in Physics & Probability Essays in honor of Edwin T. Jaynes, pp. 177-194, W. T. Grandy
and P. W. Milonni *(eds.) Cambridge University Press, England* (1993).

ON THE DIFFERENCE IN MEANS

G. Larry Bretthorst
Washington University
Department of Chemistry
St. Louis, Missouri 63130-4899

ABSTRACT. Given two sets of data that are repeated measurements of the same physical quantity, one "control" and one "trial," there are three problems of interest to the experimenter: (1) determine if something changed, (2) if something changed, what? and (3) estimate the magnitude of the change. These three problems are addressed using probability theory as extended logic. In the first section, the probability that the data sets differ is computed independent of what changed, i.e., independent of whether or not the means or standard deviations changed. In the second section, two probability distributions are computed: first, the probability that the means changed is computed independent of whether or not the standard deviations changed. Then second, the probability that the standard deviations changed is computed independent of whether or not the means changed. In the third section, the problem of estimating the magnitude of the changes is addressed. Here the probability density functions for both the difference in means and the ratio of standard deviations is computed. The probability for the ratio of standard deviations is computed independent of whether or not the means are the same, just as the probability for the difference in means is computed independent of whether or not the standard deviations are the same. This last calculation generalizes the solution of both the two-sample problem (different means and same but unknown standard deviations) and the Behrens-Fisher problem (different means and different unknown standard deviations). The calculations are illustrated with a numerical example in the fourth section.

This paper addresses one of the most fundamental problems that can occur in experimental science, that of analyzing two independent measurements of the same physical quantity under slightly different experimental conditions when the measurements are assumed uncorrelated. This problem has a long history going back to at least 1929, when Behrens [1] proposed a solution to the problem of estimating the difference in means when the standard deviations are assumed unequal and unknown. Fisher rederived the same result using fiducial probabilities in 1937 [2,3], and last Jeffreys arrived at the same distribution in 1939 using Bayesian probability theory [4]. However, the Behrens test became quite controversial because it called into question some of the basic tenets of orthodox sampling theory. For this reason the Behrens test is essentially not used today. Instead a series of *ad hoc* tests are in use. For a good discussion of the Behrens-Fisher controversy see Lee,

[5]. For a review of the Behrens-Fisher problem see Refs. [6,7,8] and for a description of how this problem is addressed in orthodox sampling theory see Refs. [9,10,11].

The problem is more complex than just determining the difference in means or the ratio of standard deviations. For example, the standard deviations might be the same, or they might be different; one simply may not know which condition applies and indeed, it is certainly possible, that the data may not strongly favor either hypothesis (as they do not in the numerical example) and so neither the two-sample test nor the Behrens-Fisher test is truly applicable. Additionally, estimating the difference in means is a parameter estimation problem, and it implicitly assumes that the means are different. Before one attempts to estimate the difference in means, it would be reasonable to ask if the means are different? The same statements apply to estimating the ratio of the standard deviations.

An early attempt to solve part of this problem was made by Dayal and Dickey [12,13], when they computed the probability for four basic hypotheses. These hypotheses, the means and standard deviations are the same, the means are the same and the standard deviations differ, the means differ and the standard deviations are the same, and the means and standard deviations differ, are of fundamental importance in this problem; but they do not directly tell one if the data sets changed, and if they changed how? To address the questions "Did the data sets change?" one needs to compute the probability that the data sets differ independent of what changed. To address the other question, "What changed?" one needs to compute two probabilities: the probability that the means changed, and the probability that the standard deviations changed. A complete list of all of the hypotheses addressed in this paper are given in Table 1. Before proceeding to discuss these hypotheses, the problem being addressed is described in more detail.

In two-sample and Behrens-Fisher like problems there are two data sets, D_1 and D_2, one "control," and one "trial." Data set D_1 has N_1 data items labeled d_{1i}, and similarly for data set D_2. These data sets are repeated measurements of the same physical quantity. This quantity is designated as A in data set D_1, and B in data set D_2. The parameters, A and B, will be referred to as the means of the data sets, although one should keep firmly in mind that this is a colloquial use. What one really means by A and B are two hypotheses. These hypotheses are of the form "the constant signal in data set D_1 had value A" and similarly for B. Each data set is contaminated by additive noise of standard deviations σ_1 and σ_2. Similarly σ_1 and σ_2 will be referred to as the standard deviations of the noise, but what is really meant is again two hypotheses of the form "the noise in data set D_1 has standard deviation σ_1" and similarly for D_2. With these definitions the hypotheses and the data are related by

$$d_{1i} = A + \text{noise of standard deviation } \sigma_1, \tag{1}$$

$$d_{2i} = B + \text{noise of standard deviation } \sigma_2. \tag{2}$$

Table 1 lists the hypotheses addressed in this paper and assigns an abbreviation, to each of them. In this paper, simple hypothesis will be labeled with a single letter. For example, from Table 1, the hypothesis v means the standard deviations are the same. Two hypotheses next to each other should be read with an 'and' between them. For example, sv stands for the means are the same and the variances are the same. Last, the negation of a hypothesis is represented with a bar over the hypothesis. For example, $\overline{v}$ means the variances are not the same. The simple hypotheses will generally be addressed first. There is one exception to this, $\overline{s} + \overline{v}$, the means or the standard deviations are not the

Table 1: The Hypotheses Addressed

Hypotheses	Abbreviation
The means are the same	s
The means are not the same	$\overline{s}$
The standard deviations are the same	v
The standard deviations are not the same	$\overline{v}$
The means and the standard deviations are the same	sv
The means are the same, and the standard deviations differ	$s\overline{v}$
The means are not the same, and the standard deviations are the same	$\overline{s}v$
The means and standard deviations are not the same	$\overline{s}\,\overline{v}$
The means or the standard deviations are not the same	$\overline{s}+\overline{v}$
The difference, $A-B$, is equal to δ	δ
The ratio σ_1/σ_2 is equal to r	r

same. This hypothesis is the one that answers the question "did something change?" and it is the first hypothesis addressed in this paper. The hypothesis δ means the difference between the constant in the first data set, and the constant in the second data is equal to δ. And similarly the hypothesis r means that the ratio of the standard deviation in the first data set divided by the standard deviation in the second data set is equal to r. The four compound hypotheses, sv, $s\overline{v}$, $\overline{s}v$, and $\overline{s}\,\overline{v}$ are the four hypotheses addressed by Dayal and Dickey [12,13].

1. Did Something Change?

The "trial" data set can only differ from the "control" data set in two important ways: either the means or the standard deviations can change; there are no other possibilities. This hypothesis is abbreviated $\overline{s}+\overline{v}$. The probability that represents this state of knowledge is denoted by $P(\overline{s}+\overline{v}|D_1D_2I)$. In words, this is the probability that the means or standard deviations are not the same given the two data sets, D_1 and D_2, and the prior information I. The prior information I is all of the assumptions that have gone into making this a well posed problem. At present I includes the separation of the data into a signal plus additive noise, and that A and B are constants.

To compute the probability that the means or the standard deviations differ, note that

$$P(\overline{s}+\overline{v}|D_1D_2I) = P(\overline{sv}|D_1D_2I) = 1 - P(sv|D_1D_2I). \tag{3}$$

It is sufficient to compute the probability that the means and the standard deviations are the same and from that one can compute the probability that the means or the standard deviations differed. This is the first hypothesis studied by Dayal and Dickey. $P(sv|D_1D_2I)$ does not depend on any parameters, it is a marginal probability density function. The hypothesis sv assumes the means and the standard deviations are the same, so two parameters (a constant A, and a standard deviation σ_1) have been removed by marginalization:

$$P(sv|D_1D_2I) = \int dAd\sigma_1 P(svA\sigma_1|D_1D_2I) \tag{4}$$

where $B \to A$ and $\sigma_2 \to \sigma_1$. The right hand side of this equation may be factored using Bayes' Theorem to obtain

$$P(sv|D_1D_2I) \propto \int dAd\sigma_1 P(svA\sigma_1|I)P(D_1D_2|svA\sigma_1I). \tag{5}$$

Assuming logical independence of the parameters and the data, this may be further simplified to obtain

$$P(sv|D_1D_2I) \propto \int dAd\sigma_1 P(sv|I)P(A|I)P(\sigma_1|I)P(D_1|svA\sigma_1I)P(D_2|svA\sigma_1I) \qquad (6)$$

where $P(sv|I)$ is the prior probability that the means and the standard deviations are the same, $P(A|I)$ is the prior probability for the amplitude, $P(\sigma_1|I)$ is the prior probability for the standard deviation, and $P(D_1|svA\sigma_1I)$ and $P(D_2|svA\sigma_1I)$ are the direct probabilities or likelihoods of the two data sets.

In this calculation uninformative prior probabilities will be used. However, care will be taken to ensure that all prior probabilities are fully normalized. For the amplitude, A, a bounded uniform prior will be used:

$$P(A|I) = \begin{cases} \dfrac{1}{R_a} & \text{If } L \leq A \leq H \\ 0 & \text{otherwise} \end{cases} \qquad (7)$$

where $R_a \equiv H - L$, and H and L are the limits on the constant A and are assumed known. A bounded Jeffreys prior will be used for the standard deviation of the noise:

$$P(\sigma_1|I) = \begin{cases} \dfrac{1}{\sigma_1 \log(R_\sigma)} & \text{If } \sigma_L \leq \sigma_1 \leq \sigma_H \\ 0 & \text{otherwise} \end{cases} \qquad (8)$$

where R_σ is the ratio σ_H/σ_L, and σ_H and σ_L are the limits on the standard deviation σ_1 and are also assumed known.

As noted earlier there are four fundamental hypotheses: sv, $s\overline{v}$, $\overline{s}v$, and $\overline{s}\,\overline{v}$, and probability theory requires us to assign a prior probability to each of them. Here it is $P(sv|I)$ that must be assigned. No assumption will be made about which of these four possibilities are present, but all four of these possibilities are assumed to occur and so a probability of $1/4$ is assigned to each.

The only quantities that remain to be assigned are the two likelihood functions. The prior probability for the noise will be taken to be Gaussian. If the limits on the A integral extend from minus infinity to plus infinity, and the limits on the σ_1 integral extend from zero to infinity, then both integrals can be evaluated in closed form. However, with finite limits either of the two indicated integrals may evaluated, but the other must be evaluated numerically. Evaluating the integral over the amplitude, and neglecting a factor of $(2\pi)^{-N/2}$, which is common to all the models, we obtain:

$$P(sv|D_1D_2I) \propto \frac{\sqrt{\pi/2N}}{4R_a \log(R_\sigma)} \int_{\sigma_L}^{\sigma_H} d\sigma_1 \sigma_1^{-N} \exp\left\{-\frac{z}{2\sigma_1^2}\right\} \left[Q\left(\frac{1}{2}, \frac{U_L^2}{\sigma_1^2}\right) \pm Q\left(\frac{1}{2}, \frac{U_H^2}{\sigma_1^2}\right)\right] \qquad (9)$$

where $\overline{d}$ and $\overline{d^2}$ are the mean and mean-square of the pooled data, $N = N_1 + N_2$,

$$U_H = \sqrt{\frac{N}{2}}(H - \overline{d}), \qquad U_L = \sqrt{\frac{N}{2}}(L - \overline{d}), \qquad z = N[\overline{d^2} - (\overline{d})^2], \qquad (10)$$

and $Q(r,x)$ is the complimentary Gamma function of index r and argument x. The sign is chosen to be minus if both U_H and U_L are of the same sign, and plus if U_H and U_L are of different sign.

Equation (6) can be used to tell one if the data sets are the same, and by using Eq. (3) one can determine the probability that the means or the standard deviations changed, and thus, answers the first question of interest, "did something change?" But as soon as one knows that something changed, ones interest in the problem changes and one wants to know "What changed?" and it is this problem that is addressed in the next section.

2. Given That Something Changed, What Changed?

Given that something changed, there are only two possibilities: either the means or the standard deviations changed. To determine if the means changed one computes $P(s|D_1D_2I)$. Similarly, to determine if the standard deviations changed, one computes $P(v|D_1D_2I)$. Using the sum rule, these probabilities may be written

$$P(s|D_1D_2I) = P(sv|D_1D_2I) + P(s\overline{v}|D_1D_2I) \tag{11}$$

and

$$P(v|D_1D_2I) = P(sv|D_1D_2I) + P(\overline{s}v|D_1D_2I). \tag{12}$$

where $P(s|D_1D_2I)$ is computed independent of whether or not the standard deviations are the same; while $P(v|D_1D_2I)$ is independent of whether or not the means are the same. Note that three of the four hypotheses studied by Dayal and Dickey (sv, $s\overline{v}$, and $\overline{s}v$) have appeared. The fourth, $\overline{s}\,\overline{v}$, appears whenever either $P(\overline{s}|D_1D_2I)$ or $P(\overline{v}|D_1D_2I)$ is computed.

Did The Means Change?

The probability that the means are the same, $P(s|D_1D_2I)$, is a marginal probability. It is computed from two terms, $P(sv|D_1D_2I)$, and $P(s\overline{v}|D_1D_2I)$. $P(sv|D_1D_2I)$ was computed in the last section and $P(s\overline{v}|D_1D_2I)$ is addressed in this subsection. Notice that $P(s\overline{v}|D_1D_2I)$ assumes the constants are the same in both data sets, but the standard deviations are different. Neither the constant nor the standard deviations appear in $P(s\overline{v}|D_1D_2I)$. Consequently, $P(s\overline{v}|D_1D_2I)$, is a marginal probability density, where the constant and the two standard deviations were removed as nuisances:

$$\begin{aligned}
P(s\overline{v}|D_1D_2I) &= \int dA d\sigma_1 d\sigma_2 P(s\overline{v}A\sigma_1\sigma_2|D_1D_2I)\\
&\propto \int dA d\sigma_1 d\sigma_2 P(s\overline{v}A\sigma_1\sigma_2|I)P(D_1D_2|s\overline{v}A\sigma_1\sigma_2 I)\\
&= \int dA d\sigma_1 d\sigma_2 P(s\overline{v}|I)P(A|I)P(\sigma_1|I)P(\sigma_2|I)\\
&\quad \times P(D_1|s\overline{v}A\sigma_1 I)P(D_2|s\overline{v}A\sigma_2 I).
\end{aligned} \tag{13}$$

The prior, $P(\sigma_2|I)$, must be assigned; and for logical consistence it must be assigned the same prior range as σ_1:

$$P(\sigma_2|I) = \begin{cases} \dfrac{1}{\sigma_2 \log(R_\sigma)} & \text{If } \sigma_L \le \sigma_2 \le \sigma_H \\ 0 & \text{otherwise} \end{cases}. \tag{14}$$

To evaluate the two integrals over σ_1 and σ_2, one substitutes Eqs. (7,8,14), a Gaussian noise prior for the likelihoods, and $P(s\overline{v}|I) = 1/4$. Then evaluating the integral over σ_1 and σ_2, one obtains

$$\begin{aligned} P(s\overline{v}|D_1D_2I) \propto & \frac{\Gamma(N_1/2)\Gamma(N_2/2)}{16R_a[\log(R_\sigma)]^2}\int_L^H U_1^{-\frac{N_1}{2}} dA U_2^{-\frac{N_2}{2}} \\ & \times \left[Q\left(\frac{N_1}{2},\frac{U_1}{\sigma_H^2}\right) - Q\left(\frac{N_1}{2},\frac{U_1}{\sigma_L^2}\right)\right]\left[Q\left(\frac{N_2}{2},\frac{U_2}{\sigma_H^2}\right) - Q\left(\frac{N_2}{2},\frac{U_2}{\sigma_L^2}\right)\right] \end{aligned} \tag{15}$$

where

$$U_1 = \frac{N_1}{2}(\overline{d_1^2} - 2A\overline{d_1} + A^2), \qquad U_2 = \frac{N_2}{2}(\overline{d_2^2} - 2A\overline{d_2} + A^2) \tag{16}$$

and $\overline{d_1}$, $\overline{d_1^2}$, $\overline{d_2}$, $\overline{d_2^2}$ are the means and mean-squares of D_1 and D_2 respectively.

Did The Standard Deviations Change?

To obtain the probability that the standard deviations changed, $P(v|D_1D_2I)$, two terms must be computed. The first of these term, $P(sv|D_1D_2I)$, has already been computed, Eq. (6), and the second term, $P(\overline{s}v|D_1D_2I)$, like the first, is a marginal probability given by

$$\begin{aligned} P(\overline{s}v|D_1D_2I) &= \int dAdBd\sigma_1 P(\overline{s}vAB\sigma_1|D_1D_2I) \\ &\propto \int dAdBd\sigma_1 P(\overline{s}vAB\sigma_1|I)P(D_1D_2|\overline{s}vAB\sigma_1I) \\ &= \int dAdBd\sigma_1 P(\overline{s}v|I)P(A|I)P(B|I)P(\sigma_1|I) \\ &\quad \times P(D_1|\overline{s}vA\sigma_1I)P(D_2|\overline{s}vB\sigma_1I). \end{aligned} \tag{17}$$

The prior probability, $P(B|I)$, must be assigned; and for logical consistency it must be assigned the same prior ranges as A:

$$P(B|I) = \begin{cases} \dfrac{1}{R_a} & \text{If } L \leq B \leq H \\ 0 & \text{otherwise} \end{cases}. \tag{18}$$

To evaluate the integrals over A and B, one substitutes 1/4 for $P(\overline{s}v|I)$, Eqs. (7,8,18), and a Gaussian noise prior is used to assign the two likelihoods. Then evaluating the integrals over A and B one obtains,

$$\begin{aligned} P(\overline{s}v|D_1D_2I) \propto & \frac{\pi}{8R_a^2\log(R_\sigma)\sqrt{N_1N_2}}\int_{\sigma_L}^{\sigma_H} d\sigma_1\sigma_1^{-N+1}\exp\left\{-\frac{z_1+z_2}{2\sigma_1^2}\right\} \\ & \times \left[Q\left(\frac{1}{2},\frac{U_{1L}^2}{\sigma_1^2}\right) \pm Q\left(\frac{1}{2},\frac{U_{1H}^2}{\sigma_1^2}\right)\right]\left[Q\left(\frac{1}{2},\frac{U_{2L}^2}{\sigma_1^2}\right) \pm Q\left(\frac{1}{2},\frac{U_{2H}^2}{\sigma_1^2}\right)\right] \end{aligned} \tag{19}$$

where

$$z_1 = N_1[\overline{d_1^2} - (\overline{d_1})^2], \qquad z_2 = N_2[\overline{d_2^2} - (\overline{d_2})^2], \tag{20}$$

$$U_{1H} = \sqrt{\frac{N_1}{2}}(H - \overline{d_1}), \qquad U_{1L} = \sqrt{\frac{N_1}{2}}(L - \overline{d_1}), \tag{21}$$

$$U_{2H} = \sqrt{\frac{N_2}{2}}(H - \overline{d_2}), \qquad U_{2L} = \sqrt{\frac{N_2}{2}}(L - \overline{d_2}), \tag{22}$$

and the sign is chosen to be minus when U_{1H} and U_{1L} are of the same sign and plus when they are of different sign, and similarly for U_{2H} and U_{2L}.

Using the calculations presented so far one can determine if something has changed, and then determine what has changed; either the means or the standard deviations. But after determining what changed, again one's interest in the problem changes. Now one wants to estimate the magnitude of the changes. To estimate the magnitude of the changes one needs to compute $P(\delta|D_1D_2I)$ and $P(r|D_1D_2I)$, and these calculations are the subject of the next section.

3. Estimating The Magnitude Of The Changes

Estimating the difference in means and the ratio of the standard deviations is where most research on this problem has been concentrated. These calculations are based on two assumptions: that one of the parameters differs, and that the other parameter is either the same or not. Then using these assumptions, one estimates the difference in means or the ratio of standard deviations. For the two-sample problem, one assumes that the means are different and that the standard deviations are the same. The difference in means is then estimated. The Behrens-Fisher problem comes about when the standard deviations are assumed unknown and different. In real problems, when a lot of data are available, one does not need probability theory or statistics. The evidence in the data is usually so overwhelming that one can draw correct conclusions without any formal statistical procedures. It is only in the case where the evidence in the data is meager, that one needs any formal statistical theory. But it is precisely in these cases, that the assumptions being made by both the two-sample calculations and the Behrens-Fisher calculations are most questionable. In the next few subsections, the problem of estimating the difference in the means and the ratio of standard deviations is addressed independent of whether or not the other parameter is the same. For the difference in means, this results is a weighted average of the two-sample calculation and the Behrens-Fisher calculation, where the weights are just the probability that the standard deviations are the same or not. A similar result holds for the ratio of the standard deviations.

Estimating The Difference In Means

To estimate the difference in means, one must first introduce this difference into the problem. Defining δ and β to be the difference and sum of the constants A and B, one has

$$\delta = A - B, \qquad \beta = A + B. \tag{23}$$

The two constants, A and B, are then given by

$$A = \frac{\delta + \beta}{2}, \qquad B = \frac{\beta - \delta}{2}. \tag{24}$$

The model equations, Eqs. (1–2), then become

$$d_{1i} = \frac{\delta + \beta}{2} + \text{noise of standard deviation } \sigma_1, \tag{25}$$

and

$$d_{2i} = \frac{\beta - \delta}{2} + \text{noise of standard deviation } \sigma_2. \tag{26}$$

The probability for the difference, δ, is then given by

$$\begin{aligned} P(\delta|D_1D_2I) &= P(\delta v|D_1D_2I) + P(\delta\overline{v}|D_1D_2I) \\ &= P(v|D_1D_2I)P(\delta|vD_1D_2I) + P(\overline{v}|D_1D_2I)P(\delta|\overline{v}D_1D_2I). \end{aligned} \tag{27}$$

This is a weighted average of the probability for the difference in means given that the standard deviations are the same (the two-sample problem) and the probability for the difference in means given that the standard deviations are different (the Behrens-Fisher problem). The weights are just the probabilities that the standard deviations are the same or different. Two of these four probabilities, $P(v|D_1D_2I)$ and $P(\overline{v}|D_1D_2I) = 1 - P(v|D_1D_2I)$, have already been computed, Eqs. (12). The other two probabilities, $P(\delta|vD_1D_2I)$ and $P(\delta|\overline{v}D_1D_2I)$, must now be addressed.

The Two-Sample Problem

$P(\delta|vD_1D_2I)$, is essentially the two-sample problem. This probability is a marginal probability where the standard deviation and β have been removed as nuisance parameters:

$$\begin{aligned} P(\delta|vD_1D_2I) &= \int d\beta d\sigma_1 P(\delta\beta\sigma_1|vD_1D_2I) \\ &\propto \int d\beta d\sigma_1 P(\delta\beta\sigma_1|vI)P(D_1D_2|v\delta\beta\sigma_1I) \\ &= \int d\beta d\sigma_1 P(\delta|I)P(\beta|I)P(\sigma_1|I)P(D_1|v\delta\beta\sigma_1I)P(D_2|v\delta\beta\sigma_1I) \end{aligned} \tag{28}$$

where $P(\delta|I)$ and $P(\beta|I)$ are assigned bounded uniform priors:

$$P(\delta|I) = \begin{cases} \dfrac{1}{2R_a} & \text{If } L - H \leq \delta \leq H - L \\ 0 & \text{otherwise} \end{cases}, \tag{29}$$

and

$$P(\beta|I) = \begin{cases} \dfrac{1}{2R_a} & \text{If } 2L \leq \beta \leq 2H \\ 0 & \text{otherwise} \end{cases}. \tag{30}$$

To evaluate the integral over σ_1 one substitutes Eqs. (8,29,30), and a Gaussian noise prior is used to assign the two likelihoods. Then evaluating the indicated integral, one obtains

$$P(\delta|vD_1D_2I) \propto \frac{\Gamma(N/2)}{8R_a^2\log(R_\sigma)} \int_{2L}^{2H} d\beta V^{-\frac{N}{2}} \left[Q\left(\frac{N}{2}, \frac{V}{\sigma_H^2}\right) - Q\left(\frac{N}{2}, \frac{V}{\sigma_L^2}\right) \right] \tag{31}$$

as the probability for the difference in means given that the standard deviations are the same, where

$$V = \frac{N}{2}\left[\overline{d^2} - 2\delta b - \beta\overline{d} + \frac{\beta^2}{4} + \frac{\delta^2}{4} + \frac{\delta\beta\Delta}{2}\right], \tag{32}$$

$$\Delta = \frac{N_1 - N_2}{N}, \quad \text{and} \quad b = \frac{N_1\overline{d_1} - N_2\overline{d_2}}{2N}. \tag{33}$$

The Behrens-Fisher Problem

The Behrens-Fisher problem is essentially given by $P(\delta|\overline{v}D_1D_2I)$, the probability for the difference in means given that the standard deviations are not the same. This probability is a marginal probability where both the standard deviations and the sum of the means, β, have been removed as nuisance parameters:

$$\begin{aligned} P(\delta|\overline{v}D_1D_2I) &= \int d\beta d\sigma_1 d\sigma_2 P(\delta\beta\sigma_1\sigma_2|\overline{v}D_1D_2I) \\ &\propto \int d\beta d\sigma_1 d\sigma_2 P(\delta\beta\sigma_1\sigma_2|\overline{v}I)P(D_1D_2|\overline{v}\delta\beta\sigma_1\sigma_2I) \\ &= \int d\beta d\sigma_1 d\sigma_2 P(\delta|I)P(\beta|I)P(\sigma_1|I)P(\sigma_2|I) \\ &\quad \times P(D_1|\overline{v}\delta\beta\sigma_1I)P(D_2|\overline{v}\delta\beta\sigma_2I) \end{aligned} \tag{34}$$

where all of the terms appearing in this probability density function have been previously assigned.

To evaluate the integrals over σ_1 and σ_2, one substitutes Eqs. (8,14,29,30) and a Gaussian noise prior is used in the two likelihoods. Evaluating the integrals, one obtains

$$\begin{aligned} P(\delta|\overline{v}D_1D_2I) \propto &\frac{\Gamma(N_1/2)\Gamma(N_2/2)}{16R_a^2[\log(R_\sigma)]^2}\int_{2L}^{2H} d\beta W_1^{-\frac{N_1}{2}}W_2^{-\frac{N_2}{2}} \\ &\times\left[Q\left(\frac{N_1}{2},\frac{W_1}{\sigma_H^2}\right) - Q\left(\frac{N_1}{2},\frac{W_1}{\sigma_L^2}\right)\right]\left[Q\left(\frac{N_2}{2},\frac{W_2}{\sigma_H^2}\right) - Q\left(\frac{N_2}{2},\frac{W_2}{\sigma_L^2}\right)\right] \end{aligned} \tag{35}$$

where

$$W_1 = \frac{N_1}{2}\left[\overline{d_1^2} - \overline{d_1}(\delta+\beta) + \frac{(\delta+\beta)^2}{4}\right], \tag{36}$$

and

$$W_2 = \frac{N_2}{2}\left[\overline{d_2^2} - \overline{d_2}(\beta-\delta) + \frac{(\beta-\delta)^2}{4}\right]. \tag{37}$$

With the completion of this calculation the probability for the difference in means, Eq. (27), is now complete. Before turning our attention to a numerical example there is one final calculation, the probability for the ratio of the standard deviations independent of whether or not the means are the same, that must be completed.

Estimating The Ratio Of The Standard Deviations

To estimate the ratio of the standard deviations, this ratio must be introduced into the problem. Defining r and σ to be

$$r = \frac{\sigma_1}{\sigma_2}, \qquad \sigma = \sigma_2 \tag{38}$$

and substituting these into the model, Eqs. (1,2), one obtains

$$d_{1i} = A + \text{noise of standard deviation } r\sigma, \tag{39}$$

and

$$d_{2i} = B + \text{noise of standard deviation } \sigma. \tag{40}$$

The probability for the ratio of the standard deviations, $P(r|D_1D_2I)$, is then given by

$$\begin{aligned} P(r|D_1D_2I) &= P(rs|D_1D_2I) + P(r\overline{s}|D_1D_2I) \\ &= P(s|D_1D_2I)P(r|sD_1D_2I) + P(\overline{s}|D_1D_2I)P(r|\overline{s}D_1D_2I). \end{aligned} \tag{41}$$

This is a weighted average of the probability for the ratio of the standard deviations given the means are the same plus the probability for the ratio of the standard deviations given that the means are different. The weights are just the probabilities that the means are the same or not. Two of the four probabilities, $P(s|D_1D_2I)$ and $P(\overline{s}|D_1D_2I) = 1 - P(s|D_1D_2I)$, have already been computed, Eqs. (11). The other two probabilities, $P(r|sD_1D_2I)$ and $P(r|\overline{s}D_1D_2I)$, must now be addressed.

Estimating The Ratio, Given The Means Are The Same

The first term to be addressed is $P(r|sD_1D_2I)$. This probability is a marginal probability where both σ and A have been removed as nuisance parameters:

$$\begin{aligned} P(r|sD_1D_2I) &= \int dAd\sigma P(rA\sigma|sD_1D_2I) \\ &\propto \int dAd\sigma P(rA\sigma|sI)P(D_1D_2|srA\sigma I) \\ &= \int dAd\sigma P(r|I)P(A|I)P(\sigma|I)P(D_1|srA\sigma I)P(D_2|srA\sigma I) \end{aligned} \tag{42}$$

where the prior probability for the ratio of the standard deviations is taken to be a bounded Jeffreys prior:

$$P(r|I) = \begin{cases} \dfrac{1}{2r\log(R_\sigma)} & \text{If } \sigma_L/\sigma_H \le r \le \sigma_H/\sigma_L \\ 0 & \text{otherwise} \end{cases} . \tag{43}$$

To evaluate the integral over A, one substitutes Eq. (8,14,43), and a Gaussian noise prior probability is used to assign the two likelihoods. Evaluating the integral, one obtains

$$P(r|sD_1D_2I) = \frac{\sqrt{\pi/8w}\; r^{-N_1-1}}{R_a[\log(R_\sigma)]^2} \int_{\sigma_L}^{\sigma_H} d\sigma \exp\left\{-\frac{x}{2\sigma^2}\right\} \left[Q\left(\frac{1}{2}, \frac{X_L^2}{\sigma^2}\right) \pm Q\left(\frac{1}{2}, \frac{X_H^2}{\sigma^2}\right)\right] \tag{44}$$

where

$$X_H = \sqrt{\frac{w}{2}}[A_H - v/w], \qquad X_L = \sqrt{\frac{w}{2}}[A_L - v/w], \tag{45}$$

$$u = \frac{N_1\overline{d_1^2}}{r^2} + N_2\overline{d_2^2}, \qquad v = \frac{N_1\overline{d_1}}{r^2} + N_2\overline{d_2}, \tag{46}$$

$$w = \frac{N_1}{r^2} + N_2, \qquad x = u - \frac{v^2}{w}, \tag{47}$$

and the sign is again chosen to be a minus if X_L and X_H are of the same sign, and plus if they are of different sign.

Estimating The Ratio, Given The Means Are Different

The second term that must be computed is $P(r|\overline{s}D_1D_2I)$, the probability for the ratio of standard deviations given that the means are not the same. This is a marginal probability where σ, A, and B have been removed as nuisance parameters:

$$\begin{aligned} P(r|\overline{s}D_1D_2I) &= \int dAdBd\sigma P(rAB\sigma|\overline{s}D_1D_2I) \\ &\propto \int dAdBd\sigma P(rAB\sigma|\overline{s}I)P(D_1D_2|\overline{s}rAB\sigma I) \\ &= \int dAdBd\sigma P(r|I)P(A|I)P(B|I)P(\sigma|I)P(D_1|\overline{s}A\sigma I)P(D_2|\overline{s}B\sigma) \end{aligned} \tag{48}$$

where all of the terms appearing in this probability density function have been previously assigned.

To evaluate the integral over A and B one substitutes Eq. (7,8,18,43), and a Gaussian noise prior is used in assigning the two likelihoods. Evaluating the indicated integrals, one obtains

$$\begin{aligned} P(r|\overline{s}D_1D_2I) \propto\ & \frac{\pi}{4R_a^2[\log(R_\sigma)]^2\sqrt{N_1N_2}} \int_{\sigma_L}^{\sigma_H} d\sigma r^{-N_1}\sigma^{-N+1}\exp\left\{-\frac{z_1}{2r^2\sigma^2}-\frac{z_2}{2\sigma^2}\right\} \\ & \times\left[Q\left(\frac{1}{2},\frac{U_{1L}^2}{r^2\sigma^2}\right) \pm Q\left(\frac{1}{2},\frac{U_{1H}^2}{r^2\sigma^2}\right)\right]\left[Q\left(\frac{1}{2},\frac{U_{2L}^2}{\sigma^2}\right) \pm Q\left(\frac{1}{2},\frac{U_{2H}^2}{\sigma^2}\right)\right] \end{aligned} \tag{49}$$

where the minus sign is chosen when U_{1H} and U_{1L} have the same sign and the plus sign is chosen when the sign of U_{1H} and U_{1L} differ and similarly for U_{2H} and U_{2L}.

4. Numerical Example

With the completion of this calculation, the probability for all of the hypotheses appearing in Table 1 have been computed. It is now time to apply these calculations in an example. The example used is taken from Jaynes [14,15]. This article is a series of examples contrasting orthodox and Bayesian methods. In one example, Jaynes applied both the two-sample and the Behrens-Fisher calculations to the same sets of data. This example will be extended here using the procedures developed in this paper. The data in this example are the mean lifetimes and standard deviations for a certain component from two different manufactures. Manufacturer A supplies 9 units for test, which turn out to have a (mean ± standard deviation) lifetime of (42 ± 7.48) hours. Manufacturer B supplies 4 units, which yield (50 ± 6.48) hours. The calculations presented here will first determine if the data sets differ; then if they differed, how; and last, given that they differed, the magnitude of the difference will be estimated.

In the analysis performed by Jaynes the nuisance parameters were removed using improper prior probabilities. This could be done, because the problem was treated as a parameter estimation problem, and the infinites introduced cancel when the distributions were normalized. However, in the calculation presented here, improper priors cannot be used; because the infinites do not cancel. The prior range on the amplitudes will be taken as $34 \leq A, B \leq 58$ and for the standard deviations $3 \leq \sigma_1, \sigma_2 \leq 10$. Because the data and the type of components are not stated, assigning these prior ranges is more difficult than normal. Consequently, at the end of this example, the calculations will be repeated using wider ranges to see what effect this has on the conclusions.

The data consists of the means and the standard deviations of each data set as well as the number of data values. But how the standard deviations were computed was not given. Here it will be assumed that $(N-1)$ was in the standard deviation calculation. With this assumption the mean-square data values may be computed, and thus, all of the calculations presented in this paper may be performed. These calculations have been implemented as a general fortran program that will analyze any two-sample or Behrens-Fisher like data sets. This program is available from the author. The output from this program is given in Table 2 for the data presented in this example.

Numerical Example – Did Something Change?

The first question of interest is whether or not the data sets are the same. Jaynes, essentially takes it as a given that the data sets differ. He indicates "I think our common sense tells us immediately, without any calculation, that this constitutes fairly substantial (but not overwhelming) evidence in favor of B." To arrive at this conclusion, one must first conclude that the data sets are different, and second that they differ in the means, independent of whether or not the standard deviations are the same. In the analysis done by Jaynes, the calculations are first done using the assumption that the standard deviations are different, from which he demonstrated that there is a 0.92 probability that $B > A$. And he went on to demonstrate that essentially the same conclusions would be drawn when the standard deviations are assumed equal.

In this paper, the probability that the data sets are the same, has been explicitly computed, Eq. (3). Similarly the probability that the means differed is given by $P(\overline{s}|D_1D_2I) = 1 - P(s|D_1D_2I)$, and $P(s|D_1D_2I)$ is given by Eq. (11) and these probabilities are given in Table 2. Consulting Table 2, the probability the data sets are different is 0.83, thus supporting Jaynes' conclusion that the data sets differed. And confirms Jaynes' intuition that it is good evidence, but not overwhelming.

Numerical Example – What Changed?

Now that one knows that the data sets are not the same, or at the very least are probably not the same, one would like to know what changed. Did the means change or did the standard deviations change? Consulting Table 2, the probability that the means are different is 0.72. So Jaynes' conclusion that the means are different is being supported.

Table 2: Numerical Example

```
Enter The Amplitude Lower Bound:  34
Enter The Amplitude Upper Bound:  58

Enter The Variance Lower Bound:  3
Enter The Variance Upper Bound:  10

   No.    Standard Deviation    Average      Data Set
    4          6.4800           50.000       Jaynes.1
    9          7.4800           42.000       Jaynes.2
   13          7.9099           44.462       Combined

---------------Model----------------   Probability
Same Constant,   Same Variance         0.1689119
Different const, Same Variance         0.4153378
Same Constant,   Different Variances   0.1077223
Different Const, Different Variances   0.3080280

The probability the constants are the same is:  0.27663
The probability the constants are different is:  0.72337
The odds ratio is 2.61 to 1 in favor of different constants

The probability the variances are the same is:  0.58425
The probability the variances are different is:  0.41575
The odds ratio is 1.41 to 1 in favor of the same variances

The probability the data sets are the same is:  0.16891
The probability the data sets are different is:  0.83109
The odds ratio is 4.92 to 1 in favor of different
    means or variances
```

Table 2 summarizes the output from the fortran implementation of this calculation. This program produces three different types of output: (1) the probability for the four hypotheses examined by Dayal and Dickey; (2) the probability that the data sets differ, the probability that the means are different and the probability that the variances are different; and finally (3) the probability for the difference in means and the ratio of the standard deviations is computed – see Figs. 1 and 2.

In the calculations performed by Jaynes both a two-sample like calculation and a Behrens-Fisher like calculation were performed to indicate the relative evidence in favor of the hypothesis that the means were different. But to perform a two-sample calculation one must assume that the standard deviations are the same. Similarly to perform a Behrens-Fisher calculation, one must assume that the standard deviations are different. The probability that the standard deviations are the same is also given in Table 2. Consulting Table 2, one finds the probability that the standard deviations are the same is 0.58; neither confirming nor denying the hypothesis. So neither a two-sample calculation nor a Behrens-Fisher calculation is justified for this data; the data do not support either hypothesis. Given how close these probabilities are to 50/50 one might expect that a weighted average of the results from a two-sample calculation and a Behrens-Fisher calculation would be a better indicator of the difference in means, and this is exactly what probability theory tells one to do.

Numerical Example – Estimating The Changes

The probability that the means are different is 0.72, indicating good but not overwhelming evidence in favor of different means. In this subsection it will be assumed that the means are different and the problem to be addressed is one of estimating the magnitude of this difference. The probability for the difference in means is given by Eq. (27). To compute this probability both a two-sample calculation and a Behrens-Fisher calculation must be performed. So it is easy to have the program report the two-sample calculation (dotted line, Fig. 1), the Behrens-Fisher calculation (dashed line, Fig. 1), and weighted average derived in this paper (solid line, Fig. 1).

The two-sample calculation assumes the standard deviations are the same. There is only a 0.58 probability of this, so the two-sample model must not fit the data much differently than a Behrens-Fisher model. Here the two-sample model estimates the standard deviation of the noise to be higher than the Behrens-Fisher calculation, because the pooled standard deviation is larger than that of either data set separately. Consequently, this distribution is more spread out, less certain, of the difference (dotted line, Fig. 1). The Behrens-Fisher calculation has a standard deviation for each data set, and can reduce the overall estimate of the noise; so the Behrens-Fisher distribution is more sharply peaked (dashed line, Fig. 1). But probability theory tells one to use a weighted average of these two distributions. Here both models fit the data about equally well. Under these conditions probability theory will prefer the simpler model. The probabilities that the standard deviations are the same is 0.58 and 0.42 that they are different. So the weighted average follows the two-sample calculation a little more closely than the Behrens-Fisher calculation (solid line, Fig. 1).

Numerical Example – The Ratio Of Standard Deviations

When estimating the difference in means, there was not much evidence in favor of different standard deviations. Consequently one would not expect either the two-sample calculation or the Behrens-Fisher calculation to be very different and this is exactly what Fig. 1 shows. But there is fairly strong evidence in favor of the means being different. If the problem is to estimate the ratio of the standard deviations, one would expect the two calculations to be substantially different. That is to say, the probability for the ratio of

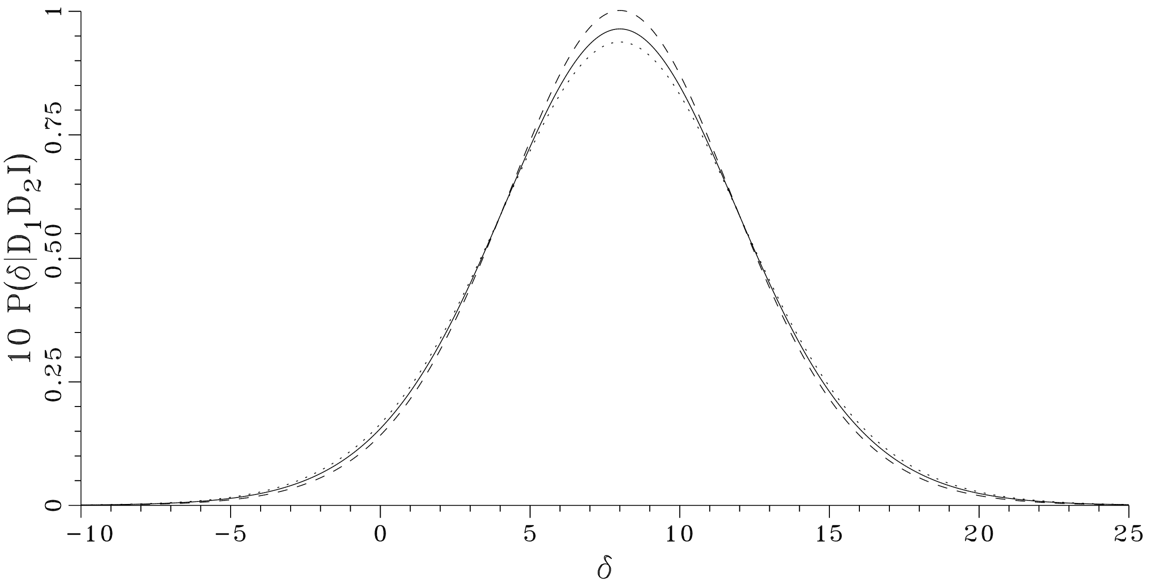

Fig. 1. Three probability density function are shown: (1) the probability for the difference in means given the standard deviations are the same (dotted line), (2) the probability for the difference in means given that the the standard deviations are different (dashed line), and (3) the probability for the difference in means independent of whether or not the standard deviation are the same (solid line).

the standard deviations given the same means should be substantially different from the probability for the ratio of the standard deviations given that the means are different.

These two distributions, as well as the weighted average are shown in Fig. 2. The probability for the ratio of the standard deviations given the means are the same is shown as the dotted line. This model does not fit the data well (the pooled data have a larger standard deviation than either data set separately). Consequently, the uncertainty in this probability distribution is large compared to the other models and the distribution is more spread out. The probability for the ratio of standard deviations given different means is shown as the dashed line. This model fits the data better, and results in a more strongly peaked probability distribution. But probability theory tells one to take a weighted average of these two distributions, solid line. The weights are just the probabilities that the means are the same or different. Here those probabilities are 0.28 and 0.72 respectively. So the weighted average follows $P(r|\overline{s}D_1D_2I)$ more closely than $P(r|sD_1D_2I)$.

Numerical Example – The Effect Of The Prior Ranges

It has been noted several times that the prior ranges do not cancel. This occurs when models are being considered that contain either different types of parameters or different numbers of parameters of the same type. Here the models all contain the same types of parameters, constants and standard deviations, but they contain differing numbers of these parameters. Consequently the prior ranges are important and will affect the conclusions.

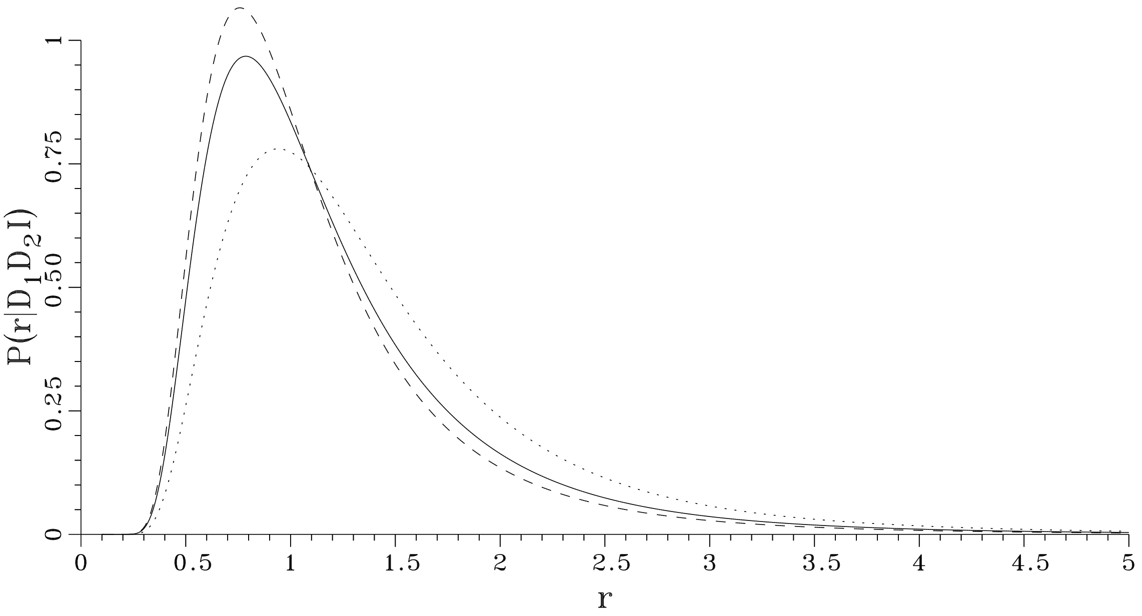

Fig. 2. Three probability density functions are shown: (1) the probability for the ratio of standard deviations given that the means are the same (dotted line), (2) the probability for the ratio of standard deviations given that the means are different (dashed line), (3) the probability for the ratio of standard deviations independent whether or not the means are the same (solid line).

In the example just given the prior range on the constants, A and B, was $42 - 8 \leq A, B \leq 50 + 8$, where 8 was a rough estimate of the standard deviation. This assumes, in effect, that all of the data values are roughly within one standard deviation of the mean. This would be an unreasonable assumption if the number of data items were large, but with only 4 and 9 data items, who knows? The question investigated here is just how strongly do the conclusions depend on the prior ranges?

Suppose that $8 \rightarrow 16$, so that the prior ranges become $26 \leq A, B \leq 66$. What effect would this have on the conclusions? Rerunning the program using these ranges (for the time being holding the prior ranges on the standard deviations constant) one finds the probability that the data sets are different is 0.765. This compared to 0.83, previously. So changing the uncertainty in the amplitudes by a factor of 2, has lowered this probability by roughly 8%; a small change. Second, the probability that the standard deviations changed is now 0.411, compared to 0.415 previously; only a 1% change. And third, the probability that the means are different is 0.61, compared to 0.72 previously; an 18% change. In all cases the results have changed slightly, but not nearly by a factor of 2 and *none of the conclusions were changed.*

Now, suppose that the prior range on the standard deviations were changed by a factor of 2: $(3 \leq \sigma_1, \sigma_2 \leq 10) \rightarrow (3/\sqrt{2} \leq \sigma_1, \sigma_2 \leq 10 * \sqrt{2})$; what would the effect of this be? Here the prior range on the constant will be returned to their original value. Rerunning the program, one obtains a probability of 0.81 that the data sets are different. This compares to 0.83 obtained previously, about a 2% change. For the probability the means are different, one obtains 0.695 compared to 0.72 previously, about a 3.6% change. And last the probability that the standard deviations are different was 0.38 compared to 0.41, about an 8% change. Again none of the major conclusions are changed.

Apparently, the effect of these prior ranges is small. Here the size of the effect is of the order of the square root of the logarithm of change or smaller. The magnitude of the effect depends on just how strongly a proposition is being supported by the data: when the evidence in the data is large, changing the prior ranges by factors of thousands has essentially no effect on the conclusions. The only reason one could see an effect here was that the evidence in the data was small and then the prior information is important.

5. Summary And Conclusions

The calculations presented in this paper are generalizations of the Behrens-Fisher and two-sample problems and the traditional F and student t-distributions. They allow the experimenter to investigate the problems of interest in a way never before possible. First, they allow one to determine if the data sets differ. Second, they allow one to determine how they differ; either in the means or in the standard deviations. And third, one can estimated the magnitude of these changes without additional assumptions. The difference in mean can be estimated independent of whether or not the standard deviations are the same or not; while the ratio of the standard deviations can be estimated independent of whether or not the means are the same. Last, these probability density functions allow one to picture the results in a way never before possible: not by stating the results as a single number, but graphically; so that one can see the evidence with ones own eyes.

ACKNOWLEDGMENTS. The encouragement of Professor J. J. H. Ackerman is greatly appreciated as are extensive conversations with Professor E. T. Jaynes. This work was partially supported by a gift from the Monsanto Company and NIH grant GM30331, J. J. H. Ackerman principal investigator.

REFERENCES

[1] Behrens, W. V., "Ein Beitrag zur Fehlerberechnung bei weinige Beobachtungen," *Landwirtschaftliche Jahrbücher*, **68**, pp. 807-837 (1929).

[2] Fisher, R. A,. "The comparison of samples with possibly unequal variances," *Ann. of Eugenics,* **9**, pp. 174-180 (1937).

[3] Fisher, R. A., *Statistical Methods and Scientific Inference,* Hafner Publishing Co., New York (1956).

[4] Jeffreys, H., *Theory of Probability,* Oxford University Press, London (1939).

[5] Lee, Peter M., *Bayesian Statistics: An Introduction,* Oxford University Press, New York (1989).

[6] Patil, V. H., "The Behrens-Fisher problem and its Bayesian solution," *Journal of Indian Statistical Assoc.,* **2** pp. 21-31 (1964).

[7] Robinson, G. K., "Properties of Student's t and of the Behrens-Fisher solution to the two means problem." *Ann. Statist.,* **4** (1976).

[8] Robinson, G. K., "Properties of Student's t and of the Behrens-Fisher solution to the two means problem." *Ann. Statist.,* **10** (1982).

[9] Roberts, Norman A., *Mathematical Methods in Reliability Engineering,* McGraw-Hill Book Co. Inc., New York pp.86-88 (1964).

[10] Smith, H. F., "The Problem of Comparing the Results of Two Experiments With Unequal Errors," *J. Sci. Ind. Research* (*India*), **9**, pp. 211-212 (1936).

[11] Scatterthwaite, F. E., "An Approximate Distribution Of Estimates Of Variance Components," *Biometrics Bull.,* **2**, pp. 110-114 (1946).

[12] Dayal, Hari H., "Bayesian statistical inference in Behrens-Fisher Problems," Ph.D. dissertation, State University of New York at Buffalo (1972).

[13] Dayal, Hari H., and James M. Dickey "Bayes Factors for Behrens-Fisher Problems" *The Indian Journal of Statistics,* **38** pp. 315-328 (1976).

[14] Jaynes, E. T. "Confidence Intervals vs Bayesian Intervals," in *Foundations of Probability Theory, Statistical Inference, and Statistical Theories of Science,* **2**, pp. 175-257 (1976).

[15] Jaynes, E. T., *Papers on Probability, Statistics and Statistical Physics,* a reprint collection, D. Reidel, Dordrecht the Netherlands, 1983)