Catalogue no. 12-001-X
ISSN 1492-0921

Survey Methodology

Conditional calibration and the sage statistician

SURVEY
METHODOLOGY

A JOURNAL
PUBLISHED BY
STATISTICS CANADA

JUNE 2005
•
VOLUME 31
•
NUMBER 1

by Donald B. Rubin

Release date: June 27, 2019

Statistics Canada
Statistique Canada

Canada

How to obtain more information

For information about this product or the wide range of services and data available from Statistics Canada, visit our website, www.statcan.gc.ca.

You can also contact us by

Email at STATCAN.infostats-infostats.STATCAN@canada.ca

Telephone, from Monday to Friday, 8:30 a.m. to 4:30 p.m., at the following numbers:

• Statistical Information Service	1-800-263-1136
• National telecommunications device for the hearing impaired	1-800-363-7629
• Fax line	1-514-283-9350

Depository Services Program

• Inquiries line	1-800-635-7943
• Fax line	1-800-565-7757

Standards of service to the public

Statistics Canada is committed to serving its clients in a prompt, reliable and courteous manner. To this end, Statistics Canada has developed standards of service that its employees observe. To obtain a copy of these service standards, please contact Statistics Canada toll-free at 1-800-263-1136. The service standards are also published on www.statcan.gc.ca under "Contact us" > "Standards of service to the public."

Note of appreciation

Canada owes the success of its statistical system to a long-standing partnership between Statistics Canada, the citizens of Canada, its businesses, governments and other institutions. Accurate and timely statistical information could not be produced without their continued co-operation and goodwill.

Published by authority of the Minister responsible for Statistics Canada

An HTML version is also available.

Cette publication est aussi disponible en français.

Survey Methodology, June 2019
Vol. 45, No. 2, pp. 187-198
Statistics Canada, Catalogue No. 12-001-X

Conditional calibration and the sage statistician

Donald B. Rubin[1]

Abstract

Being a calibrated statistician means using procedures that in long-run practice basically follow the guidelines of Neyman's approach to frequentist inference, which dominates current statistical thinking. Being a sage (i.e., wise) statistician when confronted with a particular data set means employing some Bayesian and Fiducial modes of thinking to moderate simple Neymanian calibration, even if not doing so formally. This article explicates this marriage of ideas using the concept of conditional calibration, which takes advantage of more recent simulation-based ideas arising in Approximate Bayesian Computation.

Key Words: Approximate Bayesian Computation (ABC); Bayesian inference; Fiducial inference; Fisher; Frequentist methods; Neyman.

1 Principled statisticians

There are many possible definitions for what makes a principled statistician, where by "principled" I do not necessarily imply "good" or "sage", but simply following clear principles of behavior. I think generally there are three major themes or philosophies of statistical inference. Neymanian frequentists, following ideas proposed originally by Neyman (1923, 1934), care about the operating characteristics of procedures (e.g., point estimates, interval estimates), under repeated sampling: point estimates should be approximately unbiased for their estimands (averaging over all possible samples), interval estimates should be conservative in the sense of having at least their nominal coverage of their estimands (again averaging over samples), and tests should be conservative in the sense of rejecting true null hypotheses at most at their nominal rates. These desiderata are widely viewed as being features of valid statistical inference (e.g., see Lehmann, 1959). Of course, all procedures that are valid are not equally desirable; valid point estimates with less variability are better, valid interval estimates that are shorter are better, and so forth.

Bayesian statisticians (e.g., Savage, 1954; Lindley, 1971; de Finetti, 1972), in contrast to repeated-sampling operating characteristics, care about correct conditioning on observed data under a particular probabilistic specification. Fisherian statisticians (in the sense of Fiducialists, at least as I view the most central idea of this approach, Fisher (1956)) avoid conclusions that appear to be contradicted by observed data, which is at the heart of Fisher's randomization test in experiments; I have long resonated to the wisdom of this approach and its generalizations, as expressed in Rubin (1984). Nevertheless, I also think Bayesian thinking is critical to being a wise applied statistician in practice, for example by using posterior predictive p-values and checks, which assess whether a proposed model can (that is, is able to, not must) generate data that look like the observed data set we are facing – we return to this central idea later.

There is little doubt that frequentist thinking dominates current statistical thinking even though Bayesian procedures are becoming more common largely because of current computational advances, which allow

[1] Donald B. Rubin, Professor, Yau Mathematical Sciences Center, Tsinghua University; Murray Shusterman Senior Research Fellow, Fox School of Business; Professor Emeritus, Harvard University. E-mail: dbrubin@me.com.

many complicated models to be fit routinely. Nevertheless, I believe that Bayesian and fiducial thinking are fundamental to being a sage (i.e., wise, not necessarily principled in the narrow sense of the following specific principles) statistician.

2 Should frequentists care about Bayesian procedures?

For example, why should frequentists ever use the sample mean to estimate the population mean? After all, the sample mean is essentially the center of the Bayesian posterior distribution of the population mean under a Gaussian model with relatively diffuse prior distributions on parameters, and therefore derived using an "unreliable (i.e., Bayesian) methodology". Of course this sentence is facetious, and not intended to be taken seriously, although there are serious points underlying it.

Serious Point #1: The original motivation for any statistical procedure, whether Bayesian or Fiducial or the result of some amazing dream, is irrelevant to the frequentist operating characteristics of that procedure. I used to hear this criticism directed at Multiple Imputation (MI), Rubin (1978). Because MI's initial justification was Bayesian, MI could never be trusted from a design-based (frequentist) perspective.

Serious Point #2: For creating procedures, especially in complex situations, such as those that easily arise with unintended missing data, Bayesian methods are far more generative of sensible answers than standard, frequentist arguments, such as those based on "principles" such as unbiasedness or minimizing mean squared error. Again, I think that the relative success of MI for missing data illustrates this point nicely (e.g., as argued in many places, including Rubin (1996)).

Serious Point #3: Nonetheless, frequentist evaluations (e.g., of bias of point estimates and coverage of interval estimates) are still highly relevant to the sage statistician because all idealizations, including Bayesian ones, are oversimplifications. As George Box said, "All models are wrong, but some are useful" (Box, 1976); also earlier, John von Neumann (1947) stated, "Truth is much too complicated to allow anything but approximations".

Two more Serious Points, an analogy, and some summarized points.

Serious Point #4: Frequentist criteria based on operating characteristics can be used to evaluate any procedure (really the same as Serious Point #1).

Serious Point #5: Therefore, we can, and moreover should, use Bayesian models to create procedures that appear to be appropriate under plausible assumptions, and use frequentist methods to evaluate these procedures in realistic situations, situations more general than those that were assumed when deriving the Bayesian answers.

Versions of these points have been made before, for example in Box (1980), Rubin (1984), and in Little (2008) and its discussions (e.g., Rubin, 2008), as well as earlier and later by various other authors. Many practicing statisticians would pretty much agree with all Serious Points, except perhaps Serious Points #2 and #5. Being a “calibrated” statistician generally means choosing procedures that have good operating characteristics over a broad range of circumstances. Being sage when confronted with a particular data set is more difficult to define, because it depends on the immediate context of the problem being confronted, and the consequences of resulting decisions, which formally can lead to decision theory (Wald, 1950). My own view is that although this framework is theoretically appealing, real decisions are made in contexts with many fuzzy and perpetually changing considerations, which disable the utility of the full formal structure of decision theory.

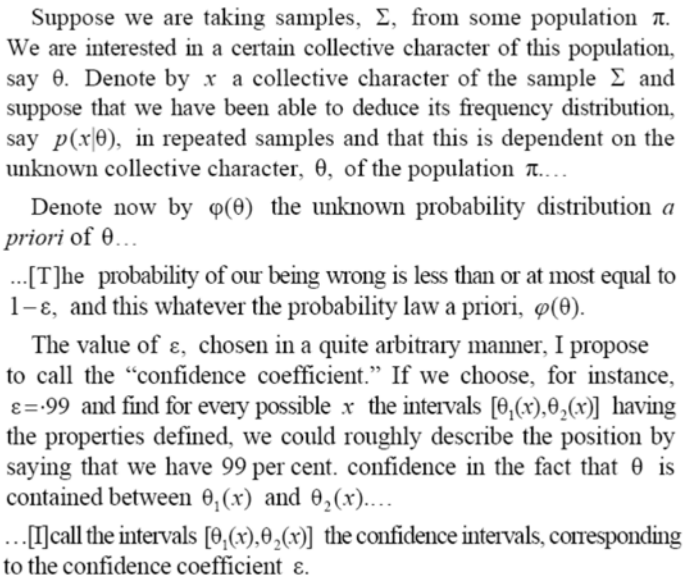
3 On the elusive goal of being calibrated and sage

Bayesians condition on what is observed, and so in principle, try to be appropriate to the data at hand. True Bayesian calibration, however, in the sense of creating interval estimates that have accurate Bayesian coverage of the true posterior distribution no matter what “Truth” generated the observed data, is essentially impossible in practice. This was illustrated to me in fairly trivial examples, first in Rubin (1983) when I was attempting to demonstrate the superiority of the Bayesian approach in the context of survey inferences, then in Rosenbaum and Rubin (1984), which documented the relevance of stopping rules on the Bayesian validity of Bayesian inferences, unless all model and prior distributions were correct, and more recently in Ferriss (Harvard PhD. Thesis, 2018), which considered the implications of re-randomization in experiments on the Bayesian validity of Bayesian inference. But despite this inability to approach the Bayesian ideal when there is the absence of knowledge of correct models, a statistician can still seek to be calibrated, in some important sense, and sage in the fiducial sense of avoiding conclusions that are contradicted by the data set actually being analyzed. I refer to this as being “conditionally calibrated” and explicate this surprisingly elusive idea here.

A personal aside relevant to this idea of being conditionally calibrated: When I was visiting the University of California, Berkeley in the 1970’s and had a visitor’s office next to, the then retired, but still intellectually vibrant and feisty, Jerzy Neyman, he clearly expressed to me his view that such conditioning for statistical inference was essentially impossible to define correctly, at least in the context of our 1970’s discussion of Fisher’s desiderata to condition on ancillary statistics when drawing inferences.

Another relevant aside: my reading is that fundamentally, both Neyman and Fisher wanted, at least in their youths, to be effectively Bayesian in that they both sought a distribution for the estimand conditional on the observed data, but took very different mathematical approaches to finding that distribution, as discussed in (Rubin, 2016). Fisher (1956) was totally forthright about this fiducial objective: “The Fiducial argument uses the observations to change the logical status of the parameter [the unknown estimand] from one in which nothing is known of it, and no probability statement can be made, to the status of a random variable having a well-defined distribution”. Values of the estimand with little support in this fiducial distribution, were those values that were stochastically contradicted by the observed data, that is, if true, they were unlikely to generate the observed data – a stochastic proof by contradiction. Despite the intuitive appeal of this approach, mathematical foundations for it have not enjoyed universal acceptance (e.g., Dempster, 1967; Martin and Liu, 2016).

Neyman was not direct as Fisher when seeking a distribution for the estimand, but consider his original definition of “confidence intervals” (Neyman, 1934), which was openly based on some Bayesian logic:



Suppose we are taking samples, Σ, from some population π. We are interested in a certain collective character of this population, say θ. Denote by x a collective character of the sample Σ and suppose that we have been able to deduce its frequency distribution, say $p(x|\theta)$, in repeated samples and that this is dependent on the unknown collective character, θ, of the population π…

Denote now by $\varphi(\theta)$ the unknown probability distribution *a priori* of θ…

...[T]he probability of our being wrong is less than or at most equal to $1-\varepsilon$, and this whatever the probability law a priori, $\varphi(\theta)$.

The value of ε, chosen in a quite arbitrary manner, I propose to call the “confidence coefficient.” If we choose, for instance, $\varepsilon=\cdot 99$ and find for every possible x the intervals $[\theta_1(x), \theta_2(x)]$ having the properties defined, we could roughly describe the position by saying that we have 99 per cent. confidence in the fact that θ is contained between $\theta_1(x)$ and $\theta_2(x)$…

…[I]call the intervals $[\theta_1(x), \theta_2(x)]$ the confidence intervals, corresponding to the confidence coefficient ε.

Figure 3.1 Neyman’s (1934, pages 589-590) definition of confidence intervals.

4 Calibration – A simulation perspective

Consider how to evaluate a proposed procedure, generically called P, which is to be applied to a data set, generically called Y, yet to be collected from a population; P is a specified function of the data Y and will be used to estimate the estimand, here a scalar quantity, Q, which describes some aspect of the population from which Y is drawn. For descriptive simplicity, suppose P is a purported 95% interval for Q; for P to be exactly calibrated means that P includes Q in exactly 95% of repeated samples. Further, suppose Y is drawn from its population using design D, which is known and fixed throughout this discussion; for concreteness, D is simple random sampling. Although D is known, at the design stage, the data set Y is not yet known. Also suppose, again for simplicity, that all “experts” interested in this problem agree on a set of K possible “Truths” for describing the unknown population to which D will be applied to obtain data set Y; call these possible Truths $T_1, T_2, \ldots, T_K$, and the values of their associated local (local to each Truth) estimands $Q_1, Q_2, \ldots, Q_K$, where $Q_k = \tilde{Q}(T_k)$, for $k = 1, \ldots, K$, for the function $\tilde{Q}$, common to all possible truths. The estimand Q is the value of the function $\tilde{Q}$ evaluated at the actual Truth.

The Q_k are here called local estimands, that is, local to the truths. As far as I can tell, Neyman never fomally considered such local estimands, but I see them as important bridges to the Bayesian perspective as well as to being a sage statistician. Only one of the possible truths is the actual truth. The inferential objective is the value of $\tilde{Q}$ for the Truth that generated the yet-to-be observed data Y.

The collection of K possible Truths can often be compactly described mathematically, so that K can be essentially infinite. One example of such Truths, and their associated local estimands, could be K Gaussian univariate populations, with unknown local means, μ_k, and with the scalar estimand Q equal to the mean of the one true population. Or the Truths could be all possible N-dimensional vectors of real numbers; this is the standard finite population set-up for survey sampling with N units and one scalar variable, as in Cochran (1963) and Kish (1965), where the estimand Q is typically the mean of the N values for the true population.

We continue by defining simple calibration using a simulation to fix ideas; this simulation will be used to define concepts throughout this manuscript, including the key concept of conditional calibration. Suppose that for each possible truth, T_k, $k = 1, \ldots, K$, with local estimand Q_k, we have drawn J data sets, labeled Y_{jk}, $j = 1, \ldots, J$, each drawn using design D. To each of these data sets, we apply procedure P to the data to create an interval estimate for Q_k, where for each k, Q_k is the same for all Y_{jk} $(j = 1, \ldots, J)$ because all such Y_{jk} arose from the same truth T_k. We then assess whether when P is applied to Y_{jk}, the resulting interval includes the local estimand Q_k. The proportion of data sets, $\{Y_{jk}, j = 1, \ldots, J\}$, for which the interval P includes Q_k is called here the local calibration (or local coverage) of the procedure P for the k^{th} Truth, notationally written C_k for $k = 1, \ldots, K$. For evaluating point estimators, rather than interval estimators, the calibration of P for Q_k could be replaced by the bias or mean squared error of the point estimate of Q_k.

This simulation is depicted in Table 4.1, where each column represents a possible truth, and the J rows represent the J data sets generated under each truth.

Table 4.1
Display of simulation (Each column represents a possible truth)

Local estimands:	$Q_1 = \tilde{Q}(T_1)$	…	$Q_k = \tilde{Q}(T_k)$	…	$Q_K = \tilde{Q}(T_K)$
	Y_{11}	…	Y_{1k}	…	Y_{1K}
	⋮		⋮		⋮
	Y_{j1}	…	Y_{jk}	…	Y_{jK}
	⋮		⋮		⋮
	Y_{J1}	…	Y_{Jk}	…	Y_{JK}
Calibration of P for Q_k:	C_1	…	C_k	…	C_K

Now we define local calibration using 95% to represent any level of coverage. A 95% interval estimate of Q, P, is called "locally (for truth T_k) conservatively calibrated" if C_k >= 95%; we could say that P is "approximately locally calibrated" (for Truth T_k) if C_k is close to 95%, but this idea was never formally defined by Neyman, although in Fisher's (1934) discussion of Neyman (1934), we can see Fisher had something like this in mind with his criticism of Neyman's formulation.

Next, following Neyman, the interval estimate P is called "confidence calibrated" across the ensemble of possible truths, $\{T_k, k = 1, \ldots, K\}$, if all C_k >= 95%, or returning to Neyman's original definition, P is then simply called a 95% confidence interval for Q. The critical point here for calibration is that all that matters to a die-hard Neymanian frequentist, when evaluating a procedure, P, for its validity, is whether the collection of C_k values for procedure P are all greater than the nominal level for P. The word "confidence" arises because when confronted with the results of Table 4.1 for procedure P and with a critic who selected one Truth from the collection of possible truths, you should be "confident" that the result of applying P to Y^* will be an interval that includes Q.

These assessments of 95% confidence calibration are well-defined no matter what the etiology of the procedure P. BUT, are they statistically apposite for evaluating P as a 95% interval estimate of the unknown Q after seeing a specific data set, call it Y^*? That is, after seeing a specific instance of Y, *now known to be* Y^*, does the 95%-attached to P necessarily reflect the judgment of a sage statistician? Maybe we should seek only procedures that are approximately calibrated for truths that plausibly could have generated the observed Y^*?

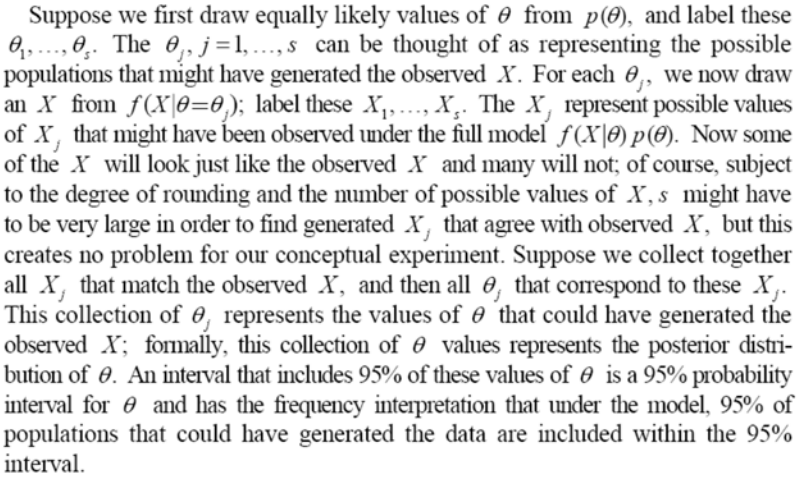
We now consider the formal Bayesian perspective because it sheds light on this concept of being sage after seeing $Y = Y^*$.

5 The Bayesian posterior distribution of Q

The Bayesian approach differs from the Neymanian approach (and from Fisher's fiducial approach) by formulating the problem so that a real conditional probability distribution for the estimand Q can be calculated, using the laws of probability theory to condition on the fact that the observed data equals Y^*- this distribution is called the posterior distribution of Q, that is, posterior after seeing $Y = Y^*$. To conduct this activity formally, Q must be a random variable, and thus Q needs to have a "starting" probability distribution, called its prior distribution, meaning prior to seeing any data; in the context of our setup, this prior distribution is a distribution over the possible local estimands, that is, a set of K probabilities (summing to one), one probability for each possible Truth. This prior distribution is essentially a set of K weights $\{W_k, k = 1, \ldots, K\}$ reflecting the prior beliefs of experts that each of the K possible local estimands is the correct one. The Neymanian frequentist has no use for such weights over the set of possible Truths, because the 95% is supposed to hold for any set of weights, and thus for each possible Truth (i.e., for all K point mass prior distributions).

Now comes the part of the argument that hints at a departure from Neyman's 1970's claim to me that conditional inference is too difficult. In the context of the simulation just described, and admitting some Bayesian or fiducial logic, when confronted with actual observed data set Y^*, attention should be focused on the parts of the simulation where the generated Y_{jk} equals Y^*; the other Y_{jk} can be ignored (at least in the context of the idealized description here, where J is essentially infinite) because, to be fully Bayesian, we want to condition on Y equaling Y^*.

In fact, let us use the simulation itself to describe the Bayesian posterior distribution of Q, i.e., the distribution of Q conditioning on the fact that $Y = Y^*$. Let M_k^* be the proportion of the J values of Y_{jk} that match Y^*, for $k = 1, \ldots, K$; that is, for truth T_k, M_k^* is the proportion of the generated data sets from truth T_k that match the actual data set Y^*. For example, if M_k^* is zero, then the a priori possible truth T_k could not be the actual truth because it could not have generated observed data Y^*. The posterior probability that the estimand Q equals Q_k, the local value of $\tilde{Q}$ for Truth T_k, is the weighted average of the proportions, M_k^*, weighted by W_k, the prior probability that T_k is the correct truth. Here, this weighted average of proportions is generally labeled π_k, where π_k for the observed data Y^* is labelled π_k^* and equals $M_k^* W_k \Big/ \sum_{k'=1}^{K} [M_{k'}^* W_{k'}]$; we could call π_k^* the estimated ability of Truth T_k to match observed data Y^*. This description of the posterior distribution of Q using simulation is from Rubin (1984); see Figure 5.1.



Suppose we first draw equally likely values of θ from $p(\theta)$, and label these $\theta_1, \ldots, \theta_s$. The $\theta_j, j = 1, \ldots, s$ can be thought of as representing the possible populations that might have generated the observed X. For each θ_j, we now draw an X from $f(X|\theta = \theta_j)$; label these $X_1, \ldots, X_s$. The X_j represent possible values of X_j that might have been observed under the full model $f(X|\theta)\,p(\theta)$. Now some of the X will look just like the observed X and many will not; of course, subject to the degree of rounding and the number of possible values of X, s might have to be very large in order to find generated X_j that agree with observed X, but this creates no problem for our conceptual experiment. Suppose we collect together all X_j that match the observed X, and then all θ_j that correspond to these X_j. This collection of θ_j represents the values of θ that could have generated the observed X; formally, this collection of θ values represents the posterior distribution of θ. An interval that includes 95% of these values of θ is a 95% probability interval for θ and has the frequency interpretation that under the model, 95% of populations that could have generated the data are included within the 95% interval.

Figure 5.1 Description of posterior distribution from Rubin (1984).

There are objections to this approach. First, where do the prior weights W_k come from and who are the experts providing these weights? Perhaps we should find some way to avoid using these potentially overly subjective prior weights? Second, perhaps the requirement for exact equality between a generated data set Y_{jk} and the observed data set Y^* should be relaxed in some way so that a generated Y_{jk} does not have to equal Y^* exactly but only "look like" it came from the same distribution as did Y^*, and so match Y^* in some way?

More on this second point first, which is clearly important when trying to conduct an actual simulation like this idealized one with a finite budget. The approximate equality between generated data Y and observed data Y^* can be achieved in situations with low-dimensional sufficient statistics, because only those statistics have to match. But this idea of generated data being "close to" observed data Y^* is the basis of all work using this description of the posterior distribution to conduct "ABC" – Approximate Bayesian Computation, apparently first described in the paragraph in Figure 5.1 (https://en.wikipedia.org/wiki/Approximate_Bayesian_computation, Tavare, Balding, Griffiths and Donnelly, 1997). We simply assume at this point that we have chosen some such metric to define the function M_k, and use it to define the ability of Truth T_k to generate data sets that match the observed data, Y^*.

6 The conditionally calibrated statistician's evaluation of procedure P

With the basic definitions of Sections 4 and 5 for C_k (the local calibration of P for Q_k under T_k) and the M_k^* (the matching ability of T_k for Y^*) now established, we are prepared to consider the concept of Conditional Calibration (CC) and to make the connection to being a sage statistician after having observed Y^*.

Conditional Calibration starts at the same place as Neymanian unconditional calibration, but is sensitive to the Bayesian and Fiducial arguments by discounting results from possible truths whose drawn data sets Y_{jk} are not close to Y^* as assessed by their match rates M_k^*. The point of doing this is: Why should we care about calibration for a priori possible Truths that could not have generated the observed Y^*, i.e., truths that are a posteriori implausible? But this CC perspective does not go to the full Bayesian extreme, which, first, ignores all aspects of the simulation except the Truths that generated data sets exactly matching the observed data set, and second, explicitly uses the often weakly justified prior distribution, W_k, to weight the local matching rates.

Hence, we are left summarizing our simulation, which evaluates a procedure P for inference about the estimand Q from observed data Y^*, using a two dimensional criterion: The local calibration of P for Q_k under truth T_k, that is C_k; and the local match rate for Truth T_k, the proportion of generated Y_{jk} under truth T_k, that are accepted as matching Y^*, M_k^* = the fraction of Y_{jk} that are considered equal to Y^*. Those possible truths with M_k^* near one are clearly more relevant to the situation with observed data Y^* than those truths with values of M_k^* near zero.

Thus procedure P applied to each possible truth with observed data Y^* can be displayed as K points in a two-dimensional graph where the horizontal axis is the average match to Y^* of the data sets generated by truth T_k, M_k^*, and the vertical axis is the local calibration of P for the data sets generated by Truth T_k, C_k. We call this the "conditional calibration plot" and it is illustrated and discussed in the Section 7. Of major relevance to drawing sage inferences for possible truths from observed data Y^*, many different procedures can be displayed on the same conditional calibration plot for a fixed data set Y^* and a fixed set of possible truths.

7 The conditional calibration plot and its use for sagely selecting procedures to use with observed data Y^*

The conditionally calibrated (CC) statistician faced with estimating Q using procedure P from data set Y^* cares about being approximately calibrated, i.e., close C_k to 95% especially for Truths with large values of M_k^*, indicating that such Truths could have plausibly generated Y^*. In other words, when comparing procedures for estimating Q from Y^*, the sage statistician, in addition to conservative unconditional calibration (i.e., confidence coverage), especially cares about accurate calibration for Truths that are plausible, and therefore implicitly ignores the calibration of procedures for Truths that are implausible given Y^*.

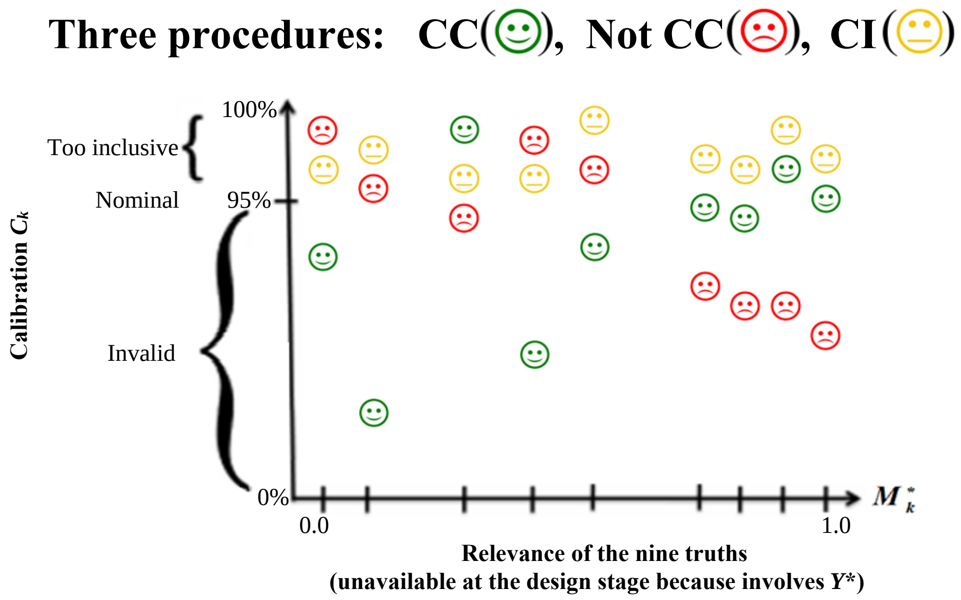


Figure 7.1 C_k versus M_k^* Plots for a fixed data set, with $K = 9$ Truths (columns).

Figure 7.1 presents hypothetical simulation results with a fixed data set Y^* and a fixed set of nine possible Truths (with nine associated local match rates to Y^*) for three procedures, indicated by faces. The vertical axis is not linear in C_k but expanded for values of C_k closer to unity, which is where our interest is focused. One procedure is labeled "Smile" because it is approximately calibrated (C_k close to 95%) for possible Truths that could have generated Y^* (M_k^* close to 1), even though poorly calibrated (C_k well below 95%) for a priori possible Truths that are implausible given the observed Y^* (M_k^* much lower than 1). A second procedure is labeled "Frown" because it is not CC, being invalid (meaning its local calibration is substantially less than 95%), including for truths that are plausible given Y^*. The third procedure is labeled as “Neutral [CI]” because, although it is a valid confidence interval in Neyman’s sense of having its minimum local calibration at least 95%, it is not approximately calibrated for Truths that are plausible given the observed data set, Y^*. This procedure could, for me, be described by a mild frown, but maybe not for Neyman, based on our 1970’s conversation.

That is, to repeat, Neymanian (conservative = confidence) calibration for each procedure formally just cares about the procedures’ minimum C_k across the entire ensemble of a priori possible truths. Also, the rigid Bayesian just cares about the weighted average of the M_k^* across the possible truths, weighted by the prior possibly unreliable distribution for the truths, W_k. The sage CC statistician cares about approximate local calibration of procedures for those Truths that are plausible; if a confidence-valid 95% procedure P displays C_k values substantially bigger than 95% for plausible Truths, this suggests that there exist better CC procedures for this situation with data set Y^*; that is, calibrated procedures that are more efficient and so result in shorter intervals. Notice for example, that the confidence-valid procedure in Figure 7.1 (Neutral face) has worse CC than Smile, and thus although a plausible competitor to Smile at the design stage should be seen as inferior to Smile after seeing data Y^* because it is too conservative for some of the relevant Truths.

8 Implementing this idea in practice

To implement this idea in practice would require work, certainly more intellectual effort than is currently expended in many statistical investigations. The implementation would begin at the same place as is standard in current carefully constructed studies. We would begin by considering a set of procedures, each of which is usually conservatively calibrated in the traditional sense, for the problem at hand. Then we would collect opinions from experts about the generally plausible Truths in the specific situation we are facing; this step is executed in some current problems, although typically informally.

If possible, then we should gather some information for W_k, the prior weights on the possible truths; these could be useful for later consideration of the construction of the matching averages M_k (no asterisk yet, because the data, Y^*, are not yet observed). We should obtain agreement on how to define M_k and whether to use the prior weights W_k. This is the ABC task. Finally, agreement is needed on how to use the CC plot to compare the various procedures being considered.

All of this effort should be conducted before the actual data set Y^* is observed. For this reason, alone, the implementation of this idea is more intellectually demanding than standard practice, but it is a component of being a sage statistician.

Acknowledgements

This is the written version of DB Rubin’s Waksberg Award address delivered 8 November 2018 in Ottawa, Canada. The author was a friend of Joseph Waksberg and was always impressed with his wise application of statistics, and hopes that this contribution continues that tradition. Other versions of this talk have been given over the past five years, most recently as the SN Roy Invited Lecture in Kolkata India 27 Decmber 2018. The author acknowledges very helpful comments from Roderick Little, Tommy Wright, as well as from Wesley Yung and other members of the *Survey Methodology* editorial board, and recently from Hal Stern and Yannis Yatracos.

References

Box, G.E.P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356), 791-799.

Box, G.E.P. (1980). Sampling and Bayes' inference in scientific modelling and robustness. *Journal of the Royal Statistical Society, Series A*, 143(4), 383-430.

Cochran, W.G. (1963). *Sampling Techniques, 2nd Edition*. New York: John Wiley & Sons, Inc.

De Finetti, B. (1972). *Probability, Induction, and Statistics*. New York: John Wiley & Sons, Inc.

Dempster, A.P. (1967). Upper and lower probabilities induced by a multivalued mapping. *The Annals of Mathematical Statistics*, 38(2), 325-339.

Ferris, T. (2018). *Topics in Casual Inference and the Law*. Senior Data Scientist, Google.

Fisher, R.A. (1934). Contribution to a discussion of J. Neyman's paper on the two different aspects of the representative method. *Journal of the Royal Statistical Society*, 97, 614-619.

Fisher, R.A. (1956). *Statistical Methods and Scientific Inference*. Edinburgh: Oliver and Boyd.

Kish, L. (1965). *Survey Sampling*. New York: John Wiley & Sons, Inc.

Lehmann, E.L. (1959). *Testing Statistical Hypotheses*. New York: John Wiley & Sons, Inc.

Lindley, D.V. (1971). *Bayesian Statistics: A Review,* SIAM.

Little, R.J. (2008). Weighting and prediction in sample surveys. *Calcutta Statistical Association Bulletin*, 60, 3-4, 147-167.

Martin, R., and Liu, C. (2016). *Inferential Models*. New York: Chapman and Hall/CRC.

Neyman, J. (1923). On the application of probability theory to agricultural experiments. Essay on principles. Translated into English in *Statistical Science*, 1990, 5(4), 463-472.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4), 558-625.

Rosenbaum, P.R., and Rubin, D.B. (1984). Sensitivity of Bayes inference with data dependent stopping rules. *The American Statistician*, 38, 106-109.

Rubin, D.B. (1978). Multiple imputations in sample surveys-A phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section*, American Statistical Association, 20-34.

Rubin, D.B. (1983). A case study of the robustness of Bayesian methods of inference: Estimating the total in a finite population using transformations to normality. *Scientific Inference, Data Analysis and Robustness*. New York: Academic Press, Inc., 213-244.

Rubin, D.B. (1984). Bayesianly justifiable and relevant frequency calculations for the applied statistician. *Annals of Statistics*, 12, 1151-1172.

Rubin, D.B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, 91, 434, 473-489.

Rubin, D.B. (2008). Discussion of "Weighting and prediction in sample surveys" by R.J. Little. *Calcutta Statistical Association Bulletin*, 60, 185-190.

Rubin, D.B. (2016). Fisher, Neyman, and Bayes at FDA. *Journal of Biopharmaceutical Sciences*, 26, 1020-1024.

Savage, L.J. (1954). *The Foundations of Statistics*. Wiley Publications in Statistics.

Tavare, S., Balding, D.J., Griffiths, R.C. and Donnelly, P. (1997). Inferring coalescence times from DNA sequence data. *Genetics*, 145(2), 505-518.

Von Neumann, J. (1947). The mathematician. In *Works of the Mind*, (Ed., R.B. Haywood), University of Chicago Press, 180-196.

Wald, A. (1950). *Statistical Decision Functions*. New York: John Wiley & Sons, Inc.