Research Article

Analysis of the Behrens-Fisher Problem Based on Bayesian Evidence

Yuliang Yin and Baoren Li

School of Economics, Beijing Technology and Business University, Beijing 100048, China

Correspondence should be addressed to Yuliang Yin; imyyl@163.com

The Behrens-Fisher problem concerns the inferences for the difference between the means of two normal populations without making any assumption about the variances. Although the problem has been extensively studied in the literature, researchers cannot agree on its solution at present. In this paper, we propose a new method for dealing with the Behrens-Fisher problem in the Bayesian framework. The Bayesian evidence for testing the equality of two normal means and a credible interval at a specified level for the difference between the means are derived. Simulation studies are carried out to evaluate the performance of the provided Bayesian evidence.

1. Introduction

The Behrens-Fisher problem may arise in the comparison of two treatments, products, and so forth. It concerns comparing the means of two normal distributions whose variances are unknown. Suppose that $X_1, \ldots, X_m$ and $Y_1, \ldots, Y_n$ are two independent random samples from two normal populations $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$, respectively, where both σ_1^2 and σ_2^2 are completely unspecified. We are interested in testing the hypothesis $H_0 : \mu_1 = \mu_2$ and giving the interval estimation for the difference between two means, $\theta = \mu_1 - \mu_2$.

The difficulty with the Behrens-Fisher problem is that the standard classical frequentist evidence is not available because nuisance parameters are present. Tsui and Weerahandi [1] introduced the concept of the generalized P value to deal with nuisance parameters in testing hypotheses. If the corresponding sample means and sample variances are denoted by $(\overline{x}, \overline{y})$ and (s_1^2, s_2^2), respectively, a generalized frequentist evidence for testing H_0 can be formulated by the approach of the generalized P value as

$$p(x) = P\left(F_{1,m+n-2} \geq \left((\overline{x} - \overline{y})^2 (m + n - 2)\right)\right.$$
$$\times \left(\tilde{s}_1^2 B_{(m-1)/2,(n-1)/2}^{-1}\right. \tag{1}$$
$$\left.\left.+ \tilde{s}_2^2 \left(1 - B_{(m-1)/2,(n-1)/2}\right)^{-1}\right)\right)^{-1},$$

where $F_{1,m+n-2}$ is an F-variable with 1 and $m + n - 2$ degrees of freedom and $B_{(m-1)/2,(n-1)/2}$ is a *Beta*-variable with parameters $(m-1)/2$ and $(n-1)/2$ that is independent of $F_{1,m+n-2}$, $\tilde{s}_1^2 = (1 - m^{-1})s_1^2$, and $\tilde{s}_2^2 = (1 - n^{-1})s_2^2$. This generalized frequentist solution is formally equivalent to the Bayesian solution given by Jeffreys [2] or the fiducial solution given by Wallace [3]. Meng [4] introduced the concept of the posterior predictive P value and provided posterior predictive evidence. In the case of Behrens-Fisher problem, this test is formulated as

$$ppp(x)$$
$$= P\left(F_{1,m+n} \geq \left((\overline{x} - \overline{y})^2 (m + n)\right)\right.$$
$$\times \left(\left[\tilde{s}_1^2 + (\overline{x} - \mu)^2\right] B_{m/2,n/2}^{-1}\right. \tag{2}$$
$$\left.\left.+ \left[\tilde{s}_2^2 + (\overline{y} - \mu)^2\right] \left(1 - B_{m/2,n/2}\right)^{-1}\right)^{-1}\right),$$

where $F_{1,m+n}$ is an F-variable with 1 and $m + n$ degrees of freedom, $B_{m/2,n/2}$ is a *Beta*-variable with parameters $m/2$ and

$n/2$ that is independent of $F_{1,m+n}$, and μ is a variable with a "combined t" distribution:

$$\pi_0(\mu \mid x) \propto \left[1 + \frac{1}{m-1}\left(\frac{\mu - \overline{x}}{s_1/\sqrt{m}}\right)^2\right]^{-m/2}$$
$$\times \left[1 + \frac{1}{n-1}\left(\frac{\mu - \overline{y}}{s_2/\sqrt{n}}\right)^2\right]^{-n/2}. \tag{3}$$

Behrens [5] gave a confidence interval for the difference between the two means in a testing context of $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ based on the pivotal quantity of

$$D = \frac{\left(\overline{X_1} - \overline{X_2}\right) - (\mu_1 - \mu_2)}{(n_1^{-1}S_1^2 - n_2^{-1}S_2^2)^{1/2}}. \tag{4}$$

Bartlett [6] revealed, from a frequentist perspective, that the coverage probability of the confidence interval given by Behrens is different from the specified confidence coefficient. Fisher [7] derived a fiducial interval for $\theta = \mu_1 - \mu_2$ which has a specified fiducial level by the method of fiducial inference. Neyman illustrated by calculation that an interval estimator with a fiducial level of $1 - \alpha$ is not necessarily a confidence interval with a confidence coefficient of $1 - \alpha$. Welch [8, 9] gave approximate solutions of the confidence intervals which are also constructed in a testing context based on the pivotal quantity D. In the Bayesian framework, Jeffreys [10], based on the objective prior

$$\pi\left(\mu_1, \mu_2, \sigma_1^2, \sigma_2^2\right) \propto \sigma_1^{-2}\sigma_2^{-2}, \tag{5}$$

constructed a Bayesian credible interval. This interval is algebraically equivalent to the fiducial interval of Fisher.

For more discussions of the Behrens-Fisher problem see Wilks [11], Chernoff [12], Chand [13], Banerjee [14], Srivastava [15], Ghosh and Kim [16], Madruga et al. [17], and McMurry et al. [18].

In this paper, we derive the Bayesian evidence for the Behrens-Fisher problem using the procedure in Yin [19] for testing point null hypotheses. Based on the provided Bayesian evidence, a Bayesian credible interval at a specified credible level for the difference of the means $\theta = \mu_1 - \mu_2$ is derived in a Bayesian testing context.

This paper is organized as follows. In Section 2, we give the main results of the Bayesian analysis of the Behrens-Fisher problem concerning the testing and interval estimation of the difference of two normal means with the variances completely unknown. Some conclusions and discussions are given in Section 3.

2. Main Results

2.1. Bayesian Evidence for the Behrens-Fisher Problem. Yin
[19] introduced a Bayesian measure of evidence for testing point null hypotheses of the form

$$H_0 : \theta = \theta_0 \quad \text{v.s.} \quad H_1 : \theta \neq \theta_0. \tag{6}$$

Let $X_1, \ldots, X_n$ be a random sample from a distribution with density $f(x \mid \theta)$, where θ is an unknown element of the parameter space Θ. The Bayesian evidence against the null hypothesis H_0 based on a prior $\pi(\theta)$ is given by

$$p^B(x) = P\left(|\theta - E(\theta \mid x)| \geq |\theta_0 - E(\theta \mid x)| \mid x\right), \tag{7}$$

where $E(\theta \mid x)$ is the posterior expectation of θ under the prior $\pi(\theta)$ and the probability is taken over the posterior distribution of θ. A smaller $p^B(x)$ means stronger evidence against the null hypothesis H_0. In his work, Yin illustrated that the Bayesian evidence given by (7) under the Jeffreys noninformative prior is just equivalent to the corresponding frequentist evidence for many classical testing situations and showed that the Lindley's paradox in Lindley [20] can be avoided by this Bayesian method of testing point null hypotheses.

Now consider the Behrens-Fisher problem of testing hypotheses

$$H_0 : \mu_1 = \mu_2 \quad \text{v.s.} \quad H_1 : \mu_1 \neq \mu_2. \tag{8}$$

Note that (8) can be reformulated as

$$H_0 : \mu_1 - \mu_2 = 0 \quad \text{v.s.} \quad H_1 : \mu_1 - \mu_2 \neq 0. \tag{9}$$

The posterior distribution for $\theta = \mu_1 - \mu_2$ under the objective prior (5) can be obtained as

$$\theta \mid x \sim \overline{x} - \overline{y} - \left(\frac{s_1 T_{m-1}}{\sqrt{m}} - \frac{s_2 T_{n-1}}{\sqrt{n}}\right), \tag{10}$$

where T_{m-1} and T_{n-1} are two independent t-variables with $m-1$ and $n-1$ degrees of freedom, respectively. Since the posterior expectation of θ is

$$E(\theta \mid x) = \overline{x} - \overline{y}, \tag{11}$$

the Bayesian evidence under the objective prior (5) can be formulated as

$$p^{\text{BF}}(x) = P\left(|\theta - (\overline{x} - \overline{y})| \geq |\overline{x} - \overline{y}| \mid x\right)$$
$$= P\left(\left|\frac{s_1 T_{m-1}}{\sqrt{m}} - \frac{s_2 T_{n-1}}{\sqrt{n}}\right| \geq |\overline{x} - \overline{y}|\right), \tag{12}$$

where the first probability is taken over the posterior distribution of θ and the second one is taken over two independent t-variables T_{m-1} and T_{n-1}.

Now we carry out a simulation study to illustrate the performance of the proposed Bayesian evidence. The simulation results listed in Table 1 show that $p^{\text{BF}}(x)$ is quite reasonable evidence for testing the Behrens-Fisher problem. For fixed values of σ_1 and σ_2, notice that the more significant the difference between μ_1 and μ_2 is, the smaller value of $p^{\text{BF}}(x)$ we may obtain, which means that the stronger Bayesian evidence for rejecting the null hypothesis of $H_0 : \mu_1 = \mu_2$ is given. Moreover, $p^{\text{BF}}(x)$ gives more reliable and efficient evidence when the population variances are small. It can also be noticed that the Bayesian evidence $p^{\text{BF}}(x)$ is very close to the corresponding generalized frequentist evidence $p(x)$ in (1) and the posterior predictive evidence $ppp(x)$ in (2).

TABLE 1: $p^{\mathrm{BF}}(x)$, $p(x)$ and $ppp(x)$ for testing the Behrens-Fisher Problem.

		$\sigma_1 = 2, \sigma_2 = 3$					$\sigma_1 = 3, \sigma_2 = 2$		
μ_1	μ_2	$p^{\mathrm{BF}}(x)$	$p(x)$	$ppp(x)$	μ_1	μ_2	$p^{\mathrm{BF}}(x)$	$p(x)$	$ppp(x)$
2.00	2.00	0.7823	0.7810	0.7816	2.00	2.00	0.7635	0.7606	0.7615
2.00	2.01	0.3133	0.3104	0.3102	2.00	2.01	0.3028	0.3031	0.3017
2.00	2.02	0.1722	0.1713	0.1711	2.00	2.02	0.1949	0.1950	0.1938
2.00	2.03	0.0854	0.0863	0.0868	2.00	2.03	0.0628	0.0627	0.0630
2.00	2.04	0.0418	0.0420	0.0428	2.00	2.04	0.0188	0.0185	0.0190
2.00	2.05	0.0117	0.0122	0.0121	2.00	2.05	0.0018	0.0017	0.0016
		$\sigma_1 = 0.1, \sigma_2 = 0.2$					$\sigma_1 = 0.2, \sigma_2 = 0.1$		
μ_1	μ_2	$p^{\mathrm{BF}}(x)$	$p(x)$	$ppp(x)$	μ_1	μ_2	$p^{\mathrm{BF}}(x)$	$p(x)$	$ppp(x)$
2.000	2.000	0.8597	0.8603	0.8596	2.000	2.000	0.8525	0.8522	0.8511
2.000	2.001	0.1329	0.1319	0.1336	2.000	2.001	0.4979	0.4998	0.4961
2.000	2.002	0.0688	0.0690	0.0696	2.000	2.002	0.1090	0.1070	0.1085
2.000	2.003	0.0384	0.0383	0.0379	2.000	2.003	0.0239	0.0239	0.0237
2.000	2.004	0.0313	0.0317	0.0318	2.000	2.004	0.0014	0.0014	0.0011
2.000	2.005	0.0033	0.0032	0.0035	2.000	2.005	0.0007	0.0008	0.0006
		$\sigma_1 = 2, \sigma_2 = 0.1$					$\sigma_1 = 2, \sigma_2 = 2$		
μ_1	μ_2	$p^{\mathrm{BF}}(x)$	$p(x)$	$ppp(x)$	μ_1	μ_2	$p^{\mathrm{BF}}(x)$	$p(x)$	$ppp(x)$
2.00	2.00	0.5239	0.5245	0.5279	2.00	2.00	0.9963	0.9963	0.9962
2.00	2.01	0.3327	0.3346	0.3357	2.00	2.01	0.2523	0.2507	0.2500
2.00	2.02	0.0224	0.0228	0.0218	2.00	2.02	0.0996	0.1006	0.0991
2.00	2.03	0.0032	0.0033	0.0034	2.00	2.03	0.0366	0.0366	0.0368
2.00	2.04	0.0017	0.0017	0.0018	2.00	2.04	0.0119	0.0122	0.0122
2.00	2.05	0.0001	0.0001	0.0001	2.00	2.05	0.0036	0.0040	0.0040

By this Bayesian evidence for the Behrens-Fisher problem, we consider two examples. One is included in Lehmann [21]. The driving times from a person's house to his working place following two different routes were measured which we list in Table 2. Another one is in Ghosh et al. [22] where the data which we list in Table 3 is from a clinical trial conducted by Sahu to compare the improvement score of surgical treatment with that of nonsurgical treatment. If it is assumed that the two independent samples in both Tables 2 and 3 are, respectively, drawn from two normal distributions $N(\mu_1, \sigma_1^2)$ and $N(\mu_2, \sigma_2^2)$ and if we are interested in the equality of the two means μ_1 and μ_2, each of these two examples reduces to the Behrens-Fisher problem of testing hypotheses (8). For both situations, the Bayesian evidence $p^{\mathrm{BF}}(x)$ and the corresponding generalized frequentist evidence $p(x)$ and posterior predictive evidence $ppp(x)$ all give very strong evidence of nearly zero for rejecting the null hypothesis that there is no difference between the two means. This agrees with our intuition from the observed data.

2.2. Bayesian Credible Interval.
Based on the proposed Bayesian evidence, a credible interval for the difference of means $\theta = \mu_1 - \mu_2$ at a specified credible level can be constructed in a testing context. For the following hypothesis testing problem of comparing two normal means:

$$H_0 : \mu_1 - \mu_2 = \theta_0 \quad \text{v.s.} \quad H_1 : \mu_1 - \mu_2 \neq \theta_0, \qquad (13)$$

where the variances are completely unspecified, the Bayesian evidence under the objective prior (5) is

$$
\begin{aligned}
p^{\mathrm{BF}}(x; \theta_0) &= P\left(|\theta - (\overline{x} - \overline{y})| \geq |\theta_0 - (\overline{x} - \overline{y})| \mid x \right) \\
&= P\left(\left| \frac{s_1 T_{m-1}}{\sqrt{m}} - \frac{s_2 T_{n-1}}{\sqrt{n}} \right| \geq |\theta_0 - (\overline{x} - \overline{y})| \right),
\end{aligned}
\qquad (14)
$$

where the first probability is taken over the posterior distribution of θ and the second one is taken over two independent t-variables T_{m-1} and T_{n-1}.

Theorem 1. *For the Behrens-Fisher problem, let $A^{BF}(\theta_0) = \{x : p^{BF}(x; \theta_0) \geq \alpha\}$, $S^{BF}(x) = \{\theta_0 : x \in A^{BF}(\theta_0)\}$, and $I_k = [\overline{x} - \overline{y} - k, \overline{x} - \overline{y} + k]$. For a fixed α, if $I_{k^{BF}}$ satisfies*

$$P\left(\theta \in I_{k^{BF}} \mid x \right) = 1 - \alpha, \qquad (15)$$

then one has

$$S^{BF}(x) = I_{k^{BF}}. \qquad (16)$$

Proof. On one hand, $I_{k^{\mathrm{BF}}}$ satisfies

$$P\left(\theta \in I_{k^{\mathrm{BF}}} \mid x \right) = 1 - \alpha, \qquad (17)$$

which means that

$$P\left(|\theta - (\overline{x} - \overline{y})| \leq k^{\mathrm{BF}} \mid x \right) = 1 - \alpha. \qquad (18)$$

TABLE 2: Measures of driving times from following two different routes.

Route	Times										
I	6.5	6.8	7.1	7.3	10.2						
II	5.8	5.8	5.9	6.0	6.0	6.0	6.3	6.3	6.4	6.5	6.5

TABLE 3: Scores of surgical and non-surgical treatments.

Treatment	Scores													
Surgical	15	9	12	16	14	15	18	13	12	11	15	9	16	9
Non-surgical	6	8	7	4	4	6	8	3	7	8	9	6	3	6	4

On the other hand, it is easy to know that $p^{\mathrm{BF}}(x; \theta_0) \geq \alpha$ is equivalent to

$$P\left(\left|\theta - (\overline{x} - \overline{y})\right| \leq \left|\theta_0 - (\overline{x} - \overline{y})\right| \mid x\right) \leq 1 - \alpha. \quad (19)$$

By (18) and (19), we know that $S^{\mathrm{BF}}(x) = I_{k^{\mathrm{BF}}}$. □

By Theorem 1, we know that the $1 - \alpha$ credible interval for $\theta = \mu_1 - \mu_2$ centered at $E(\theta \mid x) = \overline{x} - \overline{y}$ can be easily obtained by $p^{\mathrm{BF}}(x; \theta_0) \geq \alpha$. This is a Bayesian interval obtained in a testing context. Interestingly, the resulting interval by our method is just equivalent to that given by Fisher or Jeffreys.

In fact, we have another interesting result about the interval estimation of $\theta = \mu_1 - \mu_2$ on the basis of the Bayesian evidence $p^{\mathrm{BF}}(x; \theta_0)$, which shows that the $1 - \alpha$ credible interval centered at the posterior expectation for the Behrens-Fisher problem can be constructed by the α and $1 - \alpha/2$ quantiles of the posterior distribution of θ. We summarize this as the following theorem.

Theorem 2. *For the Behrens-Fisher problem, $p^{BF}(x; \theta_0) \geq \alpha$ yields the $1 - \alpha$ credible interval for $\theta = \mu_1 - \mu_2$ centered at the posterior expectation $E(\theta \mid x) = \overline{x} - \overline{y}$ as follows:*

$$I_{k^{BF}} = \left[\widehat{\theta}_{\alpha/2}(x), \widehat{\theta}_{1-(\alpha/2)}(x)\right], \quad (20)$$

where $\widehat{\theta}_{\alpha/2}(x)$ and $\widehat{\theta}_{1-\alpha/2}(x)$ are, respectively, the $\alpha/2$ and $1 - \alpha/2$ quantiles of the posterior distribution

$$\theta \mid x \sim \overline{x} - \overline{y} - \left(\frac{s_1 T_{m-1}}{\sqrt{m}} - \frac{s_2 T_{n-1}}{\sqrt{n}}\right). \quad (21)$$

Proof. We first prove that the Bayesian evidence for testing (13) can be expressed as

$$p^{BF}(x; \theta_0) = 2 \min\left\{P(\theta \leq \theta_0 \mid x), P(\theta \geq \theta_0 \mid x)\right\}. \quad (22)$$

In fact, if $\theta_0 \geq \overline{x} - \overline{y}$, we have

$$\begin{aligned} p^{BF}(x; \theta_0) &= P\left(\left|\theta - (\overline{x} - \overline{y})\right| \geq \left|\theta_0 - (\overline{x} - \overline{y})\right| \mid x\right) \\ &= P\left(\theta - (\overline{x} - \overline{y}) \geq \theta_0 - (\overline{x} - \overline{y}) \mid x\right) \\ &\quad + P\left(\theta - (\overline{x} - \overline{y}) \leq -(\theta_0 - (\overline{x} - \overline{y})) \mid x\right) \\ &= 2P\left(\theta - (\overline{x} - \overline{y}) \geq \theta_0 - (\overline{x} - \overline{y}) \mid x\right) \\ &= 2P\left(\theta \geq \theta_0 \mid x\right), \end{aligned} \quad (23)$$

where the second equation is due to the fact that the posterior distribution of θ is symmetric about $\overline{x} - \overline{y}$. Similarly, if $\theta_0 \leq \overline{x} - \overline{y}$, we have

$$p^{BF}(x; \theta_0) = 2P(\theta \leq \theta_0 \mid x). \quad (24)$$

By (23) and (24) together with the symmetry of the posterior distribution of θ, we have

$$p^{BF}(x; \theta_0) = 2 \min\left\{P(\theta \leq \theta_0 \mid x), P(\theta \geq \theta_0 \mid x)\right\}. \quad (25)$$

It then follows that $p^{BF}(x; \theta_0) \geq \alpha$ if and only if $P(\theta \leq \theta_0 \mid x) \geq \alpha/2$ and $P(\theta \geq \theta_0 \mid x) \geq \alpha/2$ hold simultaneously, which is equivalent to

$$\widehat{\theta}_{\alpha/2}(x) \leq \theta_0 \leq \widehat{\theta}_{1-(\alpha/2)}(x). \quad (26)$$

Since the posterior of θ is symmetric about $\overline{x} - \overline{y}$, $[\widehat{\theta}_{\alpha/2}(x), \widehat{\theta}_{1-(\alpha/2)}(x)]$ is a credible interval centered at $\overline{x} - \overline{y}$. This completes the proof. □

Theorem 2 provides another way of constructing the credible interval for $\theta = \mu_1 - \mu_2$. Moreover, we know easily by the proof of Theorem 2 that the $1 - \alpha$ credible interval for $\theta = \mu_1 - \mu_2$ which is centered at the posterior expectation can be given by $[\widehat{\theta}_{\alpha/2}(x), \widehat{\theta}_{1-(\alpha/2)}(x)]$ even when other priors are used so long as the posterior of θ is symmetric.

Now we return to the examples of comparing means of driving time and comparing improvement scores of treatments discussed above. We recommend the $1 - \alpha$ credible intervals of $(-0.4659, 3.2313)$ and $(5.1982, 9.3422)$ for Lehmann's and Sahu's data, respectively, which are obtained according to our procedure. The recommended intervals are essentially equivalent to the intervals given by the method of Fisher or Jeffreys.

3. Conclusions

We carry out Bayesian analysis of the Behrens-Fisher problem in this paper. The Bayesian evidence for testing the hypothesis $H_0 : \mu_1 = \mu_2$ against $H_1 : \mu_1 \neq \mu_2$ is given. Simulation results show that our evidence performs quite well and is very close to the corresponding generalized frequentist evidence and posterior predictive evidence for the Behrens-Fisher problem. Based on the proposed evidence, a method of constructing the credible interval at a specified level for

the difference of means $\theta = \mu_1 - \mu_2$ is provided in a Bayesian testing context. It is interesting that the credible interval given by our method is just in accordance with that derived by Fisher or Jeffreys. This way of constructing the credible interval via the Bayesian testing evidence is in analogy with the way of constructing the confidence interval via the frequentist evidence.

By this method of analyzing the Behrens-Fisher problem, we give an efficient way of dealing with nuisance parameters which are the source of the difficulty with this problem. This is because our inferences about $\theta = \mu_1 - \mu_2$ are based on the posterior distribution of the interested parameter, which can be easily obtained in the Bayesian framework even when nuisance parameters are present. Both the Bayesian evidence and the credible interval can be computed quite easily by the Monte Carlo method. Furthermore, by this method, even if an informative prior which is different from that in (5) is used, the corresponding Bayesian evidence and credible intervals could be obtained smoothly. In other words, this method provides an efficient way of combining the information contained in the prior and that contained in the samples. Further research would be needed to evaluate the performance of the inferences by the proposed method if an informative prior is introduced.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

References

[1] K.-W. Tsui and S. Weerahandi, "Generalized P-values in significance testing of hypotheses in the presence of nuisance parameters," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 602–607, 1989.

[2] H. Jeffreys, *Theory of Probability*, Oxford University Press, 3rd edition, 1967.

[3] D. L. Wallace, *The Behrens-Fisher and Feiller-Creasy Problems*, Edited by R. A. Fisher, Springer, New York, NY, USA, 1980.

[4] X.-L. Meng, "Posterior predictive P-values," *The Annals of Statistics*, vol. 22, no. 3, pp. 1142–1160, 1994.

[5] B. V. Behrens, "Ein Beitrag zur Fehlerberechnung bei wenige Beobachtungen," *Landwirtschaftliches Jahresbuch*, vol. 68, pp. 807–837, 1929.

[6] M. S. Bartlett, "The information available in small samples," *Proceedings of the Cambridge Philosophical Society*, vol. 32, no. 4, pp. 560–566, 1936.

[7] R. A. Fisher, "The fiducial argument in statistical inference," *The Annals of Eugenics*, vol. 11, pp. 141–172, 1935.

[8] B. L. Welch, "The significance of the difference between two means when the population variances are unequal," *Biometrika*, vol. 29, pp. 350–362, 1938.

[9] B. L. Welch, "The generalization of student's problem when several different population variances are involved," *Biometrika*, vol. 34, pp. 28–35, 1947.

[10] H. Jeffreys, *Theory of Probability*, Oxford University Press, 1961.

[11] S. S. Wilks, "On the problem of two samples from normal populations with unequal variances," *Annals of Mathematical Statistics*, vol. 11, no. 4, pp. 475–476, 1940.

[12] H. Chernoff, "Asymptotic studentization in testing of hypotheses," *Annals of Mathematical Statistics*, vol. 20, pp. 268–278, 1949.

[13] U. Chand, "Distributions related to comparison of two means and two regression coefficients," *Annals of Mathematical Statistics*, vol. 21, pp. 507–522, 1950.

[14] S. K. Banerjee, "Approximate confidence interval for linear functions of means of k populations when the population variances are not equal," *Sankhya*, vol. 22, pp. 357–358, 1960.

[15] M. S. Srivastava, "On a sequential analogue of the Behrens-Fisher problem," *Journal of the Royal Statistical Society B*, vol. 32, pp. 144–148, 1970.

[16] M. Ghosh and Y. Kim, "The Behrens-Fisher problem revisited: a Bayes-frequentist synthesis," *Biometrika*, vol. 29, no. 1, pp. 5–17, 2001.

[17] M. R. Madruga, C. A. B. Pereira, and J. M. Stern, "Bayesian evidence test for precise hypotheses," *Journal of Statistical Planning and Inference*, vol. 117, no. 2, pp. 185–198, 2003.

[18] T. L. McMurry, D. N. Politis, and J. P. Romano, "Subsampling inference with K populations and a non-standard Behrens-Fisher problem," *International Statistical Review*, vol. 80, no. 1, pp. 149–175, 2012.

[19] Y. Yin, "A new Bayesian procedure for testing point null hypotheses," *Computational Statistics*, vol. 27, no. 2, pp. 237–249, 2012.

[20] D. V. Lindley, "A statistical paradox," *Biometrika*, vol. 44, pp. 187–192, 1957.

[21] E. L. Lehmann, *Nonparametrics: Statistical Methods Based on Ranks*, Holden-Day, San Francisco, Calif, USA, 1975.

[22] J. K. Ghosh, M. Delampady, and T. Samanta, *An Introduction to Bayesian Analysis*, Springer, New York, NY, USA, 2006.