THE POSTERIOR LIKELIHOOD

Hee-Choon Shin
National Opinion Research Center, 55 East Monroe St., Suite 4800, Chicago, IL 60603

KEY WORDS: Bayes, Fisher, Inverse Probability, Likelihood, Posterior, Prior, Meta-Analysis

Introduction

The Bayesian approach began to re-gain popularity in late 1980's due to the development in computing technology. Another possible reason for this re-emergence is the unsatisfactory status of the Fisherian likelihood approach to estimation. [1]

The main point of Bayesian approach is in attempting to utilize the existing or prior knowledge on the population parameters. Meanwhile, the Fisherian likelihood approach ignores the existing knowledge, and infers population parameters solely based on the given sample of elements. Naturally the soundness of Fisherian inference completely depends on the soundness of the sample.

In this paper, I attempt to utilize existing knowledge on parameters in applying Fisherian likelihood approach to statistical estimation.

In the following, I will re-introduce the essence of Bayes' original essay to emphasize his idea, and will try to point out the problems of Bayesian applications. Then I will introduce the posterior likelihood after going over Fisherian method. Results from simulation will be discussed to support the soundness of my proposal.

Bayes' Method

To the current author, almost all the current Bayesian applications are not really "Bayesian" in strict sense. The best understanding of the Bayes' method can be garnered by examining Bayes' problem statement in his original Essay: "*Given* the number of times in which an unknown event has happened and failed: *Required* the chance that the probability of its happening in a fingle trial lies fomewhere between any two degrees of probability that can be named." (Bayes, 1763: 376).[2]

The fundamental and necessary condition for Bayes' approach is the existence of an unknown event. Bayes defines an unknown event as follows: "…any event concerning the probability of which nothing at all is known antecedently to any trials made or obferved concerning it. And fuch an event I fhall call an unknown event." (Bayes, 1763: 393-394). How many of contemporary Bayesian applications deal with unknown events? The second essential component of his approach is the "guessing" part. He was not interested in a point estimate, but in the chance between any two degrees of probability. By definition, the probability or density of a point is zero. The third point is that he was interested in the probability of its happening in a single trial.

Bayes' Rule

Bayes proposed a rule to find the probability or chance after extensively discussing the 10 Postulates. "If nothing is known concerning an event but that it has happened p times and failed q in $p+q$ or n trials, and from hence I guess that the probability of its happening in a single trial lies somewhere between any two degrees of probability as X and x, the chance that I am in the right in my guess is $\overline{n+1}\times E\times^{d}$ into the difference between the series

$$\frac{X^{p+1}}{p+1}-q\frac{X^{p+2}}{p+2}+q\frac{q-1}{2}\times\frac{X^{p+3}}{p+3}-\&cc.$$

and the series

$$\frac{x^{p+1}}{p+1}-q\frac{x^{p+2}}{p+2}+q\frac{q-1}{2}\times\frac{x^{p+3}}{p+3}-\&cc.$$

E being the coefficient of $a^{p}b^{q}$ when $\overline{a+b}\big|^{n}$ is expanded." (Bayes, 1763: 399).

---

[1] The latter reason is possible but I do not think that is the main reason for the current popularity of the Bayesian approach.

[2] This is the exact quotation from his Essay. For example, the word, "fingle", should be spelled as "single" in contemporary English.

In recent notation, the probability, $\Pr(x \le \theta \le X \setminus M = p)$, is

$$\frac{\int_x^X \binom{n}{p} \theta^p (1-\theta)^q \, d\theta}{\int_0^1 \binom{n}{p} \theta^p (1-\theta)^q \, d\theta}.$$

Bayes' Theorem

Frequently, for example, the following equation is referred as the Bayes' Theorem:

$$P(A/B) = \frac{P(B/A)P(A)}{P(B)}.$$

This is far from the truth. As Fisher (1922: 324) indicated, "the *result*, the *datum*, and the *postulate* implied by the scholium, have been somewhat loosely spoken of as Bayes' Theorem." More than often in many of modern textbooks on mathematical statistics, the Bayes' formula is erroneously referred as Bayes' Theorem. We should note that the Bayes' formula listed in the above and in any statistics textbooks (e.g., Brownlee, 1965: 18-20) is a simple re-statement on conditional probability. And applying simple conditional probabilities and conditional expectations is not Bayesian method.

The Bayesian Method: An Example

Let us look at a simple but specific example to understand the Bayes' method. Let us assume that an unknown event M happened once. What is the probability of it's happening in the second observation? Specifically, what is the probability for more than even chance that it would happen on the second observation? The probability is .75, or an odds of three to one. (See Table 1.) What if the event M happened one million times? Table 1 indicates that the probability would be still less than 1 on the 1 million first observation. Consider that the event M happened once, and we guess the chance that its happening is between 99.99% and 100.00% at the second observation. The probability is .000199990. What if the event M happened 1,000 times and we guess the chance that its happening is between 99.99% and 100.00% at the 1,001th observation. The probability is .095257589. The essence of the Bayesian inference is in that non-existence of contrary result of an unknown phenomenon does not exclude the possibility of contrary result.

Fallacy of Super Population Logic

It can be argued that the uniqueness of Bayes' method is the consideration for the prior distribution of the population parameters. They claim that the posterior distribution is proportional to the product of the likelihood and the prior distribution of the population parameters. To the current author, this is the fatal mistake of the Bayes' followers since 1,763. How can there be a distribution of the truth or the population parameters? There is one and only one truth. God does not generate the universe again and again to satisfy the need of some applied mathematicians. Bayes' "uniform" prior was on the uncertainty about his guess or estimate. In other words, Bayes' prior reflects his lack of knowledge on the truth. I think the following Bayes' phrase support my argument: "…I may juftly reason concerning it as if its probability had been at firft unfixed,…" (Bayes, 1763: 393)

Fisher's Approach

As well known, Fisher (1922; 1930; 1933; 1934) is one of most vociferous critics on inverse probability or on Bayesian approach. Fisher (1930: 528) described the Bayesian approach as follows: "I know only one case in mathematics of a doctrine which has been accepted and developed by the most eminent men of their time, and is now perhaps accepted by men now living, which at the same time has appeared to a succession of sound writers to be fundamentally false and devoid of foundation" Fisher went on to express his opinion on the inverse type of argument as follows: "We shall first cloak its fallacy under an hypothesis, and then examine it as an undisguised assumption."

Method of Maximum Likelihood

As an alternative to Bayes' inverse probability, Fisher proposed a concept of likelihood and suggested the method of maximum likelihood. As a function of the parameters (θ's) maximized, the likelihood is not a probability and does not obey the laws of probability.

It involves no differential element $d\theta_1 d\theta_2 d\theta_3 \ldots$ But it is, just as much as a probability, a numerical measure of rational belief. In any distribution involving unknown parameters $\theta_1, \theta_2, \theta_3 \ldots$, let the chance of an observation falling in the range dx be

$$f(x, \theta_1, \theta_2, \ldots) dx .$$

Then, the chance that in a sample of n, n_1 falls in the range dx_1, n_2 falls in the range dx_2, and so on, is

$$\frac{n!}{\prod(n_i!)} \prod \{ f(x_i, \theta_1, \theta_2,) dx_i \}^{n_i} .$$

The method of maximum likelihood consists simply in choosing that set of values for the parameters which makes this quantity a maximum.

For an independent sample of size n from a binomial population with parameter θ and x successes in the sample, the likelihood function can be expressed:

$$L(x \mid n, \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x} .$$

As Fisher (1922) described in his seminal paper eighty years ago, the maximum likelihood estimate, $\hat{\theta}$, can be obtained by equating the first derivative of log likelihood to 0:

$$\frac{\partial}{\partial \theta} \log L(x \mid n, \theta) = 0 .$$

The variance of $\hat{\theta}$ is obtained by

$$\frac{\partial^2}{\partial \theta^2} \log L(x \mid n, \theta) = -\frac{1}{\sigma^2_{\hat{\theta}}} .$$

The estimate is

$$\hat{\theta} = \frac{x}{n} .$$

The variance of the estimate is

$$\sigma^2_{\hat{\theta}} = \frac{\hat{\theta}(1-\hat{\theta})}{n} .$$

Using Prior Information on Parameters

Again let us consider the likelihood for an independent sample of size n from a binomial population with parameter θ and x successes in the sample:

$$L = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

Now we assume that we have prior knowledge on the parameter. What would be the likelihood function with the known θ_p for an independent sample of size n_p ? The likelihood should be

$$L_p = \binom{n_p}{x_p} \theta_p^{\,x_p} (1-\theta_p)^{n-x_p} .$$

Now consider the following estimate utilizing prior knowledge on the parameter:

$$\hat{\theta}_o = \lambda \hat{\theta} + (1-\lambda) \hat{\theta}_p .$$

The λ ($0 \le \lambda \le 1$) could be viewed as an index of confidence for the current estimate or sample, and conversely ($1-\lambda$) as an index of confidence for prior or existing estimate.
The λ could be subjectively determined by a statistician considering the method of inference on the existing knowledge, but a natural choice of λ is the ratio of the current sample size to the sum of the current, n, and the previous, n_p.
That is,

$$\lambda = \frac{n}{n+n_p} .$$

Given λ and n, the effective n_p, or n_p^E is

$$\frac{1-\lambda}{\lambda} n .$$

The n_p^E is the relative size of the sample on which the existing estimate was based, as compared to the current sample size.

Posterior Likelihood

We propose a new form of likelihood as follows:

$$L_o = \binom{n+n_p}{x_o} \theta_o^{\,x_o} (1-\theta_o)^{n+n_p-x_o},$$

where

$$x_o = (n+n_p)\left\{ \lambda \frac{x}{n} + (1-\lambda)\hat{\theta}_p \right\}.$$

We would like to call this as the posterior likelihood since it utilizes the existing or prior knowledge on the parameters.

The estimate is

$$\hat{\theta}_o = \frac{x_o}{n+n_p}.$$

Note that the traditional Fisherian estimate is

$$\hat{\theta} = \frac{x_o}{(n+n_p)\lambda} - \frac{1-\lambda}{\lambda}\hat{\theta}_p.$$

The variance of the estimate is

$$\mathrm{var}\left(\hat{\theta}_o\right) = \frac{\hat{\theta}_o(1-\hat{\theta}_o)}{n+n_p}.$$

Simulation

To evaluate the proposed method, we conducted simulation studies using a known population. The assumed population was the respondents (216,424) in the March Current Population Survey (CPS), 2003. Note that, for our purpose, we regarded the CPS sample as a given population. We were interested in the proportion of married individuals in U.S.A. The assumed true proportion (θ) of the married was 42.09%.

Two sets of samples were generated for each replication with replacement: one sample of variable size for prior or existing estimate and the other sample of 1,000 for the current estimate. We repeated the above operation 1,000 times.

Table 2 shows the effectiveness of the proposed method. Results of comparisons in terms of bias, variance, and the mean squared error (MSE) between the Fisherian MLEs and the posterior estimate are shown. From the first row, for example, we see that, with a 50% of confidence (i.e., $1-\lambda = .5$ or n_1 =1,000) for the prior estimate and 1,000 sample individuals for the current estimate, the proposed estimate ($\hat{\theta}_o$) is better than the traditional Fisherian estimate ($\hat{\theta}$) in 65.30% of replications. The MSEs of the posterior estimates are smaller than those of the Fisherian MLEs in 86.40% of replications. The variance of the posterior estimate is smaller than the one for the Fisherian MLE in all replications as expected. Even if we give a lesser degree of confidence to the existing knowledge, we are better off. With 10% of confidence for the existing knowledge, the bias is smaller in 56.60% of the replications, and the MSE is smaller in 67.50% of the replications. Again the variance of the posterior estimate is smaller in all the replications.

In Table 2, all the estimates for the prior knowledge are based on random samples generated from the population, depending on the size of confidence index. Let us look at a situation in which an arbitrary guess is used for the prior estimate. For the sake of simulation, we assume that the proportion of the married is 60%, which is very different from the true proportion, 42.09%. Table 3 shows the result. With an equal credence (50%) to both of the existing and the current estimates, the Fisherian MLE is better than the posterior estimate. When we give relatively small 10% of confidence to the existing “wrong” estimate, the biases of the posterior estimates are smaller than those of the Fisherian MLEs only in 32.40% of the replications. The MSEs are smaller only in 33.40% of the replications, even though the variance is smaller in all the replications.

Conclusion

The contribution of Bayes’ approach in statistical estimation theory is the utilization of the existing knowledge on the population parameters. I argue that the concept of distribution of population parameters or prior distribution is logically flawed one since there is one and only one truth.

Fisherian approach ignores the existing knowledge, which is contradicted with our experiences. All the current human action is based on earlier experiences. Indeed all the existing scientific and non-scientific knowledge is prior knowledge for the current action, small or large.

I propose that the existing or prior knowledge should be reflected within the likelihood function maximized. I call this likelihood as posterior likelihood. If we ignore the prior component in the posterior likelihood, the posterior likelihood is the same as the Fisherian likelihood.

Specifically we found that as long as the existing knowledge was based on a random sample, the posterior likelihood tended to generate an estimate with less bias and increased efficiency. We also saw that estimation with incorrect prior knowledge based on arbitrary or subjective conjecture generated a larger bias and decreased efficiency.

References

Bayes, T. 1763. "An Essay towards Solving a Problem in the Doctrine of Chances." *Philosophical Transactions*, Vol. 53: 370-418.

Brownlee, K.A. 1965. *Statistical Theory and Methodology in Science and Engineering, 2nd Ed.* New York: Wiley.

Fisher, R.A. 1922. "On the Mathematical Foundations of Theoretical Statistics." *Philosophical Transactions of the Royal Society of London, Series A, Containing Papers of a Mathematical or Physical Character*, Vol. 222: 309-368.

Fisher, R.A. 1930. "Inverse Probability." *Proceedings of the Cambridge Philosophical Society*, Vol. 26: 528-535.

Fisher, R.A. 1933. "The Concepts of Inverse Probability and Fiducial Probability Referring to Unknown Parameters." *Proceedings of the Royal Society of London, Series A, Containing Papers of a Mathematical and Physical Character*, Vol. 139: 343-348.

Fisher, R.A. 1934. "Two New Properties of Mathematical Likelihood." *Proceedings of the Royal Society of London, Series A, Containing Papers of a Mathematical and Physical Character*, Vol. 144: 285-307.

Table 1. The Bayesian Method: An Example

		Chance of your guess being right
p	q	$\Pr(.5000 \le \theta \le 1 \setminus M = p)$
1	0	.7500
2	0	.8750
3	0	.9375
1,000	0	$1-(1/2)^{1,001}$
1 mil.	0	Less than 1
		$\Pr(.9999 \le \theta \le 1 \setminus M = p)$
1	0	.000199990
2	0	.000299970
3	0	.000399940
1,000	0	.095257589
1 mil.	0	Less than 1

Table 2. Simulation: Relationship between the Fisherian MLE and the Posterior Estimate with the Prior Estimate from a Random Sample.

Index of Confidence for Prior Info $(1-\lambda)$	$(\theta-\hat{\theta}_o)$ < $(\theta-\hat{\theta})$	Var($\hat{\theta}_o$) < Var($\hat{\theta}$)	MSE($\hat{\theta}_o$) < MSE($\hat{\theta}$)
½	65.30%	100.00%	86.40%
1/3	60.10%	100.00%	79.30%
¼	56.80%	100.00%	74.20%
1/10	56.60%	100.00%	67.50%

Table 3. Simulation: Relationship between the Fisherian MLE and the Posterior Estimate with an Assumed "Wrong" Prior Estimate (60%).

Index of Confidence for Prior Info $(1-\lambda)$	$(\theta-\hat{\theta}_o)$ < $(\theta-\hat{\theta})$	Var($\hat{\theta}_o$) < Var($\hat{\theta}$)	MSE($\hat{\theta}_o$) < MSE($\hat{\theta}$)
½	.00%	100.00%	.00%
1/3	2.30%	100.00%	2.60%
¼	9.10%	100.00%	9.40%
1/10	32.40%	100.00%	33.40%