Why does statistics matter to philosophy?

Jun Otsuka[1]

[Note: this is a minimally edited machine translation of an article originally written in Japanese. The writing and references are optimized for Japanese readers. Please cite this article as: Otsuka, J. (2021) Why does statistics matter to philosophy? (in Japanese), *Tetsugaku-Kenkyu (The Journal of Philosophical Studies)*, 606: 1-24.]

Abstract

This article explores the intersection of philosophy and statistics by examining the philosophical assumptions underlying modern mathematical statistics from ontological and epistemological perspectives. Statistics holds interest for philosophers engaged with the problem of induction, as its mathematical apparatus serves as models for philosophical ideas. For instance, the much-discussed concepts of the uniformity of nature and natural kinds correspond to probability models and statistical models, which are fundamental to various statistical methods. Similarly, Dennett's concept of a real pattern echoes the spirit of various information criteria (such as AIC) used to determine the optimal level of complexity for maximizing a model's predictive ability. Furthermore, the recent developments in machine learning models, such as deep learning, imply that these machines possess their own 'ontology,' which is potentially more complex and efficient at understanding the world than ours. This leads to a Quinean problem of radical translation between human and machine ontologies. We suggest that this issue is key to the successful application of AI technologies in our society. The other focus of this article is epistemology, where Bayesian and classical statistics are compared to internalist and externalist epistemologies, respectively. This comparison elucidates how and in what sense the statistical methods adopted in each camp are considered to justify scientific hypotheses and also sheds light on their epistemic problems. We conclude with a plea for more research and interdisciplinary dialogues between statistics and the philosophy of various traditions.

[1] The Department of Philosophy, Kyoto University. e-mail: junotk@gmail.com

1. Introduction

Statistics is occupying an increasingly important position in today's advanced information society, where data on all kinds of matters is collected and utilized as "big data." In the 20th century, statistics played a central role in scientific methodology. In the 21st century, rapidly developing machine learning technology is now penetrating into every corner of our lives through applications such as speech and image recognition, recommendation systems based on purchase history, text or image generation, and autonomous driving.

Humanities cannot remain indifferent to the rapid development of statistics and machine learning. There have been various discussions about the legal and ethical implications of AI and computers making judgments in place of humans, or selectively presenting information that is likely to lead to certain judgments. However, apart from discussions about the ethical and/or social impact, relatively little has been said about the epistemological or ontological implications of the methods and concepts of statistics and machine learning, at least in Japan[2]. The lack is unfortunate for both philosophy and statistics, for statistics contains various problems that philosophers have traditionally discussed, as well as concrete hints for thinking about them. Conversely, philosophical analyses help clarify conceptual issues one often faces in using statistics or applying machine learning methods. In this paper, I would like to consider the bidirectional relationship between statistics and philosophy from this perspective, by introducing some of the topics discussed in my recent book on the philosophy of statistics (Otsuka, 2022).

The importance of statistics lies in the fact that it provides the scientifically standard framework for inductive inference. In brief, it is a mathematical theory that provides a means of making inferences about unobserved events from given data. On the other hand, since Hume, philosophers have long discussed the (im)possibility of inductive inference and the conditions that would authorize it. Modern mathematical statistics does not give an affirmative answer to Hume's skepticism. It does, however, allow us to formulate inductive inferences and to evaluate their accuracy under certain mathematical assumptions. When applying statistics to concrete problems, the scientist sees the world and their research target as some kind of mathematical structure that satisfies certain formal assumptions. This is a kind of ontological attitude toward the world. This observation sets up the first philosophical question that asks the ontological nature of inductive assumptions: i.e., what kind of ontological commitments are required in statistical inferences?

The second philosophical issue concerns semantics. Statistics formulates such ontological assumptions in terms of probability theory. But it is a mathematical construct, not an object of reality. From this arises the semantic problem of interpreting the former by the latter. Specifically, a group of questions such as what is meant by probability, what is

[2] See Hayashi (1960), Ode(1977), Akaike (1980), Deguchi (1998), and Taguchi et al. (2020) for a few exceptions.

causality, or what is the "p-value" widely used in statistical tests, stems from this semantic concern.

The third aspect is epistemology. The whole point of statistics is to infer from the data the mathematical structure so posited and interpreted. As an inductive inference, it does not give us a definitive answer about the subject of inference. Nevertheless, we expect statistical methods to *justify* some of our hypotheses from the data. In what sense, then, do these methods provide us with knowledge about the subject and justify our hypotheses? These questions have been traditional issues in epistemology since Plato. Here appears the third philosophical aspect of statistics, which is epistemological in nature.

The aforementioned book analyzes various statistical theories such as Bayesian statistics, classical statistics, model selection, machine learning, and causal inference from three philosophical aspects: ontological, semantic, and epistemological. However, for reasons of space limitation, this paper focuses just on the ontological and epistemological aspects and introduces a part of its contents.

2. Ontological problems in statistics

Hume pointed out that, in order to make inductive inferences, one must assume something more on the part of the world than the given phenomena/data, and called such an ontological postulate the *uniformity of nature*. But a mere uniformity is not all we need. Inferential statistics commonly introduces further ontological assumptions, and the strength of these assumptions determines the range of inferences that can be handled. In this section, we will discuss various ontological assumptions introduced in statistics and their evaluation criteria.

2.1. Probabilistic model as the uniformity of nature

Many decisions in our society are based on statistical inferences, such as judging the efficacy of a drug based on clinical trial results, weather forecasting from climate data, and a prediction of election results from exit polls. The statistical framework for inferring unobserved events from given data is called *inferential statistics*. Inferential statistics is essentially a framework for inductive inference, and thus is under the Humean predicament. Hume pointed out that in order to perform inductive inference, one must assume a uniformity of nature behind phenomena (Hume, 2000). Inferential statistics models this uniformity in terms of a *probability mode*l. A probability model formulates, in the language of probability theory, the sample space from which observed data are drawn.

As such, it includes not only the observed data, but also all the possibilities that may or may not be observed in the future. What is important is the assumption that this probability model itself will remain identical through observations. This assumption of uniformity assures the possibility of making inference about unseen events through inference of the model on the basis of observed data (Figure 1).

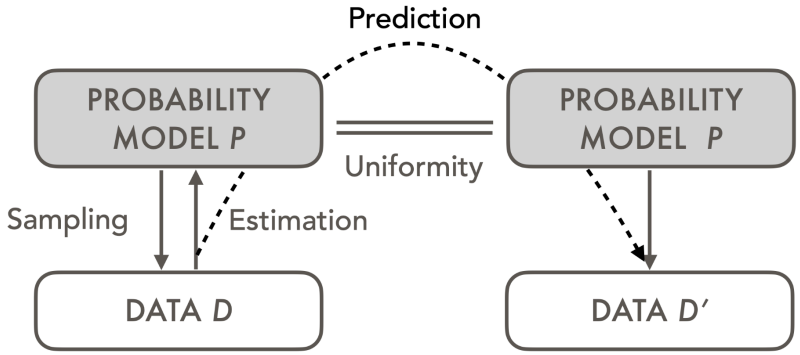

Figure 1. The dualist data/model ontology of inferential statistics.

Let's dive a little deeper into the probability model. The probability model consists of a sample space and a probability function on it[3]. The sample space is a set that contains what we call "events" as its subsets. For example, the event of "getting an even number by throwing a die" can be represented as a subset {2, 4, 6} of the sample space consisting of {1, 2, 3, 4, 5, 6}. Probability *P* is a function that measures the "size" of those events/subsets in the range from 0 to 1. The well-known axioms of probability theory can be understood as a set of minimum conditions that must be satisfied for this function to be a measure of size.

Formulating the uniformity of nature in terms of probability models allows one to draw several conclusions about inductive inference. Famous results include the law of large numbers, which states that the relative frequency of coin tosses asymptotically approaches its "true probability," and the central limit theorem, which states that the distribution of any trial, when averaged, approaches a normal distribution. These are theoretical results derived solely from the assumption that uniformity (i.e. that the object under consideration can be described as a single invariant probability model). This means that these results require only that there be a probability model, and not that we know what it is (i.e., what the "true model" is). In fact, we can never observe the entire probability model, and the "true model" is always hidden from us. But despite the sheer ignorance of the true reality, we can obtain

[3] To be precise, we need an algebraic structure (sigma algebra) that further defines what subsets should be recognized as "events," but we will not go into that here.

certain guarantees about inductive inference simply from the ontological assumption that uniformity holds. This is what the above laws and theorems show.

2.2. Statistical models as natural kinds

The asymptotic theorems guaranteed by the uniformity hold only for large samples. As such, they are often not a realistic solution in scientific practice where the amount of available data is severely limited. In such cases, we must make additional assumptions about the nature of uniformity itself. This assumption, called a *statistical model*, is expressed by characterizing the probability function with a function having an explicit form and a finite number of parameters. The famous example is the normal distribution, whose functional form can be determined two parameters, mean μ and variance σ^2, as follows:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\{-\frac{(x-\mu)^2}{2\sigma^2}\}$$

By identifying parameters μ, σ^2 and substituting various values for x in the above equation, we obtain its probability p(x), which when plotted produces the familiar bell-shaped curve. As mentioned above, the mere assumption of the uniformity (probability model) does not tell us anything about the probability of any event, or of any value of *X*. It is the introduction of the further assumption of statistical models that allows us to discuss specific probability values and have a rough idea about what the probability function looks like.

The normal distribution is just one example. Statistical models can take various functional forms, such as the Bernoulli distribution, binomial distribution, Wishart distribution, to name a few. Types of functional forms are called families of distribution. Statistical models also include regression models that express the relationship between variables in terms of equations with random noises.

The primary benefit of introducing statistical models is that it allows us to make more in-depth inferences about the uniformity of nature that, as mentioned above, is never revealed to us. By assuming that the probability function can be written in the form of an explicit function with a finite number of parameters, reasoning about the probability model boils down to reasoning about the parameters of the statistical model. For example, if we can assume that our target phenomena follows a normal distribution, we can identify the probability model simply by estimating its mean and variance. In other words, the statistical hypothesis about the whole unknown uniformity of nature is reduced to a hypothesis about a few parameters. Then by estimating parameters from the data, we could identify the nature of uniformity and predict unobserved events based on the estimated statistical model. This strategy of inferential statistics that resorts to the notion of statistical models is called *parametric statistics*.

From a philosophical perspective, a statistical model carves out a "true probability model," which is highly amorphous and elusive by nature, into a definite and recognizable form, identified by an explicit mathematical formula and a specific name. In this sense, it functions as a "natural kind." Natural kinds, such as “metal,” individuate and classify chunks of the world into units that have common properties. Some natural kinds have “parameters” that allow for a finer classification (e.g. “gold” or “iron”). In this way, natural kinds allow us not only to identify the object, but also to inductively predict how the object will behave (e.g., what will happen if we put a magnet near it or if we put it in aqua regia).

Statistical models play exactly the same role in statistical inference as these natural kinds, by serving as "kinds" to categorize various inductive problems. For example, random trials with two outcomes, such as a coin toss, will typically be modeled with a Bernoulli distribution. For other events, such as height, a normal distribution might be more appropriate. Just as a chemist identifies a reagent by its molecular formula, a statistician classifies an inductive problem by its statistical model (distribution family). And just as a chemist predicts the reaction of a reagent from the molecular formula so identified, a statistician predicts the event that will be sampled in the future from the estimated distribution. In this way, the statistical model serves as a natural kind in statistics, and for this reason I call it a *probabilistic kind*. The use of a statistical model embodies a stronger ontological assumption in the sense that it not only assumes the uniformity of nature, but also classifies and reduces it to specific discrete kinds. By making these stronger ontological assumptions, parametric statistics allow for more fine-grained inferences.

2.3 Goodness of probability kinds and real patterns

This observation leads to a question: how do we evaluate the "goodness" of the ontological assumptions of statistical modeling? It naturally depends on what we expect of natural kinds. Perhaps the most common expectation of natural kinds is that they faithfully reflect the true nature of the world. Indeed, when a chemist separates gold from silver, we expect it to reflect the objective distinctions existing in the material world. Similarly in statistics, it seems quite natural to expect that a presupposed model captures the uniformity of nature as correctly as possible. After all, Figure 1 suggests that successful prediction of unobserved events depends on the correct understanding of the uniformity.

This line of thought would lead to the reductionist conclusion that a "good" probability kind is a detailed statistical model that can describe the subject as precisely as possible. A typical measure of the level of detail of a model is the number of its parameters. Consider two linear regression models to predict a target property, say annual income, one that uses only one property, say age, and the other that also takes education into account. Then the extra parameter of the latter model provides a higher resolution. With this, one

might be tempted to conclude that a model that takes into account as many properties as possible would provide better predictions about a target variable.

In practice, however, this is not always the case. From the perspective of the Akaike Information Criteria (AIC) presented by Hirotsugu Akaike, too many parameters can in fact penalize the predictive performance of a model (Akaike, 1974; Forster & Sober, 1994; Akaike et al., 2007). This is because overly detailed models are too flexible and pick up noise in the data, resulting in overfitting to the data at hand. From this perspective, a "rough" model with a moderate granularity can still be a good model with better predictive performance.

Akaike's theory substantiates another aspect of natural kinds we expect of them, namely that natural kinds should be useful in predicting phenomena. In everyday life, we take a raven as a *bona fide* kind of its own and not as a complex aggregation of atoms and cells, simply because the former conception provides us with a wider range of predictions. Daniel Dennett called such units carved out based on our predictive interest *real patterns* (Dennett, 1991). From this perspective, although "ravens" may be a very vague and imprecise generalization from a physical perspective, it is real enough insofar as it fits well with our day-to-day predictive interests. This implies the idea that our ontology is tailored for pragmatic concerns of prediction, and that the AIC and other model selection criteria that emphasize model generalization performance can be thought of as measuring the "goodness" of natural kinds as real patterns.

2.4 Ontology of Machine Learning

Pragmatist ontology implies that the proper ontological classification (what counts as good natural kinds) at least partly depend on our predictive interests and cognitive constraints. A being such as Laplace's devil might be able to predict the future with arbitrary precision based solely on microscopic physical configurations, without resorting to somewhat vague concepts as ravens. In other words, what counts as a natural kind is determined relative to the cognizer.

The rapidly advancing field of deep learning gives an extra twist to this possibility. The models used in deep learning are extremely large, with literary billions or reaching almost trillions of parameters. As mentioned above, in the conventional framework such complex models tend to overfit and result in poor generalization performance. Recent deep learning models have solved this problem by improving the model structure, learning methods, and especially with large training data. Deep models supported by such big-data capture the uniformity of nature at a granularity far beyond our understanding, which have led to a concern that they are "black boxes". Natural kinds are usually expected to be lawful and tractable. Mendeleev's periodic table, for example, contributes to our understanding not only by classifying various chemical kinds, but also by disclosing the

regularity of the classification. This is contrasted with models obtained by deep learning, which, though captures the uniformity of the phenomenon with a high degree of accuracy, does not provide us with this kind of understanding. They do not tell us what part of the model corresponds to what, and what the consequences of changing which parameters are. Thus, deep learning confronts us with a tradeoff between predictive power and understandability. In the traditional view of science, the two aspects are expected to come hand in hand. However, deep learning seems to force us to give up understandability in exchange for its powerful predictive power. Whether such a gigantic model that lacks comprehensibility still counts as an ontological unit or "natural kinds" is a question that should be examined philosophically.

In addition to this, deep learning raises another interesting ontological question. Deep learning models themselves can be thought of as cognizers that discover patterns or "natural kinds" by their own from the data. An image recognition mode, for instance, learns and discriminates between objects such as "cats" or "cars" from data. AlphaGo, the AI program that recently defeated the world's top Go players, may have unearthed hidden patterns in the game that remain impenetrable even for the most seasoned human masters. In fact, it is an indispensable prerequisite for the practical application of, for example, automatic driving technology, that the model is able to properly identify and recognize objects from the input data. If so, understanding how deep learning models build their internal “ontology” or representation of the world isn't just an academic curiosity, but also of pragmatic importance.

An interesting problem in this regard is the phenomenon known as adversarial example (Szegedy et al., 2014). In a famous example, models are led to misclassify an image of a panda as a gibbon by adding noise that makes no difference to our eyes. Such a vulnerability raises serious concerns, such as the possibility of putting stickers on road signs to cause automated driving systems to malfunction. The existence of such adversarial examples suggests the possibility that deep models, which at first glance appear to be making decisions in the same way as we do, are actually constructing an ontology that is completely different from ours. In other words, the patterns that we extract with the concept of "panda" and the patterns that deep learning models use in similar circumstances may actually be quite different, even if they are perfectly consistent in the data so far. This is nothing but an actual example of the so-called Wittgenstein's rule-following paradox discussed by Kripke (Iida, 2016). The adversarial case demonstrates that this philosophical paradox is not just a theoretical concern, but an actual problem that casts a shadow over the social application of deep learning.

This is by no means to say that social applications of deep models must await the eventual solution of the Wittgensteinian paradox. But the possibility of such phenomena surely encourages the search for preventional measures. Such an attempt will involve deciphering the ontology of the deep model and reconciling it with our own. But how do we know whether or not the deep model uses natural kinds different from our own? This is

precisely the problem that Quine presented as the problem of indeterminacy of translation (Quine, 1960). In effect, one could say that the evaluation of a deep learning model involves radical translation. Is what the model labels a "cat" identical to that animal we usually understand by that concept? Chances are that it may simply be responding to a combination of typical “cat-like” properties and backgrounds (Xiao et al., 2020), or it may be thinking abstrusely that "catness is manifested out there." What Quine suggests by his discussion of the indeterminacy of translation is that the question as to which of these possibilities is correct may not be uniquely settled, or does not even have a correct answer in the first place. If this is the case, then various projects of "Explainable AI (XAI)" (Hara, 2018) to explain the basis of judgments of deep learning models would likewise be quests that lack definite answers.

Be that as it may, dismissing such attempts for that reason would be like throwing a baby out with the bathwater. After all, ontology is essential for understanding the other. We cannot predict the behavior of those with whom we do not share any ontology. Understanding what other human or nonhuman beings, say ravens flying low overhead, perceive is essential to predicting what they will do next. Likewise, unless we know to some extent what kind of objects the motion detection system used in automated driving technology recognizes, we will not be willing to put our lives in its hands. In this sense, the ontological issues surrounding machine learning are not only of philosophical interest, but also have important implications for its social applications.

3. Epistemological problems in statistics

Now let us turn to epistemological aspects of statistics. The combination of statistics and epistemology should not appear so far-fetched, for they both have a common role and motivation in justifying scientific hypotheses and beliefs. As Plato showed long ago in *Meno*, knowledge is not merely true belief. Similarly, we cannot equate scientific knowledge with a true hypothesis. To qualify scientific knowledge, a hypothesis must be justified in a proper way. That is why many scientific papers devote a section to "materials and methods", in which the authors specify the material, logical, or computational means to justify the results presented in the paper.

Statistics play a privileged role in this process of scientific justification, especially in drawing conclusions from collected data. All scientific hypotheses are stochastic and do not enjoy logical certainty. Even data highly favorable to a hypothesis cannot rule out the possibility that it was simply the result of chance, holding completely independently of the truth of the hypothesis. Statistical methods are needed to eliminate such "lucky guesses" and to determine whether the observations and experimental results truly support the hypothesis.

How, then, can such justification be made? The answer is not unique, for statistical methodology is not monolithic. There are various well-known schools of statistical theory, such as classical “frequentist” theory known for the notion of hypothesis testing and Bayesian statistics based on Bayes' theorem. Each school differs not only in its core mathematical methodology but also in its concept of justification, i.e., in what sense such mathematical theory can (dis)confirm an empirical hypothesis. In my book, I argued that these differences correspond to two opposing positions in philosophical epistemology: internalism and externalism. The following part of this paper summarizes this parallelism between statistics and epistemology.

3.1 Bayesian statistics as an internalist epistemology

Bayesian statistics views an empirical (dis)confirmation of a hypothesis as a process of updating the probability of a hypothesis based on evidence. We do not enter the semantic question of what the “probability of hypothesis" actually means here[4], but it is commonly interpreted as the "degree of belief" that a cognizer, e.g., a scientist conducting an inquiry, has in the hypothesis in question. Bayes' theorem is used to update the probability thus understood, by deriving the probability of the hypothesis after observing evidence (posterior probability) from its probability before obtaining evidence (prior probability) and the probability of the evidence under the hypothesis (likelihood). A hypothesis with larger prior and likelihood has a larger posterior probability and is therefore better supported.

As a matter of fact, this process of calculating the posterior probability itself is not inductive at all. Bayes' theorem is a mathematical theorem derived from probability theory, and the calculation of posterior probability using it is an outright deductive inference. Why, then, is it possible to (dis)confirm an empirical hypothesis by such deductive reasoning?

The answer lies in the epistemological assumption of Bayesian statistics. According to the above sketch, Bayesian inference is the process of deriving a conclusion in terms of posterior probability from the premises of prior probability and likelihood, using the inference rule of Bayes' theorem. In other words, Bayesian justification is the process of deriving a probability evaluation of a hypothesis in a way consistent with the premises and evidence at hand. This is akin to the so-called *internist* concept of justification in philosophical epistemology. According to internalism, a subject's belief is justified when he or she explicitly grasps the reasons or evidence from which the belief in question is derived through an appropriate inferential process (cf. Todayama, 2002). For example, if I were to claim that I know that the capital of Yamatai-koku (an ancient Japanese kingdom) was located in the Kyushu area, it is naturally expected that I have some evidence and that the reasoning I used to draw the Kyushu hypothesis from that evidence is appropriate.

[4] For more information, see chapter 2.1 of my book (Otsuka, 2022), or (Childers, 2013; Gillies, 2000; Rowbottom, 2015)

According to internalists, the subject's belief is justified if he or she has evidence that bears an appropriate inferential relationship. Similarly, when a scientist evaluates the posterior probability of a hypothesis from appropriate premises (prior probability and likelihood) using the appropriate inference rules of Bayesian inference, that evaluation (i.e., the judgment of whether the hypothesis is plausible or not) can be considered justified in the internalist sense.

The immediate question is, then, what is an appropriate assumption to begin with? For example, if the only evidence I have for the Kyushu hypothesis comes from a dubious occult magazine, my hypothesis would hardly be justified. Similarly, the Bayesian "justification" of the posterior probability would not be considered a true justification unless the assumed prior probability and likelihood are themselves appropriate.

An immediate response to this is to require that the premise of the inference be justified on its own. However, this immediately leads to the regress problem: in order to justify a conclusion one must first justify its premises, and to do so one must further justify their premises, *ad infinitum*. To prevent such regression, some internalists have presupposed the existence of foundational beliefs that require no further justification, candidates of which include supposedly infallible beliefs such as mathematical propositions or cogito, and direct perceptual experiences such as "I see a black dot right now". The former is an a priori method and the latter an a posteriori method to stop the regress. The position that tries to prevent the regress problem by means of such foundational beliefs is called *foundationalism*.

Similar ideas can be found in Bayesian statistics. Among the assumptions of Bayesian inference, traditionally more controversial has been the prior probability. This is because prior probability is peculiar to Bayesian statistics, whereas likelihood is the assumption of the statistical model/probability kinds mentioned in the previous section and is generally required by other statistical schools. Therefore, how to justify this prior probability has been a pressing issue in the debate over the validity of Bayesian statistics.

Just as two types of foundational beliefs were observed in foundationalist epistemology, two strategies are possible for justifying prior probabilities. The a priori strategy defines prior probabilities in a completely uninformative manner, so that they do not include any prejudice or personal opinion on the hypothesis. This is called a non-informative prior distribution. The other strategy is the a posteriori strategy that calibrates prior probabilities using pre-existing data. Consider, for instance, the prior probability that I have a certain disease? If I set it to 0.5 because I'm completely ignorant between two possibilities, yes or no, I would be drawing a grossly exaggerated decision from the test results. If the disease is known to be rare, the prior probability should also be set reasonably low. This method of setting prior probabilities according to empirical data is called *empirical Bayes*. From a philosophical point of view, this is nothing more than regarding some data as "given" and using it as the firm basis of inference. At first glance, this seems reasonable, but when one begins to ask what exactly amounts to calibrating prior

probabilities to data and why such a procedure is justified, one is confronted with a group of problems that are not straightforward to solve. And this is exactly the same problem structure as the "myth of the given" criticized by Sellars (Sellars, 1956) in the context of epistemology.

3.2 Classical statistics as extrinsic epistemology

While the probability of a hypothesis is the central focus in Bayesian statistics, classical statistics does not consider the probability of a hypothesis at all. According to classical statistics, the truth value of a hypothesis is fixed on the part of the world, and thus is meaningless to discuss the "degree" of its righteousness. In classical statistics, probabilities are assigned only to data. Given a certain hypothesis, what results are likely to be obtained? If the observed data are unlikely under the hypothesis, then there must be something wrong with the hypothesis. The hypothesis testing, which is the cornerstone of classical statistics, examines hypotheses based on such an idea.

At first glance, this resembles Popper's concept of falsificationism, according to which a hypothesis H is falsified if its prediction E fails to obtain ($\neg E$). For this inference to be valid, however, hypothesis H must fully imply E. If the prediction is only stochastic, then $\neg E$ does not imply $\neg H$ or even that H is unlikely, i.e., it does not imply that $P(H|\neg E)$ is low (Sober, 2008)[5] . Therefore, the concept of falsificationism cannot be directly applied to testing statistical hypotheses.

Statistical test theory avoids this by contrasting two opposing hypotheses, the null and the alternative. For example, in the development of a new drug, let H_0 be the null hypothesis that the new drug has no effect (no difference between the treatment group and the control group) and H_1 be the alternative hypothesis that the new drug has an effect (nonzero difference between the two groups). We are interested in whether we can reject the null hypothesis based on data. Common sense would dictate that the larger the observed difference, the better the basis for rejecting H_0 . But since the results are probabilistic, there is always a risk of making erroneous judgments. One type of error is called *type I*, which erroneously determines that there is an effect (rejecting H_0) when there is in fact none (H_0 is true), and the other is the *type II* error, which determines that there is no effect (failing to reject H_0)when there in fact is an effect (H_1 is true). The key to the test is to set the decision criterion so as to reduce the probability of these two errors as much as possible. In most cases, the probability of making type I error, called the significance level, is taken more seriously and is kept low (e.g., 5%). A low significance level means that the test is unlikely to reject the null hypothesis H_0 just by chance. Thus, if the null hypothesis is still

[5] As mentioned above, classical statistics does not consider the "probability of a hypothesis," so this is only from a Bayesian standpoint. But in any case, the success or failure of a stochastic hypothesis cannot be logically derived from the failure of a prediction alone.

rejected by such a test, the decision is unlikely to be a "lucky guess" and can be considered reliable. This is how statistical testing theory justifies its decision based on data.

But precisely with what epistemological standard do these procedures justify the hypothesis, like that the drug is effective? The justificatory basis of classical statistics is sought in the reliability of the test as a hypothesis-decision device. A test, after all, can be thought of as a function that returns a conclusion (rejection/non-rejection of the null hypothesis) for input data. This device has certain false positive (type I error) and false negative (type II error) rates. A test with low error rates in both can be considered highly reliable, just as we consider medical testing instruments with low false positive and negative rates to be reliable. Classical statistics takes the conclusions reached by such highly reliable tests to be justified. This idea corresponds to an epistemological position called externalism, in particular *reliabilism* (Goldman & Beddor, 2016; Todayama, 2002). Externalism, unlike internalism, does not impose the requirement that the cognizer must possess all of her inferential grounds and rules within. Justification obtains even if the cognizer is ignorant of the justificatory status of the evidence, as long as the evidence itself is valid as a matter of fact. Statistical tests provide just such an inferential device for scientists/cognizers, and the classical statistical theory evaluates its reliability in terms of two error probabilities. When the error probabilities are low, that is, when the reliability is high, the conclusions reached using such a procedure can be said to be justified in the externalist sense.

How should we think about the externalist nature of statistical testing theory? The frequently raised concern with externalism is that it renders the cognizer too irresponsible for their conclusions. In other words, by "outsourcing" their basis for inference, externalists seem to be abdicating responsibility for the accuracy of their conclusions. A similar criticism has been recently leveled against statistical test theory. During the 20th century, test theory has enjoyed its status as the standard method of scientific reasoning. At the same time, statistical decision procedures have been standardized and packaged, and now, software that allows for easy hypothesis testing just by entering data has become commonplace. In the scientific community, conclusions that meet certain criteria (e.g. low p-value) in such tests have been uncritically published in academic journals and accepted as correct. However, concerns have been raised that such uncritical use of this testing process has led to misuse and misinterpretation, resulting in the mass publication of non-reproducible research results. These issues, known for "p-value problem" or "reproducibility crisis," have become major issues in the scientific community in recent years (Wasserstein & Lazar, 2016) and invoked skepticism on the reliability of test theory that has long held hegemony in scientific reasoning.

From a philosophical perspective, one can understand these problems as arising from the externalist nature of classical statistics. As noted above, justification by test theory stems from the reliability of the tests used to make judgments. Classical statistical theory probabilistically estimates this reliability, but these estimates are obtained not "for free" but

only on the basis of various assumptions, including but not limited to the assumption of a statistical model as a "probability kind" (Section 2.2) and the correct handling of experimental outcomes. The estimates of the test reliability such as p-value are only as good as these assumptions. However, in an automated "reasoning process" in which conclusions are drawn simply by inputting obtained data into statistical software, all of these assumptions are treated as external factors and escape critical examination. Part of the reproducibility problem mentioned above may be sought in the irresponsible use of tests that push the examination of such assumptions into background. What this suggests is that the use of statistical tests must not be completely external about its premises in order to arrive at correct judgments, i.e., for its justification to be truth-conducive (Mayo, 2018; Staley & Cobb, 2011). This echoes the often-made criticism directed at externalism: that outsourcing justification to the reliability of the recognition/testing process should not exempt us from the task of putting it under scrutiny.

3.3 Epistemology and Statistics

We have compared the major positions in modern statistics, Bayesian and classical statistics, in terms of epistemological internalism and externalism. Our analysis focused on the difference in their views on the concept of justification, i.e., the procedures to be followed in order to accept a hypothesis as scientific knowledge. As we see it, this philosophical difference underlies much of the heated controversy between these camps in the 20th century. With the rise of more pragmatic statistical methods (Section 2.3) and the development of machine learning (Section 2.4) since the late 20th century, these debates have been gradually dying down in recent years, and their significance has become less apparent. However, dismissing them altogether as outdated dogmas from a modern standpoint would be a naive Whiggish historiography. Identifying the differences in philosophical standpoints is crucial for understanding and fairly assessing the discussions that have shaped statistics as we know them today.

In fact, these philosophical discussions have by no means lost their significance in modern statistics. Modern statistics as a mathematical theory has made great strides since the 20th century. Statistics, however, is not simply confined to mathematics. Why can mathematical statistics, itself a deductive system, apply to real subjects and give us guidance in reasoning about the unknown? Since mathematics itself does not provide answers to these philosophical questions, "muddy" philosophical speculations and assumptions are inevitable. Abandoning philosophical reflection does not lead to a dogma-free statistics; rather, it results in a statistics that blindly sticks to a single dogma.

The importance of philosophical reflection is even greater in contemporary statistics. For example, the rapid development of machine learning theory (section 2.4) apparently calls for a new concept of justification, different from that of traditional

statistics. How can we treat as "justified" the conclusions generated by machines using enormous amounts of data, whose complexity is beyond the grasp of our intelligence? And how will our view of science change when science is built on the basis of such machine-made conclusions? These issues are epistemological in nature, but they will also carry legal and ethical significance as machine learning technology is increasingly applied to society. Consider an example from drug approval: the current process mandates some form of statistical testing before a new drug can enter the market. This means that the legitimacy of the approval is, at least partially, based on the concept of statistical justification, as discussed in the previous section. Thus, statistical justification underpins various forms of justification in our society, including legal justification in this instance. If this is the case, reflecting on the concept of justification involved in new or yet-to-be-developed statistical methods will have important implications for their social application. For example, whether automated driving technology should be approved depends on the degree to which the decisions of automated driving systems are justified. But what kind of justification is that? Does an externalist justification, for example, one based on simulation tests, suffice, or do we require a more in-depth, internalist justification that explicates the system's judgmental foundations? Epistemological considerations like these prove essential if machine learning technology is to be truly integrated into (rather than imposed upon) our social life.

Conversely, studying the epistemological aspects of statistics should provide fruitful perspectives for philosophical interests. The root of epistemology is the long-standing question since Plato: what is knowledge (*episteme*) and how do we acquire it? One could not answer this question today without taking into account science, the epitome of contemporary knowledge, and statistics, its fundamental methodology. In fact, as discussed above, various statistical methods can serve as models for philosophical epistemology. The role of such models is not (as is the case with almost all scientific models) to provide elaborate replicas of philosophical theories in the context of scientific methodology. Rather, the model's significance lies in formulating a theory in a moderately idealized and abstract manner, and in exploring and refining the theory through its application to reality. As long as epistemology aims to elucidate human knowledge in the functioning of real society, it should be both beneficial and essential to refine its theories in light of the scientific methodologies that are actually generating knowledge in today's society, rather than merely sticking to aprioristic arguments rooted in philosophers' intuitions.

This by no means implies the Queinnian reduction of epistemology to science (Quine, 1969)). In the first place, dividing philosophy and science in such a way appears to be an arbitrary distinction. Whether we look at ancient Greece or early modern Europe, there is already a scientific part within philosophy and a philosophical part within science. This is true even today, when science has taken on a life of its own and the division of labor has become highly advanced. If this is so, philosophers should pay attention to both the scientific part of philosophy and the philosophical part of science. As we have tried to

show in this paper, statistics as a metascientific methodology provides fertile ground for such transversal speculation.

4. Conclusion

This paper discussed the connection between statistics and philosophy, focusing mainly on its ontological and epistemological aspects. For reasons of space, we have omitted the topics that are relatively familiar among philosophers, like the semantic/interpretative issues or causal inference. More details on those topics and the arguments touched above can be found in Otsuka (2022). Needless to say, the discussion in the book still is only the tiny tip of a huge iceberg of statistics. Contemporary statistics has entered into a new phase with the fusion with computer science (Efron & Hastie, 2016), but its philosophical implications are still largely untapped. Our philosophical approach here is limited to the Anglo-American tradition. However, there are many other possible approaches, including historical, legal, ethical, and "postmodern" ones. Integrating these various aspects and approaches could hopefully lead to a more fruitful analysis of the relationship between philosophy and statistics.

References

Akaike, H. (1974). A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19(6), 716–723.

Akaike, H. (1980). On the Transition of the Paradigm of Statistical Inference (in Japanese), Annual Report of the Institute of Statistical Mathematics, 27(1), 5-12. https://core.ac.uk/download/pdf/234007312.pdf

Akaike, H., Amari, S., Kitagawa, G., Kshima, Y., Shimodaira, H. (2007) *Akaike Information Criteria--Modeling, Prediction, and Knowledge Discovery* (in Japanese), Murota & Tsuchiya (eds.), Kyoritsu-Shuppan.

Childers, T. (2013). *Philosophy and Probability*. Oxford University Press.

Deguchi, Y. (1998). Quine's Theory of Science from the Viewpoint of Statistics (in Japanese). *Arcae,* 6, 60-70.

Dennett, D. C. (1991). Real Patterns. *The Journal of Philosophy*, *88*(1), 27–51.

Efron, B., & Hastie, T. (2016). *Computer age statistical inference: Algorithms, evidence, and data science*. Cambridge University Press.

Forster, M., & Sober, E. (1994). How to Tell when Simpler, More Unified, or Less Ad Hoc Theories will Provide More Accurate Predictions. *The British Journal for the Philosophy of Science*, *45*(1), 1 35.

Gillies, D. (2000). *Philosophical Theories of Probability* (1st ed.). Routledge.

Goldman, A., & Beddor, B. (2016). Reliabilist Epistemology. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2016). Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2016/entries/reliabilism/

Hara, S. (2018). Interpretability in Machine Learning (in Japanese), https://www.ai-gakkai.or.jp/resource/my-bookmark/my-bookmark_vol33-no3/

Hayashi, C. (1960). On the Basis of Statistical Methods (in Japanese), Kagaku-Kisoron-Kenkyu, *5*(1), 1-16.

Hume, D. (2000). *A Treatise on Human Nature*. Norton & Norton (eds), Oxford University Press.

Iida, T. (2016). *Rule-following paradoxes and meaning skepticism* (in Japanese), Chikuma-shobo.

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

Oide, A. (1977). Around Probability and Statistics (in Japanese). *Riso*, *528*, 173-198.

Otsuka, J. (2022). *Thinking About Statistics: The Philosophical Foundations*. Routledge.

Quine, W. V. O. (1960). *Word and Object*. The MIT Press.

Quine, W. V. O. (1969). Epistemology Naturalized. In *Ontological Relativity and Other Essays* (pp. 69–90). Columbia University Press.

Rowbottom, D. P. (2015). *Probability*. Polity Press.

Sellars, W. (1956). *Empiricism and the Philosophy of Mind*. University of Minnesota Press.

Sober, E. (2008). *Evidence and Evolution*. Cambridge University Press.

Staley, K., & Cobb, A. (2011). Internalist and externalist aspects of justification in scientific inquiry. *Synthese*, *182*(3), 475–492.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. *arXiv*, 1312.6199.

Taguchi, S. , Otsuka, J., & Saigo, H. (2020). Phenomenological Argumentation and Statistics: Toward a Fundamental Structure of Experience (in Japanese). Tetsugaku-Ronso, *47*, 20-34.

Todayama, K. (2002). *The Philosophy of Knowledge* (in Japanese). Sangyo-shobo.

Ueeda, Y. (2020). *An Introduction to Contemporary Epistemology: From the Gettier Problem to Virtue Epistemology* (in Japanese), Keiso-Shobo.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's Statement on p-Values: Context, Process, and Purpose. *The American Statistician*, *70*(2), 129-133.

Xiao, K., Engstrom, L., Ilyas, A., & Madry, A. (2020). Noise or Signal: The Role of Image Backgrounds in Object Recognition. arXiv. http://arxiv.org/abs/2006.0999