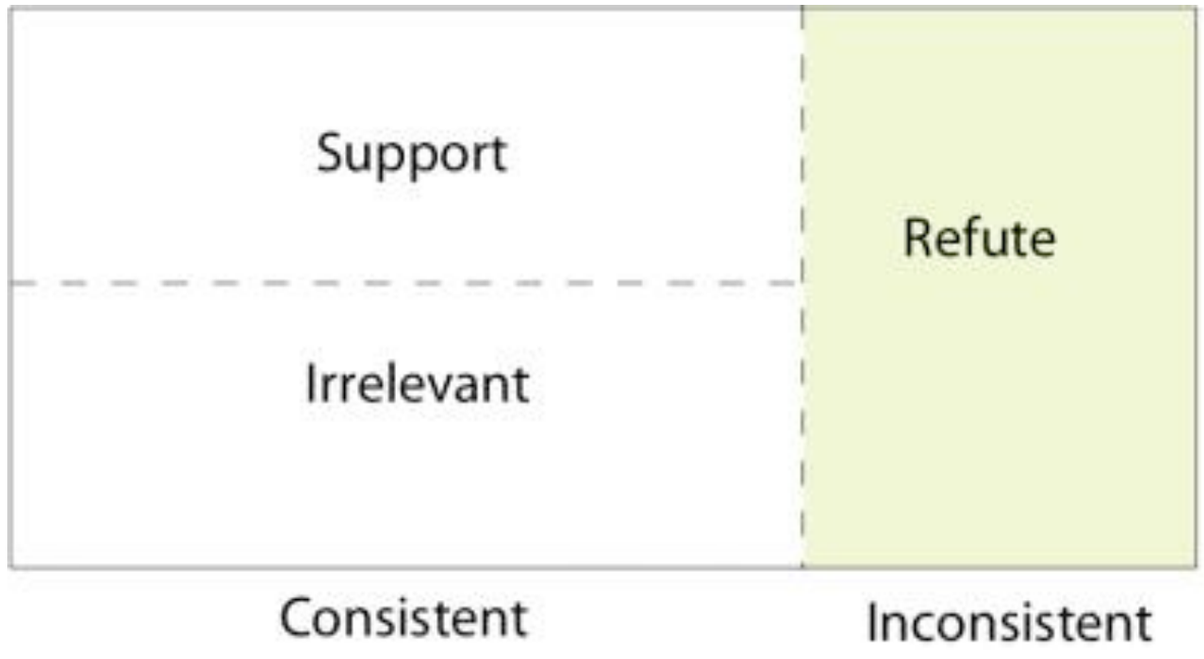
Data that support a model are consistent with it, but data may also be consistent without giving much confidence in it.

At the furthest extreme, data may be consistent with a model but be

irrelevant: the data do not address the model in any way that could have caused us to reject it

A simple picture representing the relationships of these different concepts is shown below. Support is surrounded with a dashed line because it is a fuzzy concept in some cases.

A simple picture representing the relationships of these different concepts is shown below. Support is surrounded with a dashed line because it is a fuzzy concept in some cases.



The Right Stuff -- Which Models Do You Start With?

The notion of progress in the scientific process involves rejecting models and replacing them with new models. How rapidly this progression occurs depends both on goals and on the models chosen at each step. Obviously, if one is lucky enough to start with a model that is close to the "truth" (as concerns the goal), then little progress will occur, because there is simply little room for improvement. Alternatively, starting with a bad model may ensure lots of "progress" because there is so much room for improvement.

There are different philosophies about what kinds of models to choose initially. One approach is the null model approach. A null model is a default model -- one chosen just to have an obvious starting point. A null model might be one that we think is commonly true, or might be a model that we don't think is true but we use anyway, merely to demonstrate that we can reject something. For example, in choosing a model for the probability of Heads in a coin flip, most of us suspect that this probability is close to 1/2, so we would start with P=1/2 as a null model. Or if we were investigating whether alcohol impairs coordination, most of us realize that it does, but we would nonetheless start with a null model that alcohol does not impair coordination, just to show that this simple idea is wrong. There are thus several different reasons for starting with null models, and the choice of the which null model to use will depend on those reasons.

In some cases, people start with one or more specific models. These may or may not be contrasted with a null model. For example, if a particular theory proposed that two thirds of all cancers caused by electromagnetic fields should be leukemia, then we would want to test that model of 2/3 specifically (and we might not even know what an appropriate null model should be). Or someone might propose that a particular bacterium is the cause for "sick building syndrome" (the phenomenon in which certain buildings make lots of the inhabitants feel sick), and we would test that model specifically by looking for that microbe in the ventilation ducts of SBS buildings.

The choice of models for testing is thus arbitrary to a large degree.

No data can reject all alternative models (you can't prove a negative)

There is no limit to how many models are relevant toward a particular goal. Typically, only a few models are considered in a test, but there are countless others that might be considered. For any goal, there will be infinitely many possible models that are relevant. For example, in the simple case of the probability of Heads in a coin flip, we can choose as a model any single value from the infinite set of values between 0 and 1; there is not only an infinity of such models, but the infinity is so "large" that it is considered uncoutable. But we could also choose an interval of values in this range -- the probability of Heads lies between 0.467 and 0.522. We could even choose refined models that offered details about a succession of coin flips ("2 Heads will be followed by 1 Tail" and so on). With models for the effect of radiation on cancer, there are infinitely many models which assume the relation is a straight line, infinitely many assuming a curved relationship, and so on.

In testing a model, therefore, the best we can hope for is to reject some of the models. Invariably, no matter how good of a test we conduct, there will be countless others remaining after the test. Since it is impossible even to list all of the models, so the results of a test are usually stated in terms of just the few models considered up front (which may be as few as one model -- the null model).

This inability to reject all possible alternatives is the main reason we can never prove that a model is correct -- there are always many models consistent with any set of results, so we have no basis for claiming that a particular one is correct. Thus, in a coin flip, we can never prove that the probability of Heads is exactly 1/2, because no matter how many times the coin is flipped, there will always be a range of values consistent with the data. There is an infinite number of values within that range, any of which could be true. A special case of this is the statement that we "cannot prove a negative, " which is to say that we cannot prove that a phenomenon absolutely fails to exist. In testing astrology predictions, we can never prove that there is NOTHING to them, because there will always be a range of outcomes consistent with any test, and that range will include at least a tiny bit of nonrandom prediction. In testing whether sugar in a diet influences cancer rates, we can never prove that sugar has no effect, because the data will always be consistent with a range of cancer levels. Hence a reason to rely on null models.

How Disagreement Can Persist After a Test

One would think that objective people should be able to achieve consensus with one another once the relevant models have been tested. Yet differences of opinion abound in science. These differences stem from the points described above. First, not everyone starts with the same set of models. Some people want desperately to think that trace amounts of pesticides in food are harmful; others want to think that the traces are harmless. Any particular study may fail to discriminate between two alternative models, and the proponents of each model will feel accordingly bolstered each time that their model survives a test. So a test that fails to resolve between two models can actually increase the acrimony in a debate. Furthermore, in the case of trace pesticide levels, there will always be some low level of pesticide that cannot be shown to cause harm (even if it does), simply because of intrinsic limitations of the scientific method (see the subsequent chapter "Intrinsic Difficulties" in Section V).

Criteria for rejection

Each of us personally makes many decisions daily about what to believe or accept and what to reject. The sales pitch promising high returns on your investment is obviously to be questioned. We are used to campaign promises being forgotten on the night of the election. But if our physician tells us something, or we read about a government report of a decrease in AIDS deaths, we are inclined to believe it. (In contrast, the public has come to mistrust many government statistics, especially rosy forecasts about the economy, and war casualty reports during wartime.) This is true for all of us -- we trust some sources more than others and accept some things at face value. But for something like the result of a research study or a government-approved release, somewhere back along the information hierarchy, someone has made a decision about what is true enough to be accepted and what is not. That is, someone has made a decision to accept some models and reject others.

Statistics:

The most common protocol for making acceptance/rejection decisions about a model is statistics. In some cases, results are so clear that the accepted and rejected models are obvious. But far more commonly, mathematical rigor is required to make these decisions. For example, if you want to know if a drug is helping to cure ulcers, and the tests show that 15% are cured with the drug versus 12% cured without the drug, any benefit of the drug won't be obvious. Statistical tests are mathematical tools that tell us how often a set of data is expected by chance under a particular model. (A statistical model is actually a mathematical model built on the assumptions of the abstract model we are testing, so it involves layers upon layers of models.) If the results would be expected under the model infrequently, we reject it; otherwise we accept it (which doesn't mean that it has been "proven"). Ironically, we know in advance that the statistical model is false. The question is, however, whether it can be refuted.

Scientists have agreed by "convention" what criteria to use in rejecting/accepting models. Commonly, if a set of observations (data) would be expected to occur under a particular model only 1/20 times or less often (5%), we reject the model. What this means is that, if the model is true, we will make a mistake in rejecting it 1 in 20 times. So "rejection" is imperfect. Because scientists often test many things, and they don't like to be wrong about it, they are sometimes conservative and don't get excited about a rejection unless the data would be expected less than 1 in 100 times under the model.

These criteria for rejection and acceptance are arbitrary. Yet science is often portrayed as objective and absolute. Furthermore, scientists often have difficulty relating to the public willingness to accept many things for which there is little support, when in fact, their own criteria for acceptance are subjective. There is nothing magic about using a 5% criterion for rejection. As an institution, science is fairly unwilling to adopt new ideas and abandon old ones (reluctant to reject the "null" model) -- the burden of proof is on the challenger, so to speak. But there are many facets of life for which we don't need to be so discriminating and thus don't need to wait until the 5% threshold is reached. We can be willing to try a cheap, new over-the-counter medicine without 95% confidence in its worth because the cost of being wrong is slight and the benefit is great. Conversely, when it comes to designing airline safety, we want to be extremely cautious and conservative about trying new things -- we will not tolerate a 1 in a million increased risk of a crash. Many people play the lottery occasionally; the chance of winning is infinitesimal, so it is a poor investment. Yet, the cost of a few tickets is trivial, and the hope of winning is entertainment value that it can actually make sense for people to play. After all, we pay $6 to see a movie and have no chance of recovering any money. The criteria for acceptance of a model, at least in the short run, thus depend on the cost of being wrong. Where these costs are small, we can afford to set less stringent standards than where the costs are high.

Repeated Successes:

Statistical tests are substitutes for what really counts -- whether something works time and again. We no longer need a statistical model to convince ourselves that the Sabin polio vaccine works, because it has been tried with success on billions of people. The major theories in physics, chemistry, and biology (including evolution) have held up to many different tests. Each time we conduct a specific test, we may use statistics to tell us if the results for that study are significant, but in the long run a model had better hold up time and again, or we will abandon it.

Unscrupulous Exploitation and the Limits of Evaluation

Unfortunately, however, we can't wait for dozens of trials on everything, and we must rely on statistics and other short-term evaluation methods to show us what to accept. This reliance on short-term evaluations provides a loop-hole that can be exploited to gain acceptance of a model that should be rejected. Businesses can run "scams" (legal or illegal) that take full advantage of the time lag between initial marketing of a product and widespread evaluation of its success -- with good or lucky marketing (based on hearsay, for example), a product can sell millions before it is shown to be useless. In the meantime, the consumer wastes money and time. The market of homeopathic "medicines" ("natural remedies") is full of products with suggested benefits for which there is no reliable evidence; the FDA ensures that these products are not advertised with claims of health benefits, but many counter-culture (and even mainstream) magazines provide articles touting various homeopathies. For products that do seek FDA approval, careful selection of statistical tests can obscure mildly negative effects of a health product, and careful design of clinical trials can avoid certain types of outcomes that would be detrimental to gaining approval of the product (the FDA estimates that only 1 in 5 newly approved drugs constitute real advances). For a product used by only a small percentage of the population, it may be impractical or impossible to accumulate enough data to provide long term evaluations of beneficial and detrimental effects.