

Discussion of the paper “Is Bayes posterior just quick and dirty confidence”

Tong Zhang
Statistics Department
Rutgers University, NJ 08816

1 Confidence Region Estimation

The author has written an interesting article on the relationship of confidence distribution and Bayesian posterior distribution. Confidence distribution has its origin from Fisher’s fiducial distribution, and in this discussion, we refer to it simply as the “confidence distribution approach”. It allows frequentists to assign confidence intervals (or more generally confidence regions) to the outcome of estimation procedures.

The idea can be simply described as follows. Consider a statistical model with a family of distributions $p_\theta(y)$, where y is the observation and θ is the model parameter. We assume that the observed y is generated according to a true parameter θ_* which is unknown to the statistician. If we can find a real-valued quantity $U(y; \theta)$ that depends on θ and y such that for all θ , when y is generated from $p_\theta(y)$, $U(y; \theta)$ is uniformly distributed in $(0, 1)$, then we can estimate the confidence interval of θ given an observation y as the set $I_{\alpha, \beta}(y) = \{\theta : U(y; \theta) \in (\alpha, \beta)\}$ for some $0 \leq \alpha \leq \beta \leq 1$. An interpretation of this confidence region is that no matter what is the true underlying θ_* that generates y , the region $I_{\alpha, \beta}(y)$ contains the true parameter θ_* with probability $\beta - \alpha$ (when y is generated according to θ_*).

Indeed, the above interpretation is a very natural definition of confidence region in the frequentist setting. It does not assume that θ_* is generated according to any prior, and the interpretation holds universally true for all possible θ_* in the model. This interpretation can be compared to confidence region from the Bayesian posterior calculation that assumes that θ_* is generated according to a specific prior which has to be known to the statistician. If the statistician chooses the wrong prior, then the confidence region calculated from the Bayesian approach will be incorrect in that it may not contain the true parameter θ_* with the correct probability.

The paper takes this interpretation of confidence region, and goes on to provide several examples showing that the Bayesian approach does not lead to correct confidence estimates for all θ_* . The author then argued that the confidence distribution approach is the more “correct” method for obtaining confidence intervals and the Bayesian approach is just a quick and dirty approximation.

One question that needs to be addressed in the confidence distribution approach is how to construct a statistics $U(y_0; \theta)$ with the desired property. The author considered the quantity $U(y_0; \theta) = \int_{y \leq y_0} p_\theta(y) dy$, which is well-defined if the observation y is a real valued number. This corresponds to the proposal in Fisher’s fiducial distribution. The idea of fiducial distribution received a number of discussions throughout the years, and is known to be adequate for unconstrained location families (for which the fiducial confidence distribution matches the Bayesian confidence

distribution using a flat prior). However, the general concept is controversial, and largely regarded as a major blunder by Fisher.

In this discussion article we will explain why the idea of confidence distribution with

$$U(y_0; \theta) = \int_{y \leq y_0} p_\theta(y) dy$$

has not received more attention for general statistical estimation problems, although it does give confidence region estimates that fit the frequentist intuition.

2 Sub-optimality

The purpose of confidence distribution is to provide a confidence region that is consistent with the frequentist definition. However, one flaw of this approach is that the result it produces may not be optimal. While this issue was pointed out in the article, it was not explicitly discussed. In my opinion, this is the main reason why the idea of confidence distribution hasn't become more popular in statistics. Therefore this section provides a more detailed discussion on this issue.

To understand this point, we shall first consider a simple illustration. Let $U(y, \theta)$ be a uniform random variable in $(0, 1)$ that is independent of y and θ . By definition, given any θ_* , the confidence region $I_{\alpha, \beta}(y) = \{\theta : U(y; \theta) \in (\alpha, \beta)\}$ contains θ_* with probability $\beta - \alpha$. Since this applies to the parameter that generates y , the confidence region obtained this way is consistent with the frequentist intuition of what a confidence region should mean. However, this estimate is not useful statistically because the method just randomly guesses either the entire domain of θ when $U \in (\alpha, \beta)$ or the empty region otherwise; the decision does not even depend on y .

While the above example is extreme, it does show that a confidence region merely consistent with the frequentist semantics is not necessarily a useful estimate. Statistically, this is because the confidence region obtained is sub-optimal. In fact, this claim also applies to the confidence distribution approach this article considers. More specifically, for nonlinear problems that this paper focused on, the method can produce confidence regions that are quite sub-optimal. By "optimal" (or even "good"), we mean that the confidence region a method produces should be small by some measure. In particular, if another method provides confidence regions that also fit in the frequentist semantics but is no larger on average for all θ and smaller for some θ , then it can be regarded as a better method. This corresponds to the notion of admissibility in decision theory.

Consider the following simple nonlinear location estimation model: y is generated either from $N(0, \sigma_0^2)$ when $\theta = 0$, or from $N(1, \sigma_1^2)$ when $\theta = 1$. There are only two possible positions $\theta = 0$ or $\theta = 1$ for the unknown location parameter θ , and we assume that the variance parameters σ_0^2 and σ_1^2 are known quantities that are not necessarily equal. Note that the restriction of θ to two positions is only for simplicity which is not critical for our illustration — we can extend the example to allow all locations in R .

For this example, the confidence distribution approach gives the following $U(y_0, \theta)$:

$$U(y_0, \theta) = \begin{cases} \Phi(y_0/\sigma_0) & \theta = 0 \\ \Phi((y_0 - 1)/\sigma_1) & \theta = 1, \end{cases}$$

where $\Phi(z)$ denotes the cdf of the standard Gaussian $N(0, 1)$.

Let's consider the confidence region $I_{\delta, 1-\delta}(y)$ for some $\delta \in (0, 0.25)$, which we simplify as $I(y)$. By definition, the estimated confidence region $I(y)$ contains the position $\theta = 0$ if and only if $y \in \Omega_0$

with $\Omega_0 = (\sigma_0\Phi^{-1}(\delta), -\sigma_0\Phi^{-1}(\delta))$, and $I(y)$ contains the position $\theta = 1$ if and only if $y \in \Omega_1$ with $\Omega_1 = (1 + \sigma_1\Phi^{-1}(\delta), 1 - \sigma_1\Phi^{-1}(\delta))$. For convenience, we also define

$$\mu_0 = P(y \in \Omega_1 | \theta = 0) = \int_{1+\sigma_1\Phi^{-1}(\delta)}^{1-\sigma_1\Phi^{-1}(\delta)} \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left(-\frac{y^2}{2\sigma_0^2}\right) dy.$$

In order to show that the confidence distribution approach is sub-optimal, we can for simplicity consider the case $\sigma_0 \gg 1$ and $\sigma_1 \ll 1$, so that $1 - \sigma_1\Phi^{-1}(\delta) < -\sigma_0\Phi^{-1}(\delta)$ and $\mu_0 < 2\delta$. The first condition implies that $\Omega_1 \subset \Omega_0$. Therefore when the parameter $\theta = 1$, with probability $1 - P(y \in \Omega_1 | \theta = 1) = 1 - 2\delta$ over $y \sim N(1, \sigma_1^2)$, we have $y \in \Omega_1$ and thus $|I(y)| = 2$ (that is, $I(y)$ contains both $\theta = 0$ and $\theta = 1$). Therefore we have (note that we have assumed that $\delta < 0.25$)

$$E_{y|\theta=1}|I(y)| > 2(1 - 2\delta) > 1. \quad (1)$$

Moreover, we have

$$E_{y|\theta=0}|I(y)| = P(y \in \Omega_0 | \theta = 0) + P(y \in \Omega_1 | \theta = 0) = 1 - 2\delta + \mu_0.$$

Now we would like to construct a better confidence region estimator by using the condition (which we made earlier) that $P(y \in \Omega_1 | \theta = 0) = \mu_0 < 2\delta$. Therefore we can pick Ω'_0 such that $\Omega'_0 \cap \Omega_1 = \emptyset$ and $P(y \in \Omega'_0 | \theta = 0) = 1 - 2\delta$. This means that we can choose the following confidence region estimate $I'(y)$: $I'(y)$ contains the position $\theta = 0$ if and only if $y \in \Omega'_0$ and $I'(y)$ contains the position $\theta = 1$ if and only if $y \in \Omega_1$. This estimate obeys the frequentists definition because $P(\theta \in I'(y) | \theta) = 1 - 2\delta$ both when $\theta = 0$ and $\theta = 1$. Moreover, we have

$$E_{y|\theta=0}|I'(y)| = 1 - 2\delta + \mu_0, \quad E_{y|\theta=1}|I'(y)| \leq 1.$$

The second inequality is due to the fact that $|I'(y)| \leq 1$ for all y because $\Omega'_0 \cap \Omega_1 = \emptyset$. In comparison to (1), we know that when $\theta = 1$, the confidence distribution approach gives a confidence region $I(y)$ with a larger average size. This means that for this simple problem, the confidence distribution approach gives a sub-optimal estimate of confidence region $I(y)$ that is dominated by a better method $I'(y)$. The difference can be significant when $\delta \approx 0$.

3 Conclusion

The confidence distribution approach is a rather general method to obtain confidence regions for parameter estimation problems consistent with the frequentist semantics. The method can also be easily generalized to the multivariate situation where y is a vector instead of a real number. Nevertheless, the confidence region it estimates can be rather sub-optimal in the sense that the region obtained by this method can be significantly larger than what can be done with more sophisticated methods. Although we have only illustrated this phenomenon with a relatively simple example, the conclusion holds more generally.

At the root of this sub-optimality, we note that whether a model parameter θ_0 belongs to the confidence region obtained by the confidence distribution approach only depends on the distribution $p(y|\theta = \theta_0)$ at the parameter θ_0 itself, without considering the alternative models at $\theta \neq \theta_0$. This unnatural behavior is what causes its sub-optimality for general nonlinear models. For example, in order to achieve good performance for the simple two-position location estimation example given

in the previous section, the confidence region estimate $I'(y)$ at $\theta = 0$ has to be modified in order to take advantage of the alternative model $\theta = 1$ (so that $\Omega'_0 \cap \Omega_1 = \emptyset$). Such adaptation does not occur in the confidence distribution approach. As noted by the author during the discussion of the bounded parameter example, the confidence distribution estimate does not change when we restrict the model space, and this phenomenon is rather odd. The author dismissed this problem as a secondary issue because it does not change the semantics of confidence region in the frequentist interpretation. However, if we are interested in achieving (near) optimality for the estimated confidence region, then this issue becomes a more serious concern because it means that this simple method ignores a significant amount of available information that could have been used in more complicated algorithms. In conclusion, while the confidence distribution approach is simple to apply, the simplicity is achieved by ignoring some useful information. Therefore we have to keep the limitations of this method in mind whenever it is applied to complex statistical models.