# Hypothesis Testing and the Nature of Skeptical Investigations

*Testing hypotheses is the fundamental activity of scientists and skeptics. Now, a debate within the scientific community may be about to radically change the way we think about and conduct tests in the sciences.*

MASSIMO PIGLIUCCI

*"The test of statistical significance in research may be taken as an instance of a kind of essential mindlessness in the conduct of research." (Bakan 1966)*

*"Statistical significance is quite different from scientific significance." (Cox 1977)*

In 1987 a study by Schmidt, Jahn, and Radin proved that it is possible for the human mind to act on matter at a distance. More specifically, the researchers concluded that their data were consistent with the idea that a person could deviate—by acting at a distance—a significant number of particles arriving at a "quantum gate" and that the probability of this happening by chance was about p=0.0003. Before anybody rushes to his or her word

processors and starts typing letters of outrage to the SKEPTICAL INQUIRER, let me qualify the above statement. What I should have said is that we should accept the above-mentioned conclusion *if* we apply to that data set the standard statistical methods of hypothesis testing that scientists and skeptics learn from Statistics 101. Instead, if we conduct a much more sensible statistical analysis (known as the Bayesian approach, for the curious) we would conclude that the hypothesis that the results were due to chance was many more times (53,263,000 times to be exact) more probable than the hypothesis that they really had witnessed an example of psychokinesis.

**There is a slow revolution afoot in science that might ultimately radically alter the way scientists and skeptics think about the world and conduct their all-important business of testing claims of the normal and paranormal.**

This is important because there is a slow revolution afoot in science (it has been going on for several decades) that might ultimately radically alter the way scientists and skeptics think about the world and conduct their all-important business of testing claims of the normal and paranormal. In order to understand this change let us start with a brief summary of Statistics 101 and its relevance to skeptical research.

## Null Hypotheses and P-Values

According to the standard method of statistical inference, an investigator will formulate a research question in terms of a "null" (or default) hypothesis and an alternative outcome. For example: in an experiment on telepathy, one wishes to test if in fact the subject has the ability to read another person's mind. The null hypothesis is that he doesn't, the alternative hypothesis is that he does. The investigator then sets up an experiment in which, say, twenty-five cards with five different symbols are presented to a person who acts as "transmitter" of telepathic information. The "receiver" has to guess which symbol comes up every time. Since we know that there is a certain probability (namely, one out of five) to guess the right card by chance (i.e., without any telepathic ability), the expectation for the null hypothesis is that, on average (i.e., over many trials), the allegedly telepathic subject will score only one-fifth of the times. Any *significant* deviation from that value (in the excess direction) will lead to the rejection of the null hypothesis and to the acceptance of its alternative (i.e., the subject really is telepathic).

One of the key words just mentioned is that the deviation has to be significant, but according to what criterion? Standard statistical theory tells us that *if* the distribution of the scores follows a particular distribution (say, a bell curve), then one can calculate a simple statistic (known as a *p-value*) that tells us what is the probability that a score like the one observed *or more extreme* (i.e., more favorable to the telepathy hypothesis) would be observed in an infinite run of the experiment. If that probability is lower than a preset threshold (usually 0.05, i.e., 5%), then we conclude that the hypothesis of a chance result has to be rejected and its complement (telepathy) accepted.

## The Problem with Null Hypotheses

There are two fundamental problems with the above scenario, one concerning the idea of a null hypothesis, the other regarding the calculation of p-values. Let's start with the first one. In reality, we are not interested in rejecting the null hypothesis at all. Statistically, the null is interpreted as the situation in which the outcome of the experiment is exactly the one expected when chance only is acting (in our case, exactly one-fifth positives). But scientifically, we are interested in what one could think of as an "extended" null hypothesis (figure 1). We wish to know if the actual outcome is far enough from the one predicted by chance to convince us that something is really going on.

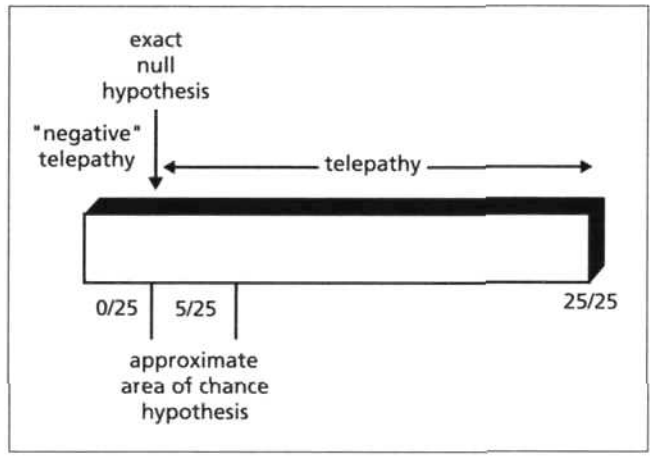


Figure 1. The structure of hypothesis testing in an hypothetical experiment on telepathy carried out with twenty-five cards and five symbols. What one is really interested in is not the rejection of the exact null hypothesis (which is what standard statistical tests do), but in seeing if the outcome of the experiment is far enough from that point to seriously raise the possibility of a systematic effect (perhaps due to real telepathy).

This is an important difference because in any real experiment there will always be some small but systematic deviation from the exact null. Some factor that the experimenter has not taken into account (other than chance) will likely cause small deviations from the expected result, especially if for some reason it is not possible to conduct a completely randomized and double-blind experiment. For example, cards may not have been shuffled well enough, or—if the input is generated by a computer—the random numbers used to produce the cards to present to the subject may not actually have been truly random, and so on. What we really want to know, in other words, is if the actual result is far enough from the range of outcomes

*Massimo Pigliucci is associate professor of ecology and evolutionary biology at the University of Tennessee and author of* Denying Evolution: Creationism, Scientism, and the Nature of Science. *His essays can be found at www.rationallyspeaking.org.*

expected by chance to raise the possibility of a telepathic phenomenon.

Furthermore, it should be clear that the two hypotheses we are considering (the null and the alternative) are actually not the only two interesting scientific possibilities. While it is true that either the outcome of the experiment is close to one-fifth or it isn't (factual result), the possible interpretations (scientific result) of either outcome may be varied. Consider some of the alternatives: there could be telepathy, but the effect is too weak to detect with the small number of runs typically carried out in these experiments (i.e., failing to reject the null hypothesis is not the same thing as accepting it!); or, a large result on the right side of figure 1 could be explained by the fact that—despite the experimenters' precautions—the subject cheated (so rejection of the null hypothesis does not automatically entail the acceptance of its complement). And so on. One could conceive of a series of different hypotheses competing for the most sensible explanation of the data at hand.

### The Problem with P-Values

The second problem with the standard statistical approach to scientific inference is the meaning of p-values. As I said earlier, they are an indication of the probability (given certain assumptions) of obtaining the same result or a more extreme one in an infinite repetition of runs. But surely this is not what we want to know. We wish to know if the observed outcome of *one particular experiment* is far enough from the null zone to raise the possibility of telepathy. This is not the same as asking what is the probability of getting a result equal to the null expectation or more extreme.

Moreover, what does it mean to say that the p-value gives us a probability valid over an infinite (or sufficiently long) series of runs? This raises serious philosophical as well as practical problems: we cannot repeat an experiment but a small number of times (sometimes only once), let alone pretend to know what would happen with an infinite series.

Finally, as I said earlier, if there is any systematic effect that we did not explicitly consider, no matter how small, this will result in a deviation from the strict null hypothesis. In that case, p-values will become smaller and smaller as a function of the sample size (the number of observations), so that for a large enough sample, anything will become statistically significant. This is exactly what happened with the psychokinesis experiment mentioned at the beginning: with 104,900,000 observations (one per particle), any slight asymmetry in the experimental conditions (for example in the building of the apparatus to count the particles) will yield false positives.

### What to Do?: The Theory

Given all these problems, what is an honest skeptic (or scientist) to do? What many statisticians and scientists have suggested for decades: Throw away the conceptual framework of standard statistics and use one of several alternative ways of thinking.

The first thing to realize is that philosophically speaking we are simply not interested in what standard statistics tells us: we do not want to know the probability of the observed data given a particular hypothesis (the null). What we wish to know are the likelihoods of different hypotheses given the observed data, that is, the exact opposite of the normal approach.

In order to get what we really want, then, we need to start by explicitly considering several (not just one or two) hypotheses and find ways to assess their relative "fitness" when compared to the data. In the case of the psychokinesis experiment, possibilities include: there really was a mind-over-matter phenomenon; the results were due to chance; there was cheating on the part of the subject; there was a small but systematic defect in the experimental apparatus that generated the particle stream; there was an error in the method of counting particles; or none of the above (the latter is a catch-all category that is always necessary in science—sometimes an experiment is simply inconclusive).

Our task then is to find ways to quantify the likelihood of these different hypotheses and decide which one(s) appear more probable given the available information. If more than one (or none) seem to win the game after this round, we obviously need to generate more data that can discriminate among the remaining competitors (or go back to the drawing board and propose new alternatives). And so on. The game of science continues.

### What to Do?: The Practice

In practice, comparing the fitness of different hypotheses is

## Why Intelligent Design Is Not Science

Another consequence of the view of hypothesis testing presented in this article is that the so-called "Intelligent Design" (ID) theory is not science, but rather an argument from ignorance. William Dembski, one of the leading proponents of ID, has claimed that what he calls the "design inference" can be made after one successively eliminates simpler hypotheses. His "explanatory filter" starts out by considering the hypothesis that, say, a complex biological structure is the result of a simple physical law (such as the law of planetary motion). If not, it asks if it could have been assembled by chance (which can yield more complicated patterns than regular laws). If the answer is again no, Dembski concludes that one is justified in inferring the action of a designer (of unknown origin and characteristics, of course).

Besides the fact that the explanatory filter does not consider several other possibilities (e.g., chaotic phenomena, emergent properties due to nonlinear interactions, etc.), it should be clear that the filter itself suffers from the same problem of the standard approach in statistics: it simply does not reflect the way science works. The design hypothesis is assumed to win by elimination, just like the opposition between null and alternative hypotheses. But we have seen that in reality scientists consider several alternative hypotheses *simultaneously* (not in sequence) and weigh their positive merits against each other. Nobody has the luxury of winning by default.

What, then, if anything, does ID represent? It is equivalent to our "catch-all" category of "none of the above." That is, ID is an unnecessarily fancy way of saying "I don't know."

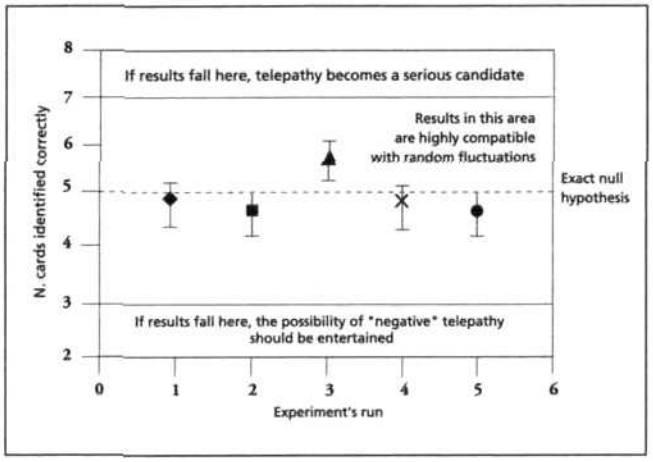


Figure 2. Hypothetical results (obtained by computer simulation) of five runs of an experiment on telepathy showing how a graphical analysis of means and standard errors (vertical bars) allows the experimenter to reach conclusions about alternative hypotheses. Notice that one of the five runs (n. 3) yielded results that are significantly different from the exact null according to the standard statistical framework. Yet there is no telepathy going on between my computer and me, you can be assured of that.

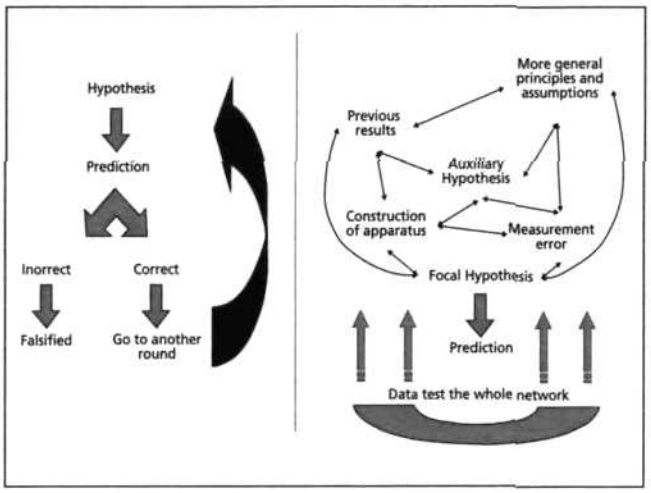


Figure 3. Two models of hypothesis testing: on the left, Popper's "naïve" falsificationism, where an hypothesis makes falsifiable predictions that are directly tested by data; on the right, a complex web of knowledge in which assumptions, hypotheses, and reliability of measurements are simultaneously tested during an ongoing process of research.

## On Falsifiability

Skeptics are often fond of invoking the concept of falsifiability, introduced by philosopher Karl Popper, to get rid of pseudoscientific beliefs. The idea is that if a theory or claim cannot possibly be refuted by any empirical evidence, it does not advance our knowledge of the world (though that doesn't necessarily mean that it is false). For example, claiming that an alleged paranormal phenomenon does not occur under controlled conditions because the very presence of a skeptic neutralizes it via some sort of negative vibration makes the claim itself unfalsifiable: no piece of empirical evidence could possibly show it false (figure 3, right).

While it is true that an unfalsifiable statement is of little use, one needs to be careful in light of our discussion of hypothesis testing in science. As I mention in this article, scientists do not reject a hypothesis just because it has failed one (or even a few) tests. The reason for this is that there may be alternative explanations for the failure, such as a faulty apparatus, or the simultaneous action of other factors not explicitly considered in the experimental design.

We can therefore think of science as testing not just a specific hypothesis at a time, but also the entire web of assumptions, corollaries, and additional hypotheses that are connected to the one that is the focus of our investigation (figure 3, right). Only if the test fails and we can reasonably exclude that other factors were at play, can we then say that the hypothesis has not withstood the test of the data. However, since there is an infinity of factors that could possibly be involved in any particular investigation, our conclusions (positive or negative) will only be probabilistic. Strictly speaking, we can never reject a hypothesis (*pace* Popper), only determine that its likelihood of being true has become vanishingly small.

actually not that difficult, although I cannot get into the technical details here (but see some of the references at the end if you wish to try). Mostly, there are two things to do: visualize the data in informative ways, and calculate simple statistics (called likelihood ratios) that quantify the probability of a given hypothesis to be a better match for the data than another one(s). (A more complex approach is to use a full-fledged Bayesian framework, but that is beyond the scope of this article; again, see references.)

An example of visualization of data is shown in figure 2. These are hypothetical data from a telepathic experiment with our familiar set of twenty-five cards with five symbols (the data were obtained by computer simulation). It is clear from the plot of means and their standard errors that the most likely of three hypotheses being considered is in fact that the results are due to random fluctuations (indeed they were, since that's how I generated them in the computer!). Had we seen one or more of the runs with a mean in the "possible telepathy" zone (and a standard error that clearly put that mean away from the boundary with the chance events), we would *not* necessarily have concluded in favor of telepathy: there are still a series of alternative hypotheses (such as cheating, or a defect in the measurement apparatus or—in the case of my simulation—that the computer did not truly generate random numbers with a mean of five) that would need to be considered before conferring a high likelihood to the telepathy hypothesis.

A common way to quantify likelihoods of different hypotheses given the data is to calculate what is appropriately called a series of likelihood ratios comparing all pairs of hypotheses. A likelihood ratio is simple to obtain (you only need a pocket calculator) once one has carried out standard statistical analyses such as t-tests, analyses of variance, or regression analyses. For the really curious (and somewhat

*Hypothesis Testing and the Nature of Skeptical Investigations*

*Continued on page 48*

They would like very much to ignore all evidence to the contrary, and they may become defensive if their selective interpretation of life comes under closer scrutiny.

Needless to say, such scrutiny often does reveal a number of pessimistic details about their allegedly rosy view of life. Consider the fact that many theists who consider themselves quite optimistic believe that a large percentage of humanity is essentially wicked, and will face eternal damnation for their sins. And then there is the content of the e-mails that many so-called optimists believe unquestioningly. Many of these e-mails are urban legends spreading false and dangerous misinformation. According to these urban legends, for example, Taiwanese people snack on human embryos and Congress is a moral cesspool filled with criminals and public leeches. These aren't exactly cheery notions, but someone who abandons critical thinking has no defense against them. They unthinkingly accept them, and consider themselves blissful optimists all the while. Such is life without skepticism.

I hope I have shown that cynicism and skepticism have nothing to do with each other. A skeptic has no obligation to unfairly judge others, and does not close his mind to contrary opinions. Most important, skepticism need not threaten human hopes, religious or otherwise.

Cynicism is an attitude about life, while skepticism is a method for uncovering facts about life. Anyone may incorporate those facts into a general philosophy of life as he or she best sees fit, but we'll all be better off if those facts have firm foundations. Skepticism is the best means of ensuring the reliability of our knowledge, and I don't think it's too optimistic to hope for its wider application in the realm of ideas.

## Note

1. See, for example, Joseph L. Daleiden's book *The Final Superstition: A Critical Evaluation of the Judeo-Christian Legacy* (Buffalo: Prometheus, 1994). As the title implies, Daleiden views theism as the last major obstacle in the path of rationalism.

## References


Hawking, Stephen. 1989. *A Brief History of Time.* New York: Bantam Books.

James, William. [1897] 1985. "The Will to Believe." Reprinted in James, William. *The Will to Believe and Human Immortality.* New York: Dover.

Morrison, Phillip. 1995. *Nothing is Too Wonderful to Be True.* Masters of Modern Physics, Volume 11. Woodbury, New York: American Institute of Physics.

Piatelli-Palmarini, Massimo. 1996. *Inevitable Illusions: How Mistakes of Reason Rule Our Minds.* New York: Wiley & Sons.


---

## *Hypothesis Testing and the Nature of Skeptical Investigations*

*From page 30*

statistically savvy) skeptics, this is the formula (I promise, the only one in the article):

$$\lambda = \left( \frac{1 - R_1^2}{1 - R_2^2} \right)^{n/2}$$

where $\lambda$ is the likelihood of hypothesis 2 when compared to hypothesis 1, $R$ is the amount of information (variance) in the data explained by each model (hypothesis), and $n$ is the sample size (number of observations). Simple and powerful, it grows on you.

### A New Philosophy of Science

So, next time you run an experiment, think in terms of several alternative hypotheses and analyze the data accordingly. This is a major shift in attitude for scientists and skeptics, and one that will become more common as more people are exposed to it. It will take decades for the textbook-entrenched standard approach to dissipate; science and scientists can be very conservative in their habits of mind. Yet, it simply makes sense to think of progress in science not as the rejection of uninformative hypotheses, but as a continuous competition among different solutions to a problem. This is the way scientists actually think, despite their frequent use of statistical methods that do not reflect such thinking. Eventually, some hypotheses will be discarded—not necessarily because we know for sure that they are wrong, but because they failed the test so many times that their likelihood is rapidly approaching zero. (That's why it is ludicrous for somebody believing in pseudoscience to pretend that skeptics actually test *every* potential case of telepathy or psychokinesis.)

By the same token, the likelihood of certain hypotheses will become so high, because of their continuously successful fit to the data, that we will regard them as true beyond reasonable doubt, at least for the time being.

One important lesson for the skeptic is that within this philosophical (in fact, Bayesian) framework, the likelihood of an hypothesis is never exactly zero or one (100%), so that one always needs to keep an open mind about things: you never know if new data will knock down the likelihood of a favorite hypothesis, or suddenly raise the probability of an almost discarded one. On the other hand, we are free to bet a beer on the (provisional) conclusion that there is no such thing as telepathy.

### Additional Readings


*Chamberlain, T.C. 1897. The method of multiple working hypotheses. Science* 15: 92–96.

Cohen, J. 1994. The earth is round ($p<0.05$). *American Psychologist* 49: 997–1003.

Dixon, P., and T. O'Reilly. 1999. Scientific versus statistical inference. *Canadian Journal of Experimental Psychology* 53: 133–149.

Jefferys, W.H., and J.O. Berger. 1992. Sharpening Ockam's Razor on a Bayesian strop. *American Scientist* 80: 64–72.

Loftus, G.R. 1993. A picture is worth a thousand p values: On the irrelevance of hypothesis testing in the microcomputer age. *Behavior Research Methods, Instruments & Computers* 25: 250–256.

Tukey, J. W. 1969. Analyzing data: sanctification or detective work? *American Psychologist* 24: 83–91.