What Constitutes Science and Scientific Evidence: Roles of Null Hypothesis Testing

Educational and Psychological
Measurement

2017, Vol. 77(3) 475–488

© The Author(s) 2016

Reprints and permissions:

sagepub.com/journalsPermissions.nav

DOI: 10.1177/0013164416667978

journals.sagepub.com/home/epm

Mark Chang^{1,2}

Abstract

We briefly discuss the philosophical basis of science, causality, and scientific evidence, by introducing the hidden but most fundamental principle of science: the similarity principle. The principle's use in scientific discovery is illustrated with Simpson's paradox and other examples. In discussing the value of null hypothesis statistical testing, the controversies in multiple regression, and multiplicity issues in statistics, we describe how these difficult issues should be handled based on our interpretation of the similarity principle.

Keywords

hypothesis testing, similarity principle, multiplicity, Simpson's paradox, causality, and false discovery rate

What Is a Science?

To put it simply, a science is mainly a set of investigations of so-called scientific laws that characterize causal relationships within a field of study. When the relationships address the underlying problems in physics, society, or psychology, we have different sciences that sometimes entail different methodologies.

Any proposed law of science or hypothesis must have testable predictions. This implies that (1) science is a study of (ideally) repeated events and (2) an essential

¹Veristat, Southborough, MA, USA

²Boston University, Boston, MA, USA

Corresponding Author:

Mark Chang, Veristat, 118 Turnpike Road, Southborough, MA 01772, USA.

Email: mark.chang@veristat.com

role of science is to make predictions, so that we can act early and wisely and verify what we believe to be true. The reason we study history and experimentation is to “discover” the law that can be used to predict the future when the cause repeats. However, I like to think (even if I would not try to prove it) that there is a single history, and everything is unique over its course. When we slice history into pieces and judge two pieces to be the same, we intentionally or unintentionally ignore their differences; at least we ignore the different “neighboring pieces” in the chain of history. Such slicing and grouping together of similar pieces artificially creates repetitions of events that we consider the same, and allows us to “discover” natural laws or causalities. In this sense, a natural law is first invention (grouping the similar pieces), and then discovery (the pattern discovery).

However, the grouping of similar things is often done implicitly and subjectively, which inevitably creates intangible controversies in scientific inference. For example, if you are a tall man, from which of the following four groups do you infer your characteristics (e.g., life expectancy, the likelihood of being successful in career, or the chance of having cancer): all men, tall people, tall men, or the entire population? Likewise, considering that I emigrated from China to the United States in 1989 and 10 year later became a U.S. citizen, should I predict my longevity based on my membership in the U.S. population or the Chinese male population, or something else?

Similarity Principle

The notion behind prediction is causality, whereas the notion of causality is the *similarity principle* (Chang, 2012, 2014): Similar conditions will likely result in the same or similar outcomes. The principle can be stated differently for better understanding depending on the problem. For example, we might state, “similar things will behave similarly.” In my view, the similarity principle is the most fundamental belief we hold in scientific communities, but is not explicitly expressed.

A scientific law states that the certain (same) condition will lead to the certain (same) consequence. But no two things/situations are exactly the same. “Same” actually means similar. When we say “the same,” we ignore the differences that might be hidden and contradict the scientific law that is held. Once such a contradiction is uncovered, an exception to the law is observed which may require some modification of the law. In other words, at some point in the future, any scientific law will be proved false or just an approximation.

As to how the similarity principle is used in scientific discovery, or even in daily life, let’s look at a few examples.

When investigating human or animal populations one often divides the study population into different groups based on age, gender, race, and the like. This is done because we think subjects in each group are similar in terms of the causal relationship we are investigating. But each subject is unique, thus such grouping is subjective. In psychological settings, we may study a person’s behavior using the knowledge we

have from similar subjects whom we have studied. In doing this, we have put the subject into a group and applied the similarity principle

In our daily life, we might rationalize someone's unpleasant behavior by placing him in a particular group and viewing his behavior as a common characteristic of that group, so that we will not take it so personally. Even mathematicians replay the similarity principle: different people may not write " $1 + 1$ " identically, but we are able to ignore differences in writing or font, and assert firmly: $1 + 1 = 2$. These are just a few simple examples of similarity grouping and the utilization of the pervasive similarity principle.

Drug discovery and development is a stepwise process: A new chemical compound is tested on animals first. When the compound shows itself to be efficacious and safe in animals it will be tested on humans in clinical trials. This approach is taken in order to protect patients and to save on costs. The reason we think this stepwise approach works is that we believe the animals are similar to humans, and therefore the drug effects on animals and human will be similar. Based on the similarity principle: The animal experiment (drug + animal) is similar to the human experiment (drug + human), and therefore its use in humans will likely lead to the same or similar drug effect.

In Bayesian statistics, data on drug effects on animals can be used as prior knowledge in deriving the posterior distribution of the drug's effect on humans. The frequentists do not use animal data to directly calculate the drug effect on the humans, but animal data are used in the drug development process to help the decision making: to stop, continue, or modify the existing drug development program.

The grouping that our minds carry out in hand with the similarity principle makes everything happen for a reason, but at the same time, due to the approximation of such grouping, every law or causal relationship has exceptions.

R. A. Fisher grappled with the meaning of probability throughout his life. Not surprisingly, his view also underwent some changes (Lehmann, 1993). His last concept is something really in agreement with the similarity principle:

In a statement of probability, the predicand, which may be conceived as an object, as an event, or as a proposition, is asserted to be one of a set of a number, however large, of like entities of which a known proportion, P , have *some relevant characteristic, not possessed by the remainder* [italics added] . . . (Fisher, 1973, p. 113).

Simpson's Paradox

Subjectivity in similarity grouping reflects an individual's belief, experience, and knowledge. The controversies caused by similarity grouping can be well illustrated using Simpson's paradox (Simpson 1951).

Suppose that all patients in your community who suffer from a certain ailment, except you, have taken one of two medications. The results are shown in Table 1. The treatment effect (in terms of response rate) is 520/1500 for B, better than 500/1500 for treatment A. Thus, you as a patient will prefer treatment B to A. However, after further looking into the data for males and females separately, you find that the

Table 1. Simpson’s Paradox.

	Drug A	Drug B
Male	200/500	380/1000
Female	300/1000	140/500
Total	500/1500	520/1500

treatment effect in males is 200/500 with A, better than 380/1000 with B, while the treatment effect in females is 300/1000 with A, better than 140/500 with B. Therefore, whether female or male, we will prefer treatment A to B. Should we take treatment A or B?

The decision is dependent on which patient category or group you put yourself into in applying the similarity principle.

People try to solve this dilemma mathematically, but that is wrong. One can find a mathematically optimal solution, yet that will not address the underlying issue. Here, no random error is involved because every patient is treated and the response is correctly recorded. You are the only new patient, do you choose drug A or B?

The issue here is not to find a single answer that everyone agrees on or one that is scientifically correct, because there is *no single correct answer*. To generalize to a situation with random errors, it is not about which statistical paradigm or method gives a consistent answer. The issue is about the fundamental belief inherent in the similarity principle, which comes before any statistical principles or paradigms. The criteria on a person used for grouping him or her with others reflect one’s personal belief about which errors can be pooled together meaningfully to minimize the total error in the statistical analysis. Some people may believe that errors about drug effects on all patients, even animal data, can be meaningfully combined; others may believe only patients with say, the same gender, can be combined.

Simpson’s paradox is not uncommon if we continue slicing the data into more and more categories until the paradox appears. In a worldwide drug development process, for example, a drug can appear to be effective globally while its effects in different countries/regions are very different. The question is how should such a drug be used in different countries or regions?

Causality

Every event is the effect of antecedent events, and these in turn are caused by events antecedent to them, and so on. Human actions are no exception to this rule, including such consequential actions as the choice of committing a crime and the choice of punishing a criminal. In other words, there is no such thing as a choice made primarily of one’s own will. However, we, as scientific researchers, always ask why we make or what causes us to make certain choices. I often ask myself: With the ultimate development of neuroscience and understanding of the human brain, how much space remains for free will?

So what does causality really mean if everything is deterministic? The construct that we call causality is a consequence of our belief in the similarity principle. Since our poor brains cannot deal with everything happening in our lives, we group similar things together and apply, consciously or not, the similarity principle. Only in this way can our brains handle the universe. That is, only with the similarity principle can causality scientifically make possibly sense, only on the basis of this principle can scientific laws be possibly derived, only with this principle can we meaningfully discuss inferences and predictions, only with this principle does the connotation of “scientific knowledge” become lucid, and only with this principle do we have the concept of probability and the concept of random errors (Chang 2014). Indeed, I believe the similarity principle is a necessary, even though not sufficient, condition for the existence and usefulness of all sciences.

Within the context of scientific discovery, a causal relationship between events A and B means three conditions to me: (1) A proceeds B in time, (2) the law of factor isolation holds: If A then B ; if not A then not B , and (3) the relationship is verifiable, at least in principle, which means that condition 2 is persistent over time, which further requires that the events in condition 1 are repeatable. The repetition of events is what gives causality its great usefulness: we use it to make predictions.

So far we have only discussed the possibility that factor A is the sole cause of B . But how do we test the causes when either factor A or C alone can be sufficient cause of B ? This issue becomes very complicated. Let us look into the “paradox of murder in desert” (Smullyan, 1978): A caravan of three (A , B , and C) were going through the Sahara Desert, where one night they pitched tents. A hated C and decided to murder him by putting poison in the water of his canteen (C ’s only water supply). B , who also wanted to murder C , didn’t realize that C ’s water was already poisoned, and he drilled a small hole in C ’s canteen so that the water would slowly leak out. As a result, several days later C died of thirst. The question is, who was the murderer, A or B ? According to one argument, B was the murderer, since C never did drink the poison put in by A ; hence, he would have died even if A hadn’t poisoned the water. According to the opposite argument, A was the real murderer, since B ’s actions had absolutely no effect on the outcome; once A poisoned the water, C was doomed, and hence A would have died even if B had not drilled the hole. Which argument do you agree with?

In the two-factor situation, to determine the cause(s), we need to know the outcome (B or not B) in all of the following situations: (1) not A and not C , (2) A , but not C , (3) not A , but C , and (4) A and C .

If A implies B , and not A implies *sometimes* not B , then A is a cause of B in some situations (with C and/or without C). Of course, the statement is equally true when factors A and C are exchanged.

Null Hypothesis Testing and False Discovery Rate

Consider the classical experiment of flipping a coin repeatedly. If all the conditions (the force used, the direction of the force, the coin will not be worn out, etc.) are

precisely the same each time I flip the coin, the result of each flip will also be identical. The so-called random variation is simply caused by many hidden or unobserved variables: The way you flip the coin is in fact slightly different each time. The point being made here is that there is in some sense no real underlying randomness at all; everything is a very fine cause and effect.

To deal with randomness, descriptive statistics may provide information about the practical significance of an observed value, but provide no information about whether such a finding is real or just a random chance event, that is, the issue of statistical significance. When there are several confounders, a scientist will find it very difficult to objectively get a sense of the totality of scientific evidence simply based on descriptive statistics. To reduce false findings, we use the measure of false discovery rate (FDR). FDR is the expected proportion of false positive findings among all findings (Benjamini & Hochberg, 1995; Chang, 2011). One way to control FDR is through null hypothesis statistical test (NHST), which can be briefly described as follows.

Suppose we want to hypothesis test H_0 : drug x is ineffective, versus H_a : drug x is effective. When $p < \alpha$, a nominal level, we reject H_0 and conclude the drug is effective. Technically, a p -value is the probability of obtaining an effect at least as extreme as the one in one's sample data, assuming the truth of the null hypothesis. Therefore, a p -value does not directly answer the question as to how effective a drug is.

A smaller p -value indicates a discrepancy between the hypothesis and the observed data. In this sense, p -value measures the strength of evidence against the null hypothesis. However, p -value is not a probability of a null hypothesis being true. Rejecting H_0 at $\alpha = .05$ level with $p = .02$ does not mean we have a 5% or 2% probability of making a mistake.

A level of significance of $\alpha = 5\%$ does not mean there are 5% false positive findings because the null hypothesis is just an assumption. For example, if all the null hypotheses under investigation are true, then all positive findings are false, no matter what value of α is used. In contrast, if all null hypotheses we investigate are in fact false, there will be zero false findings, independent of α .

If some proportion R of all null hypotheses to be investigated are true, we want to control the false findings among these true hypotheses at a rate of α , the significance level of the null hypothesis test. Therefore, α can be interpreted as the expected proportion of false findings among all true null hypotheses. The problem is, since we do not know which null hypotheses are true, we have to apply the same α to all null hypotheses. As a consequence, the power of the hypothesis testing is reduced or the ability (probability) of detecting a true positive finding is reduced unless the sample size increases. How can we use a single index to measure both probabilities of correctly identifying positive and negative findings? The false discovery rate is such a measure. The proportion of false positive findings among all findings (FDR) is expectedly,

$$\text{FDR} = (\alpha \times R) / (\alpha \times R + \text{power} \times (1 - R)) \quad (1)$$

Because the true power is unknown, not a constant, and model dependent, the above equation is only an approximation. Thus, as far as FDR is concerned,

increasing power, reducing α , and reducing R will reduce FDR. We can use a smaller α or a larger sample size to increase power. But how can we reduce R , the proportion of null hypotheses to be investigated? Practically speaking, we use a stepwise approach. In drug development, from the preclinical to the clinical stage, at each step forward we screen out more and more false positive findings or true null hypotheses; that is, we reduce R for the next step in research. Strategically, we progress from the so-called exploratory study (often using descriptive statistics only) to confirmatory studies using NHST or other similar methods, such as a Bayesian approach. We may argue that our goal is reduce the impact of errors, not the error rate itself. This is true, but when the impact of error is generally difficult to define, reducing the FDR is reasonable.

However, p -value from a NHST often does not provide the full picture about the scientific evidence. Descriptive statistics and confidence interval are also valuable to look into.

We often wish to have a direct answer to the question of the effectiveness of a drug or a clinical intervention. To this end, we can use Bayesian posterior probability, which is a combination of prior knowledge and current experimental data. One common criticism of the Bayesian method is its subjectivity in determining prior knowledge. Such subjectivity combined with a psychological bias, such as *anchoring*, makes the result more controversial (Lench, Safer, & Levine, 2011). Anchoring is a cognitive bias that describes the common human tendency to rely too heavily on the first piece of information offered (the “anchor”) when making decisions. We are constantly using the Bayesian notion of learning or reasoning in our scientific research and our daily lives, because every conscious action we take is a utilization of prior relevant experiences or knowledge.

As one example, when a pharmaceutical company seeks approval of a drug for marketing, it is difficult without some hard criterion such as “ $p < \alpha$ ” as a statistical basis for obtaining approval. This is because the outcomes are in the inventor’s and the sponsor’s interests. Using their subjective judgments or priors may lead to a biased decision. Likewise, the regulatory authorities’ subjective judgments may also be biased because officials may favor a certain outcome. The situation also arises in academic research to a certain degree. Therefore, NHST can be viewed as an important tool for controlling FDR. The question is how to choose the significance level, α . A p -value can be used as one piece of evidence in composing the totality of scientific evidence, but we ought not necessarily use a single universal, constant value of 5% for α whenever we wish to make a binary decision. As different errors have different impacts to our lives, the thresholds should be different too.

Controversies in Regression

When multiple factors are involved in a cause–effect relationship, null hypothesis testing is often used in conjunction with a regression model.

We always say that an association between variables is not necessarily causal. Indeed, association is a necessary but not a sufficient condition for causality. But more important, an association can become a causal relationship. For instance, data have shown that people who carry matches are more likely to have lung cancer than people who do not carry the matches. That is, there is an association between matches and lung cancer, because smoking cigarettes can cause cancer and smokers usually carry matches. Therefore, carrying matches is not a cause of lung cancer. However, by further inspection, we learn that the direct cause of lung cancer is carcinogens in the cigarette, not the cigarette. If we can disassociate/extract the carcinogens from cigarettes, then smoking cigarettes will no longer be a cause of lung cancer. On the other hand, a causal relationship can become an association. For example, if only cigarette smokers can carry matches and all cigarettes smokers carry matches, then by extension we can consider matches a cause (risk factor) of lung cancer.

The principle of factor isolation is often used to investigate causality with a single cause in physics. However, in life, social, and medical science, there are many other random factors, observable to hidden, simple regression can be viewed as a utilization of principle of factor isolation in the presence of randomness. In order for the regression to be effective the experiment is often conducted with a control group and randomization to balance out confounding factors. When the causality involves multiple causes, we often use multiple regression or, equivalently, analysis of covariance to investigate the main effect and covariate effects. The main effect can often be treatment interventions and the covariates are often confounders.

For a variable to be a confounder in a clinical trial it must satisfy three conditions (Chang, 2014):

1. It must be associated with the treatment (the main factor).
2. It must be a predictor (not necessarily a cause) of the outcome being measured.
3. It must not be a consequence of the treatment (the main factor) itself.

A factor is not a confounder if it lies on the causal pathway between the variables of interest. Such a factor can be a direct cause or a mediator. Of course, the causal pathway is often not observable and very much dependent on our interpretation. It is not always necessary to know the intimate pathway; however, through a hypothetical pathway, we often can often better model the mechanism of a system. Latent variable structure equation modeling is an example of such hidden pathway modeling. Such a technique is often used in sociology, psychology, and econometrics studies. Before we can deal with sophisticated situations, let us just study the challenges imposed by an observable confounder in multiple regression.

The randomized controlled experiment is a great invention, but the analyses of data from such an experiment and the interpretation of results do not lack of challenges and controversies. Whether it is an observational or a controlled experiment, multiple regression is a common tool for analyzing the data. For a linear regression

Figure 1. Paradox of multiple regression.

model (linear in constants a and b), $Y \approx aX^2 + bX$, we would not say the effect of X on Y is b because a is also a partial indicator of the X effect. However, if we write the equation as $Y = aZ + bX$, and the covariate $Z = X^2$ is not indicated but the correlation $\text{corr}(X, Z) \approx 1$ is observed, then we often mistakenly say the effect of X on Y is b . Here, Z is a measured covariate, a confounder, or a mediator. Let us discuss the controversies further through an example I have provided elsewhere (Chang, 2007, 2011).

Suppose five subjects in each of two treatment groups ($X = 0$ for Placebo and $X = 1$ for Test Drug) in a clinical trial are measured for clinical responses Y and biomarker responses Z . We first calculate Pearson's correlations between the variables: the treatment, biomarker, and clinical endpoints. The results are presented in Figure 1. We can see that the correlations between them are not transitive. In other words, a correlation between the treatment and the biomarker ($R_{XZ} = 0.45$) and a correlation between the biomarker and the clinical responses ($R_{YZ} = 0.90$) do not ensure a correlation (R_{XY}) between the treatment and the clinical response.

The average response with the clinical endpoint is 4 for each group, which indicates that there is no treatment effect. On the other hand, the average biomarker response is 6 for group B and 4 for group A , which indicates that the drug has effects on the biomarker.

After fitting a linear model with the data we obtain

$$Y \approx Z - 2X \tag{2}$$

This model fits the data well based on the model-fitting p -value and R^2 . Specifically, R^2 is equal to 1. The p -values for the model and all parameters are equal

to 0. The coefficient 2 in the model is the separation between the two lines. Based on the model, we would conclude that the biomarker has a positive effect and the treatment has a *negative effect* on the clinical response. But in fact, the treatment has no effect on the clinical response at all. The reason is that Z relates to X through their means $\bar{Z} = 3 + 2\bar{X}$. This implies that $\bar{Y} = 3 + 2\bar{X} - 2\bar{X} = 3 + 0 \cdot \bar{X}$. Therefore, it is not surprising that the coefficient for X changes when the different covariate Z is added to the model. Note that Z is not a confounder per our definition since Z is a consequence of treatment X , and thus does not meet criterion 3 in the confounder definition.

Now suppose the biomarker responses in the example are changed to baseline confounder Z before randomization. We would obtain the same model and the same conclusion that the treatment has a negative effect on the clinical response. We further learn that if we change our similarity grouping to also consider the biomarker value at baseline, then we will select 3 patients (with $z = 3, 4, 5$) in the control and 3 patients with the same biomarker values in the test group. Now there is clearly a treatment effect of -2 for every patient with same baseline value for Z . This is consistent with the multiple regression producing $Y = Z - 2X$ obtained from the data.

From this example we can see that in multiple regression one should use baseline confounders, but not variables that result from treatment (or main effect) as covariates. In general, a different set of variables included in the regression model implies a different similarity grouping for which the similarity principle applies.

The problem is when the biomarker or a similar factor is hidden or not observable. When this happens its imbalance in the two groups is unknown since randomization cannot ensure that all confounders will be balanced. This phenomenon, that adding explanatory variables may reverse the relationship between the outcome variable and explanatory variables, is called Lord's paradox by some scholars (Lord, 1967; Tu, Gunnell, & Gilthorpe, 2008). Here, we encounter the same paradox as in Simpson's paradox or when we try to predict an emigrant's longevity.

I want to point out that we should not encourage any predictor to be included in a multiple regression model if the predictor is a consequence of other predictors in the model. For, if we do, the interpretation of the outcome from the model is difficult: the effect of a predictor is not simply its coefficient. In general, a multiple regression coefficient reflects the effect of the predictor when other predictors are fixed and unchanged. When this cannot be done the meaning of the coefficients in a regression model is difficult to interpret, as in the example given in Figure 1 and Equation (2). Since most predictors in a model are correlated, multiple-regression results are generally difficult to interpret.

Recently, the journal *Basic and Applied Social Psychology* published a special issue on the topic of Mediation Analysis. Trafimow (2015) uses data from planets in the solar system to illustrate the problem in multiple regression of mediation analysis. However, his example is not very convincing because (1) the momentum $Y = \text{mass} \cdot \text{velocity}$ is not a scientific law but rather a definition and (2) finding $Y = 1.016 \cdot \text{mass} + 0.054 \cdot \text{velocity}$ fits the data well is not really an issue because the model truly reflects

the data pattern. It is possible there are more than one model that fit the data well. The reason we have multiple models is that the data show a correlation between mass and velocity: $\text{correlation}(\text{mass}, \text{velocity}) = 0.313$. When there is a high correlation we will have a collinearity problem, that is, there are infinitely many models that equally fit the data. As we stated previously, when the correlation is present, the effect of velocity on momentum is not simply characterized by its coefficient, 0.054.

We often use an entire data set for statistical modeling; the resulting model is the consequence of minimizing the total error. But that may not be desirable. For instance, a drug's effect on youth and the elderly can be very different. Thus, what we really want to minimize is the relevant error. What is relevant error? It has to be based on the similarity principle: if you think people with similar ages will have a similar treatment effect, then use that group of people to do modeling to minimize the error of the model within that group. How do we determine such subgroups? It is a tradeoff between accuracy and precision. Dividing the sample into many smaller subsamples may allow a researcher to reach accurate conclusions but with poor precision, due to small sample sizes in a subgroup, whereas dividing the sample into a few bigger subsamples may yield conclusions with good precision but inaccuracy because characteristics in the subsample may still be very different.

P-values are often obtained from multiple regressions to determine if a variable is significant to keep or insignificant to remove. When there can be many hidden confounders, as in the field of psychology, randomization cannot ensure the balance of all these confounders. Thus, the NSHT has less value because the *p*-value is not accurate. Recently, we discovered some interesting ways to deal with hidden confounders using mixed distributions, an idea similar to the method to reveal treatment effect without unblinding the randomization code (Chang, 2014)

Multiplicity Challenges

The “multiplicity issue” refers to the statistical phenomenon of increasing FDR or Type-I error in NHST when multiple analyses are conducted. Here are various examples:

Assume that a terrible crime is committed in a large city. The police decide to run a fingerprint check, according to the alphabetic order of last names in their database starting with the letter A, and the check is to be stopped if and when a match is found to the crime sample. Suppose that one confirming pair of prints is found, and that the suspect's fingerprints were the very first ones checked (his last name being Aaberg). Now, the probability of a match of two pairs of fingerprints chosen at random is $1/20,000$. So the suspect could well argue that if the agency's check had started with the letter Z, and even if the agency wants to check *everyone*, they would have identified all 0.36 million fingerprint matches in the world population, of 7.2 billion before finally checking on him. Thus, he says, the unfortunate position of his name in the alphabet and the act of stopping the search prematurely have cast unreasonable suspicion on him.

If a clinical trial involves an interim analysis for a possible early claim for efficacy, then we have to decide how much Type-I error is spent at each analysis (interim and final) to control overall Type-I error at the alpha level. Consequently, the ability to claim efficacy (reject the null hypothesis) will be very much dependent on alpha-spending at each analysis. However, from chemistry, we know that how the statistician spends his alpha will not affect the effectiveness of the drug, and there is no reason to adjust alpha at each stage. Here, we can see two different concepts of the effectiveness of a drug. One is the set of chemical properties of the test compound, which does not depend on the statistical method used. The other is the set of statistical properties which *will* change since it reflects an aggregated attribute of a group of similar patients.

Linear regression is one of the most common statistical procedures used in research. For multiple regression, the backward elimination algorithm (BEA) is often applied. The problem with this method is Type-I inflation and consequently FDR increases. For example, if there are 10 factors to investigate, in multiple regression with BEA the Type-I error is about 40%. This means that a scientist can use any data with 11 variables and perform multiple regression with BEA and have at least a 40% chance of finding at least one model that is significant. As the variables increase, the Type-I error rate increases. Therefore, we have to control, or at least assess, the Type-I error or FDR. Otherwise, there will be too many false scientific discoveries.

Multiplicity is indeed challenging in scientific discovery. Even if we use one of the most stringent criteria, NHST, we still have to face the terrifying fact: If one research study has a false discovery and other 100 research studies are conducted to confirm it, then we will expect that 5 studies will falsely verify the finding using NHST, assuming the significance level 5%. There might be a misconception that if we do the experiment very carefully, the error or false positive findings can be eliminated. But this is just an illusion. The main reason that error is unavoidable is not that our experimental design is poor or the measurements we have taken have errors, but it is because there are hidden confounders that we cannot know or are unobservable at the moment of experiment.

Data accumulate continually in the course of any long investigation or series of investigations into medical intervention or other research fields. As more public data become available, more (and more) analyses will be performed on the same data, and this will inflate the FDR. On the other hand, since meta-analyses have more reliable data, the statistical conclusions will be more reliable. How then can we simply dismiss the results from meta-analyses just because we have “spent all alpha?”

Summary and Recommendations

Science is a study of (ideally) repeated events. Since everything is unique in its own way, by grouping similar things we identify the repeating patterns of events that matter to us. On the basis of these repetitions we apply the similarity principle. When the difference among the members in a group is acknowledged, the statistical properties

that characterize the group are used to predict properties of individual members in the group. Now, the properties of a member are fixed, if unknown, and they do not vary as the grouping and statistical methods change. But the statistical properties *will* depend on the grouping and the statistical analysis methods used. Similarity grouping is fundamentally subjective, but can be combined with objective means such as unsupervised learning methods or clustering.

An important goal in scientific research is to reduce FDR, the proportion of false positive findings among all findings. To reduce FDR, we can increase the power of NHST, or reduce the level of significance in NHST. In this sense, NHST and p -value are useful. We can also reduce FDR by reducing R , the proportion of false relationships in the investigations. One effective way to reduce R is to use a stepwise approach: From exploratory studies we gradually progress to confirmatory studies. When a hypothesis reaches its confirmatory stage, the sample size is usually larger and cost is high. The stepwise approach is a cost-effective and time-saving way of eliminating most false casual relationships prior to confirmatory studies and reducing R in the confirmatory stage. Even though the FDR may still be high with NHST, but that does not automatically qualify other methods as better than NHST. Even applying the most stringent criteria when using NHST methods, the false discovery rate may be still high due to hidden confounders and multiplicity issues. Therefore, NHST should not be used alone, rather it should be used in conjunction with other scientific or statistical methods.

A point estimate cannot tell if a difference is truly significant or only a chance occurrence. P -value combines the effect of treatment difference and effect of sample size: We cannot tell if a smaller p -value is due to a larger treatment difference or a large sample size. Any other single statistical measure will suffer a similar shortcoming. We should differentiate the misuse and misinterpretation of NHST's utility from the limitations of NHST. When one knows the limitations, he will use NHST cautiously. When one misuses NHST, he or she will be led to false conclusions; when one misinterprets the utility of NHST he or she may not see the value of NHST and may even want to ban it. To my knowledge, no scientific journals use NHST or p -value solely to qualify a manuscript for publication. If that were the case we would not need any particular experiment at all; instead, we could just flip a coin 10 times, and if we get 8 or more heads (which happens 5% of the time), we will consider whatever the null hypothesis is to be true!

Statistics requires both deductive and inductive reasoning. Unlike mathematics, in which a single counterexample can dismiss a false theorem, examples of failing a statistical hypothesis are not difficult to find and counterexamples may not necessarily invalidate a statistical conclusion or method. We should not over trust statistical modeling, believing whatever a model tells us; we should also not dismiss or ban a statistical method simply because it cannot handle some situations.

Because of controversies in multiple regression, we recommend subgroup analysis using simple univariate models and great caution in using multiple regression with highly correlated predictors. As much as possible, avoid using mediators in a multiple regression.

On one hand, we have developed very advanced statistical models. On the other hand, we lack understanding of some fundamental scientific principles, such as the similarity principle, and of basic statistical models such as multiple regression. It is time for us to pause and rethink those fundamentals before putting all our efforts into creating more and more complex statistical models.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

References

- Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B*, 57, 289-300.
- Chang, M. (2007). *Adaptive design theory and implementation using SAS and R*. Boca Raton, FL: CRC Press/Taylor Francis.
- Chang, M. (2011). *Modern issues and methods in biostatistics*. New York, NY: Springer.
- Chang, M. (2012). *Paradoxes in scientific inferences*. Boca Raton, FL: CRC Press/Taylor Francis.
- Chang, M. (2014). *Principles of scientific methods*. Boca Raton, FL: CRC Press/Taylor Francis.
- Fisher, R. A. (1973). *Statistical methods and scientific inference* (3rd ed.). London, England: Collins Macmillan.
- Lehmann, E. L. (1993). The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two? *Journal of the American Statistical Association*, 88, 1242-1249.
- Lench, H. C., Safer, M. A., & Levine, L. J. (2011). Focalism and the underestimation of future emotion: When it's worse than imagined. *Emotion*, 11, 278-285.
- Lord, F. M. (1967). A paradox in the interpretation of group comparisons. *Psychological Bulletin*, 68, 304-305.
- Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B*, 13, 238-241.
- Smullyan, R. M. (1978). *What is the name of this book? The riddle of Dracula and other logical puzzles*. Englewood Cliffs, NJ: Prentice-Hall.
- Trafimow, D. (2015). Introduction to the Special Issue on Mediation Analyses: What if planetary scientists used mediation analysis. *Basic and Applied Social Psychology*, 37, 197-201.
- Tu, Y. K., Gunnell, D., & Gilthorpe, M. S. (2008). Simpson's paradox, Lord's paradox, and suppression effects are the same phenomenon—the reversal paradox. *Emerging Themes in Epidemiology*, 5, 2. Retrieved from <https://ete-online.biomedcentral.com/articles/10.1186/1742-7622-5-2>