Role of information and its processing in statistical analysis

Bryce M. Kim

07/15/2017

Abstract

This paper discusses how real-life statistical analysis/inference deviates from ideal environments. More specifically, there often exist models that have equal statistical power as the actual data-generating model, given only limited information and information processing/computation capacity. This means that misspecification actually has two problems: first with misspecification around the model we wish to find, and that an actual data-generating model may never be discovered. Thus the role information - this includes data - plays on statistical inference needs to be considered more heavily than often done. A game defining pseudo-equivalent models is presented in this light. This limited information nature effectively casts a statistical analyst as a decider in decision theory facing an identical problem: trying best to form credence/belief of some events, even if it may end up not being close to objective probability. The sleeping beauty problem is used as a study case to highlight some properties of real-life statistical inference. Bayesian inference of prior updates can lead to wrong credence analysis when prior is assigned to variables/events that are not (statistical identification-wise) identifiable. A controversial idea that Bayesianism can go around identification problems in frequentist analysis is brought to more doubts. This necessitates re-defining how Kolmogorov probability theory is applied in real-life statistical inference, and what concepts need to be fundamental.

1 Pseudo-equivalent models and limited information statistics framework, or simply decision theory

To inspire later discussions, I will start from Milton Friedman's instrumentalism [4]. To say simply, Friedmanite instrumentalism says that a model is just an instrument to analysis, and that how realistic assumptions are do not matter, as long as the model is shown to have statistical powers, including predictive powers.

The question one can ask is when Friedmanite instrumentalism is justified. Instead of trying to justify Friedmanite instrumentalism on broad circumstances, one can consider, at least theoretically, validity scope of such instrumentalism. We will see that such philosophy of instrumentalism is necessary for many statistical inferences, though possibly in ways Milton Friedman himself did not think of. The key word for justifying Friedmanite instrumentalism is by looking at what information available or unavailable has been traced out from statistical analysis.
To define this rigorously, we need to define what we mean when two models are practically equivalent. The following game is proposed to define when two models - model C and D - are pseudo-equivalent.

Definition 1.1 (Original game: Pseudo-equivalent models, Model C and Model D)**.** I will define when two models - model C and D - are pseudo-equivalent by presenting a game. Person A is the sole player of the game. A is given model D, but it is not given model C. A is additionally given some information that is not data points from the data-generating model. Data points, or samples, are selected by a fair coin toss, and if head came up, model C is the chosen data-generating model, if tail came up, model D is the chosen data-generating model. A is then given k data points generated from the chosen data-generating model and computation time to process k data points and some other given information. A, after given computation time, has to guess the outcome of the coin toss (again, it only knows that tail means Model D is the chosen data generation model). Given the information set of A, if one can prove, with knowledge of model C, unavailable to person A, that person A's choice effectively is a fair coin toss, then one can see that model C and D are pseudo-equivalent models in terms of person A's information set.

The definition concerns much more than consideration of required number of samples for valid analysis. Data points that are generated by the chosen data generation model are not only information that is available - this point will be clear, once the sleeping beauty problem is discussed in the following section - and information that is not a sample may be valuable for statistical inference. Also, it may be possible that even with infinite number of data points, one may not recover the correct data-generating model. For example, let us think of Brownian motion under classical/Newtonian physics [2]. If we know initial conditions of each particle in the system undergoing Brownian motion, then under classical physics, one does not have to rely on stochastic analysis to analyze the system - the system is deterministic. However, in reality, and even not considering the fact that classical physics is not a fundamental physics theory, we only get limited information/attributes of ensemble of particles. The fact that we do not have initial condition information (and also trajectories starting from initial conditions) of each particle affects the end result of our statistical analysis. Even with infinity of data, and even if we know the number of particles inside the system, we will never be able to recover the trajectory model of the system. We would have to be content with the description that effectively treats each particle as if it follows stochastic standard Brownian motion process.

This is further complicated by the fact that in real-life analysis, we have to get to the right model before estimating the model. In terms of the game, one may attempt to analyze data and information to get to some model, to help Player A win the original game. What if a correct and unique data generation model can be derived from analyzing infinite number of samples? I have not specified how one would learn such a model, so this is just an assumption for now. Here, required number of samples clearly matters, but if it takes too much time to learn a correct model even when given sufficient number of samples, then it is certainly possible that a pseudo-equivalent model exists. That is, unless given directly by someone else about existence of Model C, Player A may fail to consider Model C while performing statistical analysis. Recall that in the original game, what Player A is asked whether Model D is the data generation model or not, not whether Model C or Model D is the correct data generation model. When given sufficient computation time, Player A may have inferred Model C, and realize that Model C is the correct data generation model (in case it is), but it is possible that computation time required to reach that correct inference is so much that this is practically impossible. In this regard, Player A may theoretically be in the position identical to predicting the outcome of a fair coin toss before the coin toss actually occurs, even when the coin toss already happened.

If anyone is from the world of cryptography, one may realize that the original game is very close to the definition of pseudo-random functions [8]. What differentiates cryptographic case from this stochastic case is that in case of the pseudo-random function of a particular security key, if Player A is given all possible function input-output pairs of the function then the pseudo-random function can be derived, along with the key, if the cryptographic system is known. In this sense, the above original game is, in one way, a more general game, at least under suitable extensions. Also, notice that in the example of Brownian motion, we can consider initial conditions of particles in the system as a key consisting of security sub-keys, with classical physics considered as an analogue of a cryptographic system. As in the example of a pseudo-random function, if the key is known, then one can easily resurrect the trajectory of the system.

From now on, I will refer to this entire real-life statistical inference framework as limited information statistics framework.

1.1 Garden of forking paths: researcher degree of freedom

The key point in the above (demonstrated especially using the concept of pseudo-equivalent models) was that what real-life statistical analyst comes to form is more of best credence, or belief, of events given available information, rather than estimation to objective probability of events. When the word "credence" appears, it immediately puts statistical analyst under parts of decision theory that discuss how to rationally set beliefs of events.

Furthermore, this inherent nature necessarily invites researcher degree of freedom [9]. What one is interested in not only the best estimate to some model,

but also the best data generation model. And the model is not given free - one has to learn the model. And data and information are, tautologically, the only ways to improve learning of the model. This creates what one would call as Researcher degree of freedom. To one extreme, an analyst may come to collect particular types of data based on the model to improve analysis, but even without this extreme, while selecting models, inference strategies may come to choose different significance or other statistical tests depending on data [5]. This researcher degree of freedom is well-known to invalidate significance tests, often misused to support (and more appropriately, reject) some hypothesis, which includes misspecification issues as well.

Unfortunately again, as one can see in limited information statistics framework, we do need to search for models, along with parameters of models. While a different model-learning process would have different consequences, some of simplest learning ones would inevitably feature iterating over different models and ruling out models using some statistical tests picked based on data. It simply is infeasible to stick to one particular statistical test pre-chosen, when we really do need to search for at least a semi-true model that is pseudo-equivalent to a true data generation model.

It does seem that instead of trying to fix the P-value significance problem and making it work as intended, it is better if we can reduce reliance on P-value significance analysis, and increase more of information-based analysis, and what learning algorithms we can use to process information into inference about credence of models and their parameters, and provide theoretical bounds to accuracy of the algorithm to some desired level, given a true model [1]. (Significance testing is about confidence based on the given model, not the true model, and data.) To some extent, this is what is being done in machine learning literature. Even significance-test-wise, one can consider learning algorithms, or more correctly a meta-algorithm that selects and combines different learning algorithms as a learning protocol that can serve as preregistration so that valid significance analysis can be done.

To summarize, a main take-away from this section is that limited information statistics framework, which is how real statistical analyis often operates upon, is essentially probabilistic/credence-belief-setting parts of decision theory. As long as decision theory is interested in forming beliefs objectively and rationally as possible - whatever the definitions for objective and rational are for now - it directly connects to our needs of statistical analysis and are not distinguishable. **credence/belief analysis of decision theory = real-life statistical analysis**

2 Sleeping Beauty Problem, and breakdown of Kolmogorov probability theory in decision theory

In the previous sections, I described how studies of statistical analysis are most of time effectively decision theory. Thus it is not possible to isolate paradoxes and problems in decision theory from affecting statistical analysis. For this reason, it may be worth mentioning one of so-called paradoxes in decision theory: sleeping beauty problem.

First, note that in all of our previous sections, I assumed that statistic analysis is about discovering objective properties. Even when "valid" models used in statistical analysis is not the actual working of reality, these models can be derived from the actual working model by tracing out information and/or imposing information processing limit caused by computation limitation (computational complexity).

I will maintain this spirit of objectivity in the sleeping beauty problem. While the sleeping beauty problem [3] is about assigning "credence/belief," which is subjective, to events, we are searching for the most objective, or more precisely said, rational way of assigning credence.

The description of the problem goes as follows. On Sunday, a person (interviewee) in the experiment is put to sleep and a coin is tossed. If head comes up (simply referred to as "head" from now on), she is awakened only on Monday. If tail comes up, she is awakened on both Monday and Tuesday. For any day that she is awakened, she is interviewed regarding her assigned credence for head/tail at the time of the interview and given some sleeping pill that erases memories regarding interviews and induces sleep for the interviewee. P refers to probability as usual - or more correctly credence, but I will distinguish it from objective probability $\mathbb{P}$. I will describe the typical reasoning process, regardless of halfer or thirder positions below.

$P(Monday|tail) = p_1$, $P(Tuesday|tail) = p_2$.

$P(Monday\&tail) = p_1 P(tail)$, $P(Tuesday\&tail) = p_2 P(tail)$

$P(Monday\&head) = 1 - P(tail)$.

The typical reasoning process can thus be described as assumption (at least at the end of the reasoning process) on p_1 and p_2 and $P(tail)$, with the requirement that $p_1 + p_2 = 1$. This may be too obvious but this is an important point to notice. For any Bayesian inference founded upon Kolmogorov probability theory, as long as we give a credence prior to p_1, p_2 and $P(tail)$ with $p_1 + p_2 = 1$, a such prior will be self-consistent. (The typical thirder position [3] is about setting $p_1 = p_2$, by some variant of principle of indifference, and by the same logic that (via Bayes rule) $p_1 P(tail) = 1 - P(tail)$.)

Looking from the perspective of the interviewee, this is what she does when interviewed - assigning credence prior, and looking for any signal that her Bayesian analysis may use to update her prior to posterior. There of course is no external signal to produce her posterior - her prior dominates analysis. Typical Bayesian analysis can be somewhat flexible with choice of priors - though some priors are

more efficient - because it allows for asymptotic convergence by updating using data after the prior is set. But the interviewee clearly cannot update, and thus prior is completely dominant.

In this regard, sleeping beauty collapses to the problem about the best way a prior may be set in typical Bayesian analysis. And in this way, as many statisticians would agree that there will be no "real" answer to the problem, one may dismiss the problem as unanswerable.

Except, is this analytical framework really the way we should see the sleeping beauty problem? I will argue that this is not the case. In fact, the framework described above strips away any reference to objective information one has: $\mathbb{P}(head) = 1/2$. Also, $\mathbb{P}_+(head)$ is either 1 or 0 after the coin toss occurs, where $\mathbb{P}_+$ refers to the objective probability after the coin toss occurs.

This is where we can connect to the previous sections. The sleeping beauty problem, in the limited information statistics framework, can be described as follows. All other people, except the interviewee has information on $\mathbb{P}_+(head)$, and thus the experiment actually has no stochastic element after the coin toss occurs. But because the interviewee does not have that information on $\mathbb{P}_+$, to the interviewee its stochastic model of what happens is pseudo-equivalent to actual reality. (This of course requires some caveat: the interviewee already knows that a particular coin toss outcome was already determined. Thus, "Model C" and "Model D" should be each coin outcome. However, it is certainly possible to set "Model C" as reality and "Model D" as some stochastic credence model, with minimal deviations from the original set-up of the experiment.)

Now consider the experiment identical to the sleeping beauty experiment except that the interviewee is not assumed to have knowledge of experiment settings, except for the occurrence of coin toss on Sunday and that she suffers from memory loss on both Monday and Tuesday, though she remembers only Sunday events on both Monday and Tuesday. When interviewed on either Monday or Tuesday, everyone would agree that credence should be set $1/2$ for $P(head)$ and $P(tail)$. This means that the interviewee having knowledge of "head comes up meaning waking up only on Monday" and "tail coming up meaning waking up on both days" (let us refer to this information as Information α) must provide more information to change credence, if the halfer position is not taken.

The Bayes update rule essentially says $P(+) = P(\alpha|-)P(-)/P(\alpha)$, where $+$ refers to posterior, $-$ refers to prior and α refers to Information α. As $P(\alpha)$ can be easily calculated from $P(-)$ and is constant, it can be dropped from the discussion.

$$P(+) \propto P(\alpha|-)P(-)$$

At this point, it is easy to see that new information α does not give any information on $P(\alpha|-)$. Does it make sense to calculate probability of information α given that the coin toss was head? No. In fact, this is just noise information that cannot be used to update prior. Without prior already containing some assumption on prior probability of occurrence of Monday and Tuesday, there is no reasonable way to calculate value of Information α for updating prior. And posterior will directly depend on what prior says about event Monday and

Tuesday, and this also makes no sense.
However, $P(\alpha|-)$ can be calculated, if one follows, for example, the thirder solution. Thus one now realizes that something clearly went wrong: A. consensus prior $-$ is wrong, or B. the common solutions/strategies are wrong, or C. Bayesian framework is somewhat faulty.
The A option is hard to take in. It would mean, in our life, that given some fair coin toss happened in the past and we do not know the outcome, we should deviate from the assumption of fair coin toss as time goes on. The options I explore here are B and C, and I will argue that both B and C are what happen. Since the C option shows, in an obvious way why the B option is the case, I will explore the C option only. Specifically, I will argue that Kolmogorov axioms necessarily break down in decision-theoretic settings, and the bound where Kolmogorov axioms can safely be used exists.
Consider what happens when $P(head) = 1/2$ and $P(tail) = 1/2$, as we require.

$$P(head) = P(head|Monday)P(Monday) = 1/2$$

$$P(tail) = P(tail|Monday)P(Monday) + P(tail|Tuesday)P(Tuesday) = 1/2$$

$P(head|Tuesday)$ is obviously zero, so it is dropped from the equation. And $P(tail|Tuesday) = 1$. Now this framework raises the question of $P(head|Monday)$. Should it be 1/2? Yes, if we trust in our prior $-$. Prior $-$ had $P(head) = 1/2$ because on Monday, it would still consider as $P_-(head|Monday) = 1/2$. (P_- refers to prior probability.) And by the same prior updating example, $P_+(head|Monday) \propto P_-(\alpha_{hm})P_-(head|Monday)$, but Information α is noise and should not be used to update information.
This leads to the following crisis:

$$1/2 = (1/2)P(Monday)$$

$$1/2 = (1/2)P(Monday) + P(Tuesday)$$

$P(Monday) = 1$, so this means that $P(Tuesday) = 0$! Thus this double halfer position seems to fail, and the entire analysis in this section seems to fall apart. Except this is not the case.
So far, it was assumed that conventional understanding of Kolmogorov probability theory (from now on simply Kolmogorov theory) "rules" over stochastic analysis. The Bayes rule can be understood and derived from Kolmogorov probability theory. But supremacy of Kolmogorov theory is only guaranteed for actual stochastic processes, where possible events can be clearly identified with "somewhat complete" (more about the qualifier soon) information and objective stochastic probability for these events exist. The settings where stochastic inference is done almost always carry limited information and information processing, and there is no logical reason to assume that we would have a complete credence picture on every sub-part of the system, or we may even have a wrong division of the system. For experimenters, on Monday they knows that they are on Monday (and is part of Sunday-Monday-Tuesday stochastic process, where Monday and Tuesday random variables depend only on Sunday's). However,

for the interviewee, the picture of the world clearly is different, as they cannot separate Monday and Tuesday random variables from the stochastic process and consider them separately. And this incompatibility means no guarantee of ensuring Kolmogorov-theory-wise completeness for decision-theoretic analysis.
The above means that there is no reason to unconditionally accept event completeness of Kolmogorov theory. One may assert that event completeness never was the case for conventional understanding of Kolmogorov theory, and this is true. After all, probabilistic analysis can still calculate probability of some events without probability of other events identified. What I rather mean, by event completeness of conventional Kolmogorov theory - or limited event completeness - is that some concepts, particularly conditional probability $P(X \in A|Y \in B)$, are often defined by recourse to $P(A, B)$ or others. But why should we assume that $P(A, B)$ exist, and define $P(A|B)$ based on $P(A, B)$ in limited information decision-theoretic settings?
It may be said differently as follows. Conditional probability in traditional Kolmogorov understanding was not a fundamental concept. It was a constructed and derived concept, and thus the definition reflected upon that. In decision theory, conditional probability does need to come out as a fundamental concept. Otherwise, we will not be able to carry out consistent analysis when $P(A|B)$ is available, but $P(A, B)$ is not available, as it will not even make sense to talk of $P(A|B)$.

2.1 Role of information and validity scope of conventional understanding of Kolmogorov theory in decision theory

It should be clear, at this point, that this paper emphasizes information in statistical analysis. This emphasis is somewhat obvious, but yet our stochastic pedagogy, and how analysis is presented, has not kept up with this required emphasis.
Think back on the sleeping beauty problem. The problem for the thirder solution was that it applied traditional Bayesian inference tools without realizing its scope limits in decision-theoretic settings. And this somewhat echoes the debate on whether Bayesian analysis/inference can "somewhat resolve" identification issues in regressions. That $\mathbb{P}_+(head) = 0, 1$ while subjective probability for the interviewee is not never is the problem - even if how reality works may be different from what our models indicate, these models may be pseudo-equivalent to reality - in other words, we would have no way of knowing that difference.
In case of the sleeping beauty problem, several values such as $P(Monday|tail)$, $P(Monday)$, $P(Tuesday)$ are not identifiable, even at subjective credence level. One may call this uncertainty Knightian uncertainty, but there is no need to do so.
All this, however, does not mean conventional Kolmogorov theory understanding does not apply in decision theory. To use "identification" terminology, in case events are identifiable (note that probability of these events may still not

be identifiable), conventional Kolmogorov theory may be applied safely, as long as they do not directly involve events that are not identifiable.
As any statistical analysis can be understood as updating prior, even for frequentist analysis - if prior is "objectively" defined, one can also think of statistical analysis as updating upon separable informations. This view clearly highlights how important understanding information clearly is important.
The question now then is whether information analysis can be done in a systemic way. I believe the answer is yes, and that part of the answer involves Bayesian network analysis, including things such as do-calculus. Whether this assessment is justified is outside the intended scope of this paper. But a clear strategy can be discussed: the sleeping beauty problem was resolved by essentially doing permutation/"shuffling"/"re-ordering" of available information - that is, information can be re-organized in different ways. One can check on all possible shuffles of available information, and because shuffles are all consistent with each other, a consistent picture can be derived from these shuffles. This of course is an expensive operation to do, in terms of computation resources, regardless of what happens in the future, and thus in many ways, we may have to rely on approximations and oracles based on ordinary intuitions for real-life statistical analysis.
As a digression, note that for the sleeping beauty problem, we also essentially imposed the computation limit, in terms of the game in the definition of pseudo-equivalent models (let me refer to this game as the original game, in contrast to the sleeping beauty experiment). To recall the original game, Person A tackling the game is given k data points and had to choose whether model D is the data-generating model or not (this alternate model is known to be model C for experimenters). Data points were selected by a fair coin toss, and if head came up, model C is the chosen data-generating model, if tail came up, model D is the chosen data-generating model. Given the information set of A, if one can prove, with knowledge of model C, unavailable to person A, that person A's choice effectively is a fair coin toss, then one can see that model C and D are pseudo-equivalent models in terms of person A's information set. The sleeping beauty game sets $k = 0$ (which is part of computational limit), and the only information the interviewee has is how the experiment is set up and that she is currently in the experiment.

2.2 Principle of Indifference: uniform prior when indifferent?

The thirder arguement of the sleeping beauty problem applied the following principle of indifference [7] [6]: $P(Monday|tail) = P(Tuesday|tail)$, as given that the coin toss is tail, the person does not really know whether she is on Monday or Tuesday. As she has minimal information, she may apply 50-50 chance to Monday and Tuesday, given that the coin toss result is tail.
But is this really a correct application of principle of indifference? Consider the following example from basic economics. Ignore decreasing marginal utility, and consider that each unit consumption of good L and M gives utility of Z to

Consumer A. Then Consumer A is indifferent over all combinations of good L and M as long as $|L|+|M|$ remains the same. If one answers that uniform prior should be used to represent what combination will be consumed by Consumer A, then $P(Monday|tail) = P(Tuesday|tail)$ should also follow.
But there is difference between "what can be a good prior" and "what prior must be chosen." Justifying uniform prior by principle of invariance under permutation belongs to consideration of "what can be a good prior," not "what prior must be chosen." Furthermore, in light of analysis done in this section, one can see that what seem to be a good prior may not actually be a good prior.

3 Conclusion

In all of the discussions in the paper, an alternate decision could have always been made. Recall the sleeping beauty problem. Because $\mathbb{P}_+(head) = 0, 1$, one could simply conclude that any credence/belief of the interviewee is either meaningful or meaningless, and that all these discussions are meaningless. Similarly, for pseudo-equivalent models, one may conclude that because we may never be able to infer a true model, statistical analysis is meaningless.
While there is substance in this pessimistic viewpoint, they are misguided. As cryptography was mentioned: pseudo-random functions are not actually random functions, but in reality, they effectively work as random functions, except for those that know secure keys that label those pseudo-random functions. Thus, as in Friedmanite instrumentalism, a model that is not descriptively realistic may actually turn out to have some validity. But Friedmanite instrumentalism must be considered in terms of what information is traced out to justify an alternative model. Considering information traced out as part of assumption, then realism of assumptions does matter, in terms of information. For example, if a model requires dropping out too much information, then the model should simply be rejected in favor of better more information-reflective models in case information is available.
The main take-aways of this paper are: that we use a misspecified model to analyze reality does not always mean garbage statistical analysis more than we expect fundamentally. By the concept of pseudo-equivalent models, one can see that two models may effectively be equivalent in terms of information limit and computation power assigned to process such information. We always face limited information and processing/computation capacity for statistical analysis, and thus real-life statistical analysis likely may be studying a model that is not an actual data-generating model, and we may never see a sign of misspecification. Computational processing limit, exemplified by function 2^n computation steps for n data points, with online updating algorithm unavailable or infeasible, makes valid model inference almost impossible, even if informational limits are not really severe. Thus, statistical analysts are not really different from deciders in decision theory who have to form credence/belief over events, but try their best to get close to accurate description of reality as far as they see things. If decision theory matters for statistical analysis, then its puzzles and paradoxes

matter for statistical analysis/inference. The sleeping beauty problem example is studied in this spirit. A fundamental point is raised: that prior in Bayesian inference has to be chosen carefully, and that Bayes rule is valid only up to its scope. This necessitates re-evaluating how Kolmogorov probability theory is to be applied in decision-theoretic settings. That is, Bayesian inference has to be done after event/parameter identification problem is clearly understood, in terms of available information.

References

[1] L. Breiman. Bias, variance and arcing classifiers. 1996.

[2] A. Einstein. Investigations on the theory of the brownian movement. 1956.

[3] A. Elga. Self-locating belief and the sleeping beauty problem. *Analysis*, 60(2):143–147, 2000.

[4] M. Friedman. The methodology of positive economics. 1953.

[5] A. Gelman and E. Loken. The garden of forking paths: Why multiple comparisons can be a problem, even when there is no fishing expedition or p-hacking and the research hypothesis was posited ahead of time. 2013.

[6] Edwin Thompson Jaynes. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[7] John Maynard Keynes. The principle of indifference. In *A Treatise on Probability*, chapter 4. Macmillan and Company, limited.

[8] Goldreich. O., S. Goldwasser, and S. Micali. How to construct random functions. *Journal of the ACM*, 33(4):792–807, 1986.

[9] J. P. Simmons, L. D. Nelson, and U. Simonsohn. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11):1359–1366, 2011.