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

Donald B. Macnaughton
MatStat
donmac@matstat.com

Abstract

A scientific journal that publishes data-based research papers can use a threshold *p*-value as a "gateway" to publication of a paper in the journal. A research paper must successfully pass through the gateway to be selected for *consideration* for publication. The gateway helps the journal to maximize the scientific and social benefits of the papers it publishes. The journal does this by choosing a threshold *p*-value that roughly minimizes the long-run sum of the costs of the false-positive and false-negative errors that the gateway makes in selecting papers to consider. A proof is given that the optimal threshold *p*-value for a scientific journal exists and is unique to the journal.

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

The *p*-value has been discussed in a broad range of uses. However, we consider the *p*-value in a specific but important single application—in its role in a *gateway to publication* for an empirical-research paper in a scientific journal. We discuss how a journal editor can use a "threshold" *p*-value gateway to help to maximize the scientific and social benefits the journal provides.

The *p*-value is a key component of the topic of *statistical inference*, which has an intriguing history, as summarized by Kennedy-Shaffer (2019). We consider a modern view of statistical inference in the journal-publication process, using ideas that have been used by reputable journals since the 1950s. The view is useful because it provides a sensible general way of interpreting scientific research.

1. Relationships Between Variables

We focus on scientific research studies that systematically 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 and data scientists will agree, it is useful and easy to view most 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 variables reflecting the *dose* of a drug given to medical patients and the subsequent *severity* of a disease in the patients in a specified population of patients.

Science and society are interested in relationships between variables because if a researcher finds a new real 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(s) in the entities and then using our knowledge of the relationship to accomplish the prediction or control.

Most scientific studies of relationships between variables (or subunits of more complicated studies) have a single "response" variable (e.g., disease severity) and they have one or more "predictor" variables (e.g., drug dose, patient gender, patient age). The response variable is the variable that we would like to learn how to predict or control. Predictor variables are the variables that we use to enable us (we hope) to predict or control the values of the response variable.

Response and predictor variables are sometimes called "dependent" and "independent" variables respectively. However, those names are less appropriate because either the dependency or the independency may not be present.

2. 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. 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 the disease that the drug was intended to treat.

A researcher determines whether there is good evidence of a relationship between variables by analyzing appropriate research data about the entities in the population. That is, the researcher selects 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 "experimental" research studies, the researcher also *manipulates* the values of one or more of the predictor variables in the entities.

The researcher collects 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 explain how journals and researchers perform the first step of generalizing from sample data—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 (i.e., the form) of a relationship between variables unless we are first confident that the relationship between the variables likely exists. If we don't have good evidence that a relationship exists, then we may be studying not a relationship, but studying mere noise in the data, perhaps mistakenly thinking that the noise reflects a relationship.

If we find good evidence that a potentially useful relationship between variables exists, then our research paper reporting the study will usually discuss various aspects of the form of the putative relationship because readers will want to know about the form. However, apart from a few references to show linkages of ideas and a high-level summary in appendix A, the aspects of the form of a relationship between variables are beyond the scope of this paper. Instead, we focus on the vital and difficult initial question of empirical research of whether a relationship between certain variables in a population likely *exists.*

Appendix H discusses how, formally, a research study must use "random sampling" of entities from the population to make proper generalizations from the sample to the population. The appendix also explains how researchers often sensibly bypass this requirement to reduce research costs at the expense of an acceptable loss in accuracy and precision in the subsequent prediction or control.

3. Positive Results, Negative Results, and the *p*-Value

A *positive* result occurs in an empirical-research study if a proper analysis of the sample data finds good evidence that the studied relationship between variables exists in the population. For example, a medical research study may find good evidence of a relationship between a drug and a disease in a population of medical patients.

In contrast, a *negative* result occurs if a proper analysis of the data *doesn't* find good evidence that the relationship exists.

So, how can we tell whether we have good evidence in a research study that a relationship exists between certain variables? That is, how can we perform a proper analysis of the sample data in a data table to determine whether we have a positive result?

A standard way to address this question is to use a computer 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—i.e., they all go up and down in step, though in some cases in reverse step. So, the measures all operate quite similarly.

For simplicity, this paper focuses on the *p*-value because it is intuitively sensible and because it is the most popular of the measures. However, the ideas in the paper apply equally to all the other standard measures of the weight of evidence of the existence of a relationship between variables. Any of those measures could (sometimes with minor caveats or adjustments) 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 formal definition of the *p*-value, which is included for completeness, but which you needn't understand:

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

For people who aren't statisticians, data scientists, or mathematicians, the definition of the *p*-value is difficult. This is because the "at least as extreme" idea is hard to understand and because the definition uses the complicated concept of conditional probability, as indicated by the two "if" clauses at the end of the definition. Therefore, many people misunderstand the precise technical meaning of the *p*-value.

Fortunately, researchers (including statisticians and data scientists) should *not* try to understand the role of the *p*-value in empirical research in terms of conditional probability. This is because that task is an unnecessary and complicated distraction. Instead, we need only understand the role of the *p*-value in terms of its *function* as a measure of the weight of the evidence for the existence of a relationship between variables (or for the existence of some other research effect). If everything is done properly, the lower the value of the *p*-value, the greater the weight of the evidence that the studied relationship or effect exists in the population. This point of view (with no notion of conditional probability) is sensible, is arguably fully adequate for empirical-research work, and is much easier to understand than the technical definition of the *p*-value in terms of complicated (though also very sensible) conditional probability.

The idea in the preceding paragraph that “everything is done properly” is fundamental in science. There must be *no reasonable alternative explanation* for a low *p*-value before we can trust it, where the concept of “reasonable” is at the discretion of the relevant research community. *Any* alternative explanation is potentially acceptable to the community as long as it is “reasonable”. The idea of “no reasonable alternative explanation” includes the important idea that the technical assumptions underlying the computation of the *p*-value must be adequately satisfied.

The idea of “no reasonable alternative explanation” reflects, from a different point of view, Popper’s idea of “severity of tests” (1980, sec. 82) and Mayo’s idea of “severe testing” or “severe scrutiny” of an empirical-research result (2018). Of course, good empirical research is carefully designed to attempt to eliminate the possibility of reasonable alternative explanations arising of the results.

Although the mathematical theory behind the *p*-value is somewhat complicated, it is easy to compute required *p*-values with modern data-analysis software. That is, if we give modern software a data table and a few simple instructions, the software will automatically compute the required *p*-values for the table and display them in the output together with other important information about the data, including information to help us check if the underlying assumptions are adequately satisfied. (Commercial data-analysis software is generally substantially easier to use than free software due to superiority in documentation, output design, and testing.)

The definition of the *p*-value implies that *p*-values are meaningfully comparable from one research study to the next. We will see in the next section how we can use the *p*-value to sensibly distinguish between positive and negative results. We will also see how the distinction between positive and negative results is fundamental in publishing scientific journals.

4. The Perspective of a Scientific Journal

Scientific journals are the main channels for disseminating new scientific information. The publication of a research paper in a reputable scientific journal implies that the paper has passed through peer review, which implies that the paper was vetted, was usually improved as part of the vetting, and was judged to be worth publishing by the editors and referees. Researchers are eager to have their papers published in scientific journals because that makes a social contribution and may also bring the researchers recognition and reward.

Many papers that are published in scientific journals report about the analysis of the data from an empirical-research study. As noted above in section 1, these analyses can usually be sensibly viewed as studying one or more relationships between variables observed in the data. We focus on the process of publishing (in scientific journals) papers about relationships between variables observed in empirical research.

4.1 Journals Wish to Publish Positive Results

A scientific journal wishes to publish papers that report interesting *positive* results about new relationships between variables. This is because if the variables were carefully chosen, interesting correct positive results about new relationships are invariably useful. That is, the new knowledge of the relationship will give us the ability to better predict, control, or understand the variables in the entities in the studied population. For example, if medical researchers can find a beneficial new causal relationship between the amount of a certain 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 *don’t* want to 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 good evidence of a relationship between a drug and a disease, then doctors and patients can’t do much with that.

Appendix E further discusses why it is generally inefficient for scientific journals to publish negative results.

4.2 Using a Threshold p-Value to Distinguish Between Positive and Negative Results

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 view your result as a positive result and will therefore *consider* your paper for publication.” This enables the journal to automatically decide quickly and fairly whether a paper has enough weight of evidence to make the paper worth considering.

Journals that publish empirical-research papers have used statistical thresholds as a gateway to publication since at least the 1950s, as illustrated by the leading journal in experimental psychology at that time, the *Journal of Experimental Psychology* (Melton 1962). Some journals use 0.01 as the threshold *p*-value instead of 0.05, and other values may also be used, at each journal’s discretion. We discuss how a journal chooses its threshold value later below.

Metaphorically, a journal says to the researcher, “You must be at least 4 feet tall to be allowed on this ride.” Of course, in the case of an amusement park, this rule isn’t used to be arbitrary or mean to children—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. Also, the rule enables the journal to save time at the ticket booth by avoiding both quibbling and puzzling in individual cases.

Journals often use an *implicit* threshold *p*-value. This is generally because they don't know how to justify an explicit threshold. Appendix M presents an argument to justify an explicit threshold.

In the case of an implicit threshold, the threshold value isn't stated explicitly in the journal's instructions to authors, but the specific threshold value (e.g., 0.05) is well known to researchers in the journal's field through word of mouth and through observation of the *p*-values for the main results in the papers published in the journal. However, for ease of understanding, we ignore the distinction between explicit and implicit thresholds and we assume that all journals using threshold values use an *explicit* threshold value, which is arguably the preferred approach for scientific transparency. Appendix G further discusses why some journals use an implicit threshold *p*-value.

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". This label is a useful shorthand to indicate if a result has satisfied the journal's weight-of-evidence criterion.

The concept of statistical significance is often misunderstood because it is often used with different meanings. In the view of this paper, if a research result is statistically significant, this *only* means that the result has satisfied a journal's weight-of-evidence criterion for acceptance for consideration for publication. In particular, if a result is or isn't statistically significant, this *doesn't* mean that the result is a "true" or a "false" result, as explained below in section 5.

It is important to understand that a journal's 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 *p*-value because (assuming that everything is done properly) the lower the *p*-value, the greater the weight of the evidence that the studied effect is real in the population.

4.3 Necessary and Sufficient Conditions for Publication

Satisfying the threshold-*p*-value condition is a *necessary* condition for an empirical-research paper to be published in some scientific journals, but it is certainly not a *sufficient* condition. That is, a reputable journal will have two other necessary conditions that a paper must also satisfy before the journal will publish the paper.

First, the paper must be of enough *interest* to the journal's readers, as judged by the journal's editors and referees. The judgment is made in terms of the perceived scientific and social usefulness in the journal's field of the observed putative relationship between variables. Of course, judging the interest involves considering the observed *form* of the relationship between the variables because some forms of a relationship will be more interesting than others.

A second and important condition for publication of a paper in a reputable journal is that both the research and the reporting of the research in the paper must exhibit enough *quality* according to the journal's standards, as also judged by the editors and referees. The quality condition is highly multifaceted.

The interest condition, the quality condition, and the threshold-*p*-value condition must *all* be satisfied before some scientific journals will publish a submitted empirical-research paper. However, a journal usually evaluates the interest condition and the threshold-*p*-value condition of a newly submitted paper before evaluating the quality condition. This is because an experienced editor can usually evaluate the interest condition and the threshold-*p*-value condition for a new paper in less than half an hour. But it often takes more than 10 hours (perhaps many more) for a journal's editors and referees to evaluate (and often stimulate the author to improve) a paper's quality. So, if a paper fails to satisfy the necessary interest condition or fails to satisfy the necessary threshold-*p*-value condition, then the editor can reject the paper without sending the paper out for quality review, which saves editors' and referees' time.

4.4 Section Summary

A scientific journal can use a threshold *p*-value to determine whether a relationship between variables reported in a submitted paper has enough associated weight of evidence to make the result a positive result. The journal will accept a paper for *consideration* for publication if the main result in the paper is a positive result and if the ideas in the paper are of enough interest to the journal's readers. The journal will *publish* the paper if it successfully passes through a review of its quality.

The following sections discuss how the threshold-value gateway is efficient.

5. 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 the threshold *p*-value makes two types of errors. There appears to be no way to escape from these errors in empirical research.

5.1. False-Positive Errors

A false-positive error occurs if the computed *p*-value for a relationship between variables is less than or equal to a 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 in appendix D, false-positive errors are published surprisingly often in the empirical-research literature.

False-positive errors are the main cause of the so-called replication crisis in empirical 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, which is always possible.)

A false-positive error can occur due to chance, as explained in Appendix M. Also, a false-positive error can occur if the researcher breaks the rules for computing and reporting *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 conditions for publication, then the journal will publish the paper, thereby unknowingly contributing to the replication crisis. This is normal science because, as noted, some false-positive errors are inescapable in empirical research.

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

Of course, false-positive results in the literature are identified and corrected through the process of replication. Other researchers invariably attempt to use or extend *interesting* new positive results. In doing that, the researchers indirectly replicate the results. In the case of attempting to replicate a result that reflects false-positive error, the researchers will fail, and thus the false-positive error will be exposed. This reflects the continuity of science in which new research generally builds on or extends earlier research.

*Un*interesting false-positive results in the scientific literature generally aren't replicated, which reduces replication costs. So, uninteresting false-positive results generally remain uncorrected in the literature. However, that isn't a problem in a practical sense because these results are uninteresting.

5.2. 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 empirical research, though we don't know exactly how often.

Like false-positive errors, false-negative errors are *costly*, but for different reasons: False-negative errors 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 *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 incorrectly 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.

5.3. General Points

The possibility of false-positive and false-negative errors implies that we can't have strong faith in individual findings in empirical research. That is, any new positive finding might be a false-positive error, and any new negative finding might be a false-negative error. However, we *can* put our faith in the overall *system* under discussion. This is because, as we will see below, if the system is used properly (and we must always check carefully for that), it is designed to minimize (over the long run) the sum of the costs of the false-positive and false-negative errors a journal makes in selecting papers to consider for publication.

Of course, though we can't have strong faith in an individual positive result, we can certainly (in the absence of a reasonable alternative explanation) view a positive result as being *suggestive*. And if a positive result is interesting enough, someone will try to replicate or extend the result which, if successful, will help to confirm the existence of the studied effect.

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

6. Controlling the Error Rates

As explained in appendix M, 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 more papers that are reporting evidence about real relationships between variables. And vice versa.

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?

7. The Optimal Threshold *p*-Value

The choice of 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 its published papers.

The journal maximizes the benefit by finding the "sweet spot" for the threshold value that minimizes the long-run *sum of the costs* of the false-positive and false-negative errors, which helps to maximize the benefit. This cost minimization includes the cost of the replication research that fails to replicate earlier false-positive research. 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 empirical research and are costly.

Appendix M presents a graphical and mathematical argument showing that the optimal threshold *p*-value for a scientific journal exists and is unique to each journal.

8. Choosing the Optimal Threshold *p*-Value

The argument in appendix M implies that the optimal threshold *p*-value for a scientific journal exists and is unique, which is important to establish. However, the argument can't determine the actual *numeric* optimal value for a journal. So, a journal must use other means to determine the optimal value.

Ideally, a journal would choose its optimal threshold *p*-value based on formal empirical 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 empirical research or (b) the ongoing *costs* of false-positive and false-negative errors in empirical research. Thus, we can't determine the "loss function" (i.e., the sum of the costs of the false-positive and false-negative errors) for a journal as a function of the threshold *p*-value used by the journal.

Therefore, a journal (apparently) can't determine the optimal threshold value (i.e., can't determine the threshold *p*-value that minimizes the loss function) based directly on formal empirical 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.

Fisher provided the initial intuition for the idea of a threshold *p*-value of 0.05 (1925, secs 12, 20−26). However, in his view, the threshold was used by the *researcher* as a way of deciding whether to tentatively believe in the existence of an effect. That is different from (though not contradictory to) the view in this paper and in some scientific journals that the threshold value serves as a gateway to publication in a journal.

The often-mentioned threshold *p*-values of 0.05 and 0.01 may be popular because, at least for some journals, they appear to give us a roughly optimal long-run cost trade-off between the false-positive and false-negative errors made by the journals in the process of selecting papers to consider for publication.

Appendix K discusses the question whether the threshold *p*-value of 0.05 is somewhat arbitrary.

9. Conclusions

The *p*-value is a mathematically sensible measure of the *weight of evidence* provided by a data table that a studied relationship between variables exists in the studied population. If everything is done properly, the lower the *p*-value for a research result, the greater the weight of evidence that the associated relationship between variables exists.

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 balancing minimizes the sum of the error costs, which helps to maximize the long-run scientific and social benefit of the research papers published in the journal.

This concludes the main discussion in this paper, giving an overview of the key ideas.

Appendices

Appendix A: The Form of a Relationship Between Variables

If a research study finds good evidence that a potentially useful relationship between variables exists, then the research paper reporting the results of the study will usually discuss various aspects of the *form* of the observed relationship. These ideas help us to understand the (presumed) relationship, so a research paper may discuss them in substantial detail. A paper may present:

1. graphs illustrating the observed relationship
2. an estimate of the "effect size" or "strength" of the relationship and an estimate of the precision of the estimated effect size
3. a proposed "model equation" that is a mathematical model of the relationship, which typically has the general form
$$y = f(x;\ \theta) + \varepsilon$$
where y is the response variable; x is the vector (i.e., ordered set) of one or more predictor variables; $f(x;\ \theta)$ is an explicit mathematical function of x; θ is the vector containing the estimated numeric values of the one or more parameters of the function; and ε is the error term representing the random error that the function makes each time it makes a prediction, this error being different and unpredictable (though following rules) for each prediction
4. estimates of the precision of the estimates of the values of the parameters (i.e., θ) of the model equation
5. an estimate of the precision of the predictions or control made by the model equation for new entities from the population, which is an estimate of the average size of ε for new predictions from the model equation; the precision generally depends on the value of x—values of x near the middle of the range will generally lead to more precise predictions than values near the ends of the range.

The graphs in the first item of the list are highly important because a good graph can show a relationship between variables at a glance. The other items in the list vary in importance from one research study to the next.

It is noteworthy that, for technical reasons, effect sizes and parameter estimates in empirical research are often overestimates in absolute value. Similarly, real-life prediction accuracy or prediction precision may turn out to be less than implied by the analysis. Experienced researchers bear these facts in mind when they are considering the results of research. We might mathematically attempt to correct these problems, though that may be overrefinement due to (a) the required speculation and (b) the ever-present noise, which muddies things up.

Though the preceding five aspects of the form of a relationship are simple at the conceptual level, they are complicated in the details because there are many different forms of a relationship between variables, and the underlying math is often somewhat complicated. Statistics and data science textbooks explain the details of the form of a relationship between variables.

Appendix B: Scientific Hypothesis Testing

B.1 The Research and Null Hypotheses

Statisticians, data scientists, editors, and researchers sometimes sensibly refer to the procedure of checking whether a *p*-value is less than (or equal to) a threshold *p*-value as a "statistical test" of the "research hypothesis" (or of the opposing "null hypothesis"). The research hypothesis says that a relationship *exists* between the variables. In contrast, the null hypothesis says that *no relationship* exists between the variables. If the *p*-value is less than or equal to the threshold *p*-value, then the research hypothesis has, so to speak, passed the statistical test.

Following an old tradition, the research hypothesis is sometimes referred to as the "alternative hypothesis". However, that is a misnomer because it inappropriately downplays the vital importance of the research hypothesis in empirical research.

Note that the research hypothesis says nothing about the *form* of the relationship between the variables, which is discussed in the preceding appendix. The research hypothesis only says that a relationship exists.

Sometimes researchers state the research hypothesis and the corresponding null hypothesis in a research study in algebraic terms about the value of the relevant parameter of the model equation of the relationship between the variables. That is, the null hypothesis states that the value of the parameter is equal to the "null" value—the value the parameter will have if there is or were no relationship between variables. The research hypothesis states that the value of the parameter is different from the null value. This mathematically succinct point of view is important because it helps us to understand the mathematical details of the relationship and is a technical basis for computing the *p*-value and the other measures of the weight of evidence. However, this point of view tends to hide the fact that the two hypotheses are readily and sensibly viewed as being about the existence of a relationship between variables. We sometimes get buried in the algebra and lose sight of the science.

B.2 Is the Null Hypothesis or the Research Hypothesis Ever True?

The null hypothesis for a relationship between variables has a special status. This is because, due to measurement limitations, we can almost never empirically *prove* that a given null hypothesis is true. That is, we can almost never empirically prove that a given relationship between variables *doesn't* exist. The relationship might very well exist but, for one reason or another, we have failed to detect it, as discussed below in appendix E.

However, the scientific principle of parsimony (Baker 2022) tells us to keep things as simple as possible and therefore to *assume* that any null hypothesis is true until (if ever) someone provides good evidence to the contrary. This implies that we almost never need to perform analyses or do research to try to show that a null hypothesis is *true*. Instead, we efficiently *assume* that each null hypothesis is true until someone shows otherwise.

Though we rarely try to prove that a null hypothesis is true, we do need to do research and perform analyses if we wish to show that a given null hypothesis is *false*, which is almost always what we wish to show. In this case, following the principle of parsimony, we begin with the assumption that the relevant null hypothesis is true. However, this assumption is usually only a *formal* assumption because we usually don't believe it. And we usually believe the opposite or we wouldn't be doing the research.

That is, we hope that our research will provide enough good evidence to allow us to "reject" the assumed null hypothesis and to allow us to conclude that the research hypothesis is *likely* true—conclude that a relationship *likely* exists between the relevant variables. "Enough good evidence" consists of a relevant *p*-value that is less than (or equal to) a journal's threshold *p*-value and the absence of a reasonable alternative explanation for the low *p*-value.

It is useful to rephrase an important point in section 4.2 of the body of this paper in hypothesis-testing terms: If a research hypothesis passes an appropriate statistical test, this doesn't mean that the hypothesis is *necessarily* true, and the studied relationship exists. This is because the outcome might reflect a false-positive error. Instead, it simply means that the weight of evidence is enough to make the finding a positive result and to make the paper reporting the result worth considering for publication.

On a side note, if a hypothesis test about a relationship between variables can't decide whether a relationship definitely exists (i.e., can't decide whether the research hypothesis is definitely true), then how is the decision about the existence of a relationship between variables made in science? Interestingly, the decision about whether a relationship between variables exists is *never* made *formally* in science because every scientific idea is open to revision if the revision can be shown to be efficient. This is because we sometimes find we are wrong or inefficient about a scientific idea when new information or new theory comes to light. So, no idea in science is cast in stone.

However, an *informal* decision about the existence of a relationship is made implicitly and gradually by the relevant research community. The decision is reflected in the community members' written and spoken remarks about the relationship. Of course, the informal decision about the existence of a relationship between variables occurs in a scientific community after the relevant result has been believably replicated one or more times in independent research.

It is sometimes suggested that the null hypothesis may never be true in empirical research and therefore hypothesis testing may be unnecessary. To help understand this, consider the imaginary situation in which we have perfect measuring instruments, and we have a sample consisting of the entire population, and we can properly measure the relevant variables with our perfect instruments in every entity in the population. In this situation, we can say that the null hypothesis is precisely true only in cases when the relevant computed effect size from the complete data is precisely equal to zero.

If we could carry out this exercise with multiple sets of different variables in multiple populations, we would almost certainly *occasionally* find that the measured effect size is precisely zero, though it is unknown how often we would get precisely zero. Would we obtain an effect size of precisely zero in once in 20 cases, or in once in 20 trillion cases, or somewhere in between (or beyond)?

The preceding points suggest that the null hypothesis is likely sometimes (though perhaps not very often) *precisely* true in scientific research. However, having established that idea, we can see that it isn't directly relevant for the exercise of scientific hypothesis testing. This is because it doesn't matter whether the null hypothesis is ever precisely true because we aren't interested in that. Instead, researchers are almost always trying to prove that the null hypothesis is empirically shown to be *false*, and, for that, it doesn't matter whether it is ever precisely true—that is irrelevant.

This is reflected in the fact that there are three possible categories of interest, not two. The categories are:

1. The null hypothesis is precisely true.
2. The null hypothesis is "in effect" true in the sense that the size of the effect of interest is nonzero but it is too small (perhaps much too small) for us to detect with current measuring instruments and affordable sample sizes.
3. The null hypothesis is demonstrably false.

Researchers and journals *aren't* interested in distinguishing between when the null hypothesis is *precisely* true and when it is not precisely true but is in effect true because the distinction generally isn't useful because it can't be made empirically. Instead, researchers and journals are interested in distinguishing between when (a) the null hypothesis is either precisely true or in effect true and (b) the null hypothesis is almost certainly false. If we can correctly show that the null hypothesis for a relationship in a carefully chosen set of variables is almost certainly false, then we will have almost certainly found a scientifically and socially useful relationship between variables. And, for that, it doesn't matter whether the null hypothesis is ever precisely true.

Consider the borderline in empirical research between (a) when a null hypothesis is precisely true or in effect true, and (b) when the null hypothesis is demonstrably false. This borderline 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 well 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.

B.3 Some Exceptions

It is instructive to study exceptions to the general idea that we can't prove that a null hypothesis is true. A rare exception occurs if the research hypothesis specifies the strength of the relationship under consideration. In this case, appropriate data can provide "good evidence" that the research hypothesis is false and therefore the corresponding null hypothesis is true. This happens occasionally in the physical sciences when the strength of a relationship (or the smallest possible strength) if the relationship exists, can be derived from physical theory.

Another exception occurs in "equivalence testing" of a new generic drug in an attempt to show that the drug is "equivalent" on key attributes to the more expensive brand-name drug that the generic drug aspires to replace. In this case, we can't prove that the two drugs have exactly the same relationship to the disease or to side effects. That is, we can't prove that the two drugs are *identical* on the key attributes—we can't prove that the various relevant null hypotheses of no differences are true. However, we can use a measure of the weight of evidence to sometimes demonstrate that there is *no good evidence of a difference* between the two drugs on each attribute when we study the possible differences carefully. Of course, this exercise is subject to false-negative and false-positive errors, just as in standard empirical research.

B.4 Terminology

The procedure of checking whether a *p*-value is less than (or equal to) a threshold value is sometimes called "null-hypothesis significance testing", with the acronym NHST. However, that name is less appropriate because researchers generally aren't interested in the null hypotheses associated with their research. Thus, for the sake of understanding, the name of the procedure should reflect a more central idea. The present paper refers to the concepts under consideration as the "threshold-value gateway to publication" (of an empirical-research paper in a scientific journal).

Appendix C: Four Views of the Use of a Threshold *p*-Value

There are at least four different views of the use of a threshold *p*-value in empirical research. Three of the views can each be seen as a "decision procedure," with each view making a different type of decision.

First, in the view discussed in the body of this paper, each *scientific journal* chooses its own threshold *p*-value. This value is used by the 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.

Section 7 in the body of this paper says that a journal chooses its threshold *p*-value so as to minimize the sum of the costs of the false-positive and false-negative errors the threshold makes in selecting papers to consider for publication. It is noteworthy that many editors of journals that publish papers reporting empirical research *don't* view the journal's choice of the threshold value in these terms. This is because the theoretical idea of minimizing the sum of the error costs isn't well known. Instead, as suggested in section 4.2 of this paper's body, the editor views the threshold as indicating the *minimum weight of evidence* required to make a paper worth considering for publication. This concept emphasizes controlling false-positive errors and it pays less attention to false-negative errors.

However, as editors become more experienced, they recognize that the basic idea of "minimum weight of evidence" can be usefully expanded to the idea of *balancing* false-positive and false-negative errors against each other. Experienced editors wish to sensibly balance or compromise between these errors in their journals because they sense that if this compromise is properly done, it is the optimal approach to initially select papers for consideration for a journal—optimal to maximize the scientific and social benefit of the papers published in the journal.

A second view of the use of a threshold *p*-value is that the *researcher* chooses the threshold value, and the threshold somehow *decides,* or enables the researcher to decide, whether a relationship between variables (or some other studied effect) *exists* in the studied population. (Formally, 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 the threshold sometimes makes false-positive and false-negative errors.

For example, Benjamin, Berger, Johannesson, et al. call for changing the threshold *p*-value from 0.05 to 0.005 for "claims of discovery of new effects" (2018, p. 6). Here, they mean claims *by researchers* of the discovery of new effects (i.e., usually new relationships between variables). And they explicitly say in the "Concluding remarks" section of their article that they aren't discussing a threshold for publication of new findings in journals.

Similarly, Lakens et al. (2018) say that *researchers* (not journals) should "justify" the threshold values that they use, and Maier and Lakens (2022) propose some ways to help researchers do that. Similarly, Gönen, Johnson, Lu, and Westfall discuss setting a threshold value from the point of view of either minimizing the total probability of misclassification (i.e., minimizing the total probability of false-positive and false-negative errors) or minimizing the costs of these errors through a loss function, saying that "Such an approach allows researchers to differentially weight [false-positive] and [false-negative] errors if desired (2019, p. 29)". Note the reference to researchers.

Similarly, Miller and Ulrich (2019) make an argument close to the argument in the present paper about using the optimal threshold *p*-value to maximize the benefit. However, they make the argument from the point of view of the *researcher* as opposed to from the point of view of a *journal*. They refer on page 5 to the threshold value used by a journal, noting that it may constrain the researcher, but say it is "the researcher's problem" to choose the threshold *p*-value (and the sample size) to maximize the payoff. Appendix M.7 below further discusses the Miller and Ulrich point of view.

However, as observed in a survey of researchers by Białek, Misiak, and Dzieken (2023), when it comes time for researchers to choose the threshold value for an individual research result, conscientious researchers find that hard to do. This is because it is difficult or impossible to realistically know the ramifications of different threshold values for *individual* research results, any of which, if viewed as a positive result, might readily be a false-positive error.

This difficulty also applies if the researcher (as opposed to a journal) tries to set the threshold value for a set of related results in a field. This is because different researchers may call for different thresholds. Therefore, as discussed in the body of this paper, it is more sensible for a *journal* to specify a common threshold for all submitted papers and to use the threshold as a fair and efficient gateway to publication in the journal.

A third view of the use of a threshold *p*-value is that a *scientific journal* chooses its own threshold *p*-value, and the threshold somehow *decides* whether a research paper will be published in the journal. This view, though partly correct, is, on balance, incorrect because it gives the threshold *p*-value more importance than it deserves. This is because, though satisfying the threshold *p*-value condition is a *necessary* condition for publication in some reputable scientific journals, it is never a *sufficient* condition, as discussed in section 4.3 in the body of this paper.

Due to the frequent misunderstanding of the concepts of statistical significance and the threshold *p*-value, a fourth view is that science should abandon the concepts (Wasserstein, Schirm, and Lazar 2019). That would be possible, but then it would take more words to indicate whether a research result has satisfied a journal's weight-of-evidence criterion, which this paper argues is a sensible criterion. So, arguably, the concepts are useful.

Appendix D: How Often Are False-Positive Errors Published 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, high-power, near-exact replications of 21 important positive 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 empirical 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 empirical-research literature because they are unavoidable if we wish to minimize the sum of the costs of false-positive and false-negative errors.

Appendix E: When Should Scientific Journals Publish Negative Results?

Negative results (i.e., results in which the relevant *p*-value is *greater* than a journal's threshold *p*-value) occur surprisingly often in empirical research though we generally don't hear about these occurrences because, as noted, they are generally uninteresting. Negative results occur often because nature's secrets are hard to unlock, so researchers' hypotheses about the existence of relationships between variables are, unfortunately, often wrong. Negative results may also occur for other reasons, as discussed in section 5.2 of the body of this paper. So, it is quite normal that negative results occur in empirical research.

Unfortunately, we can't objectively *know* how often negative results occur in a given field of science. This is because, to know that, we would need to track negative results in the field. However, science generally doesn't track negative results in a field because doing so is judged to be too difficult and too costly to justify the perceived small payoff.

Some researchers and statisticians think that journals should regularly publish reports of negative results. This is because they think that reports of negative results tell us about relationships between variables that *don't* exist, which would be useful to know. That is, they think that a negative result tells us that the *null* hypothesis is true.

But a negative result can't tell us that a relationship doesn't exist—it can only tell us that good evidence of 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 researcher had only used slightly different research conditions or an improved research design, then the research might have properly found the sought-after relationship. For example, perhaps the relationship under study will only appear if the room temperature is above 25°C, but nobody knew that and (unfortunately) the research was performed at 20°C. So, negative results are almost never definitive—they can almost never tell us that a studied relationship between variables *doesn't* exist. So, negative results usually don't provide much useful information. (They do tell us that, from the present result, there is no good evidence that a relationship *does* exist, but that generally isn't useful in a practical sense.)

Furthermore, even if negative results *could* tell us that a relationship between variables doesn't exist, we still wouldn't need to do empirical research to demonstrate this fact. This is because we (efficiently) *assume* that any relationship between variables doesn't exist until someone provides good evidence to the contrary, as sensibly dictated by the principle of parsimony.

Also, the economics of scientific-journal publishing limit the number of papers that can be published to only the most interesting ones. There are usually more than enough interesting *positive* results submitted to a journal, so negative results, being generally less interesting and less useful, are usually immediately eliminated from consideration. If we did wish to publish negative results, and since negative results occur often, we would need many more journals, editors, and referees, which is arguably economically infeasible.

Some researchers think that if journals don't publish negative results obtained in failed replications, then these results won't be available to debunk false-positive results that are sometimes published. However, the sociology of science generally takes proper care of this issue—news about unpublished failed replications gets around quickly in a scientific field through meeting presentations, online paper archives, social media, and personal communications. This is because scientists care greatly about what is true in their field. So, an interesting published false-positive result in a field is usually quickly called into question when other researchers can't replicate the result. This informal approach is sensible because it generally casts adequate doubt on the original research, while saving journal space for new positive results.

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 such results generally aren't useful.

Journals that specialize in publishing negative results exist, as can be seen by searching the web for "journal of negative results". However, these journals don't have much readership or impact. And they usually cease publication when the founding editor retires.

It is noteworthy for completeness that occasionally a negative result is sensational, surprising, or useful (e.g., for policymakers studying the effectiveness of a new policy). 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 because they don't tell us about a new relationship, so they aren't published.

Appendix F: Could a Journal's Threshold *p*-Value Be Discretionary on a Paper-by-Paper Basis?

As discussed below in appendix M, the use of a threshold *p*-value by a journal is theoretically optimal in the sense that if a journal chooses the appropriate threshold *p*-value, then this choice minimizes the sum of the long-run costs of the false-positive and false-negative errors made by the journal. And, apparently, no other known approach can minimize the sum of the long-run costs of the errors and, apparently, there is no other viable criterion that we can minimize or maximize that would be more important than minimizing the sum of the long-run error costs. This supports the idea that the use of a threshold *p*-value is efficient.

The approach is also sensible because journal editors will almost always wish to avoid publishing negative results because, as noted in section 4.1 in the body of this paper and in the preceding appendix, these results are generally much less useful than positive results. Also, as discussed in section 5.1 of the body, editors will always wish to avoid publishing false-positive errors because such errors are scientifically costly and are somewhat embarrassing for both the journal and the editor when it later turns out that the result can't be replicated. Therefore, papers without enough weight of evidence must be screened out. And using a threshold *p*-value is a sensible way to do that, even though it makes errors.

With the preceding ideas in mind, we can note that there are essentially three approaches that a journal can take with respect to assessing the weight of evidence behind the main result in an empirical-research paper:

1. The journal can enforce a *formal* threshold *p*-value (or a formal threshold for some other sensible measure of the weight of evidence) for the main result in a submitted research paper. In this case, researchers know that satisfying this threshold is a necessary condition for publication of a submitted paper, as discussed in the body of this paper.
2. The journal's editors and referees can assess the weight of evidence for the main result in a submitted paper *informally* on a paper-by-paper discretionary subjective basis.
3. The journal can ignore the idea of initially (or even ever) assessing the weight of evidence behind the main result in an empirical-research paper and can instead assess the paper based on other criteria.

Arguably, we can immediately rule out the third approach because, as noted, a journal wishes to screen out papers that are reporting negative results or are reporting weak results that might reflect false-positive errors. So, most editors will agree that they must either formally or informally determine whether the main result in a submitted research paper has enough weight of evidence behind it to make the result a positive result. Thus, arguably, editors must choose approach 1 or 2 from the list above.

Approach 1 has the advantages that it is fair, fast, theoretically optimal in a reasonable sense, and has worked well in some top-level scientific journals since the 1950s (Melton

1962). Approach 1 also has the advantage that the choice of the threshold value is made at the aggregate *journal* level as opposed to being made at the individual *paper* level. This makes approach 1 more reliable than approach 2. This is because, as noted in appendix C, it is difficult or impossible to reliably choose a custom threshold *p*-value for each submitted paper because it is difficult or impossible for an editor to know an individual paper's ramifications. This is because if the main result in a paper is viewed as positive, it is always possible that the result is a false-positive error.

In view of the apparent lack of advantages of approach 2, this paper postulates that approach 2 has no noteworthy advantages. Therefore, approach 1 is preferred to approach 2. So, arguably, the use of a fixed threshold value for a measure of the weight of evidence as a necessary condition for publication in a scientific journal is efficient and is therefore preferred.

Despite the preceding argument, the threshold *p*-value for a journal needn't be inviolable. And it should be possible for a journal editor to overrule the journal's threshold value in a particular case if that seems useful. However, that won't happen often because experienced editors know that it is difficult to distinguish between a signal and noise when a result is close to the borderline of statistical significance. And any seemingly promising almost-significant result might very well be a manifestation of noise. So, if the main result in a paper seems promising to a researcher, but isn't statistically significant, then it is arguably better for the researcher to go back to the lab and to perform the research again, preferably with a more powerful research design to increase the chance that the studied effect will (if it exists) be convincingly detected in the population.

Appendix G: Implicit Versus Explicit Threshold *p*-Values for Scientific Journals

As noted in section 4.2 in the body of this paper, some journals use an *implicit* threshold *p*-value instead of an explicit one. In this case, the threshold *p*-value for a journal isn't stated in the journal's instructions to authors, but researchers who are familiar with the journal are well aware of the threshold through word of mouth among researchers in the field. Researchers are aware of the threshold because the editors regularly reject papers if the *p*-value for the main result is greater than the journal's implicit threshold value, usually with the editor informing the researcher that the weight of evidence for the main result in the paper, as expressed in the *p*-value for the result, isn't enough.

If a journal uses an implicit fixed threshold *p*-value, then we can quickly empirically identify the (likely) value by tabulating the *p*-values for the main results in articles published in the journal over the last few years. If the *p*-values are all less than, say, 0.05, but if some are greater than 0.01, then this suggests that the journal is using an implicit fixed threshold *p*-value of 0.05.

A sensible reason for a journal to use an implicit threshold value instead of an explicit one is that, in the past, there have been no widely accepted technical justifications of the use of a threshold value as a gateway to publication in a scientific journal. So, it was difficult for a journal editor to make a strong argument why a threshold is useful, and it was easy to think that using a threshold value was somewhat arbitrary, as discussed below in appendix K.

Of course, if a journal's use of a statistical threshold value seems arbitrary, then this makes it hard for the journal to defend its use of a threshold. So, journals recognized that they couldn't technically justify their use of a threshold. However, experienced editors also believed that the threshold was sensible as a gateway to publication to control the rate of publication of false-positive errors. So, they used an implicit threshold that they didn't discuss, which enabled them to use a threshold without having to defend it.

A second reason why a journal might use an implicit threshold value is that the editor might think that the discretionary aspect associated with an implicit threshold is to the journal's advantage. In this case, which is discussed in the preceding appendix, the editor is betting, so to speak, that his or her judgment can beat the threshold *p*-value at its own game. That is a risky bet due to the high complexity of the subjects of empirical research and because any seeming positive result might be a false-positive result, with the chance of that directly related to, though generally greater than, the associated *p*-value, as discussed below in appendix M.7.

Regardless of the reason for using an implicit threshold instead of an explicit one, many editors agree that a threshold is mandatory to help to control the rate of false-positive errors published in a journal because these errors are embarrassing and scientifically costly.

A good early example of a journal using an implicit threshold *p*-value was the highly respected *Journal of Experimental Psychology* (*JEP*) in the years between 1950 and 1962. The instructions to authors in that period appeared on the inside front cover of the journal and copies of these instructions are available online (*JEP* 1960). The instructions say nothing about the journal's use of a threshold *p*-value, though the instructions refer the reader to conventions in the American Psychological Association (APA) Publication manual. But both the 1952 APA Publication Manual (APA 1952) and the 1957 revision (APA 1957) that were relevant during the period say nothing about using threshold *p*-values in journals. However, we know that between 1950 and 1962 *JEP* mostly used an implicit threshold *p*-value of 0.01 because the editor during that period discusses this point in his final editorial (Melton 1962).

For a modern example of the use of an implicit threshold value, consider the top-tier *New England Journal of Medicine* (*NEJM*). The "Statistical Reporting Guidelines" for this journal give detailed instructions for reporting *p*-values in papers submitted to the journal (*NEJM* 2023), but the instructions don't specify an explicit threshold *p*-value for the main result (primary outcome or primary endpoint) in a submitted paper. (The guidelines indirectly suggest that *NEJM* uses a threshold *p*-value of 0.05 in section B.d.) But it seems highly likely that the editors either (a) have a jointly agreed-on implicit threshold *p*-value for the main result in submitted papers or (b) each editor uses their own personal threshold *p*-value in considering a paper for publication, with the personal threshold values possibly being discretionary on a paper-by-paper basis. We can be confident that the editors of *NEJM* use one or more formal or informal implicit thresholds for the *p*-value because these editors, like all knowledgeable editors, wish to avoid publishing embarrassing and costly false-positive errors in the journal. Using a threshold *p*-value (or a threshold for some other sensible measure of the weight of evidence) is, arguably, the best way to initially screen papers to minimize the cost of these errors.

It is interesting that the general idea of using a threshold value for a measure of the weight of evidence as a gateway to publication is rarely formally discussed in the scientific literature, though it is implicit in some discussions. Formal discussion may be sparse because, as noted, the threshold-value gateway is sometimes implicit in journals. So, the idea tends to fade into the background. We glimpse the idea from time to time in formal discussions but only in passing.

For example, an editorial in *Nature* says that the threshold *p*-value "decides whether … papers are published", acknowledging the role of the *p*-value in journal publication ("Significant debate" 2019). However, there is no further discussion about how this decision process for a journal works.

Similarly, articles by Ioannidis (2005) and Jager and Leek (2014) use the concept of a threshold value for a journal, but they don't say much explicitly about the threshold. Similarly, Campbell and Gustafson (2019) present a model of the selection of articles by a journal explicitly using a threshold *p*-value as an important concept in the model. But they take this aspect of the publication process more or less for granted and they focus on a sensible model of the publication process to help understand how researchers might game the system.

Habiger and Liang (2022) directly discuss the threshold-value gateway to publication in terms of a measure of the false-discovery rate. However, they focus on controlling false-positive errors in journals (while saying little about false-negative errors) rather than focusing on minimizing the sum of the costs of the false-positive and false-negative errors.

The present paper recommends that a scientific journal use a single fixed threshold *p*-value that is chosen by the editor or editorial board of the journal. This value should be explicitly stated in the instructions to authors with a link to a discussion that clearly justifies the journal's use of an explicit fixed threshold value. Making the threshold *p*-value explicit with a proper justification makes empirical research more transparent and therefore easier for researchers to understand.

Appendix H: Can a Research Study Generalize from a Sample to the Population If It Doesn't Use Random Sampling?

As noted, researchers make generalizations from samples to populations of entities about the existence of relationships between variables in the population and about the associated form of the relationship. It is important to understand that, strictly speaking, for accepted technical reasons, we can make such generalizations from a sample to a population only if the sample is a *random* sample from the population. A sample is a random sample if *every entity in the population* has an equal chance of being selected to be in the sample. (Certain sensible refinements to this idea arise in more complicated research.)

Proper random sampling is done regularly in empirical-research studies when precise representation of the population is required. However, the samples of entities used in most empirical-research studies are *not* random samples from the target population. Instead, the samples are so-called convenience samples that are obtained by using or recruiting entities that are readily at hand. This is because convenience samples are much easier to obtain than random samples and are often adequate for the problem under study.

In the case of a convenience sample, since the sample isn't a random sample from the population, we generally can't safely generalize the results from the sample to the full population of interest. However, we *can* safely generalize the results from the sample to other entities that are "sufficiently like" the entities in the sample. Unfortunately, the concept of "sufficiently like" is vague. However, the concept is often judged to be acceptable for initial research, in which it is more important to detect new relationships between variables (or to detect other effects) and it is less important to have complete generalizability from the sample to the target population.

For example, an experimental psychologist at, say, Stanford University in California may perform an experiment using a convenience sample consisting of a group of psychology students at Stanford. The researcher uses this sample because students in a psychology department are relatively easy to recruit to participate in psychology experiments in the department. If the researcher obtains an interesting positive result in the research, then they will likely publish a paper about the result and the researcher will likely generalize the result beyond Stanford students, perhaps suggesting that the result may apply to all similar university students in California or to all similar university students in the United States. And for some relationships between psychological variables, psychologists may think it is sensible to expect to find the relationship in all *adults*, even though the original research that discovered the relationship was performed using a sample of Stanford psychology students.

Of course, such generalizations from a convenience sample to the full population amount to speculation, which is risky because the generalizations are sometimes wrong. However, researchers sensibly make such generalizations, preferably qualifying their remarks as carefully considered speculation.

And, of course, if there is doubt about the generalizability of a new relationship between variables that was found in a convenience sample, or doubt about the generalizability of the form of a relationship observed in a convenience sample, then we can investigate the generalizability in appropriate further research, perhaps using random sampling if high precision is required.

It is noteworthy that in some cases, generalization from a convenience sample to the full population may be quite sensible. For example, suppose that a physicist draws a convenience sample of atoms of argon gas from a top-grade laboratory-quality cylinder of argon bought from a trusted chemical company. In this case, the physicist's sample of argon atoms will (almost certainly) be essentially equivalent to a true random sample of argon atoms from the complete population of atoms of argon that are local to the earth's surface. So, any inferences that the physicist draws from the convenience sample of argon atoms will (almost certainly) apply to the full population of argon atoms that are local to the earth's surface.

Appendix I: When Do Scientific Journals Need to Use Statistical Significance?

Some research papers published in scientific journals don't report about empirical research studying relationships between variables. For example, some papers discuss (a) scientific theories, (b) newly discovered entities (e.g., fossils, proteins, celestial objects), (c) methods, (d) ethics, or (e) training, with no direct study of relationships between variables in empirical research. Since we use the concept of statistical significance to help to detect relationships between variables, we generally don't need the concept in research studies that aren't studying (and can't be interpreted as studying) relationships between variables in empirical research.

Some research studies that use the concept of statistical significance don't appear to be studying relationships between variables. However, usually such research can be sensibly recast in terms of the study of a relationship between variables. For example, if a research study uses a two-sample *t*-test to compare the average values of a particular continuous variable between two groups, then one might think that this study isn't studying a relationship between variables. However, it is easy to view this study as studying a relationship in which the response variable in the relationship is the continuous variable, and the predictor variable is the binary variable that distinguishes between the two groups.

Even when we *are* studying relationships, sometimes we don't need the concept of statistical significance. For example, some fields 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 the formal system of statistical significance to confirm that there is good evidence of a relationship. This is because merely looking at an appropriate graph of the sample data for a strong relationship will tell an experienced researcher that (assuming that the underlying research and analysis were done correctly) the relationship definitely exists. And the graph will also imply that the computed *p*-value for the relationship would be extremely low if it were computed. So, in this case, the researcher needn't compute a *p*-value (or some similar measure) to measure the weight of evidence, and the journal needn't consider the concept of statistical significance.

Also, in some research studies that study relationships between variables, we already know that the studied relationship exists, and the purpose of the research is to refine our knowledge of the relationship. In this case, in theory, we don't need to check whether there is good evidence that the relationship exists though it doesn't hurt to check to confirm that the analyses are working properly.

In the case of "big data", we have a data table with a large number of rows (i.e., entities in the sample) or a large number of columns (i.e., variables) or both. Here, the "large number" may be in the millions or billions for entities and in the thousands or hundreds of thousands for variables. In this case, researchers are still generally interested in relationships between the variables, and the relationships that are found are sometimes weak.

So, if researchers studying big data wish to publish (in a reputable journal) a paper about a relationship between variables they have discovered in the data, they will need to provide good evidence in the paper that the relationship exists, just as with other empirical research. They can often do that with a properly applied measure of the weight of evidence. Alternatively, depending on the situation, they may be able to convincingly demonstrate the existence of the relationship graphically. Alternatively, researchers with big data may use the data to build a computer model of the relationship and then use the model to make accurate predictions or control in such a way that there is no doubt that the modeled relationship exists.

So, a journal only needs to use the concept of statistical significance if it is evaluating a paper that is reporting about an observed new *weak* relationship between variables in a population and if the paper has no other way of demonstrating convincing evidence of the relationship. This situation occurs often in empirical research because most of the strong relationships between variables have already been discovered and because often research studies don't have big data. So, using a measure of the weight of evidence is often the easiest way to demonstrate good evidence of the existence of a relationship.

Cox refers approvingly to the idea of a journal's 0.05 *p*-value threshold (1977, sec. 4.9).

Appendix J: What If a Research Study Reports Many *p*-Values?

The discussion in the body of this paper says that a journal's threshold-value gateway applies to the *main* result in a paper submitted to the journal. What happens in a research study when there is *no* single main result, and the study is reporting the results of study of multiple relationships between variables, thereby with multiple main *p*-values, such as 5 main *p*-values or 500,000 main *p*-values?

We refer to the research situation in which there are multiple main *p*-values as “multiple testing”—it is also sometimes called “multiplicity”. Dealing with multiple testing is important because modern researchers sometimes find it cost-efficient to study multiple closely related relationships between variables simultaneously.

In the case of multiple testing, for technical reasons, the researcher can expect some low or very low standard *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 analysis methods are available to handle research studies that perform multiple testing, such as the method discussed by Benjamini and Hochberg (1995). These methods help researchers, editors, and journal readers to decide whether the studied relationships between the variables likely exists in the population.

Less experienced researchers who perform multiple testing sometimes don’t take account of the multiple testing because taking account of it adds another layer of complexity to the data analysis. Also, taking account of multiple testing generally leads to substantially fewer statistically significant results. Sometimes, researchers don’t even *report* the fact that they did multiple testing, reporting only the statistical tests that yielded positive results. These practices, which generally happen through incomplete training, are all unethical because they lead to increased rates of published false-positive errors.

In using formal methods to perform multiple testing, the goal is usually still to balance the costs of the false-positive and false-negative errors that the tests make in a way that minimizes the sum of the error costs. Thus, each method has a procedure (with an explicit or implicit threshold value) for deciding whether a result is a positive result. And each procedure will make false-positive and false-negative errors, which both the journal and the researcher would like to optimally balance. Choosing the optimal threshold value for one of these multiple-testing procedures is difficult (for both a journal and a researcher) due to the complexity and due to the typical uncertainty about the costs of the two types of errors. Thus, choosing the optimal threshold value for this research situation is an open problem, which is currently handled by researchers and editors through careful judgment about where to set the value on a case-by-case basis.

If we study articles in the popular general scientific journals *Nature* and *Science*, we see that some articles report many *p*-values in graphs or tables with no corrections for multiple testing. No correction is made for multiple testing because these *p*-values aren’t deemed highly important and are merely viewed as being *suggestive*. If a particular one of these *p*-values is perceived as possibly indicating an important relationship between variables, then it would be sensible (if not already done) to independently repeat the specific research that obtained this *p*-value to replicate the result to confirm that the suggested effect is real in the population and isn’t merely a false-positive error.

Appendix K: Is 0.05 Somewhat Arbitrary?

The choice of 0.05 for a scientific 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, it is close to the home plate, 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, as explained in section 8 in the body of this paper, we can’t determine the optimal threshold *p*-value for a scientific journal with high precision. But, based on experience, many editors and researchers agree that the optimal value for many journals appears to lie somewhere in or close to the range between 0.05 and 0.01.

It is noteworthy that physicists sometimes use much stricter thresholds than researchers in other fields. In this case, it is generally the researchers, *not* the journals that specify strict threshold values. For example, if we make some sensible assumptions, the standard threshold of 5σ (five sigma) that is used by some physicists is equivalent to using a very strict threshold *p*-value of around 5.7×10^{-7}. Using such a strict threshold greatly decreases the chance of false-positive errors though it also greatly increases the cost of the research to detect a given relationship. These are the trade-offs that researchers and journals must make in order to balance false-positive errors, false-negative errors, and direct research costs. The physicists’ approach of using a very strict threshold is consistent with the idea of minimizing the sum of the costs of false-positive and false-negative errors, with the added idea that false-positive errors are viewed as being much costlier than false-negative errors.

Alternatively, the physicists’ choice of very strict threshold values can be viewed as an attempt to define physical “truth”. That is, if the measure of the weight of evidence satisfies the very strict threshold value, then the effect observed is viewed as *truly* occurring. This is sensible because if physicists use this approach, then (assuming everything is done properly, and by the laws of probability) they will virtually never make false-positive errors.

Though the choice of the threshold *p*-value for a journal is *somewhat* arbitrary, it is still sensible for a journal to choose and enforce its best estimate of the optimal threshold value because this gets things *approximately* right and, as noted above in Appendix F, the threshold-value gateway is fair, fast, and theoretically optimal.

Based on that, the choice of the actual threshold *p*-value used by a journal, while still based on experience, intuition, and norms, may also be 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. A journal wishes to reduce its false-positive publication rate because published false-positive errors (though inevitable) are somewhat embarrassing and scientifically costly. Of course, as noted in section 6 of the body of this paper, using a lower threshold *p*-value also *increases* the rate of false-*negative* errors that are incorrectly omitted from consideration by the journal, and these errors are also costly. But false-negative errors are invisible, so they receive less attention.

Although a more prestigious journal can use a threshold *p*-value of 0.01, a more *progressive* journal, whether prestigious or not, may choose 0.05 to cast a wider net for interesting results. A less prestigious journal will typically need to use a threshold *p*-value of 0.05 to get enough qualified submitted papers.

Appendix L: Reducing the Misuse of *p*-Values

Arguably, the problem of the misuse of *p*-values can be reduced or even eliminated by improving the training of empirical researchers. This is because proper training will show researchers that misusing *p*-values is harmful to their reputations. That is, 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 harmful to the original researcher's reputation. 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 reputations.

For teaching data-analysis concepts and procedures to students who aren't majoring or minoring in statistics or data science, I recommend that teachers focus on the proper *use* of data-analysis concepts and the underlying *scientific* concepts. I also recommend deemphasizing the associated mathematics. This is because the math can be efficiently handled by a computer if the student properly understands the scientific concepts. Of course, for students who are majoring or minoring in statistics or data science, 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 or minoring in statistics or data science, I recommend that teachers introduce students to the following topics, discussing each topic to a depth that is consistent with the students' abilities and consistent with the available time:

- the basic ideas of entities, properties of entities, and variables in human reality
- the ideas of summarizing the values of variables for the entities in a sample with dot plots, bar charts, and histograms
- relationships between variables in human reality and in scientific research, including the idea of the response variable and the predictor variable(s) and the idea of summarizing relationships with scatterplots, line graphs, multi-variable bar charts, and contingency tables
- the idea of the empirical study of a relationship between variables to achieve reliable prediction, control, and understanding
- the distinction between observational and experimental research
- how to read an empirical-research paper looking for weaknesses in the research design and weaknesses in the logic and how to identify and summarize the relationships between variables discussed in a paper
- how to design an *observational* empirical-research study with maximum statistical power under the available resources
- how to design an *experimental* empirical-research study with maximum statistical power under the available resources
- how to focus during the design phase of a research study on eliminating the possibility of reasonable alternative explanations arising of the results
- how to conduct the physical aspects of an empirical-research study
- how to analyze and interpret the results of a study, including discussion of interpreting the computer output and checking whether the assumptions underlying the statistical procedures are adequately satisfied, and
- how to write a research paper reporting the results of a study.

Note that it isn't necessary to cover any of the preceding topics in great depth, only enough depth to enable students to understand the main ideas. However, it is recommended that *all* the topics be covered because they are all helpful to understand scientific research. The details behind the ideas can come in later courses after the students have a proper overview of the main ideas. Researchers and students are eager to learn the high-level ideas of empirical research because proper knowledge increases the researcher's chance of successfully publishing research papers and thereby advancing their careers.

It is important to carefully study examples of computer output from data analyses because that helps students to understand a key step of scientific research. However, I recommend that there be no training in data-analysis computer programming in an introductory course for students who aren't majoring or minoring in statistics or data science. This is because programming takes a large block of time and there are other more important topics.

Using proper research design and proper data-analysis methods to study relationships between variables is invaluable in all branches of science. If teachers can tightly link the statistics and data-science ideas directly to empirical research, then this inclusive approach will help our field to assume its natural leadership role in methods for analyzing and interpreting the data obtained in empirical research. If we can teach students to use the general ideas of *science*, with proper coverage of the useful ideas of statistics and data science, we maximize our contribution to society.

Appendix M: A Proof that the Optimal Threshold *p*-Value for a Scientific Journal Exists and Is Unique

M.1 Introduction.

The body of this paper says that if a scientific journal uses the optimal threshold *p*-value as a gateway to publication for empirical-research papers, then this will help to maximize the scientific and social benefit of the papers that are published or are refused publication in the journal. This appendix gives a formal economics argument to show how the optimal threshold *p*-value for a journal exists and is unique to each journal.

The argument consists of an extended thought experiment in which we pretend that we know certain things that we don't know. The thought experiment uses ideas developed by Ioannidis (2005), Tabarrok (2005), Jager and Leek (2014), and Miller and Ulrich (2019). We will see how the thought experiment reveals new facts.

The argument is developed using the *p*-value as the measure of the weight of evidence. However, an equivalent argument could be developed using any other standard measure of the weight of evidence, such as the confidence interval or the Bayes factor. Using another measure would lead to the same conclusion as under the *p*-value approach—the optimal threshold value for the measure for the journal exists and is unique. (For some of the measures of the weight of evidence, the optimal value may also depend somewhat on other factors, such as the sample size or the choice of the "prior" distribution.) Though we don't demonstrate it here, when things are done properly, if the optimal threshold value for a journal is chosen for any of the measures, it will behave in the same way in selecting or rejecting papers as the optimal values for the journal for any of the other measures. This is due to the monotonic relationships between the measures.

We first consider the thought experiment graphically to illustrate the logic. We then follow with a mathematical discussion of the simple ideas behind the graphs.

Consider a group of 1000 randomly selected independent research studies in some field of empirical research that will be submitted to a particular 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 study 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: (a) the group of studies with a *positive* result for the main statistical test (i.e., $p \leq 0.05$), which will be submitted to Journal A and (b) 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 model the rates of occurrence of the four types of results as a function of the threshold *p*-value used by the journal. 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 helps to maximize 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 threshold's numeric value.

M.2. Graphical Version of the Argument

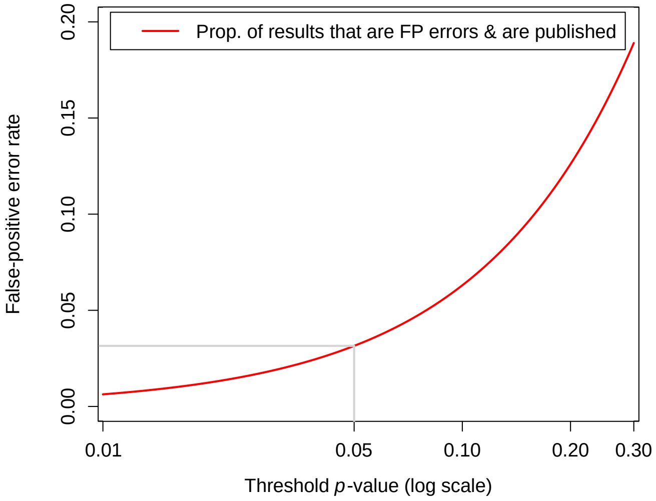

Figure M.1. The published false-positive error rate versus the threshold *p*-value for Journal A. Prop. = Proportion; FP = false-positive.

Figure M.1 shows, for Journal A, based on certain assumptions that are discussed below, 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 red 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 red line shows that if the journal uses a higher threshold *p*-value, then the rate of publication of false-positive errors will be higher.

Note that the false-positive errors shown on the graph are false-positive errors that are *due to chance*. These false-positive errors occur because in all fields of empirical research a significant percentage of the research hypotheses are false. And in the cases when the research hypothesis is false, the statistical tests will sometimes make false-positive errors. Thus, for example, if the threshold *p*-value for a journal is 0.05, then in cases when the research hypothesis is false (i.e., the null hypothesis is true or at least in effect true), false-positive errors will theoretically occur roughly 5% of the time.

Note how the error rates implied by the graph are always somewhat lower than the corresponding threshold *p*-values. The mathematical discussion below explains this phenomenon.

Of course, in real scientific research there is a second source of false-positive errors, which is errors that researchers sometimes make in performing their research, which sometimes lead to false-positive errors. We are ignoring the researcher-caused false-positive errors in the present discussion. If we were somehow able to know the rate of the researcher-caused false-positive errors, we could modify the graph to take account of these errors, which would cause the red line to be higher on the graph. That would have no effect on the main argument under discussion.

Unfortunately, it isn't possible to *empirically* derive the correct version of figure M.1 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 red 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 red line monotonically increases as the journal's 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 in section 5.1 in the body of this paper, due to the complexity of empirical 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 M.1.

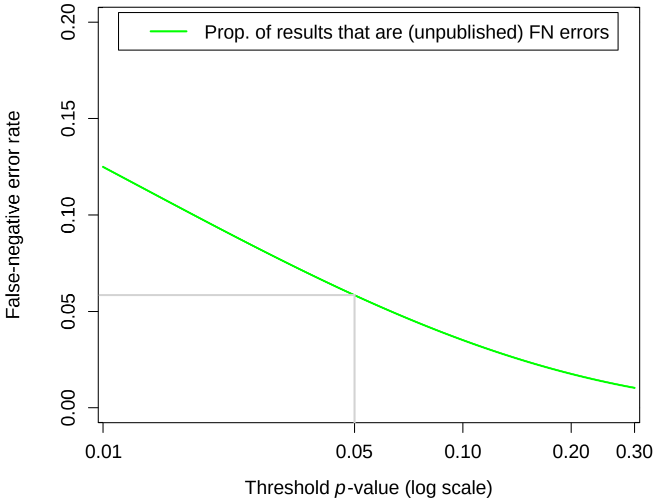

Figure M.2. The false-negative error rate versus the threshold *p*-value for Journal A. FN = false-negative.

Figure M.2 shows, for Journal A, based on assumptions discussed below, 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 M.1 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 green line on figure M.2 tells us that the proportion of the 1000 research studies that are reporting about *real* (i.e., extant and with non-trivial effect size) 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 green line shows that if the journal uses a higher threshold *p*-value, then the rate of incorrect rejections of real relationships (i.e., the rate of false-negative errors) will be lower.

As with figure M.1, the green line shows false-negative errors that are due to chance, and false-negative errors due to researcher errors are ignored. If we were somehow able to know the rate of the researcher-caused false-negative errors, we could modify the graph to take account of that, which would cause the green line to be higher on the graph. As with figure M.1, that would have no effect on the main argument under discussion.

As with figure M.1, we can't empirically derive the correct version of figure M.2 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 red line in figure M.1, we can't know the exact shape or position of the green line on the graph for a journal. However, as with figure M.1, we can model the line mathematically.

Similarly to figure M.1, we know from theory that the green line in figure M.2 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 M.2.

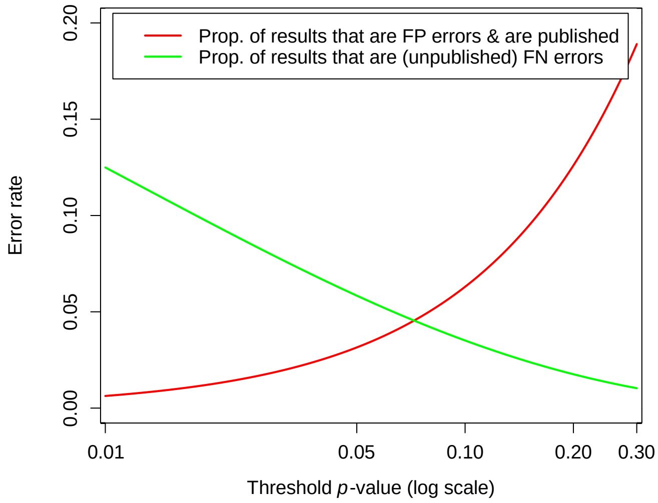

Figure M.3. Figures M.1 and M.2 overlaid.

It is helpful to plot the red and green lines on figures M.1 and M.2 overlaid on a single graph, which yields figure M.3. Overlaying the lines is sensible because both lines pertain to the same 1000 research studies.

A key idea associated with figure M.3 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 are 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 *rates* of false-positive or false-negative errors go up, then the total *cost* of these errors obviously also goes up.

However, since we are performing a thought experiment, let us suppose 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 M.3 into cost lines, as shown on figure M.4.

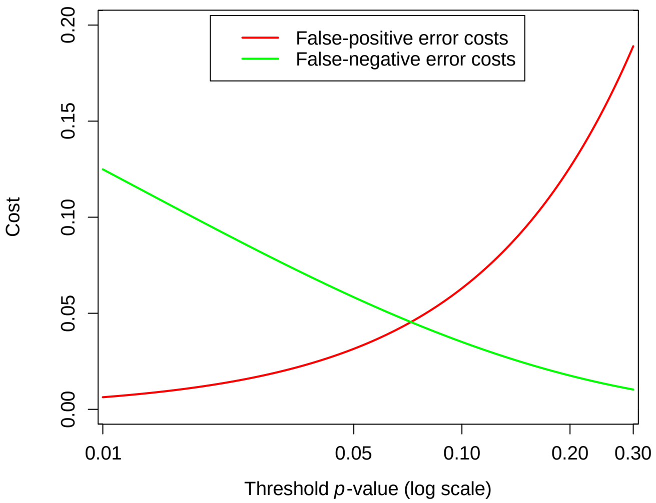

Figure M.4. The costs of false-positive and false-negative errors versus the threshold *p*-value for Journal A.

Note how the vertical axis label on figure M.4 isn't "Error rate" but is "Cost," though the numbers on the axis are unchanged. This is because this discussion is hypothetical, so we can view the numbers on the Cost axis as merely a relative scale. This scale has a true zero because if there were no errors, there would be no associated error costs.

Reflecting the simplest case, the two cost lines on figure M.4 have the same shapes as the two associated error lines on figure M.3. 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 M.4 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 M.4 to take account of those beliefs. 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 red and green lines on the graph, but they wouldn't change the fact that the two lines cross on the graph in the form of a curving *X*, which is the important point for the present discussion.

So, after we have made any necessary adjustments to figure M.4 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 the 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 curving black line on figure M.5 shows the sum of the costs of the two types of errors.

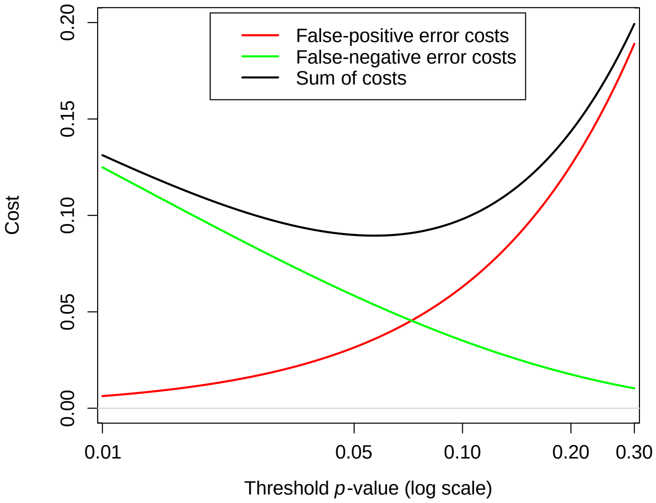

Figure M.5. Figure M.4 with a black line added showing the sum of the costs of the two types of errors versus the threshold *p*-value for Journal A.

In figure M.5, either with a ruler or with measurement by eye, it is easy to see that the height of any point on the curving black line on the figure is the sum of the heights (above the horizontal zero line) of the red and green lines at points that are vertically directly below the point on the black line. For example, if you carefully measure the vertical heights of the red, green, and black lines at 0.05 on the horizontal axis, you will see that the height of the black line is exactly equal to the sum of the heights of the red and green lines.

Of course, the black line shows the theoretical "loss function" that is discussed in section 8 of the body of this paper.

Note how the black 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.

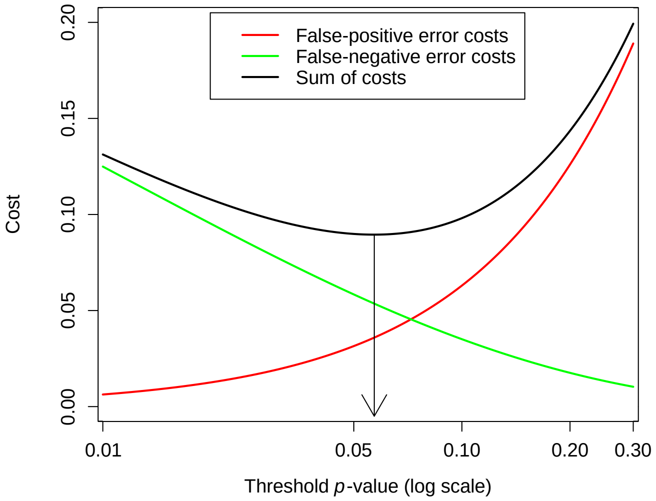

Figure M.6. Determining the optimal threshold *p*-value for Journal A.

Figure M.6 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.

Note how the bowl is somewhat flat in the vicinity of the lowest point, so one could argue that the optimal value is somewhat indeterminate. That argument would be reasonable if we were considering the *empirical* optimal value—i.e., if we had somehow obtained figure M.6 empirically—because then the black line on the graph would be subject to measurement error, which would likely make the minimum point difficult to identify, so the chosen point would likely be somewhat arbitrary.

However, that argument doesn't apply in the present *theoretical* case because in our thought experiment the theoretical *exact* optimal value is, of course, the exact lowest point on the bowl which, due to the geometry of the situation, exists and is unique (though we can't know the value numerically). Appendix K above discusses how the *actual* threshold *p*-value chosen by a journal *is* somewhat arbitrary.

The optimal threshold *p*-value is unique to each journal because different journals will generally have different false-positive and false-negative error cost lines. Therefore, different journals will have different black lines representing the sum of the costs with different minimum points leading to different optimal threshold values.

Figure M.6 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 bowl 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, as explained below. Instead, the figure illustrates how things work and illustrates that the theoretical optimal threshold *p*-value for Journal A *exists*, as defined by the point on the horizontal axis where the vertical arrow dropping from the lowest point on the bowl points.

M.3. Mathematical Version of the Argument for Figure M.1

For some readers, the graphical argument in the preceding section may be enough to convince them that the optimal threshold *p*-value for a journal exists. However, it is useful for further understanding to consider a mathematical version of the argument. We consider the math in terms of the computer program that generates the data to draw figures M.1–M.6 because the program precisely specifies the math.

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 readily converted to any other standard programming language (assuming the language can provide the needed simple statistical functions) and it will give exactly the same red, green, and black 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 in a DATA step is a DATA statement that indicates the beginning of the step and names the data set that we will generate—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 the meaning and use of the values of the preceding five variables in due time below.

Figures M.1–M.6 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;
```

This statement tells SAS to execute the statements that follow the 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 empirical 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, as discussed in Appendix B).

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 researchers who do day-to-day research. These researchers 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 the relationship between variables or other effect that the researcher thought was likely present.

Some authors suggest that the rate of true research hypotheses in psychological research may be as low as 10% (Miller and Ulrich 2016, p. 685; Johnson et al. 2017, abstract).

For readers who think 30% may be too low or too high, if we rerun the present program with PctTrue set at, say, 60% or at, say, 10% instead of 30%, the program will produce a somewhat different version of figure M.6, but there will be no change to the general bowl *pattern* of the black line 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 on the definitions of the *p*-value and the threshold *p*-value. By definition, the threshold *p*-value for a journal is roughly 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, and if everything is done properly, 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 × 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 gray lines on figure M.1 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 M.1. 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 M.1. 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 M.1.

(The program to generate the data, including the code to draw the six figures and the PDF output from the program, is available in the Supplementary Information for this paper. For interested readers, instructions in the Supplementary Information explain how to run the program using free online SAS software and how to change the values of the parameters in the RETAIN statement to see what happens if you do that.)

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 M.1. The second proportion, which is directly related to the first, is always higher than the first due to the underlying mathematics and because (with rare exceptions) only positive research results are published in scientific journals. The second proportion is discussed appendix D above.

M.4. Mathematical Version of the Argument for Figure M.2

Let us now add code to the program to generate the data for figure M.2, 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 as a function of the threshold *p*-value. The values of the PropFalseNeg variable together with the values of the ThreshP variable are the data behind figure M.2.

As with figure M.1, in generating the data for figure M.2, 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 M.1. In generating figure M.1, we considered 1000 *different* research studies. To generate figure M.2, 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 briefly discussed above in appendix I and 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 two-sample *t*-test is applicable if certain often-satisfied assumptions are adequately satisfied, which we assume are satisfied in the thought experiment. For completeness, it is noteworthy that we could also compute the *p*-value using the "before" and "after" blood pressures for each patient and testing for an interaction between Time and Treatment in a repeated measurements analysis of variance, 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—30 patients receiving the drug and 30 receiving the placebo. 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 common group standard deviation 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, which we use to provide a graphical approximation of the general case. (We generalize the approximation later below.) In adding the code, we can use the fact 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 M.2, 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 instances among the 300 cases in which the relationship exists, 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. We will do this counting in a moment after some necessary preliminary points.

(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 M.2, 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, which is why we have used the approach. In general, though, the strength of the relationship between variables under study varies 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. Of course, we usually don't know the values of the parameters for a relationship between variables in advance, but we assume we know them in our thought experiment. 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 M.2. But first we explain the concept of statistical power for readers who may be unfamiliar with it.

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 and data science 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 M.2.

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 an empirical-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 if the relationship is present in the population. That is, if a research study has a power of 0.8, then it will successfully detect the relationship 0.8 of the time that the study is performed if the relationship has the form specified in the power computations.

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 and data scientists 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 somewhat complicated in the mathematical details. Fortunately, we needn’t consider the math because a computer can look after that. 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. We will use this key fact in a moment after we briefly explain the high-level steps to compute power.

To compute the power of a statistical test, we substitute the values of the parameters of the model equation of the relationship and the required specifications of the research design into the appropriate power equations and then the computer evaluates the equations to determine the power. Power equations to do this computation are derived in statistics textbooks about power and are available for all the standard statistical tests of relationships between variables. Many general data-analysis software systems contain preprogrammed routines with power equations that can compute statistical power for a variety of standard research designs.

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 common group standard deviation is 27 units, we use ntotal to tell the equations the total 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 Institute (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 M.2. In particular, using TestPower, we can compute the number of true positive results in the 1000 research studies for the current value of ThreshP by multiplying the number of true relationships (computed earlier) by the power, as follows:

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

For example, if the number of true relationships is 300, and if the threshold *p*-value is 0.05, then, as noted, the power equations tell us that the power is approximately 0.805. In this case, the estimated number of true positive results will be 300 × 0.805, which is roughly 242. Then we can compute the number of false-negative results as

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

Thus, if we have 242 true positive results then we will have 300 − 242 = 58 false-negative results in the 1000 research studies. The two gray lines on figure M.2 reflect this case in terms of an 0.058 proportion of the 1000 studies.

Finally, we convert the count of research studies with false-negative results to a proportion as

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

This completes the code to compute the data needed to draw figure M.2. Note the relative simplicity of the computation. We discuss the generalization of the figure in a moment.

M.5. Generating Figures M.3–M.6

Of course, figure M.3 is simply figures M.1 and M.2 overlaid, and figure M.4 is simply figure M.3 with different labeling.

For figures M.5 and M.6, we need to compute the sum of the costs of the false-positive and false-negative errors to enable plotting the curving 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. Here is the line:

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

We can easily relax the assumption of equal error costs by including the proper multiplicative factor for each term on the right-hand side of the equals sign of the statement. Section M.7 below explains the mathematics of computing these factors.

Finally, there is one more line of code that is (formally) required in the DATA step after the END statement, as follows:

```
run;
```

This line tells SAS that the specification of the DATA step is complete and therefore SAS can now "run" (i.e., compile and then execute) the statements in the DATA step to generate the GraphData data set containing the data needed to draw figures M.1–M.6.

M.6. Generalization of Figure M.6

Figures M.1–M.6 illustrate the concepts. However, figure M.6 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, with a mean difference of 20, etc.) as specified in the values of the parameters 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 the generalization in our thought experiment by using the program above to draw a (imaginary) correct version of figure M.6 for the main result in *each research study that might be submitted to Journal A*. That is, for each study, we can change the two-sample *t*-test program code above into the correct code for the main result in the study. For each study, we can do this (in our imagination) 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 main result in 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 convert the program for 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 the main result in any research study, we could in theory insert (a) the appropriate values of the parameters of the model equation of the studied relationship and (b) the appropriate values of 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 could compute the path of the green false-negative cost line behind figure M.6 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 M.6 for that study.

Thus, we can in our imagination generate a correct version of figure M.6 for every research study that might be submitted to Journal A, which we can assume leads to tens of thousands of imaginary 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 our 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: In the preceding discussion we don't need to know the *true model equation* for the relationship between variables we are studying, but we do need to know the true *population values of the parameters* of the model equation that we are using. 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 an appropriate empirical-research study 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* (and if we were to do everything according to the rules behind the procedures). Thus, in principle, the true population values of parameters are empirically obtainable for any population and any model equation, though obtaining the values would generally be prohibitively expensive. So, in practice, the true population values of parameters of model equations generally aren't knowable, though 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 correct version of figure M.6 for the study.

In each of the tens of thousands of cases, we will obtain a graph that is similar to figure M.6. The red 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 green false-*negative* error line, though it will always be monotonically decreasing, will move around from graph to graph. That is, the slope and the horizontal and vertical positions of the point of inflection of the green 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-results.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 M.6 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 M.2.

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 M.6 though, as noted, the lowest point on the bowl will generally be different from graph to graph. The computer output BowlGraphs-results.pdf 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 of these imaginary values in practice because we don't know the values. The key 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 numeric 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 is correct. However, we don't need to deal with this problem because we don't intend to do the averaging because, as noted, we can't do it in practice because the values we would average are unavailable because they are unknown.

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 BowlGraphExamples-Output.pdf in the Supplementary Information. In this case, the bowl is truncated.

Summarizing this subsection, we can generalize the specific case depicted in figure M.6 by generating in our imagination an equivalent figure for each research study that might be submitted to Journal A. The "average" of the horizontal positions of the vertical arrows shown on the set of such figures is the optimal threshold *p*-value for the journal in the sense that this value minimizes the sum of the costs of the false-positive and false-negative errors made by the threshold *p*-value in selecting papers to consider for publication.

M.7 Costs and Benefits

The preceding sections focus on minimizing the sum of the *costs* of false-positive and false-negative results when these errors occur. However, the discussion doesn't refer to the *benefits* of true-positive and true-negative results, which also regularly occur. This raises the question whether we need to consider the two benefits together with the two costs in the mathematical analysis. This section addresses that question, explaining how the benefits can be handled in terms of the costs. The discussion also further explores the proposed mathematical model of statistical hypothesis testing.

As discussed above in appendix B, if we consider a standard research hypothesis, there are two possible states of affairs, which are

(a) the research hypothesis is true and thus a relationship exists between the associated variables or

(b) the research hypothesis is false and thus the corresponding null hypothesis is either true or is in effect true, and thus there is effectively no relationship between the variables.

We can represent this situation for all the research studies that are candidates for Journal A with figure M.7

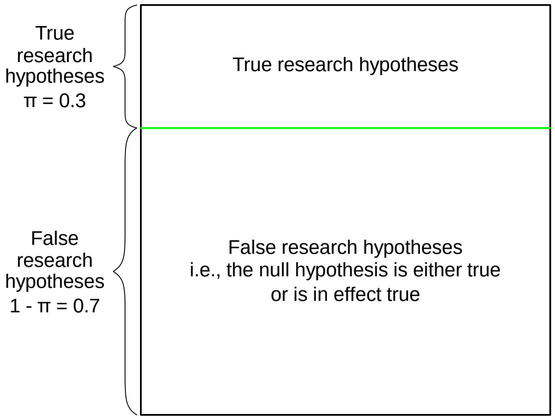
True research hypotheses π = 0.3
True research hypotheses
False research hypotheses 1 - π = 0.7
False research hypotheses i.e., the null hypothesis is either true or is in effect true

Figure M.7. A diagram representing the set of all the research studies that are potential candidates for submission to Journal A prior to performing the research. The set is broken into two natural subsets reflecting the research studies in which the main research hypothesis is true (top rectangle) and the research studies in which the main research hypothesis is false and thus the null hypothesis is true or is at least in effect true (bottom rectangle). The area of each inner rectangle indicates the proportion of research studies that are in the associated subset.

Of course, for any given new research hypothesis, we don't know ahead of time which subset it belongs to. This is because determining whether a studied new research hypothesis is true is generally a key goal of the research.

For a concrete example, the location of the green horizontal line in figure M.7 implies that we continue to assume that $\pi = 0.3$ of the main research hypotheses in papers that might be submitted to Journal A represent true hypotheses though we can readily change this proportion without harming the argument. Here, π (expressed as a proportion) is conceptually identical to the variable PctTrue (expressed as a percentage) in the SAS program above. We use the π form here because it works better in the algebraic discussion later below.

Recall that a positive result occurs in a research study if the relevant *p*-value is less than or equal to the journal's threshold *p*-value. And a negative result occurs if the *p*-value is greater than the journal's threshold *p*-value. Thus, in any properly completed scientific hypothesis test, there are four possible outcomes, which are a true-positive result (TP), a true-negative result (TN), a false-positive result (FP), and a false-negative result (FN). Thus, we can further partition figure M.7, breaking the case when the research hypothesis is true into possible outcomes of false-negative results and true-positive results. Similarly, we can break the case when the research hypothesis is false into possible outcomes of true-negative results and false-positive results. This leads to figure M.8

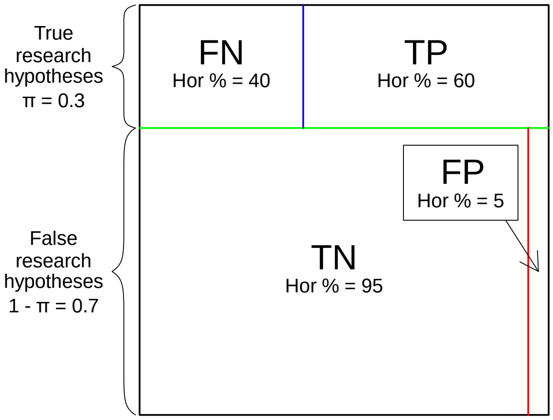

Figure M.8. A partitioned version of figure M.7 showing the set of research studies that might be submitted to Journal A after the outcomes are obtained. The true research hypotheses are broken into two “cells” (FN and TP), and the false research hypotheses are also broken into two cells (TN and FP). The figure is organized so that positive results are on the right-hand side of the figure in both the upper and lower rows. Hor = Horizontal.

As with the rectangles in figure M.7, the area of each cell in figure M.8 indicates the proportion of research studies associated with the cell. For illustration, we assume in the figure that the average power of research studies that might be submitted to Journal A (in cases when the research hypothesis is true) is 0.6. This leads to the 40/60 split in both the horizontal cell widths and the cell areas in the top row of the figure. We also assume that the threshold *p*-value used by Journal A is 0.05, and we assume that all the assumptions underlying the computations of *p*-values are properly satisfied, which leads to the 95/5 split in both the horizontal cell widths and the cell areas in the bottom row.

Similar to the uncertainty situation in figure M.7, if we obtain a positive result in a research study, we can’t definitively tell whether the result is a true-positive result or a false-positive result. Also, if we obtain a *negative* result, we can’t definitively tell whether it is a true-negative result or a false-negative result. However, in any case, we will usually have a studied opinion as to the correct interpretation. And if the situation is interesting enough, then we or other researchers will perform further research that will help to confirm or disconfirm our opinions about the postulated relationship between the variables, which is standard science.

We can calculate the area of each of the four cells in figure M.8 (as measured in terms of each cell’s row and column proportions) to give us the proportion or probability of the respective outcome. For example, the area of the upper-right cell for true-positive results is 0.3 × 0.6 = 0.18. Therefore, under the assumptions, 18% of all the research studies that might be submitted to Journal A will be reporting true-positive results.

The four possible outcomes in figure M.8 are exhaustive for the set of properly performed research studies that are candidates for Journal A. Therefore,

$$\Pr(TP) + \Pr(FP) + \Pr(TN) + \Pr(FN) = 1$$

where $\Pr(E)$ indicates the probability of event E.

We can use the areas for the two types of positive results (i.e., TP and FP) to determine the proportion of the results published in Journal A that are false-positive errors under the assumptions that *all* positive results are published, and *only* positive results are published. (Both these assumptions are generally false for a journal, but they are generally close to true—we take account of this additional complexity later below.) As noted, the area for *TP* results in the top row of figure M.8 is 0.18. The area for the *FP* results in the bottom row is 0.7 × 0.05 = 0.035. So, under the assumptions, the proportion of the published main results in Journal A that are false-positive errors is 0.035 / (0.18 + 0.035) = 16.3%.

The preceding discussion is about false-positive and false-negative errors that are due to *chance*, and the discussion ignores false-positive and false-negative errors that are due to *researcher errors*. If we take proper account of false-positive errors due to researcher errors, then this will cause the percentage of published results that are false-positive errors to be higher, which will widen the lower-right false-positive cell in figure M.8.

In any field of scientific research, it is hard to know the proportion of false-positive errors that are due to researcher errors versus the proportion due to chance because, for technical reasons, the proportions are hard to measure. However, to establish a perspective, let us assume that half of the false-positive errors published in Journal A are false-positive errors due to researcher errors. We can visualize this by noting that if half of the false-positive errors are due to researcher errors, and if the threshold *p*-value is 0.05, then the FP cell in the lower right of figure M.8 will be twice as wide as shown in the figure. Of course, this widening increases the rate of published false-positive errors. Similarly, if we were to properly take account of false-*negative* errors due to researcher errors, then this would widen the upper-left FN cell in the figure, which would reduce the rate of true-positive results, which would further increase the rate of published false-positive errors.

The idea of researcher-caused errors suggests that the 16.3% estimate of published false-positive errors computed three paragraphs above might be roughly doubled, or more than doubled, so somewhere between 30% and 50% of published positive results will be false-positive errors. Though it is speculative, this range of values is consistent with the results of direct replication studies in social science (Camerer et al. 2018).

To aid understanding, imagine a computer app that is programmed to draw figure M.8 on a screen. Imagine that the app has three user-controllable “sliders” that allow the user to specify:

(a) π, the percentage of the research hypotheses in the journal’s field that are true, with possible slider positions ranging in small steps between 0 and 100

(b) the threshold *p*-value used by the journal, with possible slider positions ranging in small steps between, say, 0.001 and 0.3, and

(c) the average power in research studies in the journal's field, with possible slider positions ranging in small steps between 0 and 1. If the user moves one of the three sliders in the app, then this will cause one or two of the blue, green, and red interior partitions in figure M.8 on the screen to move accordingly.

The value of π, i.e., the rate of true research hypotheses in the field, is sensibly viewed as a property of the field, and the variable for this property is generally out of our direct control in any given field of empirical research. So, in using the app, we would generally set the slider for π to its estimated "natural" value for papers submitted to Journal A and then not adjust it further. Interestingly, it is hard to determine the correct natural value of π for a scientific journal. This is because, to measure it, we would need to track negative results in the journal's field but, as noted above in appendix E, that difficult task is sensibly judged to be not worth the effort in most areas of scientific research. However, though π is clearly an important variable in the model, the fact that the value of π is hard to determine isn't a serious problem, as we will see below.

So, after setting the value of π in the app at our best estimate of its natural value, we focus our attention on the threshold *p*-value and the average power, both of which are more easily adjusted in empirical research. Of course, a journal can adjust its threshold *p*-value directly by simply specifying a new threshold value. And a researcher can increase the average power in research studies by using more powerful research designs, such as by (a) increasing the precision of the measuring instruments, (b) increasing the sample size, or (c) using various other effective research-design principles to increase the power of their research.

The app could be designed to emulate the two-sample *t*-test case discussed above beginning in section M.4, using the parameter values specified in the RETAIN statement. More generally, the app could be designed so that, in addition to using the power function for a specific instance of the two-sample *t*-test, the app could also accept a user-specified function to compute the statistical power as a function of the threshold *p*-value. Designing the app to accept a user-specified power function will allow us (if we know the correct function and know the correct slider settings) to use the app to represent the *average* situation in the set of research studies that might be submitted to Journal A.

In addition to displaying figure M.8, the app could also display the percentage of the published results that are false-positive errors for the given slider settings, showing us the extent of the "replication crisis" under the conditions specified by the sliders.

Note that the relationships between all the variables under discussion are somewhat complicated because if we move one of the three sliders in the app then, as noted, this will cause one or two of the interior partitions in figure M.8 to move, but it may also cause another slider to move. To help us control this situation, the app would allow us to "freeze" one of the sliders (often π), so the frozen slider's value is prevented from changing if other sliders are moved. Then (depending on our choice) if we move one of the two unfrozen sliders, then the other unfrozen slider may move accordingly.

Thus, if we move the slider for π, then the green horizontal partition in figure M.8 will move up or down, but the blue and red vertical partitions *won't* move left or right, so the cell *widths* won't change, so the sliders for the power and the threshold *p*-value won't change. Still, if we move the slider for π, the *heights* of all four cells will change, so the cell areas and hence the cell probabilities will all change.

In contrast, since, for a given research design, power depends on the threshold *p*-value, if we move the slider for the threshold *p*-value, then (assuming that π is frozen) this will cause the slider for the power to move—the lower the threshold *p*-value, the lower the power and, therefore, the narrower the true-positive cell. Thus, if we move the slider for the threshold *p*-value, all four cell probabilities will change. Of course, the exact relationship between the threshold *p*-value and the power depends on the power function discussed four paragraphs above.

Though it is mathematically necessary, we would not normally think in a practical sense that the relationship also works in the other direction—that changing the average power in research studies would somehow change the threshold *p*-value. This is because we view the journal as choosing the threshold *p*-value, and the threshold *p*-value as setting the power (through the power function) in the fixed prevailing research climate, and we don't view these events in the opposite order of causation. So, if we try to move the slider for power, it won't move because, in the fixed prevailing a research climate, the power is controlled by the journal's threshold *p*-value, and not the other way around.

So, in studying the three-slider app, we would only need to set π and the threshold *p*-value, and the value of the power would be set indirectly by the value of the threshold *p*-value in conjunction with the power function. Here, it is helpful to keep in mind that the figure and the app to draw the figure are representing the set of research studies in a field of research.

A key feature of the app is that if we set the three sliders at any (mathematically permissible) values of our choosing, then the app tells us (through the computable cell areas) the associated cell probabilities of the four possible outcomes—$\Pr(TP)$, $\Pr(FP)$, $\Pr(TN)$, and $\Pr(FN)$. These cell probabilities play a central role in the following discussion.

With figure M.8 in hand, let us now consider some important ideas discussed by Miller and Ulrich (2019). Their article is titled "The quest for an optimal alpha", where "alpha" is another name for the threshold *p*-value. Their article builds on important earlier work by Mudge, Baker, Edge, and Houlahan (2012).

Miller and Ulrich sensibly specify that the average payoff, $\mathcal{P}$, for a single study for a researcher conducting a set of research studies in a “research area” is

$$\mathcal{P} = \mathcal{P}_{TP}\Pr(TP) + \mathcal{P}_{FP}\Pr(FP) + \mathcal{P}_{TN}\Pr(TN) + \mathcal{P}_{FN}\Pr(FN) \qquad (1)$$

where the subscripted $\mathcal{P}$s are the individual average scientific and social payoffs for each of the four possible outcomes (2019, p. 4). Of course, the average payoffs for true results (i.e., $\mathcal{P}_{TP}$ and $\mathcal{P}_{TN}$) are positive numbers (or possibly zero for $\mathcal{P}_{TN}$) and the average payoffs for false results (i.e., $\mathcal{P}_{FP}$ and $\mathcal{P}_{FN}$) are negative numbers.

Note how the overall average payoff, $\mathcal{P}$ in (1) depends on the threshold *p*-value being used in the research because, as suggested by figure M.8, the four probabilities in (1) all depend on the threshold *p*-value if π is (sensibly) held constant.

Miller and Ulrich reasonably say that the optimal alpha (i.e., the optimal threshold *p*-value) for a researcher in a research area is the value that maximizes the value of $\mathcal{P}$ in (1) for the researcher in the area. So, Miller and Ulrich sensibly wish to determine the threshold *p*-value that maximizes $\mathcal{P}$.

In the present discussion, instead of using the Miller and Ulrich point of view of a researcher in a research area, we use their equation (1) from the point of view of a *scientific journal*, Journal A. So, the “researcher” in the Miller and Ulrich point of view becomes Journal A. And the “research area” in the Miller and Ulrich point of view becomes all the research studies that are potential candidates for submission to Journal A. The four payoffs under this point of view are the average scientific and social payoffs for the four types of events associated with publication or refusal of publication of a report of the indicated type in Journal A. Of course, like a researcher in a research area, Journal A wishes to choose the threshold *p*-value that will maximize the overall payoff, $\mathcal{P}$, for the journal, as specified by equation (1).

We can simplify things by noting from figure M.8 that

$$\Pr(FN) + \Pr(TP) = \pi$$

and

$$\Pr(TN) + \Pr(FP) = 1 - \pi.$$

So, we can solve the preceding two equations for $\Pr(TP)$ and $\Pr(TN)$ and then substitute in (1) to get the overall payoff in terms of the four subscripted $\mathcal{P}$s, $\Pr(FN)$, $\Pr(FP)$, and π. This yields

$$\begin{aligned}\mathcal{P} &= \mathcal{P}_{TP}[\pi - \Pr(FN)] + \mathcal{P}_{FP}\Pr(FP) \\ &\quad + \mathcal{P}_{TN}[1 - \pi - \mathrm{P\ r}(FP)] + \mathcal{P}_{FN}\Pr(FN) \\ &= (\mathcal{P}_{FN} - \mathcal{P}_{TP})\Pr(FN) + (\mathcal{P}_{FP} - \mathcal{P}_{TN})\Pr(FP) \\ &\quad + \pi\mathcal{P}_{TP} + (1 - \pi)\mathcal{P}_{TN}. \end{aligned} \qquad (2)$$

As noted, the goal of Journal A is to find the threshold *p*-value that maximizes $\mathcal{P}$ in (1), which is equivalent to maximizing $\mathcal{P}$ in (2). In this maximization, we can ignore the last two terms in (2), i.e., $\pi\mathcal{P}_{TP}$ and $(1 - \pi)\mathcal{P}_{TN}$, because those terms are constants that don’t depend on the threshold *p*-value for Journal A, so they don’t play a role in the maximization.

Therefore, we wish to find the threshold *p*-value that maximizes

$$(\mathcal{P}_{FN} - \mathcal{P}_{TP})\Pr(FN) + (\mathcal{P}_{FP} - \mathcal{P}_{TN})\Pr(FP). \qquad (3)$$

Expression (3) shows that we can do the maximization in terms of $\Pr(FN)$ and $\Pr(FP)$ which, of course, both depend on the threshold *p*-value, and we needn’t consider $\Pr(TP)$ and $\Pr(TN)$.

Note how the two terms in (3) are sensibly balancing the payoffs and costs of the four types of results.

If we study the mathematics of the discussion in this section, we see that the maximization of (1), (2), or (3), as applied to a journal, is conceptually equivalent to the ideas discussed above in sections M.1–M.6. However, the discussion in sections M.1–M.6 doesn’t proceed by maximizing (3) but proceeds by *minimizing* the *negative* of (3), which is an equivalent way to do the analysis.

That is, the negative of (3) defines the two dynamic terms of the relevant loss function. Figure M.6 shows (as the bowl in the figure) the computed loss function (for the specific two-sample *t*-test case) as a function of the threshold *p*-value.

In the graphical argument discussed above in section M.2, the payoff differences in (3) are (behind the scenes) now reversed because we are minimizing the negative of (3), so the payoff differences become $\mathcal{P}_{TP} - \mathcal{P}_{FN}$ and $\mathcal{P}_{TN} - \mathcal{P}_{FP}$. These values are implicitly respectively multiplied by $\Pr(FN)$ and $\Pr(FP)$ in the step above in section M.2 of converting from figure M.3 (which shows the *probabilities* of false-negative and false-positive errors) to figure M.4 (which shows the *costs* of the errors). The discussion shows the simplest possible case in which $\mathcal{P}_{TP} - \mathcal{P}_{FN}$ and $\mathcal{P}_{TN} - \mathcal{P}_{FP}$ are both equal to 1.0. This state of affairs will occur if, for example, the four scientific and social payoffs in the preceding sentence are respectively, 0.5, −0.5, 0.0, and −1.0. Of course, the correct values of the four payoffs for any journal will generally be different from those values, but in the thought experiment we assume that we know the four payoffs (or at least we know the two payoff differences), so we could readily set the values to the correct values.

In the SAS program that generates the data behind graphs above in sections M.3–M.5, the relevant error costs are multiplied by the relevant probabilities in the program line that assigns a value to the Cost variable in section M.5. Being the source of the graphs, the SAS program is likewise showing the simplest possible case in which both payoff differences are equal to 1.0.

For theoretical completeness and for greater realism, we could sensibly add six more sliders to the app:

(a) a slider for $\mathcal{P}_{TP} - \mathcal{P}_{FN}$ and a slider for $\mathcal{P}_{TN} - \mathcal{P}_{FP}$

(b) a slider for the percentage of the submitted positive results that are published in Journal A, ranging between 0 and 100 (and which will be somewhat less than 100 for reputable journals because journals reject some papers with positive results due to interest or quality shortcomings)

(c) a slider for the percentage of negative results that are published in Journal A, ranging between 0 and 100 (and which will generally be slightly greater than zero for most journals)

(d) a slider for the percentage of false-positive errors in Journal A's field that are due to researcher errors, ranging between 0 and 100, and

(e) a slider for the percentage of false-*negative* errors in Journal A's field that are due to researcher errors, ranging between 0 and 100.

The mathematical role of each slider value in the computations would have to be properly handled in the app's mathematical algorithm to draw figure M.8. Due to the various interactions of the sliders, that isn't a simple matter, but it could be done through careful programming. Mathematical considerations imply that, depending on the settings of the other sliders, some values of some sliders will be unavailable. We might also add other sensible sliders to the app, such as a slider for the payoff of a published surprising negative result. However, further sliders would generally have low impact on $\Pr(FN)$ and $\Pr(FP)$, which are our main interest to enable us to find the threshold *p*-value that maximizes (3).

Note that if we knew the correct settings for the 3 + 6 = 9 sliders, and if the app is programmed with the correct power function, then the app would have enough information to enable it to draw the correct version of figure M.6 in section M.2 above for Journal A. So, since the app would take correct account of many of the relevant functions and variables, we could in theory use the app to determine the exact optimal threshold *p*-value for a journal.

However, for any scientific journal (or for almost any scientific "research area"), we don't know the correct forms of the functions, and we don't know the correct numeric values of any of the sliders in the app. In theory, we could *estimate* all the functions and values required by the app, but as Miller and Ulrich (2019) note in their section 2, we must estimate these functions and values "rather subjectively". This is because for Journal A we don't know the detailed functional relationship between the threshold *p*-value and the probabilities of false-negative and false-positive errors (as exemplified for the hypothetical case above in figures M.1 and M.2). Also, though the four types of payoffs of results are clearly present in scientific research, the payoff *amounts* are difficult or impossible to reliably measure or estimate in a research area. Also, science sensibly doesn't track negative results, and science sensibly doesn't manipulate threshold *p*-values in designed experiments. Therefore, we lack the data we would need to *mathematically* determine the optimal threshold *p*-value for a journal. So, the argument in this appendix enables us to show that the optimal threshold *p*-value for a journal *exists*, but we must *determine* the numeric optimal value for a journal by other means.

Therefore, the app discussed above isn't useful in a practical sense, though it is pedagogically useful as an imaginary tool to assist understanding.

The discussion in this appendix M is in terms of minimizing false-positive and false-negative error costs. Due to symmetry, a strictly parallel discussion is possible in terms of maximizing the benefits of true-positive and true-negative results. That discussion will lead to the same conclusion that the same optimal threshold *p*-value for a journal exists and is unique to the journal. A view of this maximize-the-benefits approach appears on page 6 in the document BowlGraphMain-Output.pdf in the Supplementary Information for the present paper.

In summary, the discussion in this section has explored a mathematical model of statistical hypothesis testing. The discussion implies that we need only use the two payoff differences, $\mathcal{P}_{TP} - \mathcal{P}_{FN}$ and $\mathcal{P}_{TN} - \mathcal{P}_{FP}$, multiplied respectively by $\Pr(FN)$ and $\Pr(FP)$ to carry out the mimization in the thought experiment to show that the optimal threshold *p*-value for a journal exists. That is, we need only consider the *costs of the errors*, and we needn't *directly* consider the benefits of the correct conclusions. This is because we are appropriately indirectly including the benefits (i.e., $\mathcal{P}_{TP}$ and $\mathcal{P}_{TN}$) in composite cost terms in the loss function.

M.8 Refinements and Summing Up

It is useful to consider refinements to the mathematical model discussed above in this appendix that would make the model more closely resemble real empirical research. For example, the model assumes that a certain percentage of the research hypotheses in the journal's field are true ($\pi = \text{PctTrue}$), and the remainder of the hypotheses are false and thus the null hypothesis is true (or is in effect true) in these cases. This state of affairs is illustrated above in figure M.7 where the horizontal green line indicates the borderline between true and false research hypotheses.

However, in reality, there is no hard borderline between true research hypotheses and false research hypotheses and instead the borderline is fuzzy, as noted above in appendix B. That is, there is (presumably) a continuum, which we could model. However, such modeling may be unnecessary because π is absent from the dynamic terms of the cost function (3) in the preceding section.

There may be other refinements that we might make to the 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 conclusion of this appendix.

The thought experiment tells us that the optimal threshold *p*-value for a scientific journal exists but, as noted, the experiment can't tell us the *numeric* value of the optimal threshold *p*-value for a journal because we don't know the values of the relevant payoffs and probabilities used in the argument. So, a journal must use another method to determine the numeric value. As discussed in section 8 of the body of this paper, editors and researchers determine the numeric optimal value based on experience, intuition, and norms. This approach is sensible because editors and researchers generally agree that it is fair, fast, near optimal, and, so far, nobody has thought of a viable better approach.

Supplementary Information

Supplementary information for this paper with instructions, computer programs, and PDF output from the programs is at https://matstat.com/optp.zip

References

APA (American Psychological Association) (1952), *Publication Manual of the American Psychological Association*. Washington, DC: American Psychological Association.

APA (American Psychological Association) (1957), *Publication Manual of the American Psychological Association 1957 Revision*. Washington, DC: American Psychological Association.

Baker, A. (2022), "Simplicity." *The Stanford Encyclopedia of Philosophy (Summer 2022 Edition)*, ed. E. N. Zalta https://plato.stanford.edu/archives/sum2022/entries/simplicity/

Benjamin, D.J., Berger, J.O., Johannesson, M. et al. (2018), "Redefine Statistical Significance." *Nature Human Behaviour* 2, 6–10. https://doi.org/10.1038/s41562-017-0189-z

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

Białek, M., Misiak, M., and Dziekan, M. (2023), "The Vicious Cycle that Stalls Statistical Revolution." *Nature Human Behaviour* 7, 161–163. https://doi.org/10.1038/s41562-022-01515-3

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

Campbell, H., and Gustafson, P. (2019), "The World of Research Has Gone Berserk: Modeling the Consequences of Requiring 'Greater Statistical Stringency' for Scientific Publication." *The American Statistician* 73:sup1: 358–73. https://doi.org/10.1080/00031305.2018.1555101

Cox, D. R. (1977), "The Role of Significance Tests" (with discussion). *Scandinavian Journal of Statistics* 4: 49–70. https://www.jstor.org/stable/4615652

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

Fisher, R. A. (1925) *Statistical Methods for Research Workers.* Edinburgh: Oliver and Boyd. **The 14th edition of this book appears in Fisher (1990).** http://psychclassics.yorku.ca/Fisher/Methods/chap1.htm

Fisher, R. A. (1990) *Statistical Methods, Experimental Design, and Scientific Inference.* Oxford, UK: Oxford University Press.

Gönen, M, Johnson, W. O., Lu, Y., and Westfall, P. H. (2019), "Comparing Objective and Subjective Bayes Factors for the Two-Sample Comparison: The Classification Theorem in Action." *The American Statistician*, 73:1, 22-31. https://doi.org/10.1080/00031305.2017.1322142

Habiger, J., and Liang, Y. (2022), "Publication Policies for Replicable Research and the Community-Wide False Discovery Rate." *The American Statistician* 76 (2): 131-14. https://doi.org/10.1080/00031305.2021.1999857

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

JEP (*Journal of Experimental Psychology*) (1960), Volume 60, Issue 1 (entire issue with front and back covers). https://archive.org/details/sim_journal-of-experimental-psychology-general_1960-07_60_1/mode/2up

Johnson, V. E., Payne, R. D., Wang, T., Asher, A., and Mandal, S. (2017), "On the Reproducibility of Psychological Science." *Journal of the American Statistical Association* 112, 1–10. http://dx.doi.org/10.1080/01621459.2016.1240079

Kennedy-Shaffer, L. (2019), "Before $p < 0.05$ to Beyond $p < 0.05$: Using History to Contextualize p-Values and Significance Testing." *The American Statistician* 73:sup1: 82–90. https://doi.org/10.1080/00031305.2018.1537891

Lakens, D., Adolfi, F.G., Albers, C.J. et al. (2018), "Justify Your Alpha." *Nature Human Behaviour* 2, 168–171. https://doi.org/10.1038/s41562-018-0311-x

Maier M, and Lakens D. (2022), "Justify Your Alpha: A Primer on Two Practical Approaches." Advances in Methods and Practices in Psychological Science, 5(2): 1−14. https://doi.org/10.1177/25152459221080396

Mayo, D. G. (2018), *Statistical Inference as Severe Testing: How to Get Beyond the Statistics Wars*. Cambridge, UK: Cambridge University Press.

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

Miller J., and Ulrich R. (2016), "Optimizing Research Payoff." *Perspectives on Psychological Science* 11(5), 664–691. https://doi.org/10.1177/1745691616649170

Miller J., and Ulrich R. (2019), "The quest for an optimal alpha." *PLoS ONE* 14(1): e0208631. https://doi.org/10.1371/journal.pone.0208631

Mudge, J. F., Baker, L. F., Edge, C. B., and Houlahan, J. E. (2012), "Setting an Optimal α That Minimizes Errors in Null Hypothesis Significance Tests." *PLoS ONE* 7(2): e32734. https://doi.org/10.1371/journal.pone.0032734

NEJM (*New England Journal of* Medicine) (2023), "Statistical Reporting Guidelines" (under "Author Center > New Manuscripts"). Accessed Dec. 4, 2023. https://www.nejm.org/author-center/new-manuscripts

Popper, K. R. (1980), *The Logic of Scientific Discovery*. London: Routledge.

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_introtopss_sect017.htm

"Significant Debate [online title: It’s time to talk about ditching statistical significance],” editorial. (2019), *Nature* 567: 283. https://doi.org/10.1038/d41586-019-00874-8

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. *The American Statistician* 73:sup1: 1–19. https://doi.org/10.1080/00031305.2019.1583913