"Repeated Sampling from the Same Population?" A Critique of Neyman and Pearson's Responses to Fisher

Mark Rubin
The University of Newcastle, Australia

Citation: Rubin, M. (2020). "Repeated sampling from the same population?" A critique of Neyman and Pearson's responses to Fisher. *European Journal for Philosophy of Science, 10,* Article 42, 1-15.
https://doi.org/10.1007/s13194-020-00309-6

Abstract

Fisher (1945a, 1945b, 1955, 1956, 1960) criticised the Neyman-Pearson approach to hypothesis testing by arguing that it relies on the assumption of "repeated sampling from the same population." The present article considers the responses to this criticism provided by Pearson (1947) and Neyman (1977). Pearson interpreted alpha levels in relation to imaginary replications of the original test. This interpretation is appropriate when test users are sure that their replications will be equivalent to one another. However, by definition, scientific researchers do not possess sufficient knowledge about the relevant and irrelevant aspects of their tests and populations to be sure that their replications will be equivalent to one another. Pearson also interpreted the alpha level as a personal rule that guides researchers' behavior during hypothesis testing. However, this interpretation fails to acknowledge that the same researcher may use different alpha levels in different testing situations. Addressing this problem, Neyman proposed that the average alpha level adopted by a particular researcher can be viewed as an indicator of that researcher's typical Type I error rate. Researchers' average alpha levels may be informative from a metascientific perspective. However, they are not useful from a scientific perspective. Scientists are more concerned with the error rates of specific tests of specific hypotheses, rather than the error rates of their colleagues. It is concluded that neither Neyman nor Pearson adequately rebutted Fisher's "repeated sampling" criticism. Fisher's significance testing approach is briefly considered as an alternative to the Neyman-Pearson approach.

Keywords: Fisher; Neyman; Neyman-Pearson; replication crisis; Type I error; Type II error; Type III error

Correspondence concerning this article should be addressed to Mark Rubin at the School of Psychology, Behavioural Sciences Building, The University of Newcastle, Callaghan, NSW 2308, Australia.
E-mail: Mark.Rubin@newcastle.edu.au Web: http://bit.ly/rubinpsyc

Fisher (1945a, 1945b, 1955, 1956, 1960) criticised the Neyman-Pearson approach to hypothesis testing (Neyman & Pearson, 1933) by arguing that it relies on the idea of "repeated sampling from the same population" (Fisher, 1945a, p. 130; Fisher, 1945b, p. 388; Fisher, 1955, p. 71; Fisher, 1956, p. 77, 82, 91; Fisher, 1960, p. 480; see also Hubbard, 2004, p. 300; Johnstone, 1987, p. 492; Perezgonzalez, 2015). The goal of the present article is to provide a critical evaluation of Neyman and Pearson's responses to Fisher's criticism.

I start with a background discussion about the type of replication that is required to operationalize the Neyman-Pearson approach to hypothesis testing. I then explain Fisher's "repeated sampling" criticism. I move on to explain and critique the responses to this criticism provided by Pearson (1947) and Neyman (1977). Finally, I consider some of the implications of adopting a Fisherian approach to hypothesis testing.

What Type of Replication is Required in the Neyman-Pearson Approach?

The notion of "repeated sampling from the same population" implies the repetition or *replication* of a test. Several different types of replication are possible. Hence, to begin with, it is important to consider what type of replication is required to operationalize the Neyman-Pearson approach to hypothesis testing.

The first point to make is that the Neyman-Pearson approach does not require test users to *actually* carry out a series of replications that repeatedly sample from the same population. Instead, they can *imagine* a long run of replications in which they randomly draw a sample, conduct their test, and then, counterfactually, start afresh with a new random sample (e.g., Pearson, 1947, p. 142).

It is also important to appreciate that the Neyman-Pearson approach does not require a long run of *exact* replications that duplicate *all possible testing conditions* (i.e., the exact same sampling procedure, measures, testing environments, etc.).[1] As Neyman (1937, p. 333) explained, "the statistician may be concerned with certain experiments which, if repeated under *apparently identical conditions*, yield varying results" (my emphasis). Hence, Neyman did not require researchers to use *exactly the same* testing conditions; only testing conditions that *appear* to be the same. This point is reassuring, because exact replications are impossible in a universe that is constantly and irreversibly changing (Nosek & Errington, 2020; Rubin, 2019; Stroebe & Strack, 2014; Zwaan et al., 2018).

Two types of non-exact replication have been distinguished: *direct* (sometimes called *close*) and *conceptual* (Rubin, 2019; Stroebe & Strack, 2014; Zwaan et al., 2018). In both cases, researchers assume that they have repeated the testing conditions that are *equivalent* to those of their original test, even if those conditions are not exactly the same as those of their original test. The difference between direct and conceptual replications is that this *assumption of equivalence* is more theoretically contentious in the case of conceptual replications. For example, imagine a researcher who conducts a test of gender differences in self-esteem. In this case, a conceptual replication might entail the theoretically contentious assumption that the testing conditions remain equivalent when the measure of self-esteem focuses on *academic* issues rather than *general* issues. In contrast, a direct replication might make the less contentious assumption that the testing conditions remain equivalent when the same self-esteem measure is presented *online* rather than via a *paper-and-pencil* survey. Despite this difference, it is important to understand that, in both cases, researchers must concede that their assumption of equivalence may be incorrect, and that their replications may entail non-equivalent testing conditions that have substantively altered the nature of their test. For example, in the case of the proposed direct replication, online surveys may

increase gender differences in self-esteem because computer webcams make participants more self-aware during testing. The uncertainty about whether testing conditions are equivalent in direct and conceptual replications means that any substantively discrepant results that occur during these replications may be explained as not only (a) false positives (i.e., Type I errors) and (b) false negatives (i.e., Type II errors), but also (c) the influence of non-equivalent testing conditions that have changed the substantive nature of the test. This last possibility is sometimes referred to as a Type III error (e.g., Dennis et al., 2019). In the case of Type III errors, a substantively different result occurs because "neither the null nor the alternative hypothesis model adequately describes the data" (Dennis et al., 2019, p. 2).

Critically, Type III errors are not permitted within the Neyman-Pearson long run of imaginary non-exact replications (Dennis et al., 2019). Type III errors are only permitted *outside* of this long run (e.g., because a long run of replications is based on an unsatisfactory mathematical model; e.g., Neyman, 1952, p. 27; Neyman, 1955, p. 17). Hence, the Neyman-Pearson long run cannot refer to either direct or conceptual replications. Instead, we need to distinguish a third type of non-exact replication in order to operationalize the Neyman-Pearson long run. I describe this type of replication as an *equivalent* replication, and I contrast it with the *variable* replications that are represented by direct and conceptual replications.

Equivalent replications include only equivalent testing conditions and no non-equivalent testing conditions. Importantly, test users can only imagine undertaking equivalent replications if they are sure about which testing conditions are equivalent and which are non-equivalent. It is this *surety of equivalence* that allows test users to rule out non-equivalent replications. If test users are unsure about which testing conditions are equivalent and non-equivalent, then they may only make an *assumption of equivalence*, and they must concede that this assumption may be incorrect. In this case, they may only imagine undertaking variable replications that contain an unknown mix of equivalent and non-equivalent replications.

Both direct and conceptual replications represent variable replications because, in these cases, test users are never sure which replications are equivalent to one another and which are non-equivalent. So, for example, a test user who imagines a long run of direct replications must concede that this long run may contain some tests that may not be equivalent to their original test.

The distinction between equivalent and variable replications has important implications for the types of conclusion that can be drawn about substantively discrepant results that occur during replications. In the case of equivalent replications, test users' surety of equivalence allows them to rule out the possibility of a substantive change in testing conditions. Consequently, substantively discrepant results may only be attributed to Type I and Type II errors. In contrast, in the case of variable replications, test users' assumption of equivalence may be incorrect, and they must concede that substantively discrepant results may be due to substantive changes in their testing conditions. Hence, discrepant results during variable replications may be attributed to Type III errors as well as Type I and Type II errors.

Importantly, the Neyman-Pearson approach must operate on the basis of equivalent replications rather than variable replications. If it operated on the basis of a long run of variable replications, then it would need to consider the possibility of Type III errors during this long run, and the Neyman-Pearson approach does not permit Type III errors within its long run of replications. Hence, Neyman-Pearson replications must refer to non-exact but equivalent replications rather than to variable replications.

Understanding Fisher's "Repeated Sampling" Criticism

Fisher criticised the Neyman-Pearson approach for being limited to cases of "repeated sampling from the same population." It is worth noting a couple of potential ambiguities in the phrasing of this criticism before explaining it in depth.

First, the term "population" may be taken to encompass not only units of analysis (e.g., people, animals, plants, etc.), but also testing conditions (e.g., sampling procedures, measures, testing environments, etc.). Hence, the phrase refers to a series of equivalent replications in which the same test procedure is used to repeatedly sample from the same population (e.g., Neyman, 1937, p. 339).

Second, just as the Neyman-Pearson approach does not require exactly identical testing conditions, it does not require repeated sampling from exactly the same population. Hence, it is not necessary to imagine sampling from a fixed population that remains static over time. Test users can also imagine sampling from a dynamic population that changes into different populations. Critically, however, in order to rule out Type III errors from the imaginary long run of replications, the new populations must be conceived as being equivalent to the original population in all relevant respects. In other words, the different populations must belong to a *composite* population that contains only *admissible* equivalent *simple* populations (e.g., Neyman & Pearson, 1933, p. 294; Neyman, 1977, p. 106). In this case, although each sample may be drawn from a different simple population, the differences between the simple populations are regarded as being irrelevant with respect to the substantive test results, and so the populations are accepted as being equivalent to one another.

In summary, the Neyman-Pearson approach is not limited to a consideration of the *same* testing procedure that repeatedly samples from the *same* population. It also applies to *equivalent* testing procedures that repeatedly sample from *equivalent* populations. Given this point, it is fair to ask whether Fisher's criticism of "repeated sampling from the *same* population" (my emphasis) is applicable to the Neyman-Pearson approach. I believe it is, because Fisher's criticism applies to *both* exact *and* equivalent replications. In particular, Fisher's objection is that the idea of "repeated sampling" from either the same population or equivalent populations is inappropriate in scientific contexts because scientific researchers do not possess sufficient knowledge to adequately define their tests and populations as being either the same or equivalent. Instead, researchers' lack of knowledge limits them to considering variable replications, and variable replications do not guarantee repeated sampling from either the same population or equivalent populations. I explain this point in greater detail below.

According to Fisher, the Neyman-Pearson concept of repeated sampling from the same or equivalent populations is only appropriate in the case of "acceptance procedures" that are employed in non-scientific contexts (Fisher, 1955, p. 69; Fisher, 1956, pp. 76-77, pp. 99-100). Fisher illustrated this point by referring to quality control tests in applied settings such as industrial production (e.g., Fisher, 1955, pp. 69-70; Fisher, 1956, pp. 99-100). For example, a quality controller might sample 100 light bulbs from a given batch of 10,000 and then test each bulb's luminosity in order to make an inference about the mean luminosity of the entire batch. In this situation, the quality controller is sure about all of the equivalent and non-equivalent aspects of their luminosity test and population (batch). Hence, it is reasonable for them to imagine a long run of equivalent replications that repeatedly sample from the same population (e.g., Batch 57), because they have no doubts about how to correctly define their test and population. Consequently, it is also reasonable for them to rule out potential Type III errors during this long run (e.g., accidentally sampling from Batch 56 instead of Batch 57). Quality controllers are also given a

clear standard of quality that allows them to determine which populations should be classed as "unacceptable" (e.g., a batch of light bulbs in which the mean luminosity is either ≤ 1,425lm or ≥ 1,575lm). Consequently, it is possible for quality controllers to determine their test's smallest effect size of interest and to make firm and final decisions about whether an observed sample belongs to one of two known, adequately specified populations (i.e., an "unacceptable" alternative population or an "acceptable" null population).

Again, Fisher believed that it is appropriate to use the Neyman-Pearson approach of repeated sampling from the same or equivalent population as an "acceptance procedure" in cases of quality control. For example, he believed that the Neyman-Pearson approach can be used to identify low quality parts during the manufacture of aircraft. However, he regarded this approach as being inappropriate in the case of scientific investigations. As he explained:

> I am casting no contempt on acceptance procedures, and I am thankful, whenever I travel by air, that the high level of precision can really be achieved by such means. But the logical differences between such an operation and the work of scientific discovery by physical or biological experimentation seem to me so wide that the analogy between them is not helpful (1955, pp. 69-70).

The "logical differences" between scientific researchers and quality controllers relate to the extent of presumed knowledge about the populations under investigation and the types of conclusion that may then follow. By definition, scientific researchers must concede that they do not possess sufficient knowledge about all of the equivalent and non-equivalent aspects of their populations. Indeed, it is this lack of knowledge that motivates them to investigate the particular populations that they study. Given this self-professed lack of knowledge, researchers must always be ready to admit that they have made a Type III error in conceptualizing their populations. As Fisher (1956, p. 78) explained, for researchers, "the population in question is hypothetical,…it could be defined in many ways, and…the first to come to mind may be quite misleading." Put differently, researchers must always confront the *reference class problem*: If a researcher samples from a population of "1st year undergraduate psychology students," then is their population "1st year undergraduate psychology students" or, more narrowly, "1st year undergraduate psychology students from the researcher's university" or, more broadly, "psychology undergraduate students" or, even more broadly, "young people," etc.? In contrast, there is no reference class problem for quality controllers, because their population is consensually defined in a single, objective, and unequivocal manner (Fisher, 1956, p. 77). A quality controller accepts without question that Batch 57 is Batch 57. They do not consider reconceiving this population as, for example, part of the broader population of "lightbulbs that were manufactured after 2020."

Similarly, researchers do not know which aspects of their testing conditions are equivalent and non-equivalent. Does the time of day of testing matter? Does the ambient temperature matter? Does historical or cultural context matter? Again, by definition, researchers are always unsure about these matters. In contrast, quality controllers are sure about the limits of their tests. For example, they know that time of day of testing does not matter, but that the ambient temperature must be held between 15-30°C in order for their test to be valid.

Researchers also lack clear knowledge about when to describe one population as being substantively different to another (i.e., non-equivalent). Any researcher who has struggled to make an a priori specification of their smallest (non-zero) effect size of interest will recognize this lack of knowledge. In contrast, quality controllers are given a precise quality standard that prescribes the minimum degree to which an alternative population must be different from a standard (null)

population in order to be classed as being "unacceptable" (e.g., in the light bulb example, ≤ 1,425lm or ≥ 1,575lm luminosity).

Finally, researchers do not believe that they know all of the equivalent and non-equivalent features of their testing conditions and populations. Consequently, it is inappropriate for them to interpret their test results in relation to an imaginary series of equivalent replications that repeatedly sample from the same or equivalent population, because they concede that they are not sure what defines the same or equivalent population. Instead, it is only appropriate for them to imagine a series of variable replications that they assume to be equivalent but that they accept may be non-equivalent. In this latter case, researchers must consider the possibility of Type III errors as well as Type I and II errors during their imaginary replications, and this consideration makes it inappropriate for them to use the Neyman-Pearson approach (Dennis et al., 2019, p. 8; Hurlbert & Lombardi, 2009). In contrast, quality controllers are sure that they know all of the equivalent and non-equivalent aspects of their tests, and this surety allows them to rule out the possibility of Type III errors when they imagine a long run of replications. Hence, it is appropriate for quality controllers to use the Neyman-Pearson approach.

To be clear, scientific researchers can and must use a priori theory and evidence to make *educated guesses* about what should count as an equivalent replication (Neyman, 1977, p. 99). However, because they operate in the role of "researcher," they must also concede that these educated guesses may be flawed, that their assumptions of equivalence may be incorrect, and that imaginary replications that follow their potentially inadequate specifications may be variable rather than equivalent. If a researcher wishes to imagine a series of equivalent replications (i.e., repeated sampling from the same or equivalent population), then they must revoke their status as "researcher" and consider themselves to be more like a quality controller who is sure about the equivalence of their replications and who has no interest in discovering new knowledge about the generality and boundary conditions of their populations. Importantly, the same test user cannot have it both ways and adopt the roles of both quality controller and researcher. In other words, they cannot simultaneously claim that (a) they possess sufficient knowledge about a population to specify a long run of equivalent replications but that (b) they do not possess sufficient knowledge about that population and so must continue designing new studies to investigate its generality and boundary conditions. To avoid this epistemic inconsistency, test users must believe that they are either certain or uncertain about the adequacy of the population models that they employ.

To illustrate this point further, reconsider the researcher who investigated gender differences in self-esteem. This researcher may specify in great detail what they believe to be the defining features of their testing conditions and population, including information about their sampling procedure, measure of self-esteem, testing environments, and eligibility criteria for their research participants. For example, they might explain that their research participants should be sampled from a population of "1st year undergraduate psychology students at the University of X." They may then imagine repeatedly drawing random samples from "1st year psychology undergraduate students at the University of X." However, if they accept that they are uncertain about the equivalent and non-equivalent aspects of their population, then they must also concede that they may have unknowingly overspecified some parts of their population and underspecified other parts. For example, the fact that their participants are 1st year undergraduate students, rather than 2nd or 3rd year undergraduate students, may be irrelevant to their test. However, the fact that their participants recently attended an undergraduate course that addressed the empowerment of women in society may be very relevant to their test. Hence, the researcher's population may be something less *and* something more than "1st year psychology undergraduate students at the

University of X." Given their uncertainty on these matters, the researcher must concede that they are only able to imagine a series of direct, variable replications that are based on a fallible assumption of equivalence and that may include Type III errors. Consequently, they are not warranted to use the Neyman-Pearson approach in this case.

Finally, it is important to appreciate that Fisher's "repeated sampling" criticism does not boil down to the point that "all models are wrong…[but] some models are useful" (Box et al., 2005, p. 440). Certainly, "all models are wrong" in the strict sense of the term. But Fisher's point is about whether test users question their model's *usefulness* rather than its *correctness*. Quality controllers never doubt that each of their models represents a "useful approximation to reality" (Box, 2005, p. 440). Their question is only whether a particular population matches or does not match this useful approximation. For example, a quality controller's null model might provide an adequate (not perfect) representation of the luminosity parameters of an acceptable batch of light bulbs. They might then decide whether the luminosity parameters of other batches of light bulbs are significantly different from this null model. In contrast, researchers must concede that they do not know all of the equivalent and non-equivalent aspects of their populations, and so they must continually doubt whether their models of these populations are adequate representations of reality, both statistically and substantively.[2] Indeed, it is their job to consider whether each model is an adequate representation of reality, and, in doing so, they may often conclude that it is not adequate, and that they have made a Type III error. In this case, they must decide whether the model needs adjusting or abandoning.

In summary, quality controllers have no doubt that, although their population models are wrong, they are adequate for their specific utilitarian purposes. Hence, quality controllers ask whether a target population is significantly different from an adequately specified null population and, therefore, the same as an adequately specified alternative population. Researchers ask the same question. However, in addition, they ask whether their null and alternative population models are adequately specified. I now turn to a consideration of Pearson and Neyman's responses to Fisher's criticism.

Pearson's (1947) Response

Pearson (1947, p. 142) addressed Fisher's (1945a) "repeated sampling" criticism by explaining that,

> in other and, no doubt, more numerous cases there is no repetition of the same type of trial or experiment, but all the same we can and many of us do use the same test rules to guide our decision, following the analysis of an isolated set of numerical data. Why do we do this? What are the springs of this decision? Is it because the formulation of the case in terms of hypothetical repetition helps to that clarity of view needed for sound judgement? Or is it because we are content that the application of a rule, now in this investigation, now in that, should result in a long-run frequency of errors in judgement which we control at a low figure?

Hence, Pearson (1947) offered two potential interpretations of the alpha level. The first relates to a hypothetical, imaginary, series of repetitions of the same test. However, as explained above, this interpretation falls foul of Fisher's "repeated sampling" criticism. It is only appropriate for a test user to imagine a series of equivalent replications when they are sure about the equivalent and non-equivalent aspects of their testing conditions and populations, and, by definition, scientific researchers are not sure about these things.

Echoing Neyman (1937, p. 349) and Neyman and Pearson (1933, p. 291), Pearson's (1947) second interpretation assumes that particular researchers always use the same alpha level (e.g., $\alpha = .050$) as a "rule" for their behavior, and that this rule limits the frequency of errors in judgement that those researchers make during hypothesis testing. Fisher (1956, p. 42) was characteristically scathing of this idea of a *researcher's personal alpha rule*. He described it as:

> absurdly academic, for in fact no scientific worker has a fixed level of significance at which from year to year, and in all circumstances, he rejects hypotheses; he rather gives his mind to each particular case in the light of his evidence and his ideas.

To be fair to Neyman and Pearson, many researchers habitually adopt the convention of using an alpha level of .050 in their day-to-day work. However, this is a social norm, rather than a personal rule, and, consistent with Fisher (1956), researchers often deviate from this norm and adopt different alpha levels (e.g., .10, .01, .005, .001) depending on the relative importance of making a Type I error in the particular situation under consideration (Lakens et al., 2018; Neyman, 1977, p. 108). Hence, contrary to Pearson (1947), researchers do not tend to possess a personal alpha rule.

Neyman's (1977) Response

Neyman's (1977, pp. 108-109) response to Fisher's "repeated sampling" criticism considered the "human experience" of particular researchers who conduct a series of different tests of different hypotheses. I quote Neyman at length in order to prevent any mischaracterisation of his position, although I have excluded some unnecessary passages:

> The theory was born and constructed with the view of diminishing the *relative frequency of errors*, particularly of *'important' errors*. Thus, leaving aside the question of an error in testing some particular hypothesis, we have to contemplate a long sequence of situations, say $\{S_i\} = (S_1, S_2, ..., S_n, ...)$ in which tests of some hypotheses will be performed. This sequence, which we may label 'human experience', will be very heterogeneous. Some situations will refer to problems of astronomy [48], others to highway traffic, still others to radiation biology [49], some to problems of big cities and slums or to weather modification, etc. etc. However, there will be some elements common to all the situations of the sequence.
>
> The elements common to all the situations typified by situation S_i, will be: (1) a hypothesis H_i to be tested against an alternative $\overline{H}_i$ and (2) a subjective appraisal of the relative importance of the two kinds of error, leading to the adoption of an acceptably low level of significance α_i combined with an acceptable (hopefully 'optimal') power function. Let $\beta(H_i|\alpha_i)$ denote the value of this function corresponding to some specified simple alternative to H_i that may be judged important.
>
> Eventually, then, with each situation S_i, there will be connected a pair of numbers, α_i, and $\beta(H_i|\alpha_i)$. The question is: what can one expect from the use of the theory of testing statistical hypotheses in the above heterogeneous sequence of situations summarizing human experimence [sic] in 'pluralistic' studies of Nature? The answer is:
>
> > The relative frequency of first kind errors will be close to the arithmetic mean of numbers $\alpha_1, \alpha_2, ..., \alpha_n$…adopted by particular research workers as 'acceptably low' probabilities of the more important errors to avoid. Also, the relative frequency of detecting the falsehood of the hypotheses tested, when false, and the contemplated simple alternatives happen to be true, will differ but little from the average of $\beta(H_1|\alpha_1), \beta(H_2|\alpha_2), \ldots, \beta(H_n|\alpha_n), \ldots$.

> This answer is a simple consequence of a theorem known as the central limit theorem of probability theory….
>
> All the above is emphasized at some length for a particular reason. This is that, at a variety of conferences with 'substantive scholars' (biologists, meteorologists, etc.), accompanied by their cooperating 'applied statisticians', I frequently hear a particular regrettable remark. This is to the effect that the frequency interpretation of either the level of significance α or of power β is only possible when one deals many times with the same HYPOTHESIS *H*, TESTED AGAINST THE SAME ALTERNATIVE. Assertions of this kind, frequently made in terms of 'repeated sampling from the same population', reflect the lack of familiarity with the central limit theorem.

It is important to note the differences between Neyman's (1977) position and Neyman and Pearson's earlier concept of a researcher's personal alpha rule (Pearson, 1947, p. 142; Neyman, 1937, p. 349; Neyman & Pearson, 1933, p. 291). In contrast to the earlier position, Neyman (1977) proposed that the same researcher may select *different* alpha and beta levels in different hypothesis testing situations, depending on "a subjective appraisal of the relative importance of the two kinds of error" (Neyman, 1977, p. 108). Hence, as Lehmann (2008, pp. 63-64) observed, Neyman's (1977) later position is less limited than the earlier position of a personal alpha rule because it allows the same researcher to use different alpha levels in different situations. Consequently, Neyman's (1977) position concedes Fisher's (1956, p. 42) point that "no scientific worker has a fixed level of significance."

Neyman (1977, p. 108) argued that the distributions of different alpha levels (i.e., "α_1, α_2,...,α_n…") and beta levels (i.e., "$\beta(H_1|\alpha_1)$, $\beta(H_2|\alpha_2)$,…, $\beta(H_n|\alpha_n)$,….") that are "adopted by particular research workers" across a range of different testing situations are bound to be normally distributed, following the central limit theorem. Consequently, a particular researcher's frequency of Type I ("first kind") errors across these testing situations will be close to the arithmetic mean of their different alpha values, and their frequency of Type II errors will be close to the mean of their different beta values across these situations. Hence, we can consider a particular research worker's average alpha and beta levels as indicating their typical frequency of Type I and Type II errors respectively. For simplicity, I focus on the concept of a researcher's average alpha level, because Type I errors are often regarded as being more important than Type II errors (Neyman, 1977). However, the same arguments apply to a researcher's average beta levels.

Problems with Neyman's (1977) Response

Neyman (1977) explained that the concept of an average alpha level leaves "aside the question of an error in testing some particular hypothesis" (p. 108). To clarify, even in the case of specific alpha levels, frequentist hypothesis testing does not indicate the probability of a particular hypothesis given a set of observed data (i.e., a Bayesian, inverse probability). Instead, it indicates the probability of the observed test result, or a more extreme result, given a true null hypothesis and associated statistical assumptions. This probability can be compared with an alpha level in order to inform a decision to reject the null hypothesis as an explanation of the test result. Furthermore, a specific alpha level may be set after taking into account the particular costs of making an error during a specific test. Neyman's point about leaving "aside the question of an error in testing some particular hypothesis" (p. 108) refers to this level of specificity, and he noted that this specificity is lost when we consider an average alpha, because average alphas do not necessarily tell us the Type I error rate of a *specific* test of a *specific* hypothesis. Instead, they tell us the typical Type I error rate "adopted by particular research workers" (Neyman, 1977, p. 108).

Neyman's (1977) concept of a researcher's average alpha level avoids Fisher's "repeated sampling" criticism, because it refers to the frequency of errors that occur in a long run of some particular "human experience" (Neyman, 1977, p. 108) rather than a long run of repeated sampling from the same (or equivalent) population. However, a researcher's average alpha is scientifically uninformative. For example, if John has an average alpha level of .050, and Tina has an average alpha level of .005, then we know that John will have made more Type I errors than Tina during the course of his research career. However, these two researchers' average alpha levels tell us nothing about the Type I error rates of specific tests of specific hypotheses, either in the past, present, or future, and in that sense they are scientifically irrelevant. Hence, although Tina may have an average alpha of .005 over the course of her career, her alpha level for Test 1 of Hypothesis A may be .10, .001, or any other value. As scientists, we should be more interested in the nominal Type I error rate of Test 1 of Hypothesis A than in the typical Type I error rate of Tina.

This is not to say that researchers' average alpha levels are uninformative. They may be informative at a metascientific level that considers the behavior of particular scientists or groups of scientists. However, this metascientific issue is separate from the scientific concern of the error rate of specific tests of specific hypotheses.

Summary and Implications

Summary

In summary, neither Pearson (1947) nor Neyman (1977) provided convincing rebuttals of Fisher's (1945a, 1945b, 1955, 1956, 1960) criticism that the Neyman-Pearson approach depends on "repeated sampling from the same population" (Fisher, 1955, p. 71; Fisher, 1956, p. 77, 82). Pearson interpreted the alpha level in relation to a series of imaginary repetitions. However, this interpretation is only appropriate when test users are sure about all of the equivalent and non-equivalent aspects of their testing method and population. By definition, scientific researchers are not sure about these things. Both Neyman and Pearson also interpreted the alpha level as a researcher's personal rule for rejecting null hypotheses. However, this interpretation fails to acknowledge that the same researcher may use different alpha levels in different testing situations. Finally, Neyman proposed that we may calculate the average alpha levels used by particular researchers in order to estimate their typical Type I error rate. The concept of a researcher's average alpha level avoids Fisher's "repeated sampling" criticism, but it does so at the expense of making the alpha level scientifically irrelevant. A researcher's average alpha level may be informative at a metascientific level, but it is not informative at a scientific level.

I should note that neither Neyman nor Pearson appeared to have had strong convictions about their responses to Fisher's "repeated sampling" criticism. After offering his two interpretations of alpha levels, Pearson (1947, p. 142) explained that he "should not care to dogmatize [which is more appropriate], realizing how difficult it is to analyse the reasons governing even one's own personal decisions." Similarly, after explaining his average alpha interpretation, Neyman (1977) asked: "Is the above answer to the question of what to expect from the theory of testing hypotheses satisfactory? This is a subjective matter" (Neyman, 1977, p. 109). Certainly, different researchers may prefer different philosophical approaches to statistical inference. Nonetheless, it is curious that the architects of the Neyman-Pearson theory of hypothesis testing had no firm view about how to conceptualize such an important aspect of their approach.

Implications

What implications does the current critique hold for hypothesis testing? I would argue that Fisher's "repeated sampling" criticism of the Neyman-Pearson approach is valid, and that the alternative interpretations of alpha levels that have been offered by Pearson (1947) and Neyman (1977) are inadequate. Consequently, scientific researchers should consider alternative approaches to hypothesis testing that do not rely on the idea of "repeated sampling from the same population" or equivalent populations. The Fisherian and Bayesian approaches both fit this criterion because they condition their inferences on current testing conditions and observed sample characteristics rather than on a long run of equivalent replications (Rubin, 2017, p. 327). I briefly consider Fisher's approach here because there is already a considerable amount of literature covering the Bayesian approach to hypothesis testing (for an introduction, please see Hoijtink et al., 2019), and it is informative to consider how Fisher avoided his own criticism.

Most importantly, the Fisherian approach rejects the idea of "repeated sampling from the same population" in the context of scientific investigations. In particular, it assumes that researchers have insufficient knowledge to adequately specify a series of imaginary equivalent replications. Instead, researchers must make educated guesses about the equivalent aspects of their direct or conceptual replications. These guesses may then be judged as being provisionally correct or incorrect during a series of real variable replications that each draw single one-off samples from populations that may be either equivalent or non-equivalent to one another. In each of these real replications, researchers may make a statistical inference to a (null) "population of samples in all relevant respects like that observed," where the "relevant respects" are researchers' educated guesses (i.e., Fisher 1955, p. 72).

Importantly, Fisher's approach requires that researchers are unable to recognize any *relevant subsets* within their hypothetical populations that would yield substantively different results (Fisher, 1956, pp. 32–33, 57; Fisher, 1958, p. 268). According to Fisher, every population contains these relevant subsets (non-equivalent subpopulations). In order for a researcher to make a valid probability statement about a population, they must not recognize any of its relevant subsets. For example, a researcher cannot make the probability statement that "men have higher self-esteem than women, $p < .050$" if they know that only young men have higher self-esteem than young women in their sample, and that this gender difference is reversed for older men and women. Age must remain a hidden, unrecognised moderator in order for the researcher to make their broad claim. It is this "postulate of ignorance" about relevant subsets within a hypothetical population (Fisher, 1958, p. 268) that (a) defines test users as "researchers" who lack critical knowledge and are open to revising their conclusions (Fisher 1955, p. 74; Fisher, 1956, p. 99), (b) enables a sample to be considered "random" and equivalent to other samples that could have been drawn (Fisher, 1956, p. 33), (c) allows an inductive inference from the test results to the hypothetical null population (Fisher, 1956, p. 29), and (d) precludes an inference to a long run of repeated sampling, because new samples may be drawn from a different relevant subset to the first, making them non-equivalent (Johnstone, 1987, pp. 492-493; Rubin, 2019).

Fisher's perspective has a number of implications for the debate regarding the role of exact, direct, and conceptual replications in the context of the replication crisis in science. First, Fisher would agree that "there is no such thing as an exact replication" (Schmidt, 2009, p. 92; see also Nosek & Errington, 2020), that exact replications are "an illusion" (Stroebe & Strack, 2014, p. 59), and that "it is impossible to conduct exact replications" (Zwaan et al., 2018, p. 6). Indeed, it is pointless to consider either real or hypothetical exact replications when we inhabit a universe that is changing constantly and irreversibly. It is more appropriate to consider non-exact replications

in which testing conditions and populations are similar to, but not the same as, those of the original test. Importantly, the distinction between exact and non-exact replications does not distinguish between the Neyman-Pearson and Fisherian approaches, because both approaches may operate in relation to non-exact replications that do not repeat every single testing condition.

Second, Fisher's "repeated sampling" criticism implies a previously unrecognised distinction between equivalent and variable replications. Equivalent replications are equivalent in all respects that are relevant to the test. In contrast, variable replications contain an unknown mix of equivalent and non-equivalent tests. Fisher's approach only applies to variable replications, and it relegates equivalent replications to non-scientific contexts in which population models are regarded as being indisputably adequate, such as quality control tests in industrial settings.

Third, both direct and conceptual replications represent variable replications. Consequently, substantive discrepancies in results that occur during direct and conceptual replications may be attributed to not only Type I and II errors, but also Type III errors (i.e., a substantive change in the testing conditions and/or population). Conceptual replications are more theoretically contentious than direct replications, because there is greater doubt about the assumption of equivalence. Nonetheless, both direct and conceptual replications may discover previously unknown generality and boundary conditions (for similar views, see Machery, 2019; Nosek & Errington, 2020; Redish et al., 2018).

Finally, the distinction between equivalent and variable replications helps to distinguish between the Neyman-Pearson and Fisherian approaches. Neyman-Pearson test users must be sure which testing conditions and populations are equivalent with respect to their test. It is this surety of equivalence that warrants an imaginary series of equivalent replications in which substantively different test results may only be attributed to Type I and II errors and not to Type III errors. If test users concede the possibility of Type III errors within their long run of imaginary replications, then it would be inconsistent for them to use the Neyman-Pearson approach. In contrast, if test users disallow the possibility of Type III errors in their imaginary replications, then it would be inconsistent for them to consider themselves as scientific researchers who are interested in reconceiving their hypothetical populations in the face of new information about generality or boundary conditions. Instead, it would be more appropriate for them to consider themselves as quality controllers who are interested in checking whether a sample belongs to one of two populations whose adequate specification is indisputable (Fisher, 1956, p. 99).

In contrast to Neyman-Pearson test users, Fisherian test users are doubtful about which testing conditions and populations are equivalent to one another. Consequently, they must always concede the possibility of Type III errors as they undertake a series of real, variable replications. Furthermore, from a Fisherian perspective, Type III errors should be regarded less as "errors" and more as opportunities for "learning by observational experience" (Fisher, 1955, p. 73; Fisher, 1956, p. 99), because each "error" allows test users to reconsider the generality and boundary conditions of their putative effects. From this perspective, "failure to replicate is not a bug; it is a feature" (Barrett, 2015, p. 23), and Type III errors may be reinterpreted as scientific discoveries (Redish et al., 2018; Rubin, 2019).

Fisher was careful to build the potential for learning into his significance testing framework by stressing the tentative nature of researchers' conclusions. As he explained, "the state of opinion derived from a test of significance is *provisional*, and capable, not only of confirmation, but of *revision*" (1956, p. 99, my emphasis; see also Fisher, 1955, p. 74). In my view, Fisher's approach is more consistent with the doubtful, provisional, and changeable conclusions that scientists draw from their significance tests, especially in the context of replications. In this respect, I agree with

Shrout and Rodgers (2018), who suggested that "Fisher would have likely viewed the recent replication failures about which so much attention (and angst) has developed as 'business as usual'" (p. 137).

References

Barrett, L. F. (2015, September 1). *Psychology is not in crisis.* The New York Times, A23. https://www.nytimes.com/2015/09/01/opinion/psychology-is-not-in-crisis.html

Box, G. E. P., Hunter, J. S., & Hunter, W.G. (2005). *Statistics for experimenters: Design, innovation and discovery (2nd ed.).* Wiley.

Dennis, B., Ponciano, J. M., Taper, M. L., & Lele, S. R. (2019). Errors in statistical inference under model misspecification: Evidence, hypothesis testing, and AIC. *Frontiers in Ecology and Evolution, 7*, 372. https://doi.org/10.3389/fevo.2019.00372

Fisher, R. A. (1945a). The logical inversion of the notion of the random variable. *Sankhyā: The Indian Journal of Statistics, 7*(2), 129-132. https://www.jstor.org/stable/25047836

Fisher, R. A. (1945b). A new test for 2× 2 tables. *Nature, 156* (3961), 388. https://doi.org/10.1038/156388a0

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society. Series B (Methodological), 17*(1), 69-78. https://doi.org/10.1111/j.2517-6161.1955.tb00180.x

Fisher, R. A. (1956). *Statistical methods and scientific inference*. Oliver & Boyd.

Fisher, R. A. (1958). The nature of probability. *The Centennial Review, 2,* 261–274. https://www.jstor.org/stable/23737535

Fisher, R. A. (1960). Scientific thought and the refinement of human reasoning. *Journal of the Operations Research Society of Japan, 3*, 1-10. http://hdl.handle.net/2440/15278

Hoijtink, H., Mulder, J., van Lissa, C., & Gu, X. (2019). A tutorial on testing hypotheses using the Bayes factor. *Psychological Methods, 24*(5), 539–556. http://dx.doi.org/10.1037/met0000201

Hubbard, R. (2004). Alphabet Soup: Blurring the distinctions between *p*'s and α's in psychological research. *Theory & Psychology, 14*(3), 295-327. https://doi.org/10.1177/0959354304043638

Hurlbert, S. H., & Lombardi, C. M. (2009). Final collapse of the Neyman-Pearson decision theoretic framework and rise of the neoFisherian. *Annales Zoologici Fennici, 46*(5), 311-349. https://doi.org/10.5735/086.046.0501

Johnstone, D. J. (1987). Tests of significance following R A Fisher. *The British Journal for the Philosophy of Science, 38*(4), 481-499. https://doi.org/10.1093/bjps/38.4.481

Lakens, D., Adolfi, F. G., Albers, C. J., Anvari, F., Apps, M. A., Argamon, S. E., ... & Buchanan, E. M. (2018). Justify your alpha. *Nature Human Behaviour, 2*(3), 168-171. https://doi.org/10.1038/s41562-018-0311-x

Lehmann, E. L. (2008). *Reminiscences of a statistician: The company I kept.* Springer Science & Business Media.

Machery, E. (2019, October 10). What is a replication?. https://doi.org/10.31234/osf.io/8x7yn

Neyman, J. (1937). X—Outline of a theory of statistical estimation based on the classical theory of probability. *Philosophical Transactions of the Royal Society of London. Series A, Mathematical and Physical Sciences, 236*(767), 333-380. https://doi.org/10.1098/rsta.1937.0005

Neyman, J. (1952). *Lectures and conferences on mathematical statistics and probability*. U.S. Department of Agriculture. http://hdl.handle.net/2027/mdp.39015007297982

Neyman, J. (1955). The problem of inductive inference. *Communications on Pure and Applied Mathematics, 8,* 13-46.

Neyman, J. (1977). Frequentist probability and frequentist statistics. *Synthese, 36, 97*-131. https://doi.org/10.1007/BF00485695

Neyman, J., & Pearson, E. S. (1933). IX. On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society A, 231*(694-706), 289-337. https://doi.org/10.1098/rsta.1933.0009

Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLOS Biology, 18*(3): e3000691. https://doi.org/10.1371/journal.pbio.3000691

Pearson, E. S. (1947). The choice of statistical tests illustrated on the interpretation of data classed in a 2 X 2 table. *Biometrika, 34*(1/2), 139-167. https://doi.org/10.2307/2332518

Perezgonzalez, J. D. (2015). Confidence intervals and tests are two sides of the same research question. *Frontiers in Psychology, 6,* 34. https://doi.org/10.3389/fpsyg.2015.00034

Redish, D. A., Kummerfeld, E., Morris, R. L., & Love, A. C. (2018). Reproducibility failures are essential to scientific inquiry. *Proceedings of the National Academy of Sciences, 115*(20), 5042–5046. https://doi.org/10.1073/pnas.1806370115

Rubin, M. (2017). An evaluation of four solutions to the forking paths problem: Adjusted alpha, preregistration, sensitivity analyses, and abandoning the Neyman-Pearson approach. *Review of General Psychology, 21,* 321-329. https://doi.org/10.1037/gpr0000135

Rubin, M. (2019). What type of Type I error? Contrasting the Neyman-Pearson and Fisherian approaches in the context of exact and direct replications. *Synthese*. https://doi.org/10.1007/s11229-019-02433-0

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13*(2), 90-100. https://doi.org/10.1037/a0015108

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology, 69,* 487-510. https://doi.org/10.1146/annurev-psych-122216-011845

Spanos, A. (2006). Where do statistical models come from? Revisiting the problem of specification. *Optimality, 49,* 98-119. https://doi.org/10.1214/074921706000000419

Stroebe, W., & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science, 9*(1), 59–71. https://doi.org/10.1177/1745691613514450

Zwaan, R. A., Etz, A., Lucas, R. E., & Donnellan, M. B. (2018). Making replication mainstream. *Behavioral and Brain Sciences, 41,* e120. https://doi.org/10.1017/s0140525x17001972

Endnotes

1. The concept of an exact replication can be defined as requiring the duplication of either (a) all possible testing conditions or (b) only those testing conditions that could potentially affect the results of the study. For example, Rubin (2019) defined exact replications in the second way, as requiring "the duplication of all of the aspects of an original study that could potentially affect the results of that study." This second definition implies that researchers are sure about which aspects of their study are relevant (i.e., "could potentially affect the results") and which are irrelevant. Hence, it is similar to the concept of an *equivalent* replication that I discuss later.

In the present article, I adopt the first, more common, definition of an exact replication that requires the duplication of "all possible testing conditions," including both relevant and irrelevant conditions.

2. Following Spanos (2006), we can distinguish between *statistical* and *substantive* adequacy. Statistical adequacy occurs when a statistical model's assumptions (e.g., normal, independent, and identically distributed data for a simple normal model) are sufficiently consistent with the observed data. Substantive adequacy occurs when the characteristics of the statistical model, sample, and testing methodology (e.g., sampling procedure, measures, testing environment, etc.) are sufficiently consistent with a theoretical data generating process or "chance mechanism" (Neyman, 1977, p. 99).

Funding

The author declares no funding sources.

Conflict of Interest

The author declares no conflict of interest.