The Optimal Threshold *p*-Value for a Scientific Journal

Donald B. Macnaughton (donmac@matstat.com)

Abstract

A scientific journal that publishes data-based research papers can use a threshold *p*-value as a "gateway" to publication for submitted papers. The gateway helps the journal to maximize the long-run scientific and social benefits of the research papers it publishes. An appendix gives a constructive graphical and mathematical argument showing that the optimal threshold *p*-value for a scientific journal exists.

Keywords: relationship between variables, statistical significance, scientific hypothesis testing

1. Introduction

The *p*-value is often discussed in a broad range of uses. However, we discuss using the *p*-value in a very specific but important application—in its role in a *gateway to publication* of an empirical research paper in a scientific journal. We consider how the use of a "threshold" *p*-value gateway by a journal helps the journal editor to maximize the scientific and social benefits provided by the journal.

We begin with some basic ideas that many readers will already know, though possibly know differently from how the ideas appear here. These basic ideas are the logical foundation for the later discussion.

2. Relationships Between Variables

We focus on scientific research studies that collect and analyze data, which we refer to as "empirical" research studies. A large proportion of scientific research studies are empirical research studies.

As many statisticians will agree, it is generally useful and easy to view empirical research studies through a single unifying point of view. That is, we view a research study as studying one or more *relationships between variables* in a population of entities. For example, medical researchers often study the relationship between the *dose* of a drug given to medical patients and the *severity* of a disease in the patients in a population of patients.

Science and society are interested in relationships between variables because if a researcher can *find* a new relationship between variables in a population of entities, and if the variables were sensibly chosen, then this information is often highly useful. This is because we can use the relationship to reliably *predict* or sometimes *control* the values of one of the variables in new entities from the population. We can do this by measuring or controlling the values of the *other* variable or variables in the entities and then using our knowledge of the relationship to accomplish the prediction or control.

3. Does a Relationship Exist?

A key initial question in the scientific study of any relationship between variables is whether the relationship actually *exists* in the population or whether the relationship is merely a *figment* of the researcher's imagination. This question is important because it sometimes turns out in modern science that a postulated relationship between variables doesn't exist (or at least it doesn't *detectably* exist) in the population. This happens because, though a researcher may have a brilliant *idea* about a new relationship between variables, sometimes the idea is, unfortunately, wrong. And the postulated relationship between the variables doesn't exist in the population. For example, medical researchers sometimes find that there is no good evidence of a relationship in medical patients between the dose of what the researchers thought was a promising drug and the severity of a disease.

Of course, researchers determine whether there is good evidence that a relationship exists between variables by analyzing appropriate research data about the entities in the population. That is, researchers draw a *sample* of entities from the population and then they measure the value of each of the variables of interest in each entity in the sample. In some research studies the researchers also *manipulate* the values of some of the variables in the entities in the sample.

The researchers collect all the measured values of the measured variables in a *data table* and then they analyze the data. Then they make careful inferences from the analysis of the sample data about the existence and about the apparent form of the studied relationship between the variables in the entities in the underlying population. That is, they make careful generalizations from the sample to the population.

The following sections discuss how journals and researchers address the first step of these generalizations from a data table—how they decide whether the studied relationship between variables (likely) *exists* in the population. This step comes first because there is no practical point in considering other aspects of a relationship between variables unless we are confident that the relationship between the variables likely exists. That is, if we don't have good evidence that a relationship exists, then we may be studying not a relationship, but mere noise in the data, perhaps mistakenly thinking that the noise reflects a relationship.

4. Positive Results and Negative Results

A positive result occurs in a scientific research study if a proper analysis of the data finds "good evidence" that a

relationship between the studied variables exists in the population. In contrast, a negative result occurs if a proper analysis *doesn't* find good evidence that a relationship exists.

So, how can we tell whether we have good evidence in a research study that a relationship between variables exists. In other words, how can we perform a proper analysis of the data in a data table to determine whether we have a positive result?

A standard objective way to address this question is to apply a measure of the *weight* of the evidence to the data. Statisticians have invented more than ten different measures of the weight of evidence that a relationship exists between variables, such as the *p*-value, confidence interval, likelihood ratio, Bayes factor, and others. These measures use different measurement scales and different underlying theory, but they are all monotonically related to each other, so they all operate quite similarly.

(The measures are monotonically related because they are all monotonically dependent on the estimated effect size when other variables are held constant, and because monotonicity is transitive.)

For simplicity, this paper focuses on the *p*-value because it is intuitively sensible and because it is the most popular of the measures. But the ideas apply equally to all the other standard measures of the weight of evidence of the existence of a relationship between variables, any of which could (sometimes with minor caveats) sensibly replace the *p*-value in the following discussion.

In the case of the *p*-value, and assuming that the *p*-value is properly computed from an appropriate data table, then the *lower* the *p*-value computed from the data, the *greater the weight of evidence* in the data that the associated relationship between the variables exists in the population in a reasonable mathematical sense.

Here is a definition of the *p*-value:

> The *p*-value computed from a data table is the fraction of the time (i.e., the probability) that we will obtain a result as "extreme" (in terms of implying the presence of a relationship) or more "extreme" as the result reflected in the table if there is or were *no relationship* between the studied variables (and if certain often-satisfied assumptions are adequately satisfied).

Because the *p*-value represents a fraction of the time, the value of a computed *p*-value always lies between 0 and 1.

Although the definition of the *p*-value is easy for statisticians to understand, the definition is hard for others because it involves the concept of conditional probability. Fortunately, in its role in scientific research, we needn't think of the *p*-value using the complicated probability ideas. We need only remember that the *p*-value is a measure of the weight of evidence for the existence of a relationship—if everything is done properly, the lower the value of the *p*-value, the greater the weight of the evidence.

The definition of the *p*-value implies that *p*-values are meaningfully comparable across research studies. We will use the idea that the *p*-value measures the weight of evidence of the existence of a relationship between variables in a moment.

5. The Perspective of a Scientific Journal

A scientific journal wishes to publish papers that are reporting interesting *positive* results about new relationships between variables in populations. This is because an interesting *positive* result tells us about an observed and possibly real new relationship between variables. Knowledge of the relationship may give us the ability to better predict, control, or understand the variables in the entities in the studied population. In the medical example, if medical researchers can find a beneficial relationship between the amount of a drug given to medical patients and the amount of a certain disease in the patients, then this may enable doctors and patients to better control the disease.

In contrast, scientific journals generally *won't* publish papers that are reporting *negative* results. This is because these results usually tell us nothing new, so they are generally uninformative and uninteresting. In particular, you can't sensibly do prediction or control from a negative result. For example, if there is no evidence of a relationship between a drug and a disease, then you can't do much with that.

Some authors think that negative results should be published because they they think that negative results tell us about relationships between variables that *don't* exist, which would be useful to know. But a negative result can't tell us that a relationship doesn't exist—it can only tell us that a relationship wasn't *found* in the particular set of research conditions that were used in the research. So, in most cases of negative results, perhaps if the researchers had only used slightly different research conditions, then they would have properly found the sought-after relationship. For example, perhaps the relationship will only appear if the room temperature is above 25°C, but the researchers didn't know that and (unfortunately) performed the research at 20°C. So, negative results are almost never definitive. So, negative results usually don't provide much useful information.

Also, the economics of scientific-journal publishing limits the number of papers that can be published to only the most interesting ones. There are usually more than enough interesting submitted *positive* results, so negative results are usually immediately eliminated from consideration.

Some authors describe the journal policy to omit publishing negative results as "publication bias"—a term that suggests that the omission of publication of negative results is somehow irrational or unfair. However, arguably, the general omission of publication of negative results is sensible because these results generally aren't useful.

It is noteworthy for completeness that occasionally a negative result is sensational, surprising, or useful (e.g., for policymakers). In this case, it may be sensible to publish the result in a peer-reviewed scientific journal. However, most negative results *aren't* sensational, surprising, or useful, and are instead boring, so they aren't published.

So, a journal needs an efficient way to *distinguish* between a positive result and a negative result. Some journals make this distinction by saying to the researcher, "The *p*-value for the main result (i.e., the main reported relationship between variables) in your research paper must be less than or equal to our *threshold p*-value of 0.05 before we will *consider* your paper for publication." This enables the journal to automatically decide quickly, objectively, and fairly whether a paper has enough weight of evidence for its main result to make the result a positive result and to make the paper worth *considering* for publication in the journal.

Metaphorically, the journal says to the researcher, "You must be at least 4 feet tall to be allowed on this ride." Of course, this rule isn't used to be arbitrary or mean—it is used to ensure the safety of the ride. In the case of the threshold *p*-value, the journal wishes to ensure that there is enough weight of evidence for a relationship between variables. This is because the journal wishes to avoid mistakenly publishing a report about a claimed new relationship between variables when the relationship actually doesn't detectably *exist* in the population and the paper is reporting about mere noise in the data.

If the *p*-value for a research result is less than or equal to a journal's *threshold p*-value, then it is customary to say that the result is "statistically significant."

It is helpful to remember that the threshold *p*-value defines a *barely sufficient* condition for good evidence that the reported relationship between variables exists in the population. Researchers and editors almost always hope that the main *p*-value obtained in a research study will be well below the journal's threshold value because that means (assuming that everything is done properly) there is less chance (in an informal sense) that the result is mistaken, as discussed below.

As noted, a main *p*-value that is less than or equal to the journal's threshold *p*-value is a *necessary* condition for publication of a paper in some scientific journals. It is important to add that satisfying the threshold-*p*-value condition is certainly *not a sufficient condition* for a paper to be published. To be published, a research paper must obviously also be of enough *interest* to the journal's readers, as judged by the journal's editors and referees. Also, and importantly, the paper must exhibit enough *quality*, where quality is highly multifaceted, as dictated by the standards of the associated field of science, and again at the discretion of the editors and referees. The interest, quality, and threshold *p*-value conditions must *all* be satisfied before a research paper will be published in some scientific journals.

Some people think that a low *p*-value means that the associated relationship between variables is *important*. That is incorrect because the *p*-value doesn't measure the importance of the relationship. A low *p*-value only means (in the absence of a reasonable alternative explanation) that there is good evidence that the studied relationship between the variables *exists*.

After a journal decides that there is good evidence that a relationship exists, the journal must obviously next consider the importance of the relationship, which is a step in deciding whether the paper will be of interest to the journal's readers. Journal editors and referees judge the importance of a relationship between variables in terms of its perceived scientific and social usefulness. Of course, this involves considering the estimated "effect size" or "strength" of the relationship between the variables because if the effect size is small, then the relationship may not be of much use.

It is noteworthy that, for technical reasons, in the case of a true positive result, the initial estimate of an effect size derived in scientific research is often an *over*estimate, perhaps 30% or more greater than the true population value. Of course, this isn't a serious problem if we keep the existence of the phenomenon in mind because then we can take informal or formal account of it when we are assessing an estimated effect size.

Appendix E discusses how journals handle the situation when a paper has no *main* result and is reporting multiple *p*-values, such as 5 *p*-values or 50,000 *p*-values.

6. The Threshold *p*-Value Makes Errors

The threshold *p*-value would be perfect for detecting relationships between variables if it could always be *right* about whether the studied relationship between variables exists in the population. But, as many readers will know, the threshold *p*-value makes two types of errors. There appears to be no way to escape from these errors in scientific research.

7. False-Positive Errors

A false-positive error occurs if the computed *p*-value for a relationship between variables is less than or equal to the journal's threshold *p*-value, suggesting that the studied relationship between variables exists, but in fact the relationship *doesn't* detectably exist in the population. As discussed below in appendices B and C, false-positive errors are published surprisingly often in the scientific research literature, and they are the source of the so-called replication crisis in scientific research. That is, if you try to replicate or use a published false-positive result, you will almost always fail. (You will fail unless *your* research *also* makes a false-positive error.)

A false-positive error can occur due to chance or if the researcher breaks the rules for computing *p*-values.

In general, a paper reporting a *false* positive result looks no different from a paper reporting a *true* positive result. So, if a paper reporting a false-positive result can satisfy a journal's other conditions for publication, then the journal will publish the paper, thereby unknowingly contributing to the replication crisis.

False-positive results in the scientific research literature are *costly* because they lead to a waste of resources for other researchers who try to use or extend the false results.

Of course, false-positive results in the scientific literature are identified and corrected through the process of replication. Other researchers always indirectly replicate *interesting* new positive results as they try to use or extend the results.

*Un*interesting false-positive results aren't replicated, so they remain uncorrected in the literature. But that isn't a problem because these results are uninteresting.

8. False-Negative Errors

A false-*negative* error is the opposite of a false-positive error. A false-negative error occurs if the computed *p*-value for a relationship is *greater* than the journal's threshold *p*-value, suggesting that the relationship between the studied variables *may not* exist in the population, but in fact the relationship *does* exist in reasonable strength in the population.

A false-negative error amounts to a missed discovery. A false-negative error can occur if the relationship between the variables is weak, if the study was poorly designed, if the researcher made an error, or due to chance.

False-negative errors are, by their nature, hidden. So, we don't hear much about them. So, most of what we know about false-negative errors comes from theoretical considerations, which clearly imply that false-negative errors occur regularly in scientific research, though we don't know exactly how often.

Like false-positive errors, false-negative errors are *costly* because they lead to a loss of useful information for society and a loss of reward for the researchers who obtain the false-negative results. For example, if a medical research study somehow *fails* to detect that an effective new drug is effective, thereby committing a false-negative error, then society may lose the benefit of the drug.

Of course, false-negative errors that have been omitted from the scientific literature are identified and corrected if another (or the same) researcher performs a new research study of the relationship between the variables with a research design that can reliably detect the relationship.

For completeness, it is noteworthy that false-positive and false-negative errors are traditionally referred to as "type 1" and "type 2" errors respectively. However, those names are inferior because they are unnecessarily confusing for beginners.

9. Controlling the Error Rates

As explained in appendix B, it is easy to show mathematically that a journal's threshold *p*-value simultaneously controls the long-run rates of both false-positive and false-negative errors in the journal. That is, if a journal uses a *lower* threshold *p*-value, then it will publish *fewer* false-positive errors, *but* the journal will make *more* false-*negative* errors in the sense of refusing for consideration for publication some papers that are reporting about real relationships between variables.

So, a journal has a dilemma—where should it set its threshold *p*-value to sensibly balance the long-run rate of false-positive errors that it incorrectly publishes against the long-run rate of false-negative errors that it *should* publish but incorrectly *fails* to publish? How should a journal sensibly choose its threshold *p*-value?

10. The Optimal Threshold *p*-Value

Choosing the optimal threshold *p*-value for a scientific journal is conceptually surprisingly simple—the journal chooses the value that maximizes the scientific and social benefit resulting from the research papers that are published or are refused publication in the journal. This approach is sensible because, arguably, a journal's goal should be to maximize the long-run scientific and social benefit of the research papers it publishes.

The journal maximizes the benefit by finding the "sweet spot" for the threshold value that minimizes the *sum* of the costs of the false-positive and false-negative errors, which is equivalent to maximizing the benefit. Appendix B presents a mathematical argument showing that the optimal threshold *p*-value for a journal exists. Setting the threshold *p*-value at the optimal value to minimize the sum of the costs of the errors is useful because, as noted, the two types of errors occur regularly in scientific research and are costly.

11. Choosing the Optimal Threshold *p*-Value

Ideally, a scientific journal would choose its optimal threshold *p*-value based on formal scientific research about the scientific and social benefits realized and the costs incurred under different journal publication policies. However, for technical reasons, we can't reliably measure (a) the ongoing *benefits* of correct scientific research or (b) the ongoing *costs* of false-positive and false-negative errors in scientific research. So, a journal can't choose its threshold *p*-value based directly on formal scientific research.

So, a journal chooses its threshold *p*-value based on carefully considered experience and intuition among journal editors and researchers combined with norms that have been shaped by the multitudes of editors and researchers who have used threshold values over the past 100 years. The often-mentioned threshold *p*-values of 0.05 and 0.01 may be popular because they appear to give us a roughly optimal long-run trade-off between false-positive and false-negative errors made by a journal in the process of selecting papers to consider for publication.

12. What Use Are *p*-Values?

So, where does the *p*-value fit in the scheme of scientific research? Arguably, it has a small but useful role. Properly computed, it is a mathematically sensible measure of the *weight of evidence* that is inherent in a data table that a studied relationship between variables exists in the studied population. A scientific journal can specify that the *p*-value for the main result in a research paper must be less than (or equal to) the journal's threshold *p*-value before the journal will *consider* the paper reporting the result for publication. Using a threshold *p*-value as a gateway to publication helps the journal to optimally balance its long-run rates of published false-positive errors and unpublished false-negative errors. This, in

turn, helps to maximize the long-run scientific and social benefit of the research papers published in the journal.

I discuss these ideas in more detail in a book (2021).

Appendix A: Three Views of the Use of a Threshold *p*-Value

There are three popular views of the use of a threshold *p*-value in scientific research. Each of these views can be seen as a "decision procedure," with each view making a different type of decision.

First, in the view discussed in the present paper, the threshold *p*-value is used by a scientific journal to *decide* whether a paper reporting empirical research has enough weight of evidence for its main result to make the paper worth *considering* for publication in the journal.

A second view is that the threshold *p*-value somehow *decides*, or enables the researcher to decide, whether a relationship between variables (or some other studied effect) *exists* in the studied population. (In standard technical terms, the threshold *p*-value is thought to *decide* whether the "research hypothesis" or the "null hypothesis" in a research study is true.) This view is incorrect because a threshold *p*-value can't possibly decide whether a relationship exists because it sometimes makes false-positive and false-negative errors. This view is held by some people who have less direct experience with scientific research, so they aren't familiar with the occurrence of false-positive and false-negative errors in research.

If the threshold *p*-value (or a threshold for some other measure of the weight of evidence) doesn't decide whether a relationship between variables exists, then how is the decision made in science? The decision about whether a relationship between variables exists is *never* made *formally* in science because every scientific idea is open to revision. However, the decision is made informally, implicitly, and gradually by the relevant *research community*, as reflected in community members' written remarks, after the result has been believably indirectly or directly replicated in independent research.

A third view is that the threshold *p*-value somehow *decides* whether a research paper will be published in a journal. This view, though partly correct, is, on balance, incorrect because it gives the threshold *p*-value much more importance than it deserves. This is because though satisfying the threshold *p*-value condition is a *necessary* condition for publication in some scientific journals, it is never a *sufficient* condition, as discussed at the end of section 5 above.

Appendix B: The Existence of the Optimal Threshold *p*-Value for a Scientific Journal

The body of the present paper says that if a scientific journal uses an appropriate threshold *p*-value, then this maximizes the scientific and social benefit of the research papers that are published or are refused publication in the journal. This long and somewhat technical appendix gives a formal economics argument to show how the optimal threshold *p*-value for a journal exists.

We first consider the ideas graphically to illustrate the logic. Then we follow with a formal mathematical discussion of the simple ideas behind the graphs.

The argument consists of an extended thought experiment in which we pretend that we know certain things that we certainly don't know. The thought experiment uses ideas developed by Ioannidis (2005), Tabarrok (2005), and Jager and Leek (2014). We will see how the thought experiment reveals new facts.

Consider a group of 1000 randomly selected research studies in some field of scientific research that will be submitted to a journal (say, Journal A) in the field if they find good evidence of the relationship between variables they are looking for. That is, the report of each of these studies will be submitted to Journal A if the computed *p*-value for the main statistical test in the research is less than (or equal to) Journal A's threshold *p*-value of, say, 0.05.

From the perspective of a researcher, these 1000 research studies can be broken into two groups—the group of studies with a *positive* result for the main statistical test (i.e., $p \leq 0.05$), which will be submitted to the journal, and the group of studies with a *negative* result for the main test (i.e., $p > 0.05$), which *won't* be submitted to the journal (because they would be rejected).

From our theoretical perspective we can break the positive results into two subgroups—the true positive results and the false positive results. Similarly, we can break the negative results into two subgroups—the true negative results and the false negative results.

Based on sensible assumptions, the following discussion develops a mathematical model of the occurrence of the four types of results in the 1000 research studies that are candidates for Journal A. We include cost considerations in the model, taking direct account of the scientific and social costs of false-positive and false-negative errors. We use the model to demonstrate that a particular choice of the threshold *p*-value minimizes the total cost of the errors, which maximizes the scientific and social benefit of the papers published in the journal.

The argument demonstrates the *existence* of the optimal threshold *p*-value for Journal A though the argument can't tell us the numerical value.

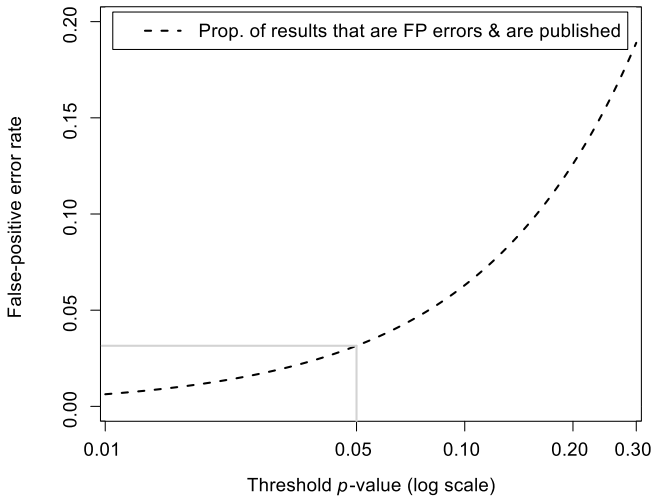

Figure B1. The published false-positive error rate versus the threshold *p*-value for Journal A. FP = "false-positive."

Figure B1 shows, for Journal A, the theoretical population rate of false-positive errors published by the journal as a function of the threshold *p*-value that is used by the journal. The horizontal axis of the graph shows a range of different possible choices for the threshold *p*-value between 0.01 at the left end and 0.3 at the right end. The axis uses a logarithmic scale to sensibly stretch things out at the lower end.

The vertical axis of the graph shows the proportion of the 1000 research studies in the field whose results are false-positive errors and are published in the journal. The dashed line on the graph shows that proportion for the different threshold *p*-values. For example, the light gray lines on the graph tell us that if Journal A uses a threshold *p*-value of 0.05, then the published false-positive error rate will be roughly 0.03 or 30 of the 1000 research studies.

The dashed line shows that if the journal uses a higher threshold *p*-value, then the rate of publication of false-positive errors will be higher.

Unfortunately, it isn't possible to *empirically* derive the correct version of figure B1 for a scientific journal. This is because, as a practical matter, we can't measure the rates of the false-positive errors in a journal under different threshold *p*-values. So, we can't know the exact shape or position of the line on the graph. However, we can model the line using mathematical principles and using reasoned guesses for the line's parameters, as discussed below.

We do know definitely that the line monotonically increases as the threshold *p*-value increases because the relatively-easy-to-understand *theory* of the *p*-value tells us that (assuming everything is done properly) the rate of false-positive errors made by a journal is in a monotonic increasing relationship with the journal's choice of the threshold *p*-value. This is because the higher the journal sets the threshold *p*-value, the more lenient the threshold is in allowing papers with weak evidence to be accepted for consideration. Papers with weak evidence are more likely to be intermixed with papers that are reporting false-positive errors. This is because false-positive errors are more likely if the threshold for a positive result is lenient and thus the threshold is easy to get past. Since more papers with *weak evidence* will be accepted for consideration, therefore more papers that are reporting *false-positive errors* will be accepted for consideration.

As noted above in section 7, due to the complexity of scientific research, editors and referees generally can't reliably distinguish between true positive results and false-positive results, so they generally don't try. Therefore, if a journal uses a *higher* threshold *p*-value, then since proportionately more papers with false-positive errors will be accepted for consideration, therefore proportionately more papers with false-positive errors will be *published*. Therefore, the false-positive error rate of papers published in a scientific journal is an increasing monotonic function of the threshold *p*-value used by the journal, as shown in figure B1.

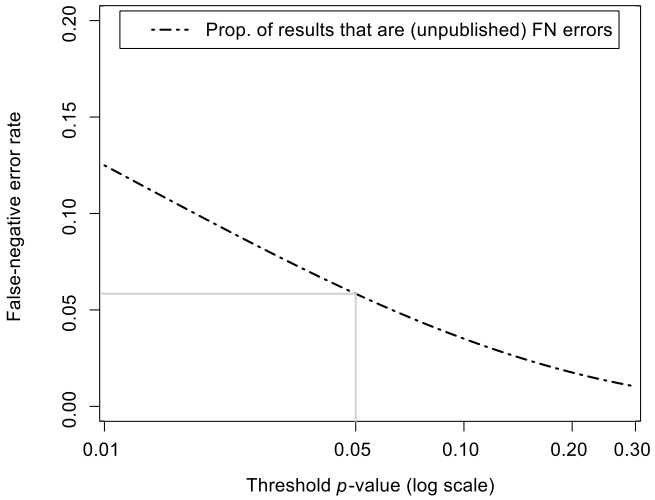

Figure B2. The false-negative error rate versus the threshold *p*-value for Journal A. FN = "false-negative."

Figure B2 shows, for Journal A, the theoretical rate of false-*negative* errors committed by the journal as a function of the threshold *p*-value used by the journal. That is, the horizontal axis is the same set of values of the threshold *p*-value as in figure B1 and the vertical axis is also the same, reporting the theoretical proportion of errors (false-*negative* errors in this case) in the 1000 research studies for the different threshold *p*-values.

The dot-dash line on the figure tells us that the proportion of the 1000 research studies that are reporting about *real* (i.e., extant) relationships between variables that were or might have been submitted to the journal *and* that were or would have been wrongly rejected because they failed to satisfy the threshold-*p*-value rule. For example, the light gray lines on the graph tell us that if Journal A uses a threshold *p*-value of 0.05, then the rate of false-negative errors that will wrongly be unpublished in the journal will be roughly 0.06 or 60 of the 1000 research studies.

The dot-dash line shows that if the journal uses a higher threshold *p*-value, then the rate of incorrect rejections of real results (i.e., the rate of false-negative errors) will be lower.

As with figure B1, we can't empirically derive the correct version of figure B2 for a scientific journal because, as a practical matter, we can't measure the rates of false-negative errors under different threshold *p*-values. So, as with the line in figure B1, we can't know the exact shape or position of the line on the graph for a journal. However, as with figure B1, we can model the line mathematically.

Similarly to figure B1, we know from theory that the line in figure B2 monotonically decreases as the threshold *p*-value increases. This is because, as noted, a higher threshold *p*-value is more lenient, which allows more true but weak results to be accepted for consideration for publication, which will reduce the rate of false-negative errors the journal makes. Therefore, the false-negative error rate of papers refused for consideration for publication in a journal is a decreasing monotonic function of the threshold *p*-value used by the journal, as shown in figure B2.

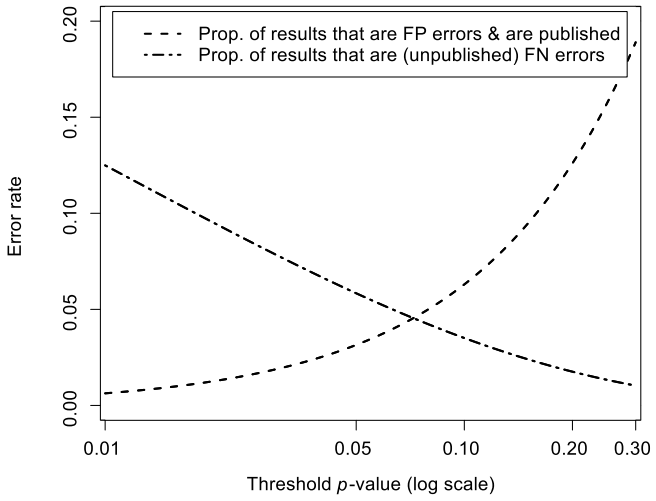

Figure B3. Figures B1 and B2 overlaid.

It is helpful to plot the lines on figures B1 and B2 overlaid on a single graph, which yields figure B3. Overlaying the lines is sensible because both lines pertain to the same 1000 research studies.

A key idea associated with figure B3 is that the false-positive and false-negative errors shown by the two lines on the figure have scientific and social *costs* associated with them. That is, every false-positive error has a cost in terms of wasted resources that will be used to try to replicate or use the false result, with every individual error having a (differing) cost. Similarly, every false-*negative* error has a scientific and social cost in terms of lost information about a new and possibly useful relationship between variables, again with every error having a (differing) cost. For technical reasons, we can't measure the error costs, but we do know that if the error rates go up, then the total cost of the errors obviously also goes up.

However, let us suppose in this thought experiment that we *can* measure the costs of both false-positive and false-negative errors in Journal A. This will allow us to convert the error lines on figure B3 into cost lines, as shown on figure B4.

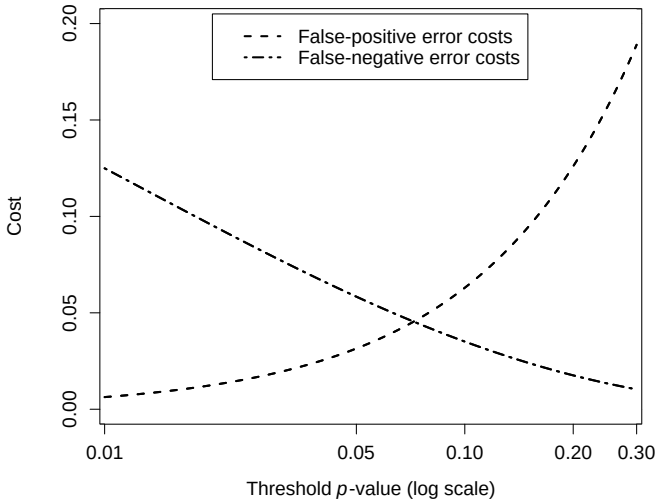

Figure B4. The costs of false-positive and false-negative errors in Journal A.

Note how the vertical axis label on figure B4 isn't "Error rate" but is "Cost," though the numbers on the axis are unchanged because this discussion is hypothetical, so we can view the numbers on the Cost axis as merely a relative scale.

Reflecting the simplest case, the two cost lines on figure B4 have the same shapes as the two associated error lines on figure B3. This is because it is sensible to assume that the overall scientific and social *cost* of each type of error is directly proportional to the *rate of occurrence* of that type of error. Figure B4 shows this simple case.

However, if we believe that the costs of the errors are more complicated than direct proportions or if we believe that the cost of a false-positive error is different from the cost of a false-negative error, then we could adjust the lines on figure B4 to take account of those facts. That would be relatively easy to do if we knew the correct lines and if we knew the correct costs, which in this thought experiment we assume we know. These adjustments would change the relative positions of the two lines on the graph but they wouldn't change the fact that the two lines cross on the graph in the form of an *X*, which is the important point for the present discussion.

So, after we have made any necessary adjustments to figure B4 to make it reflect the proper costs, we can *add together* the two costs at individual threshold *p*-values on the horizontal axis, which gives us the *total* cost of the false-positive and false-negative errors for each threshold *p*-value. (This addition is permissible because the original two error proportions behind the cost lines were computed based on all research studies in the same relevant group of 1000 research studies that might be submitted to Journal A.) Then we can plot the *sum* of the costs of the two types of errors on the graph. The solid line on figure B5 shows the sum of the two costs.

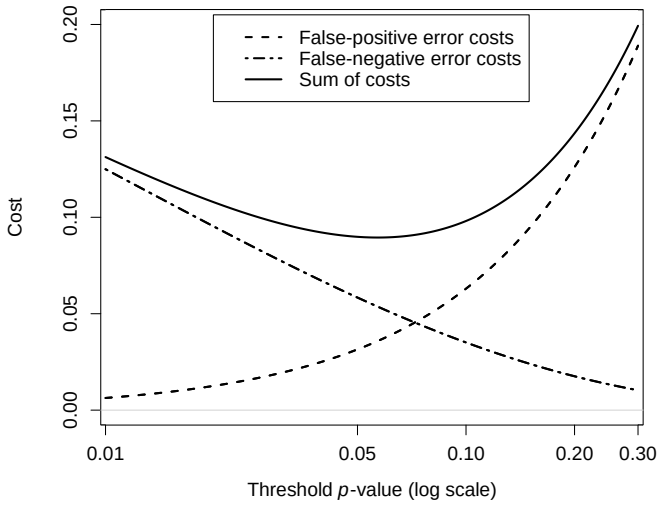

Figure B5. Figure B4 with a line added showing the sum of the costs of the two types of errors for Journal A.

In figure B5, either with a ruler or with measurement by eye, it is easy to see that the height of any point on the curving solid line on the figure is the sum of the heights (above the horizontal zero line) of the dashed lines at points that are vertically directly below the point on the solid line. For example, if you carefully measure the vertical heights of the three lines at 0.05 on the horizontal axis, you will see that the height of the solid line is exactly equal to the sum of the heights of the two dashed lines.

Note how the solid line is shaped like a bowl. The bowl has, in effect, "fallen into" the notch between the two cost lines. Of course, the lowest point on the bowl is the point where the sum of the costs of the two types of errors has the lowest possible value.

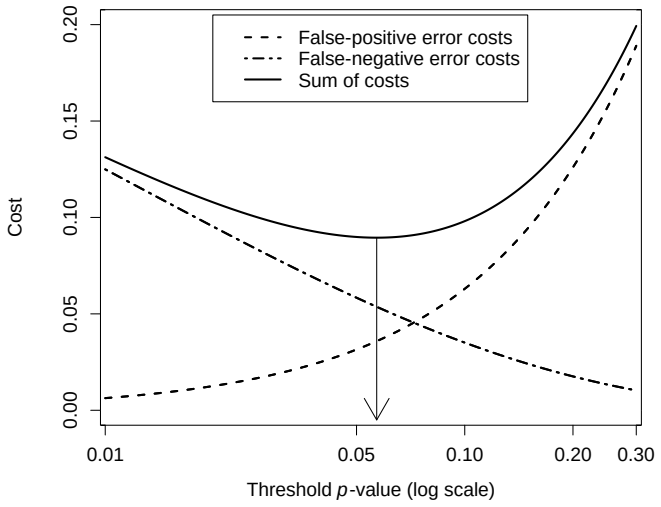

Figure B6. Determining the optimal threshold *p*-value for Journal A.

Figure B6 shows the step of drawing a vertical arrow from the lowest point on the bowl to the horizontal axis to identify the optimal threshold *p*-value for the journal—the value that gives us the lowest sum of the costs of the two types of errors. Thus, on the figure, we see that the optimal threshold *p*-value for hypothetical Journal A is around 0.06.

Figure B6 shouldn't be interpreted as suggesting that the optimal threshold *p*-value for a journal is roughly 0.05. This is because it is easy to move the minimum point of the solid line on the figure far to the left or far to the right by changing the values of the parameters of the algorithm that generates the figure. Instead, the figure illustrates how things work and illustrates that that the optimal threshold *p*-value for Journal A *exists*, as defined by the point on the horizontal axis where the perpendicular arrow dropping from the lowest point on the bowl points.

Of course, the validity of figure B6 for showing that the optimal threshold *p*-value exists depends on the validity of the algorithm that was used to generate figures B1–B6. Let us now consider the algorithm in terms of the underlying mathematics. We consider the math in terms of the computer program that generates the data behind the figures because the program precisely defines the algorithm.

The program is written in the SAS programming language though you needn't understand that language to understand the following discussion. The program can be converted to any other standard programming language (assuming the language has the required statistical functions) and it will give exactly the same lines on the graphs.

In SAS, if we wish to generate or manipulate data, we use a "data" step, which consists of multiple lines of SAS code. The first line is a DATA statement that names the data set that we will work with—we will name our data set "GraphData." Here is the statement:

```
data GraphData;
```

Next, we set the values of five variables that the program will use when it is run. We can later change these values and then rerun the program to study the behavior of the program under different conditions. We use a RETAIN statement to tell SAS to "retain" the values over consecutive "passes" through the data step because otherwise SAS will (sensibly) set the values to "missing" at the beginning of each pass:

```
retain PctTrue 30 PosPctPub 90 ntotal 60
    meandiff 20 stddev 27;
```

We discuss and use the values of the preceding five variables in due time below.

Figures B1–B6 all have the same range of threshold *p*-values on the horizontal axis, running between 0.01 and 0.3. So, to draw the figures, we must generate data at different threshold *p*-values between 0.01 and 0.3. We do this with a "DO loop" in the program, which is a set of multiple SAS statements in a data step and which, in the present case, begins with the following statement:

```
do ThreshP = 0.01 to 0.3 by 0.001;
```

SAS will execute the code that follows this DO statement down to a matching END statement once for every value of ThreshP between 0.01 and 0.03 (inclusive) using the increment of 0.001 to step between values. Of course, each time SAS executes the statements, the variable ThreshP will have the appropriate value and will be available for use in the computations.

As noted, we assume we are studying 1000 research studies in some field of scientific research that might be submitted to Journal A. Each of these studies is studying a relationship between variables. As specified in the variable PctTrue in the RETAIN statement, we assume that 30% of the research studies are studying a true relationship between variables—that is, the research hypothesis that is under study is true. Thus, the other 70% of the research studies are, unfortunately for the associated researchers, studying a situation in which the postulated relationship between the variables *doesn't* detectably exist in the studied population—that is, the *null* hypothesis is true (or is in effect true).

Of course, the null hypothesis expresses the idea that there is *no relationship* between the variables. The idea of the null hypothesis being "in effect true" enables us to take account of relationships between variables that exist but are too weak to be detected with present-day research. We take account of such relationships by noting that though the null hypothesis isn't *exactly* true in these cases, it is *in effect* true in the sense that the existence of the relationship is empirically indistinguishable from the null hypothesis with present-day measuring instruments and with affordable sample sizes.

The borderline in scientific research between (a) when a null hypothesis is true or in effect true, and (b) when the null hypothesis is false is fuzzy, which is of philosophical interest. However, the fuzzy borderline doesn't cause practical problems. This is because the nature of the situation implies that the borderline is always outside the range of our measuring instruments, so we needn't be concerned about the precise distinction between the two cases in an empirical sense.

The estimate that only 30% of the research studies in a field of science are studying true relationships between variables may seem low to some readers, but it likely won't seem low to many people who do day-to-day research. These people know that many research studies are performed that yield negative results, so the studies are abandoned, and nothing is said or written about them because they are uninteresting relative to positive results. And the negative results are slightly embarrassing because they "failed" to find what the researcher thought was likely present.

For readers who think 30% may be too low, if we rerun the present program with PctTrue set at, say, 60% instead of 30%, the program will produce a somewhat different version of figure B6, but there will be no change to the general bowl *pattern* on the graph and no changes to the key points implied by the graph.

As noted, we are working with 1000 research studies, with each study studying a possible relationship between variables. Consider the following two lines of code that immediately follow the DO statement:

```
nTrueRels = 1000 * (PctTrue / 100);
nFalseRels = 1000 - nTrueRels;
```

The first line tells us how many true relationships between the variables we have in the 1000 research studies. For example, if PctTrue is 30, then we will have 300 true relationships and, as computed in the second line, we will have 700 false relationships.

The next line of code tells us how many of the 1000 results will be false-positive results according to the current value of the threshold *p*-value:

```
nFalsePos = ThreshP * nFalseRels;
```

The preceding line of code is correct based upon the definitions of the *p*-value and the threshold *p*-value. By definition, the threshold *p*-value for a journal is the fraction of the time that the *p*-value for the main result in a paper submitted to the journal will be less than the journal's threshold *p*-value in the set of cases when *there is no relationship* (or no *detectable* relationship) between the variables in the population (and if certain often-satisfied assumptions are adequately satisfied). So, if we have 700 cases of no relationship between the studied variables, then the number of these cases in which the *p*-value will be less than the threshold *p*-value of 0.05 is estimated as $0.05 \times 700 = 35$.

Since the 35 cases of false-positive results have *p*-values less the journal's threshold *p*-value, and because researchers are almost always unaware that a positive result is a *false*-positive result, and because researchers are eager to have their research papers published, we can assume that the 35 studies containing the false-positive results will be submitted to Journal A. However, not all positive results submitted to a reputable scientific journal are published because journals have other standards that a paper must satisfy in addition to the threshold-*p*-value standard. We assume that the percentage of positive results submitted to Journal A that are published is specified in the variable PosPctPub, as given in the RETAIN statement above. Therefore, the number of false-positive results that are published in Journal A is computed as

```
nFalsePosPub = nFalsePos * (PosPctPub /
    100);
```

For example, if PosPctPub is 90 (as specified in the RETAIN statement), and if 35 false-positive results are submitted to the journal, then 35×0.9 or roughly 32 of them will be published. The two grey lines on figure B1 reflect this case in terms of an 0.032 proportion of the 1000 studies.

Now, working with 1000 research studies was an assumption to make things easier to understand, but this assumption is restrictive and unnecessary, so we can change from the *count* of the research studies with false-positive errors to a more general *proportion* with the following line of code:

```
PropFalsePosPub = nFalsePosPub / 1000;
```

At this point, the program has generated a single line of data of the multiple lines of data that we need to draw figure B1. Of course, the two *variables* in the data line that we need to draw the figure are ThreshP (plotted on the horizontal axis of the graph) and PropFalsePosPub (plotted on the vertical axis).

The next line of code tells SAS to write the values of all the variables in the line of data into a new row of data in the GraphData data set that is being created:

```
output;
```

We end the DO loop at this point with the following statement:

```
end;
```

The END statement tells SAS to go back to the DO statement above and execute the lines of code in the loop again, using the next value of ThreshP (unless, of course, ThreshP has passed 0.3). This will write the next row of data into the data set, and so on until all the rows of data are written. If we run the program, SAS tells us in the "log" of the run that it wrote 291 rows of data into the GraphData data set, which is the correct number given the specifications of the DO statement above.

As noted, the 291 values of ThreshP and PropFalsePosPub in the GraphData data set are the values we need to draw figure B1. Thus, we need only give the GraphData data set to a graph-plotting program and give it a few simple instructions and it will draw figure B1.

(The program to generate the data, including the code to draw the figure and the output from the program, is available in the Supplementary Information for this paper. For interested readers, instructions with the program explain how to change the parameters of the program and rerun it using free SAS software.)

It is important to distinguish between

- the proportion of research studies in a field that are reporting false-positive errors and that are published in a journal and
- the proportion of research studies that are published in a journal that are reporting false-positive errors.

It is the first of the above two proportions that is plotted on the vertical axis of figure B1. The second proportion, which is directly related to the first, is always higher than the first due to the underlying mathematics and due to the fact that (with rare exceptions) only positive research results are published in scientific journals. The second proportion is discussed below in appendix C.

Let us now add code to the program to generate the data for figure B2, which we do by adding six more lines of code to the program, adding the lines immediately before the OUTPUT statement. These lines generate a new variable, PropFalseNeg, which tells us the proportion of research results in the field that are false-*negative* results. The values of the PropFalseNeg variable together with the values of the ThreshP variable are the data behind figure B2.

As with figure B1, in generating the data for figure B2, we begin by working with 1000 research studies that might be submitted to Journal A because this point of view is easy to understand. However, in this case, the 1000 research studies are quite different from the 1000 research studies in figure B1. In generating figure B1, we considered 1000 *different* research studies. To generate figure B2, we consider the case in which we repeat *exactly the same* research study 1000 times, each time collecting data from a fresh sample of entities from the (same) studied population. The sensibility of this approach will become clear later below.

For a simple concrete example, consider the study of a relationship between a "binary" predictor variable and a "continuous" response variable. For example, suppose we are medical researchers, and we wish to test whether a new blood-pressure drug lowers the blood pressure in patients with high blood pressure.

Because it is efficient, we decide to use two doses of the drug in our experiment, which are a zero dose and a high but safe dose. So (using a placebo and appropriate medical "blinding"), we randomly assign the two doses to suitable volunteer patients and, after sufficient time for the drug to show an effect, we measure the *drop* in each patient's blood pressure from *before* the patient received the drug or placebo until *after* they have received it. Of course, we wish to know whether the high dose of the drug yields a significant change in the response variable, i.e., a significant drop in each patient's blood pressure in the patients who received the drug relative to the patients who received the placebo.

So, we compare the average drop in blood pressure in the patients who received the drug with the average drop (if any) in the patients who received the placebo. In this case, the accepted way to compute the relevant *p*-value is with the "two-sample *t*-test," which is the most powerful standard statistical test for evidence of a relationship between a binary predictor variable (e.g., drug dose with two levels) and a continuous response variable (e.g., drop in blood pressure).

(The *t*-test is applicable if certain often-satisfied assumptions are adequately satisfied, which we assume are satisfied in the example. For completeness, it is noteworthy that we could also compute the *p*-value using the "before" and "after" blood pressures for each patient and using the "paired" *t*-test, but this would give us exactly the same *p*-value.)

We assume that each of our 1000 research studies compares two groups of 30 patients for a total of 60 patients. Each study uses exactly the same procedures and then performs a two-sample *t*-test for the difference in the drop in blood pressure between the two groups. The only difference between the studies is that in each study we obtain a fresh random sample of patients from the population.

We also assume that we know the correct values of the parameters of the relationship between the two variables, as follows: We assume that the blood-pressure drug is effective

in 300 of the cases (as dictated by PctTrue), and in these cases the population mean difference of the response variable between the two groups under consideration is 20 units (e.g., millimeters of mercury), and the population standard deviation of the mean difference is 27 units. In the other 700 cases, we assume that there is no relationship between the two variables, so the population mean difference between the two groups is zero.

So let us add code to the program to simulate this very specific case. Recall that the key numbers in the preceding three paragraphs (i.e., 60, 30, 20, and 27) are all known to the program because they are all given in the RETAIN statement above.

In this specific situation, to draw the false-negative-error line on figure B2, we need to determine how many of the 1000 instances of the two-sample *t*-test will be false-negative errors. So, we need to count the cases when the relationship between variables of interest is present, but the research study fails to detect the relationship—i.e., the computed *p*-value is greater than the threshold *p*-value, which implies a false-negative error.

(Of course, the researcher is generally unaware that a certain negative result is a *false*-negative result because there is generally no way to know that apart from doing appropriate further research.)

In generating figure B2, we must also take account of the *strength* of the relationship between the variables because if a relationship is weak, then a false-negative error is more likely to occur than if the relationship is strong. In the *t*-test case that we are studying, we know the strength of the relationship from the information given above, and the strength is the *same* in every one of the 300 positive cases because we are doing exactly the same research studying exactly the same relationship between variables in the same population in every case. Of course, this greatly simplifies taking account of the strength. In general, though, the strength of the relationship between variables under study differs from one research study to the next, so we must take account of that fact, which we do later below.

The strength of the relationship between variables we are expecting to find in the *t*-test example is encapsulated in the two parameters of the relationship, with values 20 and 27. Since we know the strength of the relationship, this enables us to compute the "power" of the statistical test in the 1000 research studies. We will use the measured power as a key to drawing figure B2.

The "power" of a statistical test is the fraction of the time that the test will detect the studied relationship between the variables if certain sensible conditions are satisfied. The conditions are that

- we specify the form of the relationship between the variables in a way that enables us to compute the power (as we have done in the *t*-test example)
- we specify the design of the research study (as we have done in the *t*-test example)
- we use a particular specified threshold *p*-value, such as 0.05, and
- everything is done according to certain sensible rules, as explained in statistics textbooks.

We assume that the four conditions are satisfied in our 1000 research studies, though we won't use a single threshold *p*-value but will instead perform the computation with each of the 291 different threshold *p*-values to enable us to generate figure B2.

Since statistical power is the fraction of the time that a research study will detect the specified relationship between the variables, the power of a statistical test for detecting a relationship always lies between 0 and 1 (just like the values of the *p*-value always lie between 0 and 1). Ideally, a statistical test in a scientific research study should have a power of at least 0.8 for the relationship between variables that it hopes to detect because that gives the research study a good chance of detecting the relationship—it will successfully detect the relationship 0.8 of the time if the relationship has the specified strength.

An obvious question a reader might ask here is why researchers don't design research studies with a power of, say, 0.99 or even 1.0. The answer is that sensibly performing a research study with such high power would be very expensive, so the researcher must always trade statistical power against research cost. In view of this trade-off, statisticians have invented research designs that can substantially increase the power of statistical tests while only minimally increasing the costs.

We can compute the power of a statistical test using the standard theory of statistical power, which is straightforward though complicated in the mathematical details. Fortunately, for the present discussion, we needn't consider the details behind the computation of power. Instead, we need only note that power is the fraction of the time that the research study will detect the specified relationship under the specified conditions, as discussed above.

To compute the power of a statistical test, we substitute the values of the parameters of the relationship and the required specifications of the research design into the appropriate power equations and then we evaluate the equations to determine the power. Power equations to do this computation are described in statistics textbooks about power and are available for all the standard statistical tests of relationships between variables.

So, in the present example, we use the variable meandiff to tell the statistical power equations that in the 300 positive cases the population mean difference between the two groups in the *t*-test is 20 units, we use stddev to specify that the population standard deviation of the mean difference is 27 units, we use ntotal to tell the equations the number of entities in the two groups is 60, and we use ThreshP to specify the threshold *p*-value that is currently under consideration in the DO loop. Then the power equations ingest these numbers and determine the power. For example, if we use the numbers above and if

the threshold *p*-value under consideration is 0.05, then the power equations for the two-sample *t*-test tell us that the power of this statistical test is roughly 0.805 if the relationship has or were to have the specified mean difference and standard deviation.

Here are the highly obtuse three lines of SAS code that we add to the program to specify the power equations to compute the power of the two-sample *t*-test under the specified conditions:

```
Ncp = ntotal * 0.5 * 0.5 * meandiff**2 /
    stddev**2;
Critval = finv(1-ThreshP, 1, ntotal-2,
    0);
TestPower = sdf('f', Critval, 1, ntotal-
    2, Ncp);
```

You needn't understand the preceding three lines, and you need only understand that the third line properly assigns the power of the test in the situation under study to the TestPower variable according to the current value of ThreshP in the DO loop. However, for readers who are curious, the three lines of code are copied from a web page about computing the power of the two-sample *t*-test published by SAS (2021), with links to further references to standard theoretical discussions of statistical power.

In the present discussion, we view the above three lines as a black box that correctly computes the power of the two-sample *t*-test if we give the three lines the values of all the variables that appear on the right-hand side of the equals signs in the three lines.

So, if we execute the preceding three lines of code (using the values of ntotal, meandiff, stddev, and the current value of ThreshP), we obtain the value of TestPower, which we use to help to draw figure B2. In particular, using TestPower, we can compute the number of true positive results in the 1000 research studies by multiplying the number of true relationships (computed earlier) by the power, as follows:

```
nTruePos = nTrueRels * TestPower;
```

Then we can compute the number of false-negative results as

```
nFalseNeg = nTrueRels - nTruePos;
```

Finally, we convert the number of false-negative results to a proportion:

```
PropFalseNeg = nFalseNeg / 1000;
```

This completes the code to compute the data needed to draw figure B2. Note the relative simplicity of the computation. We discuss the generalization of the figure in a moment.

Of course, figure B3 is simply figures B1 and B2 overlaid, and figure B4 is simply figure B3 with different labeling.

For figures B5 and B6, we need to compute the sum of the costs of the false-positive and false-negative errors to enable plotting the black line on the two figures. We do this with one more line of code in the program immediately before the OUTPUT statement. For this line we assume that false-positive and false-negative errors are equally costly. We can easily change this assumption by including a multiplicative factor (or other sensible function) for one or both of the terms on the right-hand side of the equals sign:

```
Cost = PropFalsePosPub + PropFalseNeg;
```

So, if the preceding program is run, it will generate all of the data needed to draw figures B1–B6.

Figures B1–B6 illustrate the concepts. However, figure B6 is the correct graph for Journal A only if all the research studies in the field are exactly like the research study that was used (i.e., with a two-sample *t*-test and for the positive cases using the specific values of the parameters given in the RETAIN statement). But, of course, the research studies in any field are all different. So, we must generalize the preceding discussion.

We can easily do that in our thought experiment by using the program above to draw a correct version of figure B6 for *each research study in a field of science*. That is, for each research study, we can change the two-sample *t*-test program code above into the correct code for that study. For each study, we can do this with two steps:

1. Replace the three lines above that compute the value of the TestPower variable as a function of the threshold *p*-value with the lines of code that are appropriate to compute the test power as a function of the threshold *p*-value for the new research study.
2. Insert the correct values of the parameters referenced in the new lines of code in the RETAIN statement.

This is all we need to do to generalize the program to any other specific research study because the three lines of code and the code to set the values of the parameters are the only lines that are unique to the *t*-test case and the other lines of code in the program operate at a more general level.

Appropriate lines of code to compute the value of TestPower are available for any sensible research study either through statistical theory or, if the relevant theory isn't available, through an appropriate computer simulation. Thus, for any research study, we could in theory insert the appropriate values of the parameters of the model equation of the studied relationship and the parameters of the research design into the RETAIN statement at the beginning of the program and we could insert in the DO loop the appropriate lines of code to compute the power. Then the program would compute the path of the false-negative cost line behind figure B6 as a function of the threshold *p*-value for that research study and thus we could use the data from the program to draw the correct version of figure B6 for that study.

Thus, we could in theory generate a version of figure B6 for every research study performed in a field of science, which might lead to tens of thousands of graphs. Of course, in generating these graphs for the different research studies, we must specify (in the RETAIN statement) the true values of the relevant parameters of the model equation of the relationship

between variables under study though, of course, we invariably don't know the true values. (Generally, the purpose of a research study is, in part, to *determine estimates* of these values.) However, in this thought experiment we can assume that we know the true population values of the parameters because this advances the argument without harming the argument's validity, as we shall see.

Technical note: The *true population values* of the parameters of a specified model equation of a relationship between variables are sensibly defined as the parameter estimates that we would obtain if we were to perform appropriate empirical research to derive estimates of the parameter values, and if we were to use *perfectly accurate* and *perfectly precise* measuring instruments to measure the values of the relevant variables, and if we were to use a sample that includes *every entity in the population*. Thus, conceptually, the true population values are empirically obtainable for any population and any model equation, though obtaining the values would generally be impossibly expensive. So, in practice, the true population values of parameters of model equations aren't knowable, but they are estimable through appropriate research.

Note how the point of view has changed from the view of performing the same research study with a two-sample *t*-test 1000 times to the view of performing *every* research study that might be submitted to Journal A 1000 times. We imagine performing these multiple sets of 1000 research studies to set the scene. Then, for each of these research studies, we imagine running the above program simulating the study (with the proper modifications) to draw the relevant version of figure B6.

In each of the tens of thousands of cases, we will obtain a graph that is similar to figure B6. The false-positive line will always be the same because, as discussed above, the false-positive line doesn't depend on properties of the research studies but depends only on two properties of Journal A—the percentage of research hypotheses that are true in Journal A's field (e.g., 30%) and the percentage of positive results that are submitted to Journal A that are published (e.g., 90%).

In contrast, the false-*negative* error line, though it will always be monotonically decreasing, will move around from graph to graph. That is, the slope and position of the line will change depending on the effect size under study (which will sometimes be zero or essentially zero) and will change depending on the properties of the statistical test. (The left and right endpoints of the false-negative error line are always fixed at particular values, as explained and illustrated in the computer program output BowlGraphFinal.pdf in the supplementary information.) The fact that the false-negative line moves around implies that the optimal threshold *p*-value for each specific research study (as indicated by the lowest point on the bowl on each graph) will generally be different from graph to graph.

Will all the tens of thousands of graphs be similar to figure B6 in the sense of having a bowl with a minimum point? Yes. This is because (assuming a non-zero effect) conceptually in each case all that is changing from graph to graph is the form of the monotonically increasing relationship between the threshold *p*-value and the power of the test (as specified by the power equations and by the parameters of the situation). And, regardless of the exact form of this relationship, it will (because it is monotonically increasing) generate a monotonically decreasing relationship between the threshold *p*-value and the false-negative error rate, as illustrated in the specific case in figure B2.

Thus, the geometry of the situation implies that when the false-positive and false-negative error costs (as computed from the error rates) are added together, the sum will be shaped like the bowl in figure B6 though, as noted, the lowest point on the bowl will generally be different from graph to graph. The computer output from BowlGraphs.sas in the supplementary information illustrates some different bowls.

For the purpose of the present discussion, a sensible way to interpret the tens of thousands of graphs is: The vertical arrow on each graph indicates the optimal threshold *p*-value for Journal A if all the research studies in the field were the same as the study behind the graph.

So, after we have generated the tens of thousands of imaginary graphs, let us imagine computing the "average" of the optimal threshold *p*-values shown on the graphs. Arguably, the "average" of the minimum points on the bowls for all the graphs defines the optimal threshold *p*-value for the journal because this value minimizes the sum of the overall costs of the two types of errors across all the papers submitted or potentially submitted to the journal.

We might wonder which measure of central tendency we should use to compute the average—whether it should be a simple mean or some other function of the optimal threshold *p*-values from the tens of thousands of graphs, possibly even *weighting* each result to reflect its importance. Of course, our goal here would be to choose the averaging function that best minimizes the sum of the costs of the errors across all the research studies in the field.

However, as a practical matter, the issue of the precise way to compute the average is less important because we can't compute the average in practice because we don't know nearly enough for that. The important point is that the thought experiment implies that there *is* a sensible optimal average threshold *p*-value for a journal though the experiment can't tell us what the value is.

It is noteworthy that many studied relationships between variables don't exist (or at least don't *detectably* exist) in a population, and thus the null hypothesis is true (or in effect true) in these cases. If we run the preceding program for a research study that is studying one of these nonexistent relationships, and if we correctly tell the program that the effect size in the study is zero or very close to zero, then the program will correctly tell us that the optimal threshold *p*-value in this case is zero or very close to zero.

The threshold *p*-value of zero is intuitively sensible for these studies (which are limiting cases) because if any one of

them reports a positive result, then it will be a *false*-positive result, and a threshold *p*-value of zero will correctly prevent the false-positive result from being published. (A false-*negative* error can't happen in these cases because the relationship doesn't exist.) This leads to the question of how to handle the averaging discussed above in the cases when relationships between the variables don't exist in the population which, though we never know this to be the case in practice, we do know in the thought experiment. Might the preponderance of these cases somehow improperly disturb the balance? Seemingly not, because the algebra seems correct. However, we don't need to deal with this problem because we don't intend to do the averaging because we can't do it in practice.

Technical note: The preceding two paragraphs are correct if the percentage of research hypotheses that are true in Journal A's field—PctTrue—is less than 50%, which seems likely the case in most fields of science because nature's secrets are hard to unlock. So, researchers are generally correct in their research hypotheses less than half the time. However, if PctTrue is greater than 50% in some field of science, and if the effect size in a particular research study is zero or very close to zero, then the program tells us that the optimal threshold *p*-value for this case is 1.0 or very close to 1.0. This somewhat surprising outcome is a consequence of the underlying mathematics behind minimizing the sum of the costs of the errors and is illustrated in the output from the program BowlGraphs.sas in the Supplementary Information. In this case, the bowl is truncated.

It is easy to imagine refinements that would make the mathematical model discussed above more closely resemble real scientific research. For example, the argument assumes that a certain percentage of the research hypotheses in the field are true (PctTrue) and the remainder of the hypotheses are false and thus the null hypothesis is true (or in effect true) in these cases. In reality, there is no hard borderline between true research hypotheses and false research hypotheses and instead there is a continuum, which we might model. However, it seems likely (though not certain) that every realistic refinement will lead to total-cost bowls with minimum points, and the "average" of the minimum points across all the research studies in the field sensibly defines the optimal threshold *p*-value for the journal, which is the main point of this discussion.

The thought experiment tells us that the optimal threshold *p*-value for a scientific journal exists but, as noted, the experiment doesn't tell us the *value* of the optimal threshold *p*-value for a journal. So, a journal must use another method to determine its optimal threshold *p*-value. As noted above in section 11, editors and researchers determine the value based on experience, intuition, and norms. This approach is sensible because editors and researchers generally agree that it works well, and nobody has thought of a better approach.

Appendix C: The Rate of Publication of False-Positive Errors in Scientific Journals

For technical reasons, it is difficult to measure the ongoing rate of occurrence of published false-positive errors in a field of science. However, in carefully performed near-exact replications of 21 important research results in social science, replication failures occurred 38% of the time, that is, in 8 of the 21 studies (Camerer et al. 2018). This and other direct replication research suggests that somewhere between 20% and 60% of the published positive results in social-science journals are *false*-positive results.

The high rate of false-positive errors isn't limited to social-science research and is also recognized in biomedical research (Ioannidis, 2005; Errington et al. 2021). False-positive errors are also likely present in the physical sciences, though they aren't well documented.

When some people hear about the high rates of false-positive errors in scientific research they are either alarmed or embarrassed—thinking that this state of affairs may be a crisis. However, there is no need for alarm or embarrassment, and there is no crisis because the false-positive errors are normal science—there are false-positive errors in the scientific research literature because they are unavoidable if we wish to minimize the sum of the costs of false-positive and false-negative errors.

Appendix D: When Do We Need to Use Statistical Significance in Scientific Research?

In general, we can use the concept of statistical significance whenever we wish to determine whether there is good evidence that a relationship (or an extension to a known relationship) exists between variables in a population. However, some fields of science, especially in the physical sciences, don't regularly use the concept of statistical significance because they usually study *strong* relationships between variables. In this case, the researcher and the journal don't need a formal system like statistical significance to confirm that there is good evidence of a relationship. This is because merely looking at an appropriate graph of the data for a strong relationship will tell an experienced researcher that (assuming that the underlying research is done correctly) the relationship definitely exists. And the graph will also imply that the computed *p*-value for the relationship would be very low if it were computed. So, in this case, the researcher needn't compute a *p*-value to measure the weight of evidence and the journal needn't consider the concept of statistical significance.

Also, some research studies have only a single entity from the population in their sample, so the research paper can describe the entity, but the paper can't report about the study of a relationship between variables because you need multiple entities in a sample to study a relationship in a population. So, these studies also don't need the concept of statistical significance.

So, a journal only needs to use the concept of statistical significance if it is evaluating a paper that is reporting about

a weak relationship between variables in a population. However, nowadays the situation occurs often because most of the strong relationships have already been discovered.

Appendix E: What If a Research Study Computes Many *p*-Values?

The discussion in the body of this paper says that the threshold-value gateway to publication in a journal relates to the *main* result in a paper submitted to the journal. (In medical research, the main result is often referred to as the "primary outcome.")

What happens in a research study when there is *no* main result, and the study is reporting the results of study of multiple relationships between variables, thereby possibly with multiple *p*-values, such as 5 *p*-values or 50,000 *p*-values?

This question is important because increasingly in modern research, researchers find that it is sometimes cost-efficient to simultaneously study multiple closely related relationships between variables. In this case, for technical reasons, the researcher can expect some very low *p*-values even when the corresponding relationships between variables don't exist. So, false-positive errors are easy to make. So, the researcher must take proper account of this fact. Various sensible statistical methods are available to handle research studies that simultaneously study multiple relationships between variables, such as the approach discussed by Benjamini and Hochberg (1995). These methods help researchers, editors, and journal readers to decide whether the studied relationships between variables likely exist in the population.

Of course, these methods are sensible logical extensions of the basic methods to study a single main relationship between variables. A knowledgeable editor will require that a researcher studying multiple relationships between variables properly use one of these methods.

Appendix F: What If Journals Stop Using Thresholds as Gateways?

If a journal stops using the concept of statistical significance as a gateway to publication, then it will no longer automatically reject papers that are reporting weak results. Therefore, papers with weak results will be submitted to the journal that would have been rejected before. Most of the papers reporting weak results will cast their results as being *positive* results because that is generally easy to do if there is no threshold-value requirement and because positive results are more interesting than negative results because if they are correct, they lead to reliable prediction or control. Since the evidence in these papers is weak, a large proportion of these claimed positive results will be *false*-positive results.

As noted in section 7, a paper reporting a *false*-positive result generally looks no different from a paper reporting a *true* positive result. So, a large proportion of the false-positive results will be published. This will make the so-called replication crisis worse, which is scientifically and socially inefficient.

Due to the opposition to threshold *p*-values by some statisticians (Wasserstein, Schirm, and Lazar 2019), scientific journals are now less likely to use a *formal* threshold value for a measure of the weight of evidence. However, it seems unlikely that journals will ever abandon the concept because the concept plays what many editors and researchers recognize is a useful role. Certainly, the opposition to statistical significance by some groups may drive the concept underground in some journals. But the concept will likely always be present because experienced editors and researchers know they must guard against false-positive errors—guard against being fooled by mere noise in the data. And a sensible objective way to do that is to formally or informally use a sensibly chosen threshold value for a measure of the weight of evidence that a relationship between the studied variables exists.

Appendix G: Could Journals Use Discretionary Threshold Values?

The body of this paper suggests that a journal can use a single fixed threshold *p*-value, such as 0.05. It is instructive to consider an alternative approach in which the editor sets the threshold *p*-value for papers in the journal on a paper-by-paper discretionary basis. For example, an editor might decide to use a *stricter* threshold if a research finding is highly important because this will increase the chance that the finding is correct. Or an editor might decide to use a *more lenient* threshold if a research finding is highly important, but the evidence is weak, thinking that the information (being important) should be published even though the evidence is weak.

The discussion of "importance" in the preceding paragraph is a converse of the discussion of "importance" in section 5 above.

A journal editor is always free to use discretionary threshold values because he or she (together with the editorial board and the publisher) is the final authority in the operation of the journal. However, setting the threshold on the basis of the importance of the studied effect (i.e., a studied relationship between variables) is intermixing two quite different concepts that are arguably better kept separate. Keeping the *weight of evidence* for an effect separate from the *importance* of the effect makes these concepts easier to understand in a divide-and-conquer sense.

Also, it is difficult and time consuming to use discretionary threshold values because it is difficult for an editor to know the *ramifications* of an effect, which makes it difficult to know the importance of the effect, which makes it difficult to set the proper threshold value. This difficulty arises because an interesting positive result may turn out to be a false-positive result or may turn out to be unusable for some other reason, such as when a drug is effective in treating a disease, but the drug can't be used because it has unacceptable side effects.

Also, using a fixed threshold value is generally more reliable than using a discretionary value because the fixed value

eliminates human subjectivity, which is often a main source of unreliability.

In sum, the discretionary approach is permissible but arguably less sensible. This is because the fixed approach saves time and treats all empirical research papers submitted to a journal objectively and fairly, requiring every paper to satisfy the same easy-to-apply minimum weight-of-evidence criterion for the paper to be accepted for consideration for publication. You must be at least 4 feet tall to be allowed on this ride.

Appendix H: Isn't 0.05 Somewhat Arbitrary?

The choice of 0.05 for a journal's threshold *p*-value is somewhat arbitrary in the sense that 0.04 or 0.06 would be roughly equally as good as 0.05. The threshold *p*-value of 0.05 is chosen because it is in the right ballpark and it is a "round" number, being rounder, so to speak, than 0.04 or 0.06. But the number 0.05 itself isn't in any sense substantively important.

The reason why the threshold *p*-value is somewhat arbitrary is that we can't determine the optimal threshold *p*-value for a journal with high precision. But many editors and researchers agree that the optimal value for most journals appears to lie somewhere in or close to the range between 0.05 and 0.01. Though the choice of the threshold *p*-value is somewhat arbitrary, it is still sensible for a journal to choose and enforce a threshold value because that makes things decisive and saves time.

Based on that, the choice of the actual value used by a journal, while still based on experience, intuition, and norms, is also determined somewhat by the prestige of the journal. A more prestigious journal can use a strict threshold *p*-value of 0.01, which enables the journal to reduce its false-positive publication rate but still get a good number of qualified submitted papers. In contrast, a less prestigious journal may need to use a more lenient threshold *p*-value of 0.05 to get enough qualified submitted papers.

Appendix I: Who Invented the Idea of Balancing the Error Rates?

The idea of balancing the rates of false-positive and false-negative errors was introduced in two papers written jointly by Jerzy Neyman and Egon Pearson and published in 1933. Neyman and Pearson believed that the choice of the threshold value should be at the discretion of the *researcher*, and they didn't associate their idea with publishing scientific journals. However, their original idea evolved to the point where scientific journals began using threshold values as a way to sensibly and fairly control the rate of publication of false-positive errors, while not being so strict as to make it impossible for a paper to be published.

Editors and researchers are generally less aware that the threshold *p*-value controls the rate of false-*negative* errors because, as noted above in section 8, we don't hear much about false-negative errors in scientific research. But they occur and the threshold *p*-value controls the rate.

Arthur Melton was editor of the prestigious *Journal of Experimental Psychology* during its heyday before it was split up. Writing on his retirement in 1962, he discusses his view of the role of the threshold *p*-value in scientific publishing in experimental psychology. Though confident in his approach, he acknowledges criticism of his presumed "slavish worship" of the threshold *p*-value of 0.01.

Appendix J: What Can We Do about *p*-Hacking and Other Misuses of *p*-Values?

Arguably, the problem of the misuse of *p*-values can be reduced or even eliminated by improving the training of scientific researchers. This is because proper training will show a researcher that the misuse of *p*-values is detrimental to his or her career. This is because if a researcher publishes a false-positive result (whether due to misuse of *p*-values or not) and if the result is important, then other researchers will try to use or extend the false result. And because the original result is a *false*-positive result, these researchers will fail, and the failures will be known in the scientific community, and the failures will be detrimental to the original researcher's career. So, researchers who understand the use of *p*-values are careful to use them properly because that is best for science and best for their careers.

For teaching statistics to non-statisticians, I recommend that teachers focus on the proper *use* of the statistical procedures and the *scientific* concepts behind the procedures, as opposed to the underlying mathematics. (The math can be efficiently handled by a computer if the user properly understands the scientific concepts.) Of course, for students who are majoring in statistics, the math is fundamental but, for other students, the scientific concepts are more important and therefore deserve the focus.

For students who aren't majoring in statistics, I recommend that instruction focus on how to read a scientific research paper, how to design a scientific research study, how to conduct a study, how to analyze and interpret the results of a study, including discussion of interpreting computer output, and how to write and publish a research paper reporting the results of a study. I recommend that there be *no* computer programming in an introductory course because programming takes a large block of time and there are other more important topics—the focus should be on the basic steps of scientific research. This inclusive approach can help the field of statistics to assume its natural leadership role in methods for analyzing and interpreting the results of scientific research.

Supplementary Information

The supplementary information for this paper is at https://matstat.com/optp.zip

References

Benjamini Y., and 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 (1): 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T.-H., Huber, J., Johannesson, M., Kirchler, M., et al. 2018. "Evaluating the Replicability of Social Science Experiments in *Nature* and *Science* between 2010 and 2015." *Nature Human Behaviour* 2: 637–44. https://doi.org/10.1038/s41562-018-0399-z

Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., and. Nosek, B. 2021. "Investigating the Replicability of Preclinical Cancer Biology." *eLife* 10:e71601. https://doi.org/10.7554/eLife.71601

Ioannidis, J. P. A. 2005. "Why Most Published Research Findings Are False." *PLoS Medicine* 2 (8): e124. https://doi.org/10.1371/journal.pmed.0020124

Jager, L., and Leek, J. T. 2014. "An Estimate of the Science-Wise False Discovery Rate and Application to the Top Medical Literature" (with discussion). *Biostatistics* 15 (1): 1–45. https://doi.org/10.1093/biostatistics/kxt007

Macnaughton, D. B. 2021. *The War on Statistical Significance:* The American Statistician *vs.* the New England Journal of Medicine. Toronto: Author.

Melton, A. W. 1962. "Editorial." *Journal of Experimental Psychology* 64 (6): 553–557. https://doi.org/10.1037/h0045549

Neyman, J., and Pearson, E. S. 1933a. "IX. On the Problem of the Most Efficient Tests of Statistical Hypotheses." *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* 231: 289–337. https://doi.org/10.1098/rsta.1933.0009

Neyman, J., & Pearson, E. 1933b. "The Testing of Statistical Hypotheses in Relation to Probabilities A Priori." *Mathematical Proceedings of the Cambridge Philosophical Society* 29 (4): 492–510. https://doi.org/10.1017/S030500410001152X

SAS Institute 2021. "Introduction to Power and Sample Size Analysis: Customized Power Formulas (DATA Step) [web page]." https://documentation.sas.com/doc/en/pgmsascdc/9.4_3.2/statug/statug_intropss_sect017.htm

Tabarrok, A. 2005. "Why Most Published Research Findings are False." Marginal Revolution [website]. https://marginalrevolution.com/?s=why+most+published+research

Wasserstein, R. L., Schirm, A. L., and Lazar, N. A. 2019. "Moving to a World Beyond '$p < 0.05$'," editorial. *American Statistician* 73:sup1: 1–19. https://doi.org/10.1080/00031305.2019.1583913