What is new in Statistics?

HIEN D. TRAN[1]

[1] *School of Engineering, Tan Tao University, Long An, Vietnam*

Online: April 25, 2016

The debate between two schools of thought about statistical inference, namely frequentist and Bayesian approach, has been taking for decades. It is unecessary to try to argue which one is better, simply since *you can't win!*. It is not just a matter of taste. It is so since each approach has advantages and weaknesses. While you might get the impression that Baysian statistics is gaining ground in the 21st century, as testified by David Draper (2009) (see [1]). In addition, in view of the ban of using p-values by the Basic and Applied Scocial Psychology (BASP) (see [2]) together with the Statement by the American Statistical Association recently to warn statistical practitioners about using p-values (see [3]), it is essential to have new statistical methods to compromise disadvantages of those two "traditional" approaches. The Inferential Model (IM) introduced by Ryan Martin and Chuanhai Liu (see [4]) is a potentially candidate.

Key words: *Confidence region, Inferential model, p-Value, Random sets, Statistical inference*

http://review.ttu.edu.vn/2016/010106

1. INTRODUCTION

We are talking about using statistical science in making decisions under uncertainty.

To facilitate the discussions, let's consider the simplest situation: we are uncertain about the probability of "success" $\theta \in (0,1)$ of some operation, since we do not know its "true" value, say θ_o (in the "parameter space" $\Theta = (0,1)$). Let X denote the outcome of the operation, which could be only a "success" (that we write $X = 1$) or a failure (that we write $X = 0$). We are interested in $P(X = 1) = \theta_o$.

Discovering θ_o is a problem of estimation, whereas "guessing" a future outcome of the operation is a problem of prediction. A scientific theory for both is called statistical science.

At the turn of the century, we have two main approaches to statistical science: the frequentist and Bayesian schools.

In both approaches, we need data drawing from the population X, say, a (i.i.d.) random sample $X_1, X_2, ..., X_n$, denoted by $\mathbf{X}$. Based upon probability theory, we view the population X as a random variable, with probability density

$$f_\theta(x) = \theta^x(1-\theta)^{1-x}$$

where $x \in \mathcal{X} = \{0,1\}$, and $\theta \in \Theta = (0,1)$, or more generally with its distribution function

$F_\theta(.) : \mathbb{R} \to [0,1]$:

$$F_\theta(x) = P(X \le x) = \begin{cases} 0 & \text{for} \quad x < 0 \\ 1-\theta & \text{for} \quad 0 \le x < 1 \\ 1 & \text{for} \quad x \ge 1 \end{cases}$$

Let's focus on the estimation of θ_o.

Clearly the observation $\mathbf{X}$ should shed light on the whereabouts of θ_o, i.e., provide a localization information on it. How to quantify that information? That was Fisher's concept of likelihood, suggesting a principle to estimate θ_o in a scientific way (at least for regular models). Of course, there are other estimation methods based only on the observed data.

The Bayesian approach is different. Its framework is this. A prior information about the whereabouts of θ_o is postulated in the form of a probability, say,

distribution $\pi(.)$ on Θ. Thus, viewing θ_o as a random variable (so that an epistemic uncertainty can be "converted" into a stochastic uncertainty where localization information is described by a probability distribution). Second, viewing the sampling model as a conditional model, i.e., the density $f_\theta(x)$ is the conditional density of X given θ, so that the joint density of the random vector (X, θ) is $f(x, \theta) = f_\theta(x)\pi(\theta)$. From this, all statistical inference about θ_o is carried out by using the posterior distribution of θ given $\mathbf{X}$ via Bayes' formula

$$f(\theta|\mathbf{X}) = \frac{f_\theta(\mathbf{x})\pi(\theta)}{\int f_\theta(\mathbf{x})\pi(\theta)d\theta}$$

Specifically, in a general decision framework, an optimal estimator of θ_o is obtained as the posterior mean $\int_\Theta \theta f(\theta|\mathbf{X})d\theta$.

As far as inference about unknown parameters (in statistical models) is concerned, statisticians use observed data to estimate them, together with quantified uncertainty about their estimators, resulting in confidence regions or plausible regions, depending on which approach to statistics they take. For these outcomes to be "trusted", the meaning of "confidence", "plausibility" need to be "explained".

If they take the frequentist approach (i.e., using probability theory based on long run frequency interpretation of probability, as a measure of uncertainty), then the notion of "confidence" has a frequentist interpretation, and hence easy to understand and to be acceptable.

If they take the Bayesian approach where subjective probability (not frequency-based) is used, at the prior level, then the "plausibility" (at the posterior analysis) does not have a frequentist interpretation, but is interpreted as "degrees of belief".

It is unecessary to try to argue which interpretation (or school of thought) is better, simply since *you can't win!* It is not just a matter of taste. It is so since each approach has advantages and weaknesses. While you might get the impression that Baysian statistics is gaining ground in the 21st century, as testified by David Draper (2009) in *Bayesian statistical reasoning: an inferential, predictive and decision-making paradigm for the 21st century* (see [1]).

While the Bayesian approach seems more powerful than the frequentist one, due to the umbrella of decision theory, statisticians are often uneasy with the choices of prior distributions. Thus, the state-of-the-art of statistical science (as referred to its very foundational basis) is like "a matter of taste"!

Is there any reasonable "compromise"?, say, a "prior-free" Bayesian approach which satisfies both schools? for example, a data-based posterior analysis?

Several attempts in the last century seem to fail; And this seems to be the claim that the Inferential Model could provide a solution!

2. INFERENTIAL MODELS

Now is what you are waiting for! A new framework for doing statistics. With the book

Inferential Models: Reasoning with Uncertainty
R. Martin and C. Liu, Chapman and Hall/CRC Press, 2015

(see [4])

handy, containing history/motivation, various areas of statistical applications, we choose to elaborate on it, at the fundamental level, to ease your reading of the entire book. This will consist of describing the framework in two basic areas: confidence region estimation (of parameters) and prediction (of future outcomes).

A. A prelude to inferential models

A framework for a scientific study is a formulation of the study together with necessary ingredients to investigate it. As we will see shortly, the IM framework is not quite a new "approach" to statistical inference, as you can figure out yourself by comparing it with the frequentist and Bayesian approaches. That is why we call it a "framework" rather an approach.

It consists of laying out the general statistical inference problem in a "new" way, suitable for investigation, different than traditional ones. It sits somewhat between classical and Bayesian frameworks: on one hand, it relies only on seen data, and on the other hand, it introduces a "semi-data driven" as a "prior" before looking at the observed data. As such, the framework can still be termed "prior-free". The subsequent analysis is like a "posterior" one, but without priors!

Inferential models? Models here do not mean "statistical models" in the usual sense (sampling models are models of the distributions generating the observations, e.g., normal models), but "models for conducting inference", as it will become clear shortly. This is so since, there are various ways (models) to conduct inference on a given sampling model. Thus,

an inferential model (for a given sampling model) is a specific way to conduct statistical inference. In a sense, an inferential model is like a Bayesian framework with a chosen prior distribution. As we will see, an inferential model is a specific link between the variable (population) of interest with its distribution, where "specific" means "as an equation", rather than just a formal "description" of their relationship.

The following simple situation helps to start out. Let our variable of interest be a real-valued random variable X (defined on some probability space $(\Omega, \mathcal{A}, P)$). We observed data drawn from X, say, $\mathbf{X} = (X_1, X_2, ..., X_n)$, and wish, among other things, to discover the law governing its random evolution, and to predict its future values.

The distribution function of X is the function

$$F(.) : \mathbb{R} \to [0,1], \quad F(x) = P(X \leq x)$$

The "relation" between X and its distribution F (or F_X when needed) is just that!

A more "formal" relation between X and F is obtained via its quantile function $F^{-1}(.) : (0,1) \to \mathbb{R}$, which is defined as

$$F^{-1}(\alpha) = \inf\{x \in \mathbb{R} : F(x) \geq \alpha\}$$

namely

$$X \overset{D}{=} F^{-1}(U)$$

where U is uniformly distributed on the interval $(0,1)$.

The point is this. There exist explicit "equations" relating X, its distribution F, and some "auxiliary" unobservable random variable U, where the above "canonical" one $X = F^{-1}(U)$ is an example, so that statistical inference could be based on them. For example, in parametric sampling models where the distribution function of the observable X is $F_\theta(.)$, $\theta \in \Theta$, the above "canonical" equation take the form $X = F_\theta^{-1}(U) = a(\theta, U)$, where the function $a(.,.)$ ("a" for **association** between three variables X, θ, U) is known, as well as the distribution of the unobservable U.

There are various "association" functions $a(.,.)$ for a given problem. For example, if we are investigating a "model" such as the (sampling) normal $N(\mu, \sigma^2)$ for X, we write

$$X = \mu + \sigma Z$$

with Z being $N(0,1)$ (a known distribution), then $X = a(\theta, Z)$, where $\theta = (\mu, \sigma)'$.

As another example, the Bernoulli model can be written as $X = a(\theta, U) = 1_{[0,\theta]}(U)$, with U being uniformly distributed on $[0,1]$, since $1_{[0,\theta]}(U)$ is equal to X in distribution:

$$P(1_{[0,\theta]}(U) = 1) = P(0 \leq U \leq \theta) = \theta = P(X = 1)$$

$$P(1_{[0,\theta]}(U) = 0) = P(U > \theta) = 1 - P(U \leq \theta) = 1 - \theta = P(X =$$

How precisely to "base" statistical procedures on $X = a(\theta, U)$ is what the IM is all about.

Remarks

(i) The "equation" $X = F^{-1}(U)$ was known since the invention of computer simulation, for generating simulated data from known $F's$. It is not systematically used for developing statistical theory. The use of this equation for statistical inference in IM is a kind of reverse problem. In Engineering terminology, it looks like an **identification** problem.

(ii) A "system" interpretation: When $F(.)$ is known, say, as a model specified for a "stochastic system", then the above connection $X = F^{-1}(U)$ can be viewed as an input-output system: the (random) input U produces the output X via the "transfer" function $F^{-1}(.)$. This is the essence of the Monte Carlo method for creating simulated data from a known system: imitating the way nature "draws" random samples.

A standard "identification" problem in system theory (a sort of regression in statistics) will consist of having the observables as both input U and output X to discover (identify) the unknown function $F^{-1}(.)$.

Another (more difficult) "identification" problem is this. While the function F (and hence F^{-1}) is unknown, we can only observe the output X (the "input" U is unobservable), and the goal is the same: discover F

You recognize that, while the setting is exactly that of a statistical inference one, it is not usually formulated this way, let alone deriving statistical procedures based on it.

This seems to be the motivation of the so-called Inferential Models. It remains of course to find out whether this "new" way of formulating statistical problems together with its *random set-based inference* procedures offers any improvements on contemporary statistical knowledge?

B. The IM framework

We start out by first looking closely at the IM framework for *set estimation of parameters*, as in the paper

Random sets and exact confidence regions by Martin, R., Sankhya, 76-A, 288-304 (2013) (See [5])

Consider the basic problem of parameter estimation in parametric models. Let X be a real-valued random variable with distribution $F_\theta(.), \theta \in \Theta \subseteq \mathbb{R}$. Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ denote our observed data from X. Usually, to construct confidence intervals for the true (but unknown) θ_o, we could proceed as follows. Let G_θ be the (sampling) distribution of a, say, sufficient statistic T. For $\alpha, \beta \in (0,1)$, let $u(\theta)$ and $v(\theta)$, in the range $\mathcal{T}$ of T, such that $G_\theta(u(\theta)) = \alpha$, and $G_\theta(v(\theta)) = 1 - \beta$, so that

$$P_{T|\theta}(u(\theta) \leq T \leq v(\theta)) = 1 - \alpha - \beta$$

If we could "solve" for θ from

$$u(\theta) \leq T \leq v(\theta)$$

i.e., there exist $a(T), b(T)$ such that

$$u(\theta) \leq T \leq v(\theta) \Leftrightarrow a(T) \leq \theta \leq b(T)$$

then clearly, $[a(T), b(T)]$ is a confidence interval of at $100(1 - \alpha - \beta)\%$ level.

Note that the transformation is from the range of T to the parameter space Θ, that is the random set $[a(T), b(T)]$ in Θ is obtained as a "propagation" (an Engineering terminology for pushing uncertainty from the inputs (one space) to the outputs (another space) of systems) of uncertainty in $\mathcal{T}$ to Θ.

Confidence regions in inferential model framework

Let X be a population with values in $\mathcal{X}$ with probability law (measure) $P_{X|\theta}$ on $\mathcal{X}$, assuming to be a parametric model, $\theta \in \Theta \subseteq \mathbb{R}^d$ (or more generally, a locally compact, Hausdorff, second countable space). Let Y be a sample from X (to be an observed data) with corresponding law $P_{Y|\theta}$. Let $T(Y)$ be a sufficient statistic for θ, with law $P_{T|\theta}$. We wish to construct a set-estimator for the true (unknown) parameter θ_o, i.e., a random set $S(Y)$ on Θ, with some given specifications.

The following approach to achieve this goal is this. It is possible to express (write) $T(Y)$ explicitly as a function of θ and some auxiliary random element U: $T(Y) = a(\theta, U)$ where the function $a : \Theta \times \mathcal{U} \to \mathcal{T}$ is specified, the random element U (with values in $\mathcal{U}$) is unobservable, but having a known law P_U, in such a way that when $U \sim P_U$ (distributed as), we have $T(Y) \sim P_{T|\theta}$.

A simple example will clarify what we have just said. Suppose X has a parametric distribution function $F(x|\theta)$ on $\mathbb{R}$, with $\Theta \subseteq \mathbb{R}$, where θ is the (unknown) mean of X. Let $X_1, X_2, ..., X_n$ be a random sample (i.i.d.) from X. Set $Y = (X_1, X_2, ..., X_n)$ the sample to be observed. The distribution of Y is referred to as the **sampling model** $P_{Y|\theta}$ which here is $n-$ fold product of $dF(x|\theta)$. The sufficient statistic for the mean is $T(Y) = X_1 + X_2 + ... + X_n$ having distribution $P_{T|\theta}$ obtained by convolution of the $dF(x|\theta)$. Let $F_{T|\theta}(.)$ denote the distribution function of T, and its quantile function $F^{-1}_{T|\theta}(.) : (0,1) \to \mathbb{R}$, defined as

$$F^{-1}_{T|\theta}(\alpha) = \inf\{x \in \mathbb{R} : F_{\theta|T}(x) \geq \alpha\}$$

then, it is well known that $T \overset{D}{=} F^{-1}_{T|\theta}(U)$, where U is uniformly distributed on $(0,1)$, a basic fact for simulations.

The "auxiliary" random variable U here has a known distribution, namely, uniform distribution on $(0,1)$. In simulation context, we known the distribution of T, that is we know the true parameter θ_o, and we then create simulated data, e.g. a random sample from T, by simulating U and use the relation $T \overset{D}{=} F^{-1}_{T|\theta}(U)$, in fact $T = F^{-1}_{T|\theta}(U)$ to obtain a simulated sample. The situation here is the reverse: the true parameter is unknown, and we have real observe data, say, $T = t$, and we wish to "infer" θ from the observed t.

Thus, firstly, the inferential models came from the exploration of the reverse problem of simulations. Next, since U is unobservable (performed by nature), but having a known distribution, and is connected to the observable X in some specific way (depending on unknown parameters), it is possible to "guess" U (guessing, and not predicting, since the unknown values of U were already "realized"), how? Its answer seems to be the second "component" of the Inferential Models: **using random sets.** But why do we need to guess unobserved values of U? when our goal is to "guess" the correct population parameter? The connection $X = F^{-1}_\theta(U)$ relates three quantities x, θ and u, with only x is observed. Thus, if $U = u$ were also observed (giving rise to the actual observed value $X = x$, regardless of the unknown, true parameter), then clearly θ can be located, as solutions of

$x = F_\theta^{-1}(u)$. For this reason, it seems intuitive to try first to guess the unobservable u (based on x and the sampling model) to arrive as an approximate "confidence set" for the realized (but unobservable) u from which inference about θ can be initiated.

Here is how this procedure is carried out.

Guessing an unobservable

To be specific, consider U uniformly distributed on $(0,1)$, and the problem is to guess a realized value, say, $U = u_o \in (0,1)$. Recall that u_o is the realized value of U giving rise to our actual observed value of $X = x$, i.e., $x = F_\theta^{-1}(u_o)$.

Remark. Once a value u_o of a random variable U is realized, it is fixed (no longer random). Thus, if it is unobservable, we are uncertain about its whereabouts: It's an epistemic uncertainty! However, there is "something else" involved here: Yes, u_o is an unknown value in $(0,1)$, but it came from a random variable U with a known distribution.

Should we try to guess u_o by some value $u \in (0,1)$?

Here, by "guessing u_o", we mean "locate u_o in some subset of $(0,1)$".

This is precisely the spirit of **coarsening** (an intelligent attribute of humans): When we cannot guess a value in a space, we coarsen that space, i.e., replacing the space by some collection of its subsets and use this coarsened space for guessing.

Clearly, for a coarsening scheme to be meaningful, it has to have some "fixed point" property! exactly like the situation of **coarse data** (data with low quality): Suppose we perform U but we cannot observe the value $U = u$, instead we observe a **random set** S on its domain, containing u, i.e., the random set S is such that, for any u, $S \ni u$, with probability one, i.e., $P(U \in S) = 1$. Specifically, we wish to construct a random set $S(U)$ on $(0,1)$ such that $u_o \in S(u_o)$ (and use $S(u_o)$ as our guess about u_o), but we don't know u_o, so we require this fact to be true for all possible $u \in (0,1)$, leading to the condition: for any u, $S(u) \ni u$.

A random set $S(U)$ is called a coarsening of U if $P(U \in S(U)) = 1$. Note that the distribution of the random set S should be known from that of U. In this sense, the above in the reverse of coarse data setting: To guess a realized value of a random variable U, we seek a random set S such that S admits U as one of its (a.s.) selection. Recall that, if S is a random set with values in, say, $2^{\mathcal{U}}$, then a measurable selection of S is a random variable X with values in $\mathcal{U}$, such that $P(X \in S) = 1$.

In coarse data analysis in statistics (see, e.g., [6] or [7]), the situation is different: The population X has an unknown distribution, and we observe only a sample of sets (a sample from a random set) containing the sample drawn from X. The goal is discovering the distribution of X.

When S is a coarsening of U, we can use S to "shed light" on unseen values of U: without observing u_o, we use a realization of S as our guess for u_o, i.e., u_o is guessed to be in that realization (a subset of $(0,1)$) of S). For that to be meaningful, S cannot be chosen arbitrarily, but instead, has to be related to U: Specifically, we need to create a random set S as a function of U, denoted as $S(U)$, noting that U has a known distribution.

The "moral" is this. While nature performed U but did not reveal its value u_o to us, we create a random set $S(U)$ which is also performed by nature, but we can observe $S(u_o) \ni u_o$.

In view of such property, $S(U)$ is called a **predictive random set** in IM. It is a random set used to guess ("predict") the realizations of some random variable.

Here is an example:

$$S(U) = \{u \in (0,1) : |u - \frac{1}{2}| \le |U - \frac{1}{2}|\}$$

Clearly, $U \in S(U)$ (with probability one).

As stated above, if u_o were observed, then the whereabouts of θ_o are in

$$\Theta_x(u_o) = \{\theta \in \Theta : x = a(\theta, u_o)\}$$

But u_o is not observable, and we have a predictive random set $S(U)$, we can consider the (larger) subset of Θ

$$\Theta_x(S) = \cup_{u \in S} \Theta_x(u)$$

which is a random set on Θ, noting that $S(U)$ is a random set on the range $\mathcal{U}$ of the auxiliary random variable U.

It is this random set $\Theta_x(S)$ (depending on the observed data x, the sampling model, and our choice of a predictive random set $S(U)$) that we will consider as our "confidence region" for θ_o.

If $\Theta_x(S)$ is to be "considered" as a confidence region for θ_o, we need to figure out its coverage probability, i.e., $P_\theta(\Theta_x(S) \ni \theta)$, for any $\theta \in \Theta$. In fact,

since $\Theta_x(S)$ is a random set which is a "function" of the random set $S(U)$, its "distribution" $P_{\Theta_x(S)}$ on the power set of Θ, is a function of the probability law of the random set $S(U)$, denoted as P_S, so that what we should seek is $P_S(\Theta_x(S) \ni \theta_o)$.

Now $\Theta_x(S)$ is a $100(1-\alpha)\%$ confidence region for θ if

$$P_\theta(\Theta_x(S) \ni \theta) \geq 1-\alpha$$

By fixing $\alpha \in (0,1)$ in advance, clearly the above inequality is only possible if we can choose some predictive random set $S_\alpha(U)$ appropriately, i.e., one such that

$$P_\theta(\Theta_x(S_\alpha) \ni \theta) \geq 1-\alpha$$

If that can be done, then

$$\Theta_x(S_\alpha) = \cup_{u \in S_\alpha}\{\theta \in \Theta : x = a(\theta, u)\}$$

will be our exact confidence region at nominal level $1-\alpha$.

Thus, the basic question is: What is S_α?

Recall that the sampling model is written as $X = a(\theta, U)$ for some known function $a : \Theta \times \mathcal{U} \to \mathcal{X}$, where the auxiliary (unobservable) random variable $U : (\Omega, \mathcal{A}, P) \to (\mathcal{U}, \mathcal{B}(\mathcal{U}), P_U)$ has a known law P_U. Let $S(.) : (\mathcal{U}, \mathcal{B}(\mathcal{U}), P_U) \to 2^U$ be a predictive random set for U, with probability law P_S. The coverage function of S is $\pi_S(.) : \mathcal{U} \to [0,1]$, defined as $\pi_S(u) = P_S(S \ni u)$. We write $\pi_S(U)$ for the random variable taking values in $[0,1]$: $u \in \mathcal{U} \to \pi_S(u)$.

First, we need to choose S appropriately to obtain exact confidence levels. The random set S is said to be **valid** if it dominates stochastically (first order) the uniform distribution on $[0,1]$ (see [8], for more details), i.e., $F_{\pi_{S(U)}}(\alpha) \leq F_{U(0,1)}(\alpha) = \alpha$; equivalently, $1 - F_{\pi_{S(U)}}(\alpha) \geq 1-\alpha$, for any $\alpha \in [0,1]$.

A construction of valid random sets was suggested in Martin's (see [5], pp. 294): If S is nested (with support $\mathbb{S}$) such that for any $u \in \mathcal{U}$,

$$\pi_S(u) = 1 - \sup_{\{u \notin A \in \mathbb{S}\}} P_U(A)$$

then S is valid.

Now, if the predictive random set S is valid, then the associated random set $\Theta_T(S)$ on Θ is also valid (see Martin, 2004, p. 297), i.e., its coverage function (as a random variable: $T = t \to P_S(\Theta_t(S) \ni \theta)$)

$$\pi_{\Theta_T(S)}(\theta) = P_S(\Theta_T(S) \ni \theta)$$

is stochastically larger than the uniform distribution on $(0,1)$.

A standard procedure is this. A $100(1-\alpha)\%$ confidence region based on the valid random set $\Theta_t(S)$ (depending on the data $T = t$) is obtained by taking the $\alpha-$ level set of its coverage function

$$C_\alpha = \{\theta \in \Theta : \pi_{\Theta_t(S)}(\theta) \geq \alpha\}$$

since then, by validity, for any $\theta \in \Theta$,

$$P_{T|\theta}(C_\alpha \ni \theta) = P_{T|\theta}(\pi_{\Theta_T(S)}(\theta) \geq \alpha) \geq 1-\alpha$$

You can obtain also a confidence region directly from the random set $\Theta_T(S)$ by choosing $A_\alpha \in \mathbb{S}$ (range of S) such that $P_U(A_\alpha) \geq 1-\alpha$. Your $100(1-\alpha)\%$ confidence region is then $\Theta_T(A_\alpha)$. Indeed,

$$P_{T|\theta}(\Theta_T(A_\alpha) \ni \theta) = P_U(A_\alpha) \geq 1-\alpha$$

and the uncertainty of an assertion $A \subseteq \Theta$ is quantified as

$$[P(\Theta_X(S) \subseteq A), P(\Theta_X(S) \cap A \neq \varnothing)]$$

which are computed from the capacity functional of $\Theta_X(S)$.

An example

Let X be $N(\theta, 1)$, and given $X = x$ (for simplicity).

Then, $X = \theta + Z$, where Z is $N(0,1)$ with distribution function Φ. Thus, an association function could be $a(\theta, U) = \theta + \Phi^{-1}(U)$ for the auxiliary (unobservable) random variable U which is uniformly distributed on $(0,1)$. Set

$$\Theta_x(u) = \{\theta \in (0,1) : \theta = x - \Phi^{-1}(u)\}$$

Next, consider the valid predictive random set on $(0,1)$

$$S(U) = \{u \in (0,1) : |u - .5| \leq |U - .5|\}$$

and set

$$\Theta_x(S) = \cup_{u \in S}\Theta_x(u) = \cup_{u \in S}\{\theta : \theta = x - \Phi^{-1}(u)\}$$

$$= \{\theta = x - \Phi^{-1}(u) : |u - .5| \leq |U - .5|\}$$

$$= \{\theta = x - \Phi^{-1}(u) : .5 - |U - .5| \leq u \leq .5 + |U - .5|\}$$

$$= [x - \Phi^{-1}(.5 + |U - .5|), x - \Phi^{-1}(.5 - |U - .5|)] = [Y_*, Y^*]$$ Otherwise,

The coverage function of the random set $\Theta_x(S)$ on $\Theta = (0, 1)$ is

$$\pi_{\Theta_x(S)}(\theta) = P_U(\Theta_x(S) \ni \theta) =$$

$$P_U(Y_* \leq \theta \leq Y^*)$$

Now, if we let $A = \{\omega : Y_*(\omega) \leq \theta\}$, $B = \{\omega : Y^*(\omega) \geq \theta\}$, then

$$\{\omega : Y_*(\omega) \leq \theta \leq Y^*(\omega)\} = A \cap B$$

Thus,

$$P(A \cap B) = 1 - P(A^c \cup B^c) = 1 - P(A^c) - P(B^c) + P(A^c \cap B^c)$$

$$= 1 - P(A^c) - P(B^c)$$

since $Y_* \leq Y^* \Longrightarrow A^c \cap B^c = \varnothing$.
Therefore,

$$\pi_{\Theta_x(S)}(\theta) = 1 - P(Y_* > \theta) - P(Y^* < \theta)$$

Now,

$$P(Y_* > \theta) = P_U(x - \Phi^{-1}(.5 + |U - .5|) > \theta) =$$

$$P_U(\Phi^{-1}(.5 + |U - .5|) < x - \theta) = P_U(.5 + |U - .5|) < \Phi(x - \theta)) =$$

$$P_U(.5 - \Phi(x - \theta) < U - .5 < -.5 + \Phi(x - \theta)) =$$

$$P_U(1 - \Phi(x - \theta) < U < \Phi(x - \theta)) = 2\Phi(x - \theta) - 1$$

if $1 - \Phi(x - \theta) < \Phi(x - \theta)$, i.e., if $2\Phi(x - \theta) - 1 > 0$; Otherwise, it is 0.
Similarly,

$$P(Y^* < \theta) = P_U(x - \Phi^{-1}(.5 - |U - .5|) < \theta) = 1 - 2\Phi(x - \theta)$$

if $2\Phi(x - \theta) - 1 < 0$; Otherwise it is 0.
Thus, when $2\Phi(x - \theta) - 1 > 0$

$$\pi_{\Theta_x(S)}(\theta) = 1 - P(Y_* > \theta) - P(Y^* < \theta) = 1 - (2\Phi(x - \theta) - 1)$$

$$\pi_{\Theta_x(S)}(\theta) = 1 - (1 - 2\Phi(x - \theta))$$

so that,

$$\pi_{\Theta_x(S)}(\theta) = 1 - |2\Phi(x - \theta) - 1|$$

A $(1 - \alpha)100\%$ confidence interval for θ is

$$C_\alpha = \{\theta : \pi_{\Theta_x(S)}(\theta) \geq \alpha\} = \{\theta : 1 - |2\Phi(x - \theta) - 1| \geq \alpha\} =$$

$$\{\theta : |2\Phi(x - \theta) - 1| \leq 1 - \alpha\} = \{\theta : \alpha - 1 \leq 2\Phi(x - \theta) - 1 \leq 1$$

$$\{\theta : \frac{\alpha}{2} \leq \Phi(x - \theta) \leq 1 - \frac{\alpha}{2}\} = \{\theta : \Phi^{-1}(\frac{\alpha}{2}) \leq x - \theta \leq \Phi^{-1}(1 -$$

$$[x - \Phi^{-1}(1 - \frac{\alpha}{2}), x - \Phi^{-1}(\frac{\alpha}{2})] = [[x - \Phi^{-1}(1 - \frac{\alpha}{2}), x + \Phi^{-1}(1 -$$

Remark. This example brings out the fact that the IM framework *generalizes* the approach to construct confidence interval by pivoting the distribution function of the sampling distribution of the (sufficient) statistics involved (when we can actually pivot it) (see [9], pp. 430-435)

3. COMMENTS ON IM

Before taking a closer look at the IM framework, let's summarize the essentials.

Philosophical issues aside, the technical stuff is this. Writing the sampling model in a chosen form $X = a(\theta, U)$, the "new" step is the choice (or construction) of a random set S on the range space $\mathcal{U}$ of the auxilary (unobservable) random variable U. That random set S is constructed from U which has a known distribution. The choice of $S(U)$ is important. It must be valid, i.e., having its (random) coverage function $\pi_{S(U)}(U)$ stochastically larger than the uniformly distributed random variable on $(0, 1)$.

After this essential step, the confidence regions obtained are simply level sets of the "expended" random set $\Theta_X(S(U))$ on the parameter space Θ. We get EXACT confidence regions, for ANY sample size.

A. What is new in IM?

A beginning problem where we need statistics is this. We are "interested" in discovering the distribution $F(.)$ of some random variable of interest X by looking at a random sample (observed data) $X_1, X_2, ..., X_n$ drawn from it. What is unknown to us is the distribution function $F(.)$, so that discovering $F(.)$ could be viewed as estimating it.

How to proceed to estimating $F(.)$ from observed data?

Well, first, what we have is this.

(a) X is distributed as $F(.)$, observed data $X_1, X_2, ..., X_n$. This is the natural "framework" of the **frequentist** (standard, classical) **approach to statistics** (your well-known statistics).

(b) Another approach to do statistics is the **Bayesian approach to statistics** where one additional piece of information is provided (by the statistician): a prior distribution $\pi(.)$ on the unknown $F(.)$ (whether parametrically or nonparametrically), i.e., a subjective probability distribution (with no frequentist interpretation), noting that the uncertainty about $F(.)$ is epistemic in nature. Thus, the famework of Bayesian statistics is:

X is distributed as $F(.)$, observed data $X_1, X_2, ..., X_n$, and $\pi(.)$.

Now, the new IM framework is just like the "standard" one, i.e., with

X is distributed as $F(.)$, observed data $X_1, X_2, ..., X_n$

but specifying more about the "qualitative" information "X is distributed as $F(.)$", as an "quantitative" equation, e.g.,

$$X \overset{D}{=} F^{-1}(U)$$

Noting that, while such an "equation" (where equality is in distribution sense) is familiar in Monte Carlo simulations (for producing simulated data from a known distribution), it was not used, in the reverse direction (i.e., from X to U) in statistical analysis so far.

With this "inferential model" $X \overset{D}{=} F^{-1}(U)$, the IM framework adds to the standard framework, as

X is distributed as $F(.)$, observed data $X_1, X_2, ..., X_n$, and U, an unobservable random variable, with known distribution.

You could view U (arising from the sampling model as $X \overset{D}{=} F^{-1}(U)$) as the counterpart of the Bayesian prior $\pi(.)$ (in the sense that, like $\pi(.)$, although unobserved, it provides (localization) information about the unknown parameter of interest, here $F(.)$, since our observed X came from U. However, the essential difference is that this additional information (the known distribution P_U of U) is not a subjective probability. The "additional" information (to do statistics) is from the sampling model itself (just "objective"), and not from "outside" (like in Bayesian statistics).

Thus, the IM framework sits in between classical and Bayesian approaches to statistical analysis, and could provide a compromise between two schools of thought. Viewing P_U as an **objective prior** (before seeing the data), the subsequent analysis (using the observed data) is like a posterior analysis. Thus, the IM framework (not an "approach") is termed "posterior analysis without prior" or "prior-free" analysis!

B. What are the advantages of IM framework?

(i) Avoiding the "dispute" between the classical and Bayesian views of doing statistics,

(ii) Having almost all "convenient" stuff of Bayesian statistics (to be elaborating below), although an investigation into *decision theory* is needed.

What are the "convenient" stuffs of Bayesian statistics?

In a sense, Bayesian statistics concerns fixed samples, and not asymptotics (e.g., not approximate inference). With a "available" prior, and treating the unkown population parameter as a random element (following that known distribution), the subsequent analyses are soly based upon its posteritor distribution (known through the sampling distribution and data). Point estimators are taken simply as means of the posterior distribution, set estimators are directly based on the known posterior distribution (no need of sampling distributions, i.e., distributions of point estimators), and tests of hypotheses (based on Bayes factors) are derived also from the posterior distribution, given the observed data from the "postulated" sampling model. Of course, all these statistical procedures are "protected" (when "reasoning with uncertainty") under the umbrella of decision theory.

Without using a prior, you are in the realm of "standard" statistics, and you are surely familiar with "problems" such as : which estimation methods to use? and why? how to find sampling distributions of statistics involved? how to find optimal tests?....

Appealing to limit theorems in probability theory, you only obtain **approximate inference**. When a population is not normal and the sample size is "small", even inference about its mean is a big problem, since we do know the (sampling) distribution of the sample mean. It is precisely here that a solution was proposed: **the bootstrap** (which still provides you with approximate results, in view of its own asymptotics).

Now, the IM framework, like Bayesian statistics, is *systematic*. For any sample sizes, it provides *exact inference*. It does not require sampling distributions. This IM framework could be used in help to resolve the starting "crisis" on statistical methods in using p-values for *Null Hypothesis Significance Testing procedure* initially raised by BASP in 2015 and then officially followed by ASA (the largest statistical society) in March 2016. (See [2] and [3])

ACKNOWLEDGMENTS

The author specially thanks Prof. Hung T. Nguyen for technical supports, advices and comments.

REFERENCES

1. Draper, D., "Bayesian statistical reasoning: an inferential, predictive and decision-making paradigm for the 21st century," online (2009). TIAM-PIMS-MITACS Distinguished Colloquium -Bayesian Statistical Reasoning.
2. D. Trafimow and M. Marks, "Editorial," Basic and Applied Social Psychology **37**, 1–2 (2015).
3. R. L. Wasserstein and N. A. Lazar, "The asa's statement on p-values: context, process, and purpose," The American Statistician **0**, 00–00 (0).
4. R. Martin and C. Liu, *Inferential Models: Reasoning with Uncertainty*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability (CRC Press, 2015).
5. R. Martin, "Random sets and exact confidence regions," Sankhya A **76**, 288–304 (2013).
6. H. Nguyen, *An Introduction to Random Sets* (CRC Press, 2006).
7. H. T. Nguyen, "On random sets and belief functions," Journal of Mathematical Analysis and Applications **65**, 531 – 542 (1978).
8. S. Sriboonchita, W. Wong, S. Dhompongsa, and H. Nguyen, *Stochastic Dominance and Applications to Finance, Risk and Economics* (CRC Press, 2009).
9. G. Casella and R. Berger, *Statistical Inference*, Duxbury advanced series in statistics and decision sciences (Thomson Learning, 2002).