Applying Perspectival Realism to Frequentist Statistics:

The Case of Jerzy Neyman's Methodology and Philosophy

Adam P. Kubiak

April 2, 2021

Abstract:

I investigate the extent to which perspectival realism (PR) agrees with frequentist statistical methodology and philosophy, with an emphasis on J. Neyman's views. Based on the example of the stopping rule problem I argue that PR can naturally be associated with frequentist statistics. Then I analyze Neyman's conception of statistical inference to conclude that PR and Neyman's conception are incongruent. Additionally, I show that Neyman's philosophy is internally inconsistent. I conclude that Neyman's frequentism weakens the philosophical validity and universality of PR as analyzed from the point of view of statistical methodology.

1. Introduction

Perspectival realism ("PR" hereafter) is a currently developing trend that can be recognized as one of the post-Kuhnian theories of science, within which a remarkable emphasis is put on the fact that cognitive and social dynamics are inseparable elements of the cognitive act and the dynamics of scientific knowledge development (see Collins, Evans 2002). In particular, proponents of PR "share the general idea that there is no 'view from nowhere' and that scientific knowledge cannot transcend a human perspective" (Ruyant 2020), which means the truth condition of a hypothesis depends on an epistemic vantage point, but "it is in part mind-independent facts that make our theories true or false" (Ruyant 2020).

It has been argued that PR harmonizes with many facts and methodological practices in the formulation and development of scientific theories (see, e.g., Massimi 2018b). Although PR is sound when applied to cases of problems from exact sciences, its relation to statistical methodology appears to be not fully established. In particular, the literature lacks a compelling comparison of PR to frequentist statistics, which is one of the major methodological approaches in the sciences. The way scientific statements are formed and accepted with the use of specific statistical language and methods can convey specific metaphysical commitments similarly to how the choice of the language, in general, conveys specific metaphysical assumptions, which has been famously argued by Russell (1905). Therefore, it seems scholarly justified to take a closer look at the interplay between frequentist statistics and PR. Such analysis shall bring about a new perspective from which questions about universality, normativity, and philosophical potential of PR can be posed.

Among the frequentist conceptions, one that could be close to PR is Jerzy Neyman's theory of statistical inference. Neyman was a 20th-century statistician who is recognized as one of the co-founders of the frequentist statistical paradigm, which dominated the methodology of natural and social sciences in the 20th century (Lehmann 1985). His theory has the potentiality of having common grounds with PR because it having both realistic and perspectivistic elements. The first bearing is because sought-out quantities are assumed to be unknown constants that relate to the independently existing world (Neyman 1937, 343-44) and the assertions are based on the conception of avoiding errors of false assertions (Neyman 1952, 55). The perspectivistic element is related to the emphasis on generalizing theory-ladenness, in which “a model has to fit into the methodological framework that is conceived of as more fundamental, or prior, to modeling” (Lenhard 2006, 81). The second purportedly perspectival characteristic of Neyman’s methodological-philosophical conception is the dependence of the conclusive statements on the experimental scheme adopted (see, e.g., Neyman 1934). The perspectival nature is perhaps a more general feature of frequentist statistics as best exemplified in the problem of the rule that governs how sampling is terminated: the conditionality of the shape of the statistical hypotheses (and of the outcome to be drawn from definite evidence) on the rule of termination of data collection (Savage 1962).

The above reasons indicate the need for verifying PR’s consistency with currently persisting scientific methodology and philosophy thereof as proposed by Neyman. The goal of this paper is to investigate whether PR can be consistent with the assumptionsfrequentist statistics with an emphasis on the case study of Neyman’s conception.

The structure of the article is as follows. Firstly, in Section 2. I present the PR assumptions (2.1) and analyze their potential applicability to frequentist testing methods based on the example of the problem of the optional stopping rule (2.2.). Next, in Section 2 I reconstruct Neyman’s conception of statistical inference with an emphasis on his philosophical views and compare his stance with PR. In Section 3 I discuss aspects in which Neyman's methodological and philosophical views are consistent with realism (3.1.) and perspectivism (3.2) and then, in Section 4, I discuss antirealistic (4.1.) pragmatistic (4.2) and antipluralistic (4.3) aspects of his theory. Finally, in Section 5 I offer some solutions for problems raised within the three aspects (5.1-5.3) and offer a generalized philosophical comment in 5.4. In Section 6 I summarize the results.

2. PR as Applied to Frequentist Statistics

2.1. Assumptions of PR

Perspectival realism is a stance that mediates between the extremes of the objective realist philosophy of science on one end and social constructivist on the other. Scientific claims are not non-relatively true but are not merely constructs of social interaction. They are products of interactions with mind-independent reality taken from several perspectives, thus they are true relative to a given perspective, and "not true simpliciter" (see Crețu 2019, 1-2). In addition to that, perspectival realism advocates epistemic pluralism: these perspectival truths are descriptions of mind-independent states of affairs from different points of view that can be incompatible, but, still, equally valid epistemically because any knowledge of dispositional, objectively existing facts concerning objects or processes can only be acquired within a perspective (see Massimi 2012). Finally, that these perspective-relative claims are true regarding the same objectively existing state of affairs implies that they retain, cross-perspectively, their performance adequacy as evaluated from the points of view of the internal standards set by each of the perspectives (see Massimi 2018a, 172).

2.2. The Optional Stopping Case-study

Before I compare PR to Neyman's views on statistical inference, it is essential to show that this conception of philosophy of science can be sensibly applied to explain some features of frequentist statistical methodology and so that PR and Neyman's methodological-philosophical conception have some common ground by sharing, at least partially, the same subject of reference. Otherwise, the value of comparing PR to Neyman's philosophy would be little, just like little value would be in the comparative analysis of, for example, mathematical predicativism and virtue epistemology as the two philosophical theories do not have a shared subject of reference.

An illustration of the application of the perspectival realist stance to the statistical methodology of testing hypotheses could be an analysis of the problem of *optional stopping rules* (see, e.g., Savage 1962; Lindley, Phillips 1976) embodied in the research example of testing a hypothesis about the sex ratio of the pouch young of koala mothers in poor physical conditions (see McCarthy 2007, 31-33).

Assume that the ecological hypothesis in question states that the proportion of males in the population of pouch young is 0.5 (the number of males and females is equal), and the reasonable alternative hypothesis states that the proportion is less than 0.5 (females prevail).

The researcher surveyed 12 koala mothers, each with an offspring in its pouch—three of the offspring were males and nine were females. The data could be obtained in at least two ways: the researcher could sample until the 12th individual was recorded ($S1$), or until the 3rd male was recorded ($S2$). Regardless of the sampling strategy, the data seem to be equivalent and the two alternative statistical inferences are as follows.

Sampling in the case $S1$ is modeled by the binomial distribution that represents the probabilities of collecting n number of females until the number of trials in a sample reaches a fixed value of 12; the sum P_1 of the probability of the observed data (number of females= 9) and more extreme data (in this case of having 10, 11, or 12 females in the sample) equals 0,073 thus the observed female ratio in the sample (0.75), given 0.05 cut-off error rate, is not significantly far from (greater than) the hypothesized population ratio (0.5). The outcome of the test is not to reject the hypothesis.

Sampling in the case $S2$ is represented by a different model—the negative binomial distribution that represents the probability of collecting n number of females until the number of males in a sample reaches a fixed value of 3. The p-value P_2 in this case is the sum of the probability of observation and less probable outcomes: having 10 female records, 11, 12, 13, 14, and so on. The p-value equals 0.033 in this case, so with the conventional 0.05 error rate it is significantly low, thus the conclusion is to reject the hypothesis that the population ratio is 0.5.

The troublesome result is that two different sampling strategies, associated with different statistical models of an experiment, lead to different conclusions about the acceptance/rejection of, allegedly the same hypothesis, in the light of, allegedly, "equivalent" (McCarthy 2007, 37) set of data (evidence) in both cases, consisting of the observation of 9 females and 3 males in a sequence of 12 trials.

Below I put forth four arguments why the above methodological issue could be explained via reference to PR.

I. Both possible observational points of view determine two different, equally valid ways statistical hypothesis is defined but both hypotheses are descriptions of the mind-independent state of affairs, objectively existing population's characteristic—the proportion of pouch young males. This satisfies the PR's assumption of realism.

II. From the perspective of the methodology of natural sciences, a statement obtains scientific meaning once it is framed in such a way that it is possible to empirically verify it with the use of statistical tools. The statement about the characteristic in question becomes scientifically (empirically) meaningful only as framed in the empirical perspective of one or the other observational setup and the related statistical hypothesis (the probabilistic model of observation) that can be true or false. This is related to the fact that the parameter that represents the population ratio starts to bear a scientific meaning only within a particular probabilistic model of potential observations. The parameter as such is a constitutive element of a model and has no mathematical meaning standalone. This means the ratio in question, represented by the parameter, cannot be scientifically tested simpliciter than within a model.

III. The two perspectives differ in terms of knowledge claims. This is because they assume a different possible set of statistical hypotheses (different statements about probability distributions) that are alternative ways of representing an observable manifestation of the same physical reality. That is because the sampling spaces and models that serve to formulate the statistical hypothesis tested are different in both cases. The evidence is also not the same in these two hypothetical cases. The last is due to both cases assuming a different set of relevant information (evidence) used for inferential purposes. In the case of $S2$, a piece of partial information about the order, i.e., information about the location of the third male record in the sequence of trials, is encoded in the (negative binomial) model's random variable. Evidence that was taken into account in $S1$ can be expressed through the proposition: "exactly three males and nine females were recorded in the sample until (and including) the twelfth trial was recorded in the sample". In the sampling framework $S2$, the evidence considered can be expressed in the proposition: "exactly three males had been recorded in the sample until (and including) the eleventh trial and the twelfth trial recorded in the sample was male". It is easy to see that the second evidence implies the first, but not vice versa, therefore the evidence is not equivalent for both cases (see Kubiak 2014, 138-139).

IV. The method assumes the performance adequacy of perspectival statements about a state of the world if this state is true. In the considered example with particular data obtained, conclusions were different in both cases, but this is not inconsistent with PR. PR states that two perspectival statements shall retain performance adequacy if the common state of affairs they refer to is true. The method assumes that if the proportion of males in the population of pouch young is 0.5, then the conclusion from $S2$ will retain high performance adequacy. That is because if an observation with the use of a sampling strategy from $S2$ would be repeated iteratively, then the method anticipates (correct) acceptance of the hypothesis that the population ratio is 0.5 with performance close to the standards set in this method (error risk close to 5%). The same is true for the sampling strategy $S1$. Therefore, if the proportion of males in the population of pouch young is 0.5, then both distributions that express the hypothesis tested are true, namely, the value 0.5 of the parameter p is true in both cases of application of different stopping rules and the statistical hypotheses tested in both cases will retain, cross-perspectively, their performance adequacy relative to standards set for both models.

Points I-IV show that the same objectively existing state of affairs can be scientifically defined via different statistical models that encompass different, incommensurable, observational perspectives. Despite conclusions from testing are to be different for specific evidence possible to be obtained, like the one from the case just considered, the two different tested hypotheses will have the same performance adequacy if the objective state of affairs represented by them is true. The upshot is that PR can have its exemplification in frequentist statistical methodology and by that can be a potentially plausible explanation of some of its troublesome features, like the problem of optional stopping. Then the corollary is that Neyman's conception can be sensibly juxtaposed with PR. In the subsequent two sections of this paper, we investigate Neyman's frequentist methodology taken jointly with his philosophical interpretation thereof as compared to PR. I start with explicating in Section 3 two elements of Neyman's view that are consistent with PR. Next, in Section 3 I discuss elements inconsistent with PR. Some elements of Neyman's thought presented in those two sections have not yet been presented to the debate in the philosophy of statistics.

3. Neyman's Theory—Elements Coherent with PR

Jerzy Neyman was not a professional philosopher, therefore in communicating his philosophical views he did not use the terminology commonly used in the relevant philosophical debates. Nonetheless, part of his philosophical stance has been explicated and disputed in the philosophical literature (e.g., Hacking, 1965; Mayo Spanos 2006), In this section, we structurize those parts of his conception that could be viewed as realism-like and perspectivism-like.

3.1. Neyman's Views and Realism

Some of Neyman's basic methodological and meta-methodological conceptions appear to match with realists' ideas. Firstly, Neyman did not reject the assumption of the existence of independent reality (an ontological aspect of realism). Virtually every time he talked about the conceptually unknown *true value* of the hypotheses' parameters. The values of hypothesis' parameter(s) that a researcher asks about, were to Neyman "generally unknown constants" (Neyman 1937, 343). The constant value of the statistical model's parameter(s) is as such a mathematical concept, but Neyman writes that "there are real objects that correspond to these abstract concepts in a certain sense" (Neyman 1952, 24). Therefore, the truthfulness of the value of the hypotheses' parameter would mean that this value somehow corresponds to, or denotes an unknown, but independently existing state of affairs in the real world. By that, Neyman seems to be assuming at least ontological realism. What is the nature of the said correspondence?

The general idea of applying statistical schemes to experiments/observations is to "assume that the real value of the sought-after quantity exists […] and—based on laws of large numbers—to seek for calculable measurement results' functions that can be considered approximations of the 'true value' and mean error" (1923a[1], 19, auth. transl.). Therefore, it appears that to Neyman the ideal is to come up with conclusions, in the form of the values of these functions, where "numerical values of mathematical formulas more or less agree with the results of the actual measurements" (Neyman 1952, 24). values of these functions of actual measurements are expected to be approximately the same as the "real values" that exist independently in the real world, which assumes an epistemic realist approach. What will be important in my further analysis is that Neyman speaks in the plural when he refers to "functions", which indicates possibly different functions to be used to yield the outcome based on the empirical evidence obtained. Still, all these possibly different outcomes are thought to both agree with the evidence to a certain extent and to approximate the objective truth.

The conception of the reliability of the method of statistical inference is anchored in the conceptions of the probability of two types of error: the probability of rejection of the tested hypothesis if it is true and the probability of acceptance of an alternative if it is true (Neyman 1952, 55). A true hypothesis is one in which the stated parameter range covers this unknown, true, real value.   The method's reliability is based on performance in yielding true conclusions in the long run. Therefore, a kind of epistemic realism seems to be something that drives the method's reliability in the long run of use.

Finally, Neyman required the research schemes to be adjusted to the real-world factors that exist objectively and independently of the research scheme. Ignoring these factors might affect the correspondence between a physical (substantial) and a statistical hypothesis. Neyman's illustration of this issue refers to the famous Fisher's toy example of a hypothesis that a lady cannot tell whether the tea or milk was poured in the cup first based on the taste of the tea. An independent factor would be, for instance, an association of the lady's impression of a definite sequence of pouring with the thickness of the cup, which the lady can feel with her lips. If the experiment scheme does not take this into account and it happens that one of the two pouring methods is predominantly used with thinner cups and the other with thicker ones, then the substantial hypothesis of lack of ability may be true, while the corresponding statistical hypothesis—the distribution of probabilities of possible experiment's outcomes under the assumption of lack of ability—will be false (Neyman 1950, 282-291).

It can be concluded that some of the very foundational methodological and meta-methodological conceptions of Neyman appear to match with realists' ideas. Hypotheses are speaking of independently existing reality and they are true or false concerning this reality. Moreover, the whole research scheme is expected to be adequate in respect to independently existing, real factors. This confirms the presence of some ideas of ontological and epistemic perspectivism in Neyman's thought.

3.2. Neyman's Views and Perspectivism

Despite true statistical hypothesis represents the real value of a quantity existing independently in the world it does so via rendering the empirical meaning to this real value. A statistical hypothesis is a statement about the probability (density) distribution of a random variable where a random variable is a function of a set of random phenomena obtained in the effect of performing a random experiment. This means the distribution, and so the hypothesis is partially a product of specificity of the observational (experimental) set-up. Specifically, the notion of probability as used in the statistical hypothesis is that it does not refer to physical objects, or the properties of physical objects, but to the properties of physical events that correspond to an observational setup; in other words, the probability is ascribed not to objects, but events related to an observational setup (Neyman 1952, 10-12). This is visible, for example, in Neyman's comment on Jeffrey's toy example of a case of two boxes. One containing one white and one black ball, and the other—one white and two black ones. Firstly, a box is to be randomly selected and then a ball at random from that box. Consider Neyman's definition of probability "the probability, $P(B|A)$, of an object A having the property B will be defined as the ratio $P(B|A) = m(B)/m(A)$" (Neyman 1937, 337) When applied to this toy example, it is not the probability of the ball selected having the property of being white:

> "the objects A are obviously not balls, but pairs of random selections, the first of a box, and the second of a ball [thus], the probability sought is that of a pair of selections ending with a white ball" (Neyman 1952, 11)

Thus, in the eyes of Neyman, probabilities directly refer to properties of observational designs or procedures. Statistical hypotheses are statements about probability (density) distribution and by that, they are relativized to those designs in the same way. Even if they are to represent substantial hypotheses about the mechanisms or other characteristics of an objectively existing reality, they do so only through the perspectives of experimental constructs that determine what can be experienced. This Neyman's view is in line with the consequences of the stopping rule problem discussed in Section 2.

It appears then, that Neyman found scientific statements formulated with the use of statistical tools to be always relative to the perspectives of idealized assumptions and experimental constructs but at the same time to refer to the perspective-independent, true states of affairs: real parameter values and real experimental setups and circumstances. He thought a fraction of these statements to be true (in the classical sense) to the extent defined by the error rates.

Additionally, he appeared to accept the possibility of equally valid perspectives on the foundational assumptions of scientific methodology:

> "[…] in theoretical work, the choice between several equally legitimate theories is a matter of personal taste. In problems of application, the personal taste is again the decisive moment, but it is certainly influenced by considerations of relative convenience and empirical facts" (Neyman 1937, 336 footnote *).

The theories here refer to methodological frameworks. If they can be "equally valid", then one can speak of the epistemic pluralism of perspectives, which was told to be an element of PR.

The presented elements of conceptions of Neyman that can be regarded as realism-like and perspectivism-like make his views fairly consistent with perspectival realism thus far. Nevertheless, other important elements of Neyman's approach seem to be inconsistent with PR and also make Neyman's views internally inconsistent. These are the theses about the fictional character of scientific concepts, the pragmatistic (non-epistemic) interpretation of a scientific assertion, and the idea of normative anti-pluralism. I discuss these three topics in the follow-up section.

4. Neyman's Theory—Elements Potentially Inconsistent with PR

4.1. Fictional Nature of Scientific Concepts

Due to Neyman statistical hypotheses are stated under idealized assumptions which are false regarding the real world and empirical evidence:

> "The objects in the real world, or rather our sensations connected with them, are always more or less vague and since the time of Kant, it has been realized that no general statement concerning them is possible. The human mind grew tired of this vagueness and constructed a science from which everything that is vague is excluded—this is mathematics. […] there are many mathematical theories that are successfully applied to practical problems. However, this does not mean that these

theories deal with real objects [...] the theory [of mathematical statistics] itself deals with abstract concepts not existing in the real world" (Neyman, 1952, 23-24).

This might suggest that Neyman believed that there is no truth-correspondence between scientific models and the real world. This seems to explicitly contradict the elements of Neyman's views presented in 3.1. At this point, Neyman appears to be ambiguous on whether he identifies the "real" world with the world of physical "objects" (or systems) or the world of "sensations connected with them"—empirical observations, or equals the two. Nonetheless, the problem of weak correspondence between hypotheses (probability distributions defined with the use of abstract concepts) and the evidence seems to be the major issue for him as "no observations are capable of producing the value of a given probability" (Neyman 1957a, 15). Below I develop this point by referring to Neyman's writings.

The connection of scientific evidence from particular research with a statistical hypothesis is problematic because of the limiting theorems of statistics and the regularity-type character of the hypothesized statements. In Neyman's view, probabilities must be understood as describing these regularities, so the existence of a (sufficient) correspondence between evidence and hypothesis would require sufficiently long sequences of replications of the research on a particular matter, which is nowadays regarded as usually an unfulfilled condition (perhaps except for quality-control cases of research) and an ideal difficult to be achieved (see Rubin 2020). Therefore, the correspondence between an observation from single research and a statistical hypothesis must be regarded as very weak.

The problem of the correspondence between evidence and scientific statements is not restricted only to statistical hypotheses understood as idealized empirical models of natural mechanisms, or of population characteristics. The same problem prevails at the level of a single trial and empirical evidence obtained from it. This can be explained by referring to one of Neyman's first papers (1923a[1])[1], where the author introduced a general design for a field experiment conducted for the sake of comparison of different varieties of crops concerning their potential yields.

He considered there the design of the experiment based on the random assignment of seeding to plots in an experimental field. Each seeding ends up with what he called *true yield*. However, the outcome of the measurement of a yield of certain yeast varieties measured (with high accuracy) at a particular plot is not the true yield of that variety at that plot, which is itself an unknown, fixed value (Neyman 1923a[2], 465-67). This divergence is due to the technical error of the measurement. The true yield itself is an idealized conception, namely, the mean value from indefinite repetitions of the measurement with all conditions being equal except for the differences in random technical error that causes inaccuracy of an experimental technique. This kind of error is different than the error in statistical inference about the hypotheses and "no statistical methods can improve the

[1] Originally published in Polish (English translation of the original title is: An Attempt to Justify The Application of Probability Theory to Field Experiments (transl.—AK). The paper's core section was edited and translated to English in 1990 by D.M. Dąbrowska and T.P. Speed. In this thesis reference to the Polish original will have the form of "1923a[1]" and to the fragment translated in 1990 the form of "1923a[2]".

accuracy of the experiment beyond the limits fixed by the technical random error" (Neyman et al. 1935, 110). Therefore, there is no equivalence between the true yield from a particular trial and an observed yield regarding this trial.

The differences between the two conceptions become striking when one realizes that the true yield at a particular trial (plot) is essentially a priori counterfactual state of affairs because of the infinite number of counterfactual unrealized measurements involved in the conception of the true yield (see Rubin 1990). To stress the lack of equivalence between a scientific concept and observable facts, Neyman distinguished two different meanings of terms (such as *yield*) when used in two different aspects of the scientific process: in describing empirical data (Neyman called it "pure empiricism"), and in making inferences to a scientific scheme (Neyman 1923a[1], 18). In the first case, one is speaking of the result(s) of empirical observations (measurements)[2] and in the second—of scientific concepts that put these observations into more general frames. The specificity of using a term in a sense of it being a scientific concept is that "all scientific terms, which are defining properties and relations between investigated objects, are fictions" (Neyman 1923a[1], 18). The *true yield* in Neyman's conception is an example of such a scientific, and therefore fictional expression. Letting alone linguistic oddity, this upshot is ambiguous to the extent that Neyman seems to determine this fictional character of the concept as related to the outcome of observation, not to the unknown, real-world property or propensity of an object or experimental setting. By that, he appears to conflate the notion of the real world

[2] In contemporary apparatus this kind of activity would be labeled *descriptive statistics*, as opposed to inferential statistics.

with the notion of results of empirical observations. Anyhow, the idea that scientific concepts are fictional somewhat contradicts his realistic views as presented in 3.1.

4.2. The Pragmatistic (Non-epistemic) Interpretation of a Scientific Assertion

The second element of Neyman's theory possibly inconsistent with PR is the stance that acceptance of a scientific statement does not yield any belief about the truthfulness of a particular scientific statement:

> "The terms 'accepting' and 'rejecting' are very convenient and are well-established. It is important, however, to keep their exact meaning in mind, and to discard various additional implications which may be suggested by intuition. Thus, to accept [or reject respectively] a hypothesis H means only to decide to take action A rather than action B. This does not mean that we necessarily believe that the hypothesis is true [or false respectively]" (Neyman 1950, 259).

Neyman and Pearson even stressed that acceptance/rejection of a particular hypothesis could not—for methodological reasons—be understood epistemically:

> "[…] as far as a particular hypothesis is concerned, no test based upon the theory of probability can by itself provide any valuable evidence of the truth or falsehood of the […] hypothesis" (Neyman, Pearson 1933, 291).

This impossibility came to agree with Neyman's moral-like postulate: "The beliefs of particular scientists are a very personal matter and it is useless to attempt to norm them by any dogmatic formula" (Neyman 1957b, 16).

The stated impossibility and undesirability of the epistemic interpretation of outcomes of the application of the frequentist procedures is reflected in Neyman's pragmatist interpretation of the goal of the method of scientific investigation with the use of statistical tools. Although "[…] theory was born and constructed with the view of diminishing the relative frequency of errors, particularly of 'important' errors" (Neyman 1977, 108), an acceptance of a hypothesis is an act of will to behave as if the hypothesis was true, based on the assumption that the method, that we are using to do this, is reliable enough not to lead us astray from the truth in a sufficiently large fraction of practically important cases. That is why the final stage of accepting a hypothesis

> "[…] amounts to taking a 'calculated risk', to an act of will to behave in the future (perhaps until new experiments are performed) in a particular manner, conforming with the outcome of the experiment. It is this act of adjusting our behavior to the results of observations, that is the overlooked element of the final stages in scientific research and that is covered by the term 'inductive behavior'" (Neyman 1957b, 12).

This act of will is already present at the stage of choosing a test (decision rule) that has the desired properties (performance characteristic) based on pragmatic considerations:

> "The adoption of hypothesis H when it is false is an error qualitatively different from the error consisting of rejecting H when it is true. This distinction is essential because, with rare exceptions, the importance of the two errors is different, and

this difference must be taken into consideration when selecting the appropriate test" (Neyman, 1950, 261).

Realism seems to presuppose that scientific conclusions are accepted based on the ideal of them being at least approximately, or probably true. It appears that it would be hard to assimilate epistemic realism with the fact that particular scientific conclusions (outcomes of performing a statistical test) are effects of the need to fulfill pragmatic goals, that they cannot be evaluated in terms of their truthfulness and cannot pose a basis for beliefs.

4.3. Anti-pluralistic Elements in Neyman's Conception

In Section 2, I have shown that frequentist statistical methodology is consistent with perspectivism. In subsection 3.2. I indicated that Neyman's understanding of scientific claims was also that they have a perspectival nature. Additionally, I indicated his advocacy of a certain degree of pluralism of methodological perspectives. In what follows I show that his views on this matter are at least ambivalent and are not fully consistent. I argue that from the viewpoint of Neyman's conceptions pluralism should be understood as a description of the circumstances that a researcher faces but which should be avoided whenever possible.

Neyman's thought has two anti-pluralistic aspects. Both relate to taking somewhat *God's eye's* view. One is the perspective that could be called the in-theory perspective, and the other is the perspective of justifying the theory from a meta-level point of view. I first discuss the inside the theory perspective. It can be further divided into bottom-up and top-down kinds of anti-pluralism. The first is related to the epistemic adequacy of models/setups the second to the epistemic efficacy of statistical inference.

The bottom-up kind is the one that aims at searching for (selecting) the model of an experiment that is optimally adjusted to physical reality (see Neyman 1950, 282-291). Neyman pinpointed some crucial aspects by referring to the) example of tea tasting lady (see 3.1). To perform the experiment, one has to determine an adequate set of admissible hypotheses. For example, should the alternative hypothesis to the one that the lady does not have the ability (she makes random guesses) point in the direction of a perfect guess or perfect misguidance? The lady may be able to discriminate between pouring methods, but simply conflate one method with the other. Another issue is whether one is asking about the lady being able to discriminate between the two methods or identify each. In the second case, the cups should not be judged by comparison between the two in the pair, but independently. However, what if she can identify one of the methods, but is uncertain about the other? Does she know how many cups made with one of the methods she will be given? If so, then the trials should be treated as dependent. Finally, it is essential to arrange a proper technique for a random experiment, in which any factor that may affect the correspondence between a physical and a statistical hypothesis is neutralized by its randomization, of which the example—that refers to the order of pouring—I provided already in 3.1. As the reader can see, taking into account manifold factors prompts one to seek the best experimental set-up rather than to treat different possible setups as equally valid.

Another example of the bottom-up anti-pluralist approach of Neyman is his theory of utilizing sampling design for the sake of optimal estimation (1934; 1938) Speaking of estimation in this context is equally valid as speaking of hypothesis tests, as there is a duality between hypothesis tests and Neyman's technique of estimation by intervals: an estimation technique is tantamount to performing a series of hypothesis tests (see Neyman 1937, 372; Lehmann, Romano 2005,164-168). Neyman develops techniques of how to maximize the accuracy of estimation by taking into account some additional facts about the structure of the population studied in terms of some auxiliary factors. The technique assumes consideration of several, mathematically equally valid, ways of how a sample could be drawn from the population, to choose the one that is, from the perspective of this knowledge, the most accurate sampling design.

All the above show that Neyman advocated achieving the optimal adequacy between the theoretical models of observation and all known aspects of the investigated reality by fulfilling several specific conditions like those above, thus by narrowing down the possible observational perspectives from which a test of a hypothesis could be performed to the one that best corresponds to reality.

The top-down type of Neyman's anti-pluralistic view on the choice of research perspective is perhaps best exemplified by the normative requirement in using a test of which the probability of correctly rejecting a hypothesis would be maximal for a preassigned error of the first type:

> “if two different critical regions w_1 and w_2 are suggested, both insuring the same probability of error of the first kind, then the choice between these regions depends on their effectiveness in controlling the error of the second kind” (Neyman 1950, 304).

Originally, the rule was presented as applied to choosing among several equally acceptable test statistics, but as such, the idea of minimizing the error of the second kind when choosing between equally acceptable, equivalent mathematical models, can be applied when the choice is to be made between equivalent ways of collecting the data. From this point of view, the two alternative perspectives adopted in the discussed example of testing the hypothesis of the number of males and females of pouch young being even will not be equally valid. If the consideration of a test’s power function “seems to be the proper rational basis for choosing the test” (Neyman, 1952, 58), then the perspective of sampling design related to $S2$ is methodologically preferable as it guarantees test’s power to be higher than the perspective of sampling design related to $S1$. For example, if the true value of the ratio in question was 0.75, then the power to detect it in the case of adopting the stopping rule and the model related to $S1$ would equal 0.39 and in the case of $S2$ it would equal 0.455.

Neyman’s meta-methodological views also appear to contrast with the plurality of perspectives by suggesting that, in principle, some methodologies will be more optimal for particular cases than others. For example, Neyman admitted that the Bayesian methodological framework for testing or estimation can be mathematically perfectly valid, and yet he believed that regarding the usage of the method, the Bayesian approach "may be applied in practice only in quite exceptional cases" (1937, 343).

5. Philosophical Consequences

I have argued, that PR can have its exemplification in frequentist statistical methodology and be a potentially fruitful explanation of some of its troublesome features. Nevertheless, when applied to Neyman's frequentism, PR turns out to be only partially consistent with it.

Neyman speaks about the existence of real numerical values and real objects but is equivocal on how scientific concepts refer to the real world. The method's reliability is entrenched in its performance in yielding true outcomes, but, diversely, the acceptance of hypotheses is dependent partially on pragmatic considerations, and what is more, the truthfulness of particular scientific assertions cannot be assessed. Neyman appears to stress the perspectival nature of scientific investigations and statistical concepts, but, diversely, his methodological advice is implying that different perspectives do not have equal validity and perspectivism should be avoided at different stages and levels of this investigation.

The incoherences of Neyman's methodological philosophical views, which emerged as a result of the analyzes in Sections 3 and 4, prompt the need for an attempt to reinterpret his methodology from the perspective of its possible congruency with PR. In this section, I scrutinize the incoherence with this attempt.

5.1. Overcoming the Unclear Status of Scientific Concepts

I indicate in 5.1. that in discussing the relations between the real world and hypothesis, Neyman was most clear and unequivocal in explaining what is the relation between the hypothesis and evidence. Although the consequence of a weak connection between evidence and statistical models is that the connection between evidence and independent reality is weak too, this does not mean models (and thus statistical hypotheses) are fictitious if it comes to their relation to independent reality ("real world"). This is because the real world does not reduce to the world of empirical data. Neyman seems to mistakenly equal the two notions (the real world with empirical data) when speaking of the fictional character of scientific concepts. He indicated that their fictional nature can be expressed by contrasting them with "empiricism". If a model of the empirical outcome does not conform to a single empirical outcome, this does not mean the model is false or fictitious in terms of it representing the independent reality of more generalized mechanisms and propensities that are the (probabilistic) cause of obtaining an outcome of a certain type. Neyman's statement about the fictionality of scientific terms should be understood as pertaining to a weak correspondence evidence—model. Such a view does not contradict the realistic interpretation of statistical terms and hypotheses. The unknown value of true yield, which can only be estimated, or assumed to be equal to the observed value for practical simplicity, can be representing the propensity of physical (system of) objects to behave in certain, observable, ways under repeated observations in certain conditions. Although it is impossible to have an exact empirical realization of this hypothesized behavior, the accepted scientific statement can still be regarded as approximately true as regards the features of the real world with which the observation on which basis the statement was accepted agrees to a sufficient degree.

In 4.1 I indicated Neyman's view that the real world is "vague" while the mathematical models that describe these ontic states are idealizations that deviate from these ontic states and that are "fictitious" which means false. Take, for example, the hypothetical example of a wheel of fortune that was meant by the designer to be fair (which is known to be rather uncommon). The ontic mechanism/character of its propensity must differ from the mathematical model that describes it, as there will never be such a perfect symmetry in it as described by the designer's model. However, idealization does not exclude the approximate truthfulness of such models therefore realism remains in force.

5.2. Overcoming The lack of Epistemic Interpretation of Single Outcome

Neyman seems to be ambiguous as far as realism is concerned, in his insisting that a statistical method is a tool for making pragmatic decisions rather than acquiring true beliefs. This is because he seems to simultaneously assume the method's reliability in yielding those pragmatically useful conclusions, in the long run, is based on the method's reliability in yielding conclusions that are sufficiently often true in a realistic sense, for which I argued in 3.1.

Epistemic realism seems to be in force regarding a body of assertions being an effect of the uses of N-P. This is due to error probabilities based on which the procedure may be deemed reliable when it is iteratively used in manifold research contexts with different hypotheses and error rates set at different levels. According to Neyman, the Central Limit

Theorem allows us to conclude that the relative frequency of error will be close to the arithmetic mean of the error regardless of context (Neyman 1977, 108-109). This means the average error is an indicator of how a big part of the assertions from a body of outcomes of statistical tests is true, although the question of the truthfulness of any particular one must be abandoned.

It appears then that a special case of realism may apply to Neyman's methodology. This specificity of a realistic interpretation of scientific outcomes would here mean the applicability of realism to them being understood as a collective of outcomes that jointly forms a body of scientific knowledge. One could also imagine the perspective of the same true hypothesis being iteratively tested through replications of an experiment that are never identical, thus could be regarded as perspectival. When the error risk is set to be the same, the idea of long-run performance in accepting the hypothesis across tests is somehow connected to the idea of cross-perspectival performance adequacy.

5.3. Bogging Down in Neyman's Anti-pluralistic Inclinations.

In the already cited (in 3.2.) Footnote from the 1937' paper Neyman indicates that when choosing statistical theories (methodological framework) in the context of problems of application, the "personal taste" remains "decisive", but in the fragment from the same paper cited at the very end of 4.3. he suggests that the choice is determined by the objective context of research; this could be for example prior knowledge of the population studied, or of conditions of the experiment. In one of the later works, he is even more emphatic by claiming that there should be no "dogmatism" regarding application aspects:

> "What I am opposed to is the *dogmatism* which is occasionally apparent in the application of Bayes' formula when the probabilities *a priori* are not implied by the problem treated, and the author attempts to impose on the consumer of statistical methods the particular *a priori* probabilities invented by himself for this particular purpose" (Neyman 1957b, 19).

Neyman allows methodological choices to be based on "personal taste". Methodological choices can determine the outcome of the application of statistical procedures.[3] However, "inventing" the prior also determines the outcome, and can also be understood as a methodological choice based on "personal taste". Why the influential personal taste is acceptable when Neyman speaks of the statistical methodology adopted by him and is not acceptable when Neyman speaks of the Bayesian statistical methodology remains unexplained and seems inconsistent.

This is a striking example of the incompatibility of Neyman's statements concerning statistical constructs as applied to reality: there is a tension between the idea of the decisiveness of "personal taste" in choosing the mathematical construct, which entails the acceptance of the pluralism of perspectives, and the tendency to eradicate the equivalence of perspectives both at the meta-methodological and methodological level.

This inconsistency can be explained by narrowing down this decisive element to only one aspect—the pragmatic choice of standards for the two types of error, which is indeed inherently pluralistic since "this subjective element lies outside of the theory of

---

[3] I exposed this in Section 2.

statistics" (Neyman 1950, 263). Therefore, Neyman's method can be regarded as advocating pluralism in respect of the pragmatic-value driven differentiation of error risks and only in this aspect. However, this does not solve the problem of Neyman's inconsistency. Both setting error risks in frequentism and setting the prior probability of a hypothesis in Bayesianism are factors that influence the outcomes of the application of the respective procedure. The choice of which type of error is more important (and to what extent) than the other can be—due to Neyman—burdened with subjective, personal taste, but the choice of the prior cannot. It appears that the problem boils down to the question of what can be assumed to be "outside of the theory of statistics", and what—to belong to it. Neyman seemed to be pushing pluralism aside to this outer world and to decide that setting the nominal value of the error rate is of this outside-world type while setting the value of the prior probability would be of the inside-world type. Yet, he did this in a somewhat arbitrary way.

Anyhow, the discussed Neyman's pluralism-like declaration happens to be an empty declaration. What I have depicted in 4.3, Neyman is promoting the avoidance of pluralism of perspectives in several respects except for one specific aspect—the setting of error risks. In the follow-up subsection, I will go back to this issue to reveal that this is also an anti-perspectival feature. Neyman assumes, despite his anti-pluralistic inclinations, that full eradication of pluralism of perspectives is not achievable not only because of the assumed pluralistic nature of pragmatic perspectives represented by different settings of error rates. The other difficulty is related to the top-down kind of methodological anti-pluralism. As I indicated in 4.3, it comprises of preference for the mathematical function of the data that decreases the error of the IInd type. However, as Neyman and Pearson (see, e.g., 1933, 298) specify, it is not always possible to find among several possible functions, the one that will minimize the risk of error of the IInd type for each possibility of realization of an alternative hypothesis.

All things considered, Neyman's view has to be regarded as incongruent with perspectivism because of Neyman's conditional acceptance of the pluralism of perspectives only for the cases in which he seems not to see any solution within the theory of statistics that would help to eradicate it.

5.4. Perspectival Realism? Yes, but No.

The case of stopping rule issue considered in Section 3 suggests that from a general point of view PR might be found fairly consistent with frequentist statistical methodology. This is because if different perspective-relative claims—statistical hypotheses—are true regarding the same objectively existing state of affair, they will retain, cross-perspectively, their performance adequacy—long-run acceptance rate—as evaluated from the points of view of the internal standards—error risks—set by each of the perspectives., if a hypothesis tested is false and an alternative hypothesis is true the difference in power to detect this truth, and so the difference in performance adequacy of the alternative can be rationalized by PR the same way.

Nonetheless, when one considers the version of this methodology as presented by Neyman things gets complicated. Realism remains a plausible stance. The assumption of the existence of real-world natural mechanisms, or a population of objects with their natural attributes is consistent with Neyman's stance. When a body of assertions is taken into account it can be assumed that an expected number of them can be regarded approximately true as regards those mechanisms or population characteristics. Perspectivism is much less in line with Neyman's views. Hypotheses are statistically defined relative to perspectives of observational design and mathematical framework adopted, but Neyman is in principle against pluralism on a meta-methodological and methodological level. He advocates certain methodological norms for discriminating those possible methodologies and research designs that would be epistemically best. This means to seek frameworks and designs that would be most empirically adequate and that would minimize error risk (or maximize accuracy), even though this is not always possible, he admits.

Interestingly, the alleged methodological philosophical stopping rule riddle that can be explained by applying the philosophical principles of PR could be alternatively resolved by—competitive in this context—the principle of optimization of power that is of methodological, not philosophical, nature. This suggests that the philosophical conceptions that encapsulate, rationalize, and harmonize some of the known cognitive problems of the sciences may become superfluous once proper methodological solutions are devised.

Perhaps PR could be modified to better correspond with Neyman's views. Firstly, instead of it being perceived as the globally unavoidable and best philosophical description of the nature of scientific process ad its outcomes, PR's explanatory value could be relativized to some situations or aspects. Secondly, PR could be understood not as a normative guideline for scientists and methodologists, but as a temporarily plausible explanation of some states of affairs in science and scientific methodology that may in some respects become invalidated in time with the growth of knowledge or development of scientific methodology. As Neyman's frequentism is not the only approach to statistics and philosophy thereof an interesting research quest would also be to juxtapose PR with other methodological paradigms.

6. Conclusions

By using the example of the stopping rule problem, I argued that PR naturally conforms to frequentist statistical methodology. Nonetheless, PR and frequentism are incongruent as far as Neyman's methodological-philosophical frequentist conception is concerned.

What makes Neyman's stance close to perspectival realism is that scientific statements are designed to be expressed through observational and conceptual perspectives and at the same time they refer to the perspective-independent, states of affairs in the real world.

Nonetheless, in line with his ideas, epistemic realism can only be accepted when a body of scientific outcomes is taken into consideration. Additionally, Neyman is ambiguous in his distinction of the real world from empirical observations. The particularly troublesome upshot of this is that statistical hypothesis can be "true", he believes, but at the same time it is always "fictitious".

A solution would perhaps be to tell that only a part of it can be strictly true about the real world—the value of the model's parameter, which accurately represents the mechanism or characteristic of some part of the real world. Moreover, to tell that the whole hypothesis, which is a statement not only about the value of the parameter but also about empirical predictions relative to empirical setup given this parameter value, is approximately empirically adequate. This approximate and not complete empirical adequacy of idealized description could be then understood as what Neyman means when he indicates the fictional character of a scientific concept.

A bigger problem with reconciling PR with Neyman's frequentism is that Neyman avoids pluralism in adopting perspectives, whereas the assumption of this pluralism is a precondition for perspectivism. An interesting fact is that Neyman is internally inconsistent by appreciating the element of decisiveness in individual choices that influence the outcome when he speaks of frequentism, but condemns making choices that affect the outcome when he criticizes Bayesianism. Surprisingly, the only acceptable by Neyman aspect of appreciating pluralism of perspectives—the aspect of adopting error risks based on practical considerations—turns out to oppose PR. Nevertheless, there are some aspects where pluralism of perspectives seems to be unavoidable.

The analysis of frequentist methodology and philosophy as of Neyman's shows that PR is not a universally valid philosophical stance. If one adopts Neyman's methodology, assumptions of PR become problematic or even incongruent with what the methodology implies (pluralism of perspectives is discarded). This fact could perhaps be well formulated by rephrasing existentialists' famous maxim and saying that methodology precedes ontology. Perhaps a solution to make PR and Neyman's frequentism more coherent could be a case (or aspect)-dependent relativization of PR. But Neyman's frequentist philosophy is ambiguous, and frequentism is not the only paradigm in the philosophy of statistics. Therefore the inconsistency between the two is not decisive. To verify the general level of (dis)agreement between PR and statistical methodology would require further reflection from the viewpoint of other, alternative to Neyman's, methodological-philosophical approaches.

References

Collins, Harry M. and Evans, Robert. 2002. "The third wave of science studies: Studies of expertise and experience." *Social Studies of Science* 32: 235–296.

Creţu, Ana-Maria. 2020. "Perspectival Realism." In Encyclopedia of Educational *Philosophy and Theory*, ed. M. A. Peters. Singapore: Springer.

Fisher, Ronald A. 1956. *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd.

Hacking, Ian. 1965. *Logic of Statistical Inference*. London: Cambridge University Press.

Kubiak, Adam P. 2014. "A Frequentist Solution to Lindley & Phillips' Stopping Rule Problem in Ecological Realm." *Zagadnienia Naukoznawstwa* 50 (200): 135-145.

Lehmann, Erich L. 1985. "The Neyman-Pearson Theory After Fifty Years." In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, *Volume 1*, ed. L.M. Le Cam, 1047–1060. Michigan: Wadsworth Advanced Books & Software.

Lehmann, Erich, L., Romano, Joseph, P. 2005. *Testing Statistical Hypotheses*.Third Edition. New York: Springer-Verlag.

Lenhard, Johannes. 2006. "Models and Statistical Inference: The Controversy between Fisher and Neyman–Pearson." *The British Journal for the Philosophy of Science* 57: 69–91.

Lindley, Dennis V. and Phillips, Lawrence D. 1976. "Inference for a Bernoulli Process (a Bayesian View)." *The American Statistician* 30 (3): 112–119.

Massimi, Michela. 2012. "Scientific perspectivism and its foes." *Philosophica* 84(1): 25–52.

——— 2018a. "Perspectivism." In *The Routledge Handbook of Scientific Realism*, ed. J. Saatsi, 164-175. Oxford: Routledge.

——— 2018b. "Three problems about multiscale modeling in cosmology" *Studies in History and Philosophy of Modern Physics,* 64**:** 26–38.

Mayo, Deborah, Spanos, Aris. 2006. "Severe Testing as a Basic Concept in a Neyman-Pearson Philosophy of Induction." *The British Journal for Philosophy of Science*, 57 (2): 323–357.

McCarthy, Michael A. 2007. *Bayesian Methods for Ecology.* Cambridge: Cambridge University Press.

Neyman, Jerzy. Spława-Neyman, Jerzy. 1923a[1]. "Próba uzasadnienia zastosowań rachunku prawdopodobieństwa do doświadczeń polowych." Roczniki Nauk Rolniczych i, Leśnych 5: 1–51.

——— Spława-Neyman, Jerzy. 1923a[2]. "On the Application of Probability Theory to Agricultural Experiments. Essay on Principles. Section 9." Part of (Neyman 1923a[1]) transl. by T. M. Dabrowska, T. P. Speed. Statistical Science 5: 465–472.

——— 1934. "On the Two Different Aspects of the Representative Method: The Method of Stratified Sampling and the Method of Purposive Selection." *Journal of the Royal Statistical Society* 97 (4): 558–625.

——— 1937. "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.

——— 1938. "Contribution to the Theory of Sampling Human Populations." *Journal of the American Statistical Association* 33:101–116.

——— 1950. *First Course in Probability and Statistics*. New York: Henry Holt and Co.

——— 1952. *Lectures and conferences on mathematical statistics and probability*. Washington: U.S. Department of Agriculture.

——— 1957a. "The Use of the Concept of Power in Agricultural Experimentation." *Journal of the Indian Society of Agricultural Statistics* 9: 9–17.

——— 1957b. "'Inductive Behavior' as a Basic Concept of Philosophy of Science." *Revue De L'Institut International De Statistique* 25 (1/3): 7–22.

——— 1977. "Frequentist probability and frequentist statistics." *Synthese* 36 (1): 97–131.

——— and Pearson, Egon, S., 1933. "On the Problem of the Most Efficient Tests of Statistical Hypotheses." Philosophical Transactions of the Royal Society of London. Series A 231: 289-337.

Russell, Bertrand. 1905. "On Denoting." *Mind*, 14(56): 479-493.

Rubin, Mark. "Repeated sampling from the same population?" A critique of Neyman and Pearson's response to Fisher. Euro Jnl Phil Sci 10, 42 (2020). https://doi.org/10.1007/s13194-020-00309-6

Ruyant, Quentin. 2020. Perspectival realism and norms of scientific representation. Euro Jnl Phil Sci 10, 20 (2020). https://doi.org/10.1007/s13194-020-00285-x

Savage, Leonard. ed. 1962. *The foundations of statistical inference: A discussion.* London: Methuen.