Statistica Sinica 1(1991), 65-92

SIGNIFICANCE LEVELS FROM REPEATED p-VALUES WITH MULTIPLY-IMPUTED DATA

Kim-Hung Li, Xiao-Li Meng, T. E. Raghunathan and Donald B. Rubin

Chinese University of Hong Kong, Harvard University,
University of Washington and Harvard University

Abstract: Multiple imputation is becoming a standard tool for handling nonresponse in sample surveys. A difficult problem in the analysis of a multiply-imputed data set concerns how to combine repeated p-values efficiently to create a valid significance level. Here we propose, justify, and evaluate the validity of a new procedure, which is superior to the current standard. This problem is inherently difficult when the number of multiple imputations is small, as it must be in common practice, as made clear by its close relationship to a multivariate version of the classic Behrens-Fisher problem with small degrees of freedom.

Key words and phrases: Multiple imputation, missing data, nonresponse, surveys, Bayesian inference, Behrens-Fisher problem, incomplete data, hypothesis testing.

1. Introduction and Background

Multiple imputation, first proposed in Rubin (1978), is a general statistical technique for handling missing data. It is particularly suited to handling nonresponse in large sample surveys, especially those surveys that are designed to produce public-use data bases to be shared by many users. In such surveys, it is tempting to impute (or fill in) the missing values due to nonresponse so that all users can work with a complete data base. The key idea of multiple imputation, in contrast to single imputation, is to replace the set of missing values with $m(\geq 2)$ sets of plausible values. Each of these m resulting completed data sets is then analyzed using standard complete-data methods. These analyses are combined to create one repeated-imputation inference, which takes proper account of the uncertainty due to missing data. In this and other ways multiple imputation retains the major advantages and rectifies the major disadvantages inherent in single imputation. For a comprehensive treatment of multiple imputation and comparisons with single imputation, readers are referred to Rubin (1987), and specifically for comparisons of p-values using single and multiple imputation, to Li, Raghunathan and Rubin (1990).

All of the multiple imputation procedures described there perform much better than single imputation procedures. Of these procedures, the least successful, although adequate in most situations, is the one for combining m p-values (or χ^2 test statistics) to obtain one valid significance level. To illustrate, suppose we wish to test a ten-component regression coefficient when there is 30% missing information (defined precisely in Section 1.4). Using single imputation, the rejection rate of a nominal 5% test is approximately 45% whereas the rejection rate of the current standard procedure is 10%–not perfect but clearly an improvement over single imputation.

Here we derive and describe a new way to combine p-values that is as simple yet better than the previous standard. In particular, in the illustrative example just discussed, the rejection rate is 6%. Related improvements for obtaining p-values from repeated estimates and variance-covariance matrices are more straightforward and are reported in Li, Raghunathan and Rubin (1990). In the remainder of Section 1, we provide the necessary background material and notation, including discussion of the inherent difficulty of this problem and its relationship with the well-known Behrens-Fisher problem. Section 2 presents our procedure, and gives its theoretical justification. Section 3 evaluates its validity. Finally, in Section 4 we discuss other approaches to this problem, and summarize our conclusion that the new procedure has the simplest form among all previous procedures that give accurate levels. We also provide practical advice on its use, and discuss related research.

1.1. The complete-data case

Let $X = [x_1, \ldots, x_n]^t$ be the $n \times p$ complete-data matrix, with the associated density $f(x|\theta)$, where the parameter θ is a k-dimensional vector. Assume that with the complete data the asymptotically valid and efficient inference for θ would be based on the statement

$$U^{-1/2}(\hat{\theta} - \theta) \sim N(0, I), \tag{1.1}$$

where $\hat{\theta} = \hat{\theta}(X)$ is an efficient estimate of θ and $U = U(X)$ is the associated variance-covariance matrix. For example, in a frequentist analysis, one may take $\hat{\theta}$ to be the maximum likelihood estimate of θ and U to be the inverse of the observed information matrix. Then (1.1) can be interpreted as asserting that the sampling distribution of $\hat{\theta}$ is approximately normal with mean θ and variance-covariance U. In contrast, in Bayesian analysis, $\hat{\theta}$ and U are the posterior mean and variance-covariance of θ, respectively, where (1.1) is the large sample normal approximation to the posterior distribution of θ. Thus, approximation (1.1) is commonly acceptable to both frequentists and Bayesians.

In practice, especially in multiparameter cases, the evidence about the likely values of θ is often summarized by a p-value for a specified "null" value of θ, say θ_0. As a direct consequence of approximation (1.1), this p-value can be calculated as

$$P_x = \Pr[\chi_k^2 > kD_x], \tag{1.2}$$

where χ_k^2 is a chi-square random variable with k degrees of freedom, and D_x is the observed value of the quadratic form

$$D(X) = (\theta_0 - \hat{\theta})^t U^{-1} (\theta_0 - \hat{\theta})/k, \tag{1.3}$$

which is proportional to the Wald test statistic. Again, this p-value has both frequentist and Bayesian interpretations. A frequentist would interpret it as the probability of observing D_x or more extreme values of $D(X)$ when $\theta = \theta_0$, in an imagined long sequence of identical experiments that produce new values of $\hat{\theta}$ and U. It is also a Bayesian p-value in the sense that the $(1-p)100\%$ HPD (highest posterior density) region will just include θ_0.

The use of p-values rather than interval estimates is especially useful and common in highly multiparameter models, such as large log-linear models for contingency tables, where the parameter θ consists of all high-dimensional interactions, and the p-value summarizes evidence about the acceptability of a parsimonious model that includes, for example, only main effects and two-way interactions. In such cases, the dimensionality of θ, k, can be in the hundreds.

An important point is that whenever the complete-data likelihood ratio test has the form of (1.2), our approach can be used. Therefore, even if the test is not against a single null value of θ, but rather against a collection of null values, as with a composite hypothesis, our method of combining significance levels still can be applied. The reason is that the only distributional results explicitly used are the χ_k^2 reference distributions.

1.2. The incomplete-data case

In common survey practice, it is very likely that we can only observe part of the complete data X for various reasons (see, for example, Rubin (1987), Chapter 1). Denote this observed part by $X_{\text{obs}} = \{X_{ij} | R_{ij} = 1\}$, where $R_{ij} = 1$ if the (i,j)th element of X is observed, and zero otherwise. We also assume that the missing data mechanism is ignorable:

$$\Pr(R|X) = \Pr(R|X_{\text{obs}}),$$

where we suppress possible dependence of this distribution on unknown parameters distinct from θ.

This ignorability assumption is not strictly necessary (Rubin (1987), Chapter 4), but for simplicity of presentation, we will assume it. Thus, the relevant likelihood function for drawing inferences about θ is

$$L(\theta|X_{\text{obs}}) \propto \int f(X|\theta)dX_{\text{mis}},$$

where $X_{\text{mis}} = \{X_{ij}|R_{ij} = 0\}$ is the collection of missing values.

In analogy with the complete-data inference (1.1), we assume the following large-sample approximation holds

$$T^{-1/2}(\hat{\theta}_{\text{obs}} - \theta) \sim N(0, I), \tag{1.4}$$

where $\hat{\theta}_{\text{obs}} = \hat{\theta}(X_{\text{obs}})$ is an efficient estimate of θ based on the observed data (e.g., observed-data MLE or posterior mean), and $T = T(X_{\text{obs}})$ is the associated variance-covariance matrix (e.g., inverse of the observed-data information matrix or the posterior variance of θ). Consequently, the p-value for summarizing evidence about a null value θ_0 can be calculated as

$$P_{\text{obs}} = \Pr[\chi^2_k > kD_{\text{obs}}], \tag{1.5}$$

where

$$D_{\text{obs}} = (\hat{\theta}_{\text{obs}} - \theta_0)^t T^{-1}(\hat{\theta}_{\text{obs}} - \theta_0)/k \tag{1.6}$$

is proportional to the observed value of the Wald test statistic based on the observed data.

With the complete data X, the computation of $\hat{\theta}(X)$ and $U(X)$ usually is straightforward, either analytically or numerically. But with the incomplete data, it can become troublesome and even intractable, which is especially burdensome in contexts of public-use files.

1.3. Basic distributional results for multiple imputation

Multiple imputation is an efficient and valid way of handling incomplete data problems using only standard complete-data techniques once data are imputed. More specifically, m imputations for the missing values X_{mis} are created, $X^{(\ell)}_{*\text{mis}}$, $\ell = 1, \ldots, m$, and these are used by the data analyst to create m corresponding completed data sets

$$\{X^{(\ell)}_*; \ell = 1, \ldots, m\} = \{(X_{\text{obs}}, X^{(\ell)}_{*\text{mis}}); \ell = 1, \ldots, m\}.$$

The data analyst then conducts m standard complete-data analyses to compute the corresponding values for $\hat{\theta} = \hat{\theta}(X)$ and $U = U(X)$: $\hat{\theta}_{*\ell} = \hat{\theta}(X^{(\ell)}_*)$ and $U_{*\ell} = U(X^{(\ell)}_*)$ $(\ell = 1, \ldots, m)$. Letting the set of completed-data moments be

$$S_m = \{(\hat{\theta}_{*\ell}, U_{*\ell}); \ell = 1, \ldots, m\},$$

the data analyst can combine the statistics in the set S_m to obtain one inference for θ.

When using "proper" imputation methods (for the definition of "proper", see Rubin (1987), Chapter 4), the following approximations concerning the distribution of $(\hat{\theta}_{*\ell}, U_{*\ell})$ $(l = 1, \ldots, m)$ are justifiable (Rubin (1987, Chapter 4), Raghunathan (1987), and Schenker and Welsh (1988)):

$$\hat{\theta}_{*\ell}|X_{\text{obs}}, \theta \sim \text{i.i.d.}\, N(\hat{\theta}_{\text{obs}}, B) \tag{1.7}$$

and

$$U_{*\ell}|X_{\text{obs}}, \theta \approx \bar{U}, \tag{1.8}$$

where

$$B = B(X_{\text{obs}}) = V[\hat{\theta}(X)|X_{\text{obs}}, \theta], \tag{1.9}$$

$$\bar{U} = \bar{U}(X_{\text{obs}}) = E[U(X)|X_{\text{obs}}, \theta], \tag{1.10}$$

and $\approx$ in (1.8) means equal in the sense of lower order variability. Thus, in view of (1.4), (1.7), and (1.8), the sufficient statistics for inference about θ are

$$\bar{\theta}_m = \frac{1}{m}\sum_{\ell=1}^{m} \hat{\theta}_{*\ell}, \tag{1.11}$$

$$B_m = \frac{1}{m-1}\sum_{\ell=1}^{m} (\hat{\theta}_{*\ell} - \bar{\theta}_m)(\hat{\theta}_{*\ell} - \bar{\theta}_m)^t \tag{1.12}$$

and

$$\bar{U}_m = \frac{1}{m}\sum_{\ell=1}^{m} U_{*\ell}. \tag{1.13}$$

Here B_m measures "between imputation" variability and $\bar{U}_m$ measures "within imputation" variability. Bayesian justification for these statistics can be found in Rubin (1987, Chapter 3).

1.4. Fractions of missing information

The crucial measure of the extent to which inference is hindered by the occurrence of missing data is the fraction of missing information, which is the proportionate increase in variance due to missing values. It is especially easy to make this idea clear from a Bayesian perspective. Specifically, the posterior variance of θ based on the observed data, X_{obs}, can be written as

$$V(\theta|X_{\text{obs}}) = E[V(\theta|X)|X_{\text{obs}}] + V[E(\theta|X)|X_{\text{obs}}], \tag{1.14}$$

where the first term on the right hand side is the expected posterior variance of θ when there are no missing data, equal asymptotically to $E(U|X_{\text{obs}})$, and the second term on the right hand side is the expected increase in posterior variance due to missingness. Hence, the eigenvalues of $V[E(\theta|X)|X_{\text{obs}}]$ relative to $V(\theta|X_{\text{obs}})$ can be interpreted as measuring the increases in posterior variance due to missing data – the fractions of missing information. Small values of these eigenvalues imply less loss of precision. In particular, in the scalar case, the fraction of missing information is simply the ratio of $V[E(\theta|X)|X_{\text{obs}}]$ to $V(\theta|X_{\text{obs}})$.

In large samples, the posterior variance $V(\theta|X_{\text{obs}})$ will equal the inverse of the negative second derivative of the observed-data log-likelihood, defined in (1.4) to be $T = T(X_{\text{obs}})$. Similarly, the complete-data posterior variance $V(\theta|X)$ will equal the inverse of the negative second derivative of the complete-data log-likelihood, defined in (1.1) to be $U = U(X)$. Thus, the first term on the right hand side of (1.14) is asymptotically equal to $\bar{U}$ defined in (1.10).

From the frequentist perspective with θ_t equal to the true value of θ, asymptotically we have

$$(\bar{U}|\theta = \theta_t) \approx U_t \tag{1.15}$$

where

$$U_t = V(\hat{\theta}|\theta = \theta_t).$$

Finally, the second term on the right hand side of (1.14) is approximately B defined in (1.9); in large samples

$$(B|\theta = \theta_t) \approx B_t,$$

where

$$B_t = T_t - U_t$$

with

$$T_t = V(\hat{\theta}_{\text{obs}}|\theta = \theta_t).$$

The eigenvalues of B_t relative to T_t give the population fractions of missing information, which we denote by $\gamma = (\gamma_1, \ldots, \gamma_k) \in [0,1)^k$. Thus, γ gives the increases in variance of parameter estimates due to missing data when $\theta = \theta_t$. Because of the lower order variability of B and $\bar{U}$, γ may also be taken to be the eigenvalues of B with respect to $T = (\bar{U} + B)$, with average fraction of missing information $\bar{\gamma} = \sum_j \gamma_j/k$. Notationally, we also let $\lambda_j = \gamma_j/(1-\gamma_j)$, $j = 1, \ldots, k$, be the eigenvalues of B_t (or B) relative to U_t (or $\bar{U}$), where $\lambda = (\lambda_1, \ldots, \lambda_k)$ and $\bar{\lambda} = \sum_j \lambda_j/k$. A final set of measures are the ratios of complete to observed information, that is, $\xi_i = 1 + \lambda_i = (1-\gamma_i)^{-1}$, with $\xi = (\xi_1, \ldots, \xi_k)$ and $\bar{\xi} = \sum_j \xi_j/k$.

1.5. p-values based on S_m

Having obtained the collection of completed-data moments $S_m = \{(\hat{\theta}_{*\ell}, U_{*\ell}), \ell = 1, \ldots, m\}$, we can obtain a p-value for $\theta = \theta_0$,

$$P_m = \Pr(F_{k,w} > D_m), \tag{1.16}$$

where the test statistic D_m is a scaled distance between the estimate $\bar{\theta}_m$ of the parameter θ and the hypothesized value, θ_0, based on the sufficient statistics (1.11-1.13), and $F_{k,w}$ is an F reference distribution with k and w degrees of freedom, with $w \to \infty$ as $m \to \infty$. An excellent choice for D_m, which works well for all values of m, is

$$D_m = (\bar{\theta}_m - \theta_0)^t \bar{U}_m^{-1} (\bar{\theta}_m - \theta_0)/[k(1 + r_m)], \tag{1.17}$$

where

$$r_m = (1 + m^{-1}) \operatorname{tr}(B_m \bar{U}_m^{-1})/k. \tag{1.18}$$

The F reference distribution can be justified from both the frequentist and Bayesian perspectives, and various specific choices of denominator degrees of freedom, w, have been proposed (Li (1985), Rubin (1987) and Raghunathan (1987)). Among them, the best one so far is proposed in Li, Raghunathan and Rubin (1990):

$$w = 4 + (\nu - 4)\left[1 + (1 - \frac{2}{\nu})/r_m\right]^2, \tag{1.19}$$

where $\nu = k(m - 1)$.

Notice that this procedure requires $\nu > 4$. When $\nu \leq 4$, an alternative denominator degrees of freedom $(m - 1)(k + 1)(1 + r_m^{-1})^2/2$, proposed by Rubin (1987), can be used. The major conclusion of Li, Raghunathan and Rubin (1990) is that this procedure is very well calibrated for all $m \geq 3$ in cases of practical interest, and the loss of power due to finite m is quite modest in cases likely to occur in practice.

When $m \to \infty$, D_m given by (1.17) tends to

$$D_m^{\infty} = (\hat{\theta}_{\text{obs}} - \theta_0)^t \bar{U}^{-1} (\bar{\theta}_{\text{obs}} - \theta_0)/(k\bar{\xi}). \tag{1.20}$$

When all $\xi_i = \bar{\xi} = 1 + \bar{\lambda}$, it is easy to show that $D_m^{\infty} = D_{\text{obs}}$, the ideal test statistic. When the ξ_i vary, Li, Raghunathan and Rubin (1990) show that D_m^{∞} is very close to D_{obs} under most practical circumstances, the difference being governed essentially by the coefficient of variation of the ξ_i, that is, the coefficient of variation of the ratios of the complete to observed information.

1.6. p-values based on S_d

Clearly, the procedures given in Section 1.5 require access to the collection of completed-data moments $S_m = \{(\hat{\theta}_{*\ell}, U_{*\ell}),\ \ell = 1, \ldots, m\}$. In practice, the dimension k is often large, as when social scientists attempt to find parsimonious models from public-use data bases with hundreds of variables and thousands of sampling units. In such cases, the standard complete-data analysis may only provide the collection of completed-data χ^2 statistics (or distances) $S_d = \{d_{*1}, \ldots, d_{*m}\}$, where, asymptotically,

$$d_{*\ell} = (\theta_0 - \hat{\theta}_{*\ell})^t U_{*\ell}^{-1} (\theta_0 - \hat{\theta}_{*\ell}); \tag{1.21}$$

the corresponding p-value in the ℓth complete-data set is $\Pr(\chi_k^2 \geq d_{*\ell})$ $(\ell = 1, \ldots, m)$. Consequently, finding the p-value for the null value θ_0 given S_d rather than S_m is an important practical problem.

The problem of directly combining $\{d_{*\ell}, \ell = 1, \ldots, m\}$ is tricky since each $d_{*\ell}$ typically leads to a too significant p-value because the $U_{*\ell}$ tends to underestimate the total variability T in equation (1.4). The representation of D_m that makes progress possible (Rubin (1987)) is to note that (1.8) implies

$$D_m \approx \hat{D}_m = \frac{\bar{d}_m k^{-1} - (\frac{m-1}{m+1}) r_m}{1 + r_m}, \tag{1.22}$$

where $\bar{d}_m$ is the sample mean of $\{d_{*\ell}, \ell = 1, \ldots, m\}$ and r_m is given by (1.18).

Two penalties for the overestimation inherent in the $d_{*\ell}$ exist in (1.22). First, a positive quantity is subtracted from $\bar{d}_m k^{-1}$ (the extra factor k^{-1} is due to our using a mean square, rather than a chi-square as reference distribution), and then the result is divided by a quantity that is larger than 1; both penalties are monotone increasing functions of r_m.

Replacing r_m in $\hat{D}_m$ with estimates obtained from the set S_d yields procedures for calculating p-values when only S_d is available. For example, the existing standard procedure, as described in Rubin (1987), is to replace r_m by a method of moments estimate

$$\hat{r}_m = (1 + \frac{1}{m}) s_d^2 / \{2\bar{d}_m + [4\bar{d}_m^2 - 2k s_d^2]_+^{1/2}\}, \tag{1.23}$$

where s_d^2 is the sample variance of the $d_{*\ell}$ and $[a]_+ = \max\{a, 0\}$. The resulting procedure for calculating the p-value is

$$\hat{P}_m = \Pr\{F_{k, a_k \hat{\zeta}} \geq \hat{D}_m\}, \tag{1.24}$$

where

$$\hat{\hat{D}}_m = \frac{\bar{d}_m k^{-1} - (\frac{m-1}{m+1})\hat{r}_m}{1+\hat{r}_m}, \tag{1.25}$$

$$\hat{\zeta} = (m-1)(1+\hat{r}_m^{-1})^2, \tag{1.26}$$

and

$$a_k = (1+k^{-1})/2. \tag{1.27}$$

Evaluations of this procedure are summarized in Rubin (1987) as well as in Li (1985); Raghunathan (1987); Weld (1987); Treiman, Bielby and Cheng (1988); and Schenker, Treiman and Weidman (1988). These results suggest that whenever the fractions of missing information are small, or modest but $m \geq k$, it provides reasonably accurate levels. In other cases, an improved procedure is needed, especially when both the fractions of missing information and k are large.

1.7. Relationship with classic Behrens-Fisher problem

Our problem is closely related to the multivariate version ($k > 1$) of the Behrens-Fisher problem. In particular, when all λ_j equal $\bar{\lambda}$, the numerator of D_m on the right hand side of (1.17) is distributed proportional to a χ^2_k random variable, and the denominator as a positive affine transformation of an independent $\chi^2_{k(m-1)}$ random variable. In this case, D_m is distributed as the square of a k-variate Behrens-Fisher random variable with ∞ and $k(m-1)$ degrees of freedom, where r_m can be used to estimate the nuisance parameter $\bar{\lambda}$ with $k(m-1)$ degrees of freedom.

In traditional applications of the Behrens-Fisher problem, the two degrees of freedom are typically modest and relatively similar (e.g., 8 and 12), rather than ∞ and $k(m-1)$ where m is small (e.g., 3). In traditional cases, simple approximations often work well because the nuisance parameter, $\bar{\lambda}$, although not known precisely, is well-enough known. References include: Aspin (1948), Cochran (1964), Jeffreys (1940), Johnson and Neyman (1936), Robinson (1976), Smith (1936), Wallace (1978), and Welch (1937, 1947). Although the distribution of D_m becomes more complicated when the λ_j vary, r_m is still unbiased for $\bar{\lambda}$. It is thus not too surprising that for modest $k(m-1)$, a satisfactory reference distribution can be found for D_m, as in Li, Raghunathan and Rubin (1990).

The situation changes rather dramatically, however, when r_m is not available and must be estimated – the case that occurs when inference must be based on the set S_d rather than the set S_m. Although $\hat{D}_m$ given by (1.22) is asymptotically equivalent to D_m, when r_m is replaced by an estimate from S_d, three things happen. First, the numerator and denominator of the resulting estimate of $\hat{D}_m$

are no longer independent; second, the degrees of freedom available to estimate $\bar{\lambda}$ are reduced from $k(m-1)$ to at most $(m-1)$, at most because the $d_{*\ell}$ only provide an absolute magnitude of the difference between θ_0 and $\hat{\theta}_{*\ell}$ and not the direction of that difference as with $\hat{\theta}_{*\ell} - \theta_0$; and third, it is not obvious how to combine the $(m-1)$ apparent degrees of freedom in S_d to estimate $\bar{\lambda}$.

Clearly with small m (e.g., 3), inferences will be sensitive to values of the nuisance parameter $\bar{\lambda}$. Because in practice m is modest, we are dealing with a situation where theoretical small samples issues in the Behrens-Fisher problem are practically important.

2. The Proposed Procedure

2.1. The test statistic $\hat{D}_d$

The proposed test statistic is of the form D_m with r_m replaced by an estimate $\hat{r}_d$, as with $\hat{D}_m$. Specifically, we propose

$$\hat{D}_d = \frac{\bar{d}_m k^{-1} - (\frac{m+1}{m-1})\hat{r}_d}{1+\hat{r}_d} \tag{2.1}$$

where

$$\hat{r}_d = (1+\frac{1}{m})\left[\frac{1}{m-1}\sum_{\ell=1}^{m}(\sqrt{d_{*\ell}} - \overline{\sqrt{d}})^2\right] \tag{2.2}$$

is the sample variance of $\sqrt{d_{*1}}, \ldots, \sqrt{d_{*m}}$ times $(1+m^{-1})$. This estimate of r_m makes intuitive sense because the smaller the fraction of missing information γ, the closer each of the $d_{*\ell}$ $(\ell = 1, \ldots, m)$ should be to the ideal test statistic and thus to each other. For example, in the extreme case when there is no missing information, all $d_{*\ell}$'s must be equal to the ideal test statistic, and hence $\hat{r}_d = 0$; then the corresponding $\hat{D}_d$ is also equal to the ideal test statistic.

2.2. The derivation of $\hat{D}_d$ assuming equal eigenvalues

The main difficulty in deriving a replacement for r_m based on the set S_d is that the joint distribution of $(d_{*1}, \ldots, d_{*m})$ is very complicated – a product of m Bessel functions (Raghunathan (1987), Meng (1988)). Various estimates have been proposed in the literature. Under the equal eigenvalue assumption, Li (1985) and Rubin (1987) derived the method of moments estimates given in (1.23); Raghunathan (1987) gave an approximate MLE using a simple approximation to the Bessel function; and Meng (1988) showed how to obtain the exact MLE by applying the EM algorithm. But none of these estimates are in simple form, and the corresponding distributions of the approximations are not tractable.

Here, we adopt a different approach. We first simplify the distribution of $d_{*\ell}$ using a normal approximation to $\sqrt{d_{*\ell}}$ ($\ell = 1, \ldots, m$), and then obtain a simple estimate of $\bar{\lambda}$ by a further approximation. Although we derive $\hat{D}_d$ under the equal eigenvalue assumption, we show in Section 2.5 that $\hat{D}_d$ can be motivated without this restriction.

From (1.7) – (1.8), one can show, for all θ and θ_0, that

$$d_{*\ell}|X_{\text{obs}}, \theta \sim \text{i.i.d. } \bar{\lambda}\chi^2_{k,\delta/\bar{\lambda}}, \tag{2.3}$$

where the noncentrality parameter of the χ^2 random variable is $\delta/\bar{\lambda}$ with

$$\delta = (\hat{\theta}_{\text{obs}} - \theta_0)^t \bar{U}^{-1} (\hat{\theta}_{\text{obs}} - \theta_0). \tag{2.4}$$

Using a normal approximation to the non-central chi random variable (Patnaik (1949)), we obtain

$$\sqrt{d_{*\ell}}|X_{\text{obs}}, \theta \sim \text{i.i.d. } N(\mu, \sigma^2), \tag{2.5}$$

where

$$\mu = \{(k-1)\bar{\lambda} + \delta + \tau\}^{1/2}, \tag{2.6}$$

$$\sigma^2 = \bar{\lambda} - \tau, \tag{2.7}$$

and

$$\tau = \frac{\bar{\lambda}^2}{2(\bar{\lambda} + \delta/k)} = \frac{\bar{\lambda}^2}{2(\bar{\lambda} + (1+\bar{\lambda})D_m^\infty)}, \tag{2.8}$$

where the last step follows because $D_m^\infty = \delta/[k(1+\bar{\lambda})]$, as given in (1.20). Thus for large k and small $\bar{\lambda}$, the sample mean and sample variance of the $\sqrt{d_{*\ell}}$ are sufficient statistics for μ and σ^2.

From (2.8), τ is relatively small compared to $\bar{\lambda}$ in most the cases of importance because (i) in common practice, $\bar{\lambda}$ is less than 1 (corresponding to a maximum of 50% missing information), and (ii) D_m^∞ is large when the ideal test statistic is close to traditional values for significance. In fact one can easily show that $\tau/\bar{\lambda} < \bar{\lambda}/(2(1+2\bar{\lambda})) < 1/6$ when $\bar{\lambda} \leq 1$ and $D_m^\infty \geq 1$. If we set τ to zero, as we shall further justify in the next section, then by (2.5) – (2.7), we obtain

$$\sqrt{d_{*\ell}}\,|\,X_{\text{obs}}, \theta \sim \text{i.i.d. } N([(k-1)\bar{\lambda} + \delta]^{1/2}, \bar{\lambda}), \tag{2.9}$$

which provides the very simple estimate of $(1 + m^{-1})\bar{\lambda}$ given by $\hat{r}_d$ in (2.2). Replacing r_m in (1.22) by $\hat{r}_d$ gives $\hat{D}_d$ in (2.1).

2.3. Behavior of $\hat{D}_d$ when $m \to \infty$ – Theory with equal eigenvalues

Although the case of $m = \infty$ is of no practical interest, it can be used to ascertain the degradation in performance of a procedure with finite m. Also, theoretically, the consistency of a general test procedure is desirable, in the sense that $\hat{D}_d$ should ideally be close to D_{obs} when $m = \infty$. We will show that D_d^∞ is very close to D_m^∞ and hence, as discussed in Section 1.5, very close to D_{obs}.

From (1.7), (1.8) and (1.21), one can show that

$$E(d_{*\ell}|X_{\text{obs}},\theta) = \bar{\lambda}(k + \delta/\bar{\lambda}) = k(\bar{\lambda} + (1+\bar{\lambda})D_m^\infty). \tag{2.10}$$

Thus, when $m \to \infty$, from the strong law of large numbers, $\hat{D}_d$ converges almost surely to (conditional on X_{obs} and θ)

$$\begin{aligned} D_d^\infty &= \frac{\frac{1}{k}E(d_{*\ell}|X_{\text{obs}},\theta) - \text{Var}(\sqrt{d_{*\ell}}|X_{\text{obs}},\theta)}{1 + \text{Var}(\sqrt{d_{*\ell}}|X_{\text{obs}},\theta)} \\ &= D_m^\infty + R, \end{aligned} \tag{2.11}$$

where

$$R = \frac{(\bar{\lambda} - \text{Var}(\sqrt{d_{*\ell}}|X_{\text{obs}},\theta))(1 + D_m^\infty)}{1 + \text{Var}(\sqrt{d_{*\ell}}|X_{\text{obs}},\theta)}. \tag{2.12}$$

From equations (2.11) and (2.12) it is clear that if R is small, then D_m^∞ will be close to D_m^∞ and hence close to D_{obs}. We now show that R is indeed negligible in most cases of practical importance.

Approximation (2.7) gives $\text{Var}(\sqrt{d_{*\ell}}|X_{\text{obs}},\theta) = \bar{\lambda} - \tau$, where τ is given in (2.8). Thus, from (2.12), straightforward algebra shows

$$R \approx \frac{\bar{\gamma}^2(1 + D_m^\infty)}{2D_m^\infty + 1 - (1-\bar{\gamma})^2}, \tag{2.13}$$

where $\bar{\gamma} = \bar{\lambda}/(1+\bar{\lambda})$ is the average fraction of missing information. When D_m^∞ is equal to or larger than its expectation under the null hypothesis (i.e., 1), which is certainly true when the ideal test provides any evidence that $\theta \neq \theta_0$, it follows that

$$R \leq \frac{2\bar{\gamma}^2}{3 - (1-\bar{\gamma})^2} \equiv B(\bar{\gamma}) \quad \text{(say)}.$$

Notice that $B(\bar{\gamma})$ is a monotone increasing function with maximal value $B(1) = 2/3$. In common practice, $B(\bar{\gamma})$ usually is very small because $\bar{\gamma}$ is typically less than 30%. In fact, $B(0.3) = 0.07$, which is clearly negligible compared to D_m^∞.

2.4. Checking the theoretical approximation with $m = \infty$ using Monte Carlo

The actual level of a nominal α test based on D_d^∞ with the χ_k^2/k reference distribution is

$$\Pr\{kD_d^\infty > \chi_k^2(1-\alpha)|\theta = \theta_0\} \tag{2.14}$$

where $\chi_k^2(1-\alpha)$ is the $100(1-\alpha)$ percentage point of the chi-square distribution with k degrees of freedom. Now, under the null hypothesis $\theta = \theta_0$, and assuming $\lambda_j = \bar{\lambda}$,

$$(\delta|\theta = \theta_0, \lambda_j = \bar{\lambda}) \sim (1+\bar{\lambda})\chi_k^2.$$

Hence, the probability (2.14) can be evaluated numerically for any fixed $\bar{\lambda}$ or equivalently fixed $\gamma_j = \bar{\gamma}$.

Table 1 provides the actual level when $\lambda_j = \bar{\lambda}$ for nominal 1%, 5% and 10% tests for various combinations of $\bar{\gamma}$ and k obtained by numerically evaluating (2.14) using Monte Carlo techniques. To accomplish this, we generated δ from $(1+\bar{\lambda})\chi_k^2$ and computed D_d^∞ and D_{obs} (which equals to D_m^∞ under the equal eigenvalue assumption) for 10,000 draws of δ for each choice of $\bar{\gamma} = 0.1$, 0.2, 0.3 and 0.5 and $k = 2, 3, 5, 7, 10, 15, 20$ and 25. Overall, the levels seem to be well calibrated, although slightly on the conservative side for small $\bar{\gamma}$ and anti-conservative for large k and large $\bar{\gamma}$. Table 2 provides correlation coefficients between D_{obs} and D_d^∞ for various choices of $\bar{\gamma}$ and k. Overall, it seems that the correlations are high for small k and decrease as k increases.

2.5. Extensions to unequal eigenvalue cases

Equations (2.10) – (2.12) are valid even when the fractions of missing information are not equal, since they are derived directly from the general distributional assumptions (1.7) and (1.8) of Section 1.3. An analogue of (2.13) can be obtained by the delta method (or by applying Patnaik's (1949) technique) as follows. Using the notation of Section 1.3, let Γ be an orthogonal transformation matrix such that $\Gamma\bar{U}^{-1/2}B\bar{U}^{-1/2}\Gamma' = diag(\lambda_1, \ldots, \lambda_k)$, and $\Delta \equiv \Gamma\bar{U}^{-1/2}(\hat{\theta}_{\text{obs}} - \theta_0)$, a function of X_{obs}. Let $\beta_{\Delta,\lambda}$ be the slope of the regression of Δ on λ and CV_ξ be the coefficient of variation of the ξ. Then we have

$$\text{Var}(\sqrt{d_{*\ell}}|X_{\text{obs}}, \theta) \approx \bar{\lambda} - \tau^*,$$

where

$$\tau^* = \frac{\bar{\lambda}^2 - (1+\bar{\lambda})^2(1+2\beta_{\Delta,\lambda})CV_\xi^2}{2(\bar{\lambda} + (1+\bar{\lambda})D_m^\infty)},$$

which reduces to r of (2.8) when all $\lambda_i = \bar{\lambda}$. Substituting these two expressions into (2.12), we obtain

$$R \approx \frac{[\bar{\gamma}^2 - (1 + 2\beta_{\Delta,\lambda})CV_\xi^2](1 + D_m^\infty)}{2D_m^\infty + 1 - (1 - \bar{\gamma})^2 + (1 + 2\beta_{\Delta,\lambda})CV_\xi^2}. \tag{2.15}$$

This expression indicates that R is still negligible under the null hypothesis, because $E(\beta_{\Delta,\lambda}|\theta = \theta_0) = 0$, and CV_ξ^2 typically is small. In fact one can show that $CV_\xi^2 \leq \bar{\gamma} - 2\bar{\gamma}^2$ when all λ_i are less than 1. Assuming $\beta_{\Delta,\lambda} = 0$ and $D_m^\infty \geq 1$, we obtain from (2.15) that

$$|R| \leq \frac{|\bar{\gamma}^2 - CV_\xi^2|}{3 - (1 - \bar{\gamma})^2 + CV_\xi^2} \equiv \tilde{B}(\bar{\gamma}, CV_\xi^2) \quad \text{(say)}.$$

It is interesting to notice that $\tilde{B}(\bar{\gamma}, CV_\xi^2)$ is usually smaller than $B(\bar{\gamma})$ $(= \tilde{B}(\bar{\gamma}, 0))$, since CV_ξ^2 is positive and small. When $\bar{\gamma} \leq 30\%$, using the facts $B(\bar{\gamma}) \leq 0.07$ and $CV_\xi^2 \leq \bar{\gamma} - 2\bar{\gamma}^2$, one can show that $\tilde{B}(\bar{\gamma}, CV_\xi^2) \leq 0.07$, that is the same bound holds for $\tilde{B}$ as for B. This result suggests that the approximation in (2.15) is often more accurate than the approximation in (2.13).

Monte Carlo simulations are difficult to perform with unequal eigenvalues because $E(\sqrt{d_{*\ell}}|X_{\text{obs}}, \theta)$ is hard to evaluate. It is not clear, however, whether performing such computationally arduous comparisons of D_d^∞ and D_{obs} when the eigenvalues are unequal is worthwhile since D_d^∞ is primarily of interest as a theoretical procedure, and modest values of m must usually be used in practice. Therefore, we move on to our comparison of $\hat{D}_d$ and D_{obs} for the cases of practical importance with modest m and both equal and unequal eigenvalues.

2.6. The reference distribution for $\hat{D}_d$ with small m

When $m \to \infty$, the obvious reference distribution is χ_k^2/k since it is the correct reference distribution for the ideal test statistic, D_{obs}. For small m, some approximations are inevitable, since the exact distribution is intractable and even if it were available, it would depend on nuisance parameters $(\lambda_1, \ldots, \lambda_k)$. The reference distribution we use here is an F distribution on k and $a_{k,m} \cdot w_s$ degrees of freedom, where

$$w_s = (m - 1)\{1 + \hat{r}_d^{-1}\}^2 \tag{2.16}$$

and

$$a_{k,m} = k^{-3/m}. \tag{2.17}$$

This form for w_s is obtained by replacing r_m in $(m - 1)\{1 + r_m^{-1}\}^2$, which is the denominator degrees of freedom of an F reference distribution for D_m (Rubin

(1987)), with $\hat{r}_d$, in the same way as we did for obtaining $\hat{D}_d$; analogous substitution for r_m in w given by equation (1.19) leads to essentially the same Monte Carlo results. The extra adjustment factor $a_{k,m}$ is obtained through simulation, and essentially reflects the loss of degrees of freedom due to the fact that we are using a scalar quantity instead of a k-dimensional quantity to estimate $\bar{\lambda}$; $a_{k,m}$ was chosen to be especially good when $m = 3$, since this appears to be the most common value in current practice. Notice that, as $m \to \infty$, $a_{k,m} \to 1$, and for $m \geq 3$

$$k^{-1} \leq a_{k,m} \leq 1. \tag{2.18}$$

That is, the adjusted number of degrees of freedom is between the minimum and maximum denominator degrees of freedom.

3. Evaluation of Our Procedure for Finite m

3.1. The level of $\hat{D}_d$

The actual level of a nominal α level test based on our procedure is

$$\Pr\{\hat{D}_d > F_{k,\eta}(1-\alpha) | \theta = \theta_0\}, \tag{3.1}$$

where $\eta = a_{k,m} w_s$, and $F_{k,\eta}(1-\alpha)$ is the $100(1-\alpha)$ percentage point of the $F_{k,\eta}$ distribution. The procedure is said to have the correct level α if the probability in (3.1) is equal to α. Ideally one would hope this is true for every α between 0 and 1, or at least for common values of α such as 10%, 5% and 1%. From the nature of our procedure, it should be apparent that it will be exceedingly difficult to evaluate (3.1) analytically. Therefore, we use Monte Carlo simulation to evaluate (3.1).

3.2. Monte Carlo conditions

Note that without loss of generality, we can let $\theta_0 = 0$, $\bar{U} = I$ and $B = diag(\lambda)$. As in Section 2.4, we consider average fractions of missing information, $\bar{\gamma}$, equal to 0.1, 0.2, 0.3 and 0.5, where 50% missing information represents an extreme case for multiple imputation. We also consider equal and unequal eigenvalues, the latter chosen so that for each value of k considered (2, 3, 5, 7, 10, 15, 20, 25), the standard deviations of the γ_i nearly equal 0.12(min=0.099, max=0.124). Specifically, in the unequal case,

$$\gamma_i = (\bar{\gamma} - 0.1) + \gamma_i^*, \quad i = 1, \ldots, k,$$

where the γ_i^* are given in sequence by

0.010 0.190 0.102 0.181 0.021 0.174 0.032 0.024
0.250 0.031 0.056 0.032 0.351 0.021 0.052 0.203
0.053 0.071 0.182 0.051 0.052 0.103 0.024 0.305
0.036.

Although the standard deviation of 0.12 may appear modest, especially for large $\bar{\gamma}$, our selection of the $\gamma_1, \ldots, \gamma_k$ represents extreme situations. For example, when $k = 15$ and $\bar{\gamma} = 0.5$, the minimum and maximum fractions of missing information are 41% and 75% respectively. Some limited experience with real data (e.g., Rubin and Schenker (1987), Treiman, Bielby and Cheng (1987), Heitjan and Rubin (1990), Heitjan and Little (1991)) suggests that with real data, most fractions of missing information are less than 30%. The values of m are 2, 3, 5 and 10, which cover all values likely to be used in practice, and let $\alpha = 10\%$, 5% and 1%.

3.3. Steps of the simulation

The basic setup for the simulation is as follows:

Step 1. Generate $\hat{\theta}_{\text{obs}}$ from a normal distribution with mean zero and covariance matrix $I + B = T$. Compute $\Pr(\chi_k^2 > D_{\text{obs}})$, the ideal p-value, where $D_{\text{obs}} = \hat{\theta}_{\text{obs}}^t T^{-1} \theta_{\text{obs}}$. This mimics the situation where our analysis based upon the observed data, X_{obs}, results in $(\hat{\theta}_{\text{obs}}, T)$ and the subsequent computation of the ideal p-value; P_{obs} from D_{obs}. This is used as a simulation covariate.

Step 2. Generate $d_{*\ell}$ from its repeated imputation distribution described in Sections 1.3 and 1.6, where $U_{*\ell}$ is fixed at I. This mimics the situation where we have used the complete-data analysis on m completed data sets resulting in chi-square or Wald statistics, $d_{*\ell}$, $\ell = 1, \ldots, m$.

Step 3. Compute the p-value, P_d, based upon $\hat{D}_d$, $\Pr\{F_{k,\eta} > \hat{D}_d\}$, where η is given in Section 3.1.

Step 4. Record the p-values, (P_{obs}, P_d) obtained in Steps 1 and 3.

Step 5. Repeat Steps 1 through 4 N times and estimate (3.1) for α=10%, 5% and 1%.

The simulation was done for N=5,000 repetitions with both equal and unequal fractions of missing information. The results of our simulation are described in Tables 3 and 4, which provide the actual levels for various situations when the fractions of missing information are equal and unequal respectively.

3.4. Results of the simulation

In general, the exact levels in Tables 3 and 4 suggest that $\hat{D}_d$ tends to be conservative for both equal and unequal fractions of missing information. This is more pronounced for $\bar{\gamma} = 0.1$ and large k (≥ 10). Furthermore, the extent of conservativeness seems to be a function of the nominal level. For example, for the nominal 5% tests, the exact levels for $k = 10$ range from 3.3% to 4.3% when m changes from 2 to 10, whereas for the nominal 10% tests, the levels range from 6.9% to 9.1%. By the results in Section 2, the levels will approach the nominal levels as m increases; however, our results seem to be best when $m = 3$, which is as anticipated, since the adjustment factor $a_{k,m}$ was chosen with particular attention to the case $m = 3$.

For $\bar{\gamma} = 0.2$, the test seems to be somewhat anticonservative for the 1% nominal level and conservative for the 10% nominal level with 5% in between. For $\bar{\gamma} = 0.3$ and 0.5, the test is anticonservative, especially for large k. For example when $k = 25$, $m = 2$ and $\bar{\gamma} = 0.5$, the nominal 1% test corresponds to an exact 5.2% test, whereas the nominal 5% test corresponds to an exact 7.6% test. There is some nonmonotonicity in the exact levels as m increases, especially for 5% and 10% nominal levels. Nevertheless, the performance of $\hat{D}_d$ is distinctly superior to the current standard.

To help assess the effect of the various factors on the difference between the nominal and exact levels, we constructed an ANOVA in Table 5 for the $8 \times 4 \times 4 \times 3 \times 2$ factorial design. To measure the difference between the exact and nominal levels we consider

$$\frac{[z \text{ score for the exact level}] - [z \text{ score for the nominal level}]}{z \text{ score for the nominal level}},$$

where the z score is the standard normal deviate corresponding to the p-value. Since the total sum of squares is small (6.05) relative to the number of differences (768), the variability in differences between the exact and nominal levels is relatively small. The effect of various factors on the differences can be ascertained by the sum of squares associated with various assignable sources.

The results in Table 5 suggest that the major source of differences between the exact and nominal levels is $\bar{\gamma}$, for it explains 58% of the total variability. The next most important main effect is α, closely followed by m. Sums of squares associated with various interactions involving $\mathrm{Var}(\gamma)$ also suggest that the test $\hat{D}_d$ is somewhat sensitive to the equal eigenvalue assumption, although these interactions are dwarfed by the main effect due to $\bar{\gamma}$.

Correlations between $\hat{D}_d$ and both D_m and D_{obs} were examined, and, not surprisingly, are quite low, indicating a substantial loss of power when using

$\hat{D}_d$ rather than D_m. This issue is explored next by examining the conditional performance of $\hat{D}_d$ relative to D_m.

3.5. Some evaluations of conditional performance

Despite the approximate attainment of nominal level, our procedure leaves much to be desired because the set S_d of p-values has so much less information than the set of moments, S_m. From the frequentist perspective, there will be a substantial loss of power when using S_d rather than S_m, and from the Bayesian perspective, results will be quite sensitive to prior assumptions. The consequences with our procedure, which are visible from our simulations, are that it is poorly calibrated conditionally. In particular, consider any fixed λ and repeated samples with (a) estimated fractions of missing information indexed by $\mathrm{Var}(\sqrt{d_{*\ell}})$ and (b) p-values P_m and P_d. Assuming P_d attains nominal levels (unconditionally), when $\mathrm{Var}(\sqrt{d_{*\ell}})$ is small, we expect to see too liberal p-values from P_d relative to P_m, whereas when $\mathrm{Var}(\sqrt{d_{*\ell}})$ is large, we expect to see too conservative p-values relative to P_m. When considering a variety of values of λ, we expect to see a similar trend.

Figures 1a and 1b display the difference between P_d and P_m for all $\bar{\lambda}$ values from our simulation with $m = 3$ (which is approximately calibrated unconditionally), for $k = 5$ and 20, respectively. Attempts can be made to adjust these results to create a conditionally conservatively calibrated procedure, but such attempts typically lead to worse calibration unconditionally, as with the fiducial solution to the Behrens-Fisher problem.

The real culprit is the extreme loss of information when going from S_m to S_d, and some remedies are suggested in the next and final section.

4. Discussion

4.1. Conclusions and practical guidance on the use of our procedure

We have described a procedure for computing the p-value when only completed-data test statistics, $\{d_{*\ell}, \ell = 1, \ldots, m\}$, are available. The procedure is simple and performs reasonably well for a variety of situations described by the values of m, the number of imputations, k, the dimensionality of the parameter, and $\bar{\gamma}$, the average fraction of missing information, and for both equal and unequal fractions of missing information. The simulation study shows that the fraction of missing information is the most important factor in the performance of this procedure.

Our procedure is, nevertheless, far from what might be hoped for. Its calibration is only approximate, and moreover, not ideal conditionally given the estimated fraction of missing information. Also, because of its noisy relation

with D_m, the moment-based procedure, we know there will be a substantial power loss compared with D_m. As a result, we recommend that $\hat{D}_d$ be used primarily as a screening test statistic, thinking of it as providing a range of p-values between one-half and twice the observed P_d. If this range is not sharp enough, then ideally the researcher should use a more accurate procedure, for example D_m based on S_m, the set of moments. If S_m is not available, but only S_d, other work reported in Section 4.2 suggests that extensive efforts may be required to do much better than our procedure – the loss of information going from S_m to S_d is great. Of these, a fully Bayesian procedure is the most theoretically satisfactory (e.g., Raghunathan (1987)).

Another class of methods is based on obtaining information in addition to S_d but short of S_m. These methods are briefly discussed in Section 4.3 and have promise.

4.2. Other methods based on S_d

Li (1985) describes other procedures primarily motivated by the frequentist perspective. Note that when the fractions of missing information are equal, the ideal p-value is

$$P_{\text{obs}} = \Pr(\chi_k^2 > \delta(1-\bar{\gamma})) \tag{4.1}$$

where $\delta = (\hat{\theta}_{\text{obs}} - \theta_0)^t U^{-1}(\hat{\theta}_{\text{obs}} - \theta_0)$ and $\bar{\gamma}$ is the common fraction of missing information. Hence, a version of P_{obs} based only on $\{d_{*\ell}, \ell = 1, \ldots, m\}$ can be obtained by first obtaining estimates of δ and $\bar{\gamma}$ and then substituting these estimates in (4.1). Li (1985) provides several estimates of δ and $\bar{\gamma}$, none of which are superior to the procedure described in Section 2; in some regions of the space formed by values of $(k, m, \bar{\gamma})$, they are overly conservative or overly liberal. A main reason for the poor performance of such methods is their failure to account for the uncertainty introduced by the substitution of estimates for δ and $\bar{\gamma}$. The procedure described in this paper corrects this by considering an F distribution rather than a chi-square distribution.

Raghunathan (1987) provides procedures motivated by the Bayesian perspective. The posterior density of $(\delta, \bar{\gamma})$, for a uniform prior, can be shown to be

$$P(\delta, \bar{\gamma}|S_d) \propto (\bar{\gamma}^{-1} - 1)^{m/2} \delta^{-m(k-2)/4} \exp\{-\frac{1}{2}(\bar{\gamma}^{-1} - 1)(\bar{d}_m + \delta)\} \cdot \prod_{i=1}^{m} I_{(k-2)/2}[\sqrt{d_{*\ell}\delta}(\bar{\gamma}^{-1} - 1)],$$

where

$$I_q(x) = \sum_{j=0}^{\infty} \frac{(x/2)^{q+2j}}{\Gamma(j+1)\Gamma(j+q+1)}$$

is the modified Bessel function of the first kind. The Bayesian p-value based on S_d is obtained as

$$\int \Pr(\chi_k^2 > \delta(1-\bar{\gamma})) \cdot \Pr(\delta, \bar{\gamma}|S_d) d\delta \, d\bar{\gamma}. \tag{4.2}$$

Numerical integration techniques can be used to evaluate the above integral. Raghunathan (1987) develops various approximations to the above integral and investigates their frequency properties. Overall, they tend to be conservative for large k and liberal for small k. A serious disadvantage of this approach is its sensitivity to the prior specifications for $(\delta, \bar{\gamma})$. Since usually the number of imputations m is small, say less than 5, the prior distribution for $(\delta, \bar{\gamma})$ plays a very prominent role and can drastically alter the performance in terms of frequency properties for a slight change in the prior specifications.

An easily implemented Monte Carlo algorithm is described in Meng (1988) for approximating the above integral. It is derived based upon the decomposition of $d_{*\ell}, \ell = 1, \ldots, m$, with $\bar{\lambda} = \bar{\gamma}/(1-\bar{\gamma})$,

$$d_{*\ell} = \bar{\lambda}\chi_{k-1}^2 + [N(\sqrt{\delta}, \bar{\lambda})]^2. \tag{4.3}$$

It is also shown that its frequency properties are sensitive to the choice of prior for $\bar{\gamma}$. Meng (1988) also discusses several approximations based on the maximum likelihood estimate of δ and $\bar{\lambda}$ (obtained using the EM algorithm) and the associated observed Fisher information matrix. Bayesian versions are also considered. These procedures work well for not too small m.

Meng (1988) also proposes a translated F reference distribution for $\hat{D}_d$, derived using the same normal approximation for the distribution of $\sqrt{d_{*\ell}}$. This reference distribution is more accurate than the F reference distribution given here for large values of $\bar{\gamma}$. Unfortunately, this reference distribution does not converge to the ideal reference distribution when $m \to \infty$, although they are quite close.

4.3. Methods based on more information than S_d but less than S_m

As we mentioned before, the loss of information from S_m to S_d is extreme, especially when k is large and m is small. This should be clear since S_m contains all k-dimensional vectors and their normalizing matrices, whereas S_d consists of only the normalized scalar distances of these vectors from θ_0. This severe loss of information is responsible for the poor performance of the existing procedures, including $\hat{D}_d$, and makes the problem a very difficult one.

Clearly, the only way to overcome this inherent difficulty is to obtain more information. Ideally, of course, we would have the collection of the moments,

S_m, but this is impossible in cases when the covariance matrices are not available. Fortunately, however, it is possible in some cases of practical importance to obtain some extra information, which is less than S_m, but is enough to approximate D_m well. The key idea is that all we want is the scalar test statistic D_m of (1.22), and the only quantity there that is not directly obtainable form S_d is r_m of (1.18).

One such kind of procedure has been proposed recently in Meng and Rubin (1990). The basic requirement of this new procedure is that besides S_d, one also has (i) the collection of the m estimates $\hat{\theta}_{*\ell}$ and (ii) computer code for evaluating the complete-data likelihood ratio test statistic as a function of parameter estimates. This is often feasible in practice since all it requires is the complete-data computations for scalar quantities. It is shown in Meng and Rubin (1990) that in large samples, the r_m of (1.18) is proportional to the difference between the average of m complete-data log-likelihood ratios, $\bar{d}_m$, and the average of m complete-data log-likelihood ratios, each evaluated at the average of m estimates, where the proportionality constant is a simple function of k and m. Based on this result, an approximate likelihood ratio test is constructed, and it is shown to be asymptotically equivalent to D_m for any number of multiple imputations.

A common situation in practice is that the standard complete-data analyses provide not only the significance levels, but also the estimates $\hat{\theta}_{*\ell}$ and their standard errors (the diagonal elements of $U_{*\ell}$), but neither the entire covariance matrix $U_{*\ell}$ nor the code for evaluating the complete-data likelihood ratio test statistic as a function of parameter estimates. These estimates and standard errors certainly provide more information than does S_d alone, and one would expect, therefore, to be able to obtain better test procedures using them. The construction of such test procedures and their evaluation are still open questions, although they are under current investigation.

Acknowledgements

We wish to thank the reviewers for exceptionally helpful comments on an earlier draft of this paper. This work was partially supported by U.S. N.S.F. grants SES-83-11428 and SES-880543, and partially by Joint Statistical Agreements 87-07, 88-02, 89-08, and 90-23 between the U.S. Bureau of the Census and Harvard University.

Table 1. Levels of $\hat{D}_d$ when $m = \infty$ as a function of the nominal level, α; the dimension, k; and the fraction of missing information $\bar{\gamma}$. (Based on 10,000 draws of δ.) Equal fractions of missing information.

α	$k \backslash \bar{\gamma}$	0.1	0.2	0.3	0.5
1%	2	0.9	0.9	0.8	1.0
	3	0.9	0.9	0.9	1.1
	5	1.0	1.0	0.9	1.2
	7	1.0	1.0	0.9	1.0
	10	1.0	1.0	1.0	1.3
	15	1.0	1.0	1.0	1.0
	20	1.0	1.0	1.0	1.4
	25	1.0	1.0	1.0	1.5
5%	2	4.8	4.8	4.8	5.6
	3	5.2	4.8	4.8	5.8
	5	5.1	4.8	5.0	5.9
	7	5.1	5.0	5.1	6.1
	10	5.0	5.0	4.9	5.8
	15	5.0	5.0	4.9	6.1
	20	5.0	5.0	4.9	6.1
	25	5.0	5.0	4.8	6.1
10%	2	9.8	9.8	10.0	10.8
	3	9.9	9.8	10.2	11.2
	5	9.8	9.8	10.3	11.3
	7	9.7	9.9	10.4	11.4
	10	9.6	9.8	10.3	11.8
	15	9.5	9.7	10.4	11.8
	20	9.5	9.6	10.8	11.7
	25	9.4	9.6	10.9	11.1

Table 2. Correlation coefficients between D_{obs} and $\hat{D}_d$ when $m = \infty$ as a function of the dimension k and the fraction of missing information, $\bar{\gamma}$. Equal fractions of missing information.

$k \backslash \bar{\gamma}$	0.1	0.2	0.3	0.5
2	0.998	0.998	0.998	0.997
3	0.971	0.981	0.998	0.921
5	0.962	0.988	0.996	0.905
7	0.892	0.976	0.988	0.896
10	0.891	0.892	0.921	0.821
15	0.876	0.872	0.904	0.831
20	0.772	0.802	0.872	0.781
25	0.792	0.821	0.872	0.761

Table 3. Level (in %) of $\hat{D}_d$ with F reference distribution as a function of the nominal level, α; the dimension, k; the number of imputations, m; and the fraction of missing information, $\bar{\gamma}$. Equal fractions of missing information.

$m = 2$

α	k \ $\bar{\gamma}$	0.1	0.2	0.3	0.5
1%	2	0.9	1.4	1.7	2.4
	3	0.7	1.4	1.8	2.8
	5	0.8	1.2	1.8	3.4
	7	0.8	1.1	2.1	4.0
	10	0.7	1.3	2.3	4.9
	15	0.7	1.3	2.3	5.4
	20	0.7	1.7	2.7	5.3
	25	0.6	1.9	3.5	5.5
5%	2	5.1	5.1	5.7	6.5
	3	4.5	5.4	5.8	8.0
	5	4.4	4.9	6.0	7.9
	7	4.3	4.8	6.5	8.6
	10	3.7	5.1	6.9	9.1
	15	3.6	5.1	6.1	9.1
	20	3.2	5.4	6.5	8.0
	25	3.0	5.0	7.3	8.0
10%	2	10.3	9.8	10.9	11.5
	3	9.3	10.1	11.1	12.2
	5	9.1	9.6	10.7	12.6
	7	8.2	9.5	10.9	12.7
	10	7.9	9.4	11.3	12.8
	15	7.8	8.9	9.9	12.0
	20	6.5	8.5	9.5	10.3
	25	6.0	8.5	10.4	10.3

$m = 3$

α	k \ $\bar{\gamma}$	0.1	0.2	0.3	0.5
1%	2	0.9	1.4	1.6	1.5
	3	0.7	1.3	1.5	1.9
	5	0.8	1.5	1.7	2.4
	7	0.8	1.1	2.0	3.0
	10	0.7	1.3	1.9	4.2
	15	0.8	1.2	2.1	4.5
	20	0.8	1.4	2.5	4.2
	25	0.8	1.7	2.5	4.3
5%	2	5.4	5.2	5.6	6.1
	3	5.1	5.5	6.1	6.7
	5	4.6	5.2	5.9	7.4
	7	4.4	5.2	6.2	7.4
	10	4.4	4.9	6.2	8.7
	15	3.8	4.8	5.9	8.7
	20	3.6	5.1	6.6	8.1
	25	3.4	4.9	6.7	7.5
10%	2	10.6	10.2	11.4	11.4
	3	10.1	10.1	11.3	12.4
	5	9.6	9.8	11.4	12.4
	7	9.1	10.1	11.0	12.0
	10	9.2	9.4	10.8	12.5
	15	8.9	9.5	10.0	12.3
	20	7.7	8.7	10.5	11.8
	25	7.4	8.7	10.6	10.6

$m = 5$

α	k \ $\bar{\gamma}$	0.1	0.2	0.3	0.5
1%	2	0.8	1.2	1.3	1.6
	3	0.8	1.2	1.3	2.8
	5	0.9	1.4	1.6	2.3
	7	0.8	1.2	1.7	2.2
	10	0.9	1.2	2.0	3.3
	15	0.8	1.1	1.9	4.0
	20	0.8	1.4	2.2	4.2
	25	0.7	1.4	2.4	4.1
5%	2	5.2	5.5	5.5	5.8
	3	5.1	5.7	6.2	6.7
	5	4.9	5.0	5.8	7.1
	7	4.7	5.3	6.3	7.5
	10	4.5	5.4	6.4	8.0
	15	4.3	5.1	7.0	8.7
	20	4.2	5.3	7.0	9.1
	25	4.2	5.3	6.8	8.5
10%	2	10.8	10.5	11.2	10.9
	3	10.3	11.0	11.4	12.8
	5	9.8	10.6	11.4	13.2
	7	9.6	10.7	11.5	13.1
	10	9.6	10.4	11.5	12.8
	15	9.4	10.8	12.1	13.1
	20	9.0	10.3	11.4	13.5
	25	8.5	10.3	11.7	12.5

$m = 10$

α	k \ $\bar{\gamma}$	0.1	0.2	0.3	0.5
1%	2	0.9	1.3	1.2	1.1
	3	0.8	1.4	1.1	0.9
	5	1.0	1.5	1.4	1.3
	7	0.9	1.3	1.4	1.6
	10	0.9	1.3	1.4	2.3
	15	0.9	1.4	1.9	3.1
	20	1.2	1.6	2.0	4.3
	25	1.0	1.8	2.1	4.4
5%	2	5.3	6.0	5.5	5.2
	3	5.2	6.0	5.7	6.3
	5	5.1	6.1	6.1	6.6
	7	5.1	5.8	6.8	7.3
	10	5.1	6.3	6.8	8.6
	15	5.0	6.6	7.4	10.1
	20	5.1	7.2	8.3	11.4
	25	5.2	7.0	8.5	12.5
10%	2	10.8	11.3	11.2	11.1
	3	10.2	11.2	11.4	12.2
	5	10.1	11.3	11.9	13.7
	7	10.4	11.8	12.5	13.6
	10	10.3	11.8	12.5	15.0
	15	10.2	12.4	13.5	18.2
	20	10.3	13.2	14.9	19.0
	25	10.0	12.7	14.6	20.3

Table 4. Level (in %) of $\hat{D}_d$ with F reference distribution as a function of the nominal level, α; the dimension, k; the number of imputations, m; and the average fraction of missing information, $\bar{\gamma}$. Unequal fractions of missing information.

$m = 2$

α	k \ $\bar{\gamma}$	0.1	0.2	0.3	0.5
1%	2	1.0	1.1	1.3	2.6
	3	0.7	1.1	1.5	3.1
	5	0.7	1.3	1.4	3.7
	7	0.8	1.2	2.1	4.2
	10	0.6	1.4	2.1	4.5
	15	0.6	1.6	2.4	5.0
	20	0.7	1.4	2.8	5.1
	25	0.6	1.9	3.2	5.2
5%	2	5.0	4.8	5.3	7.1
	3	4.7	5.0	5.5	7.7
	5	3.7	4.7	5.6	8.1
	7	3.9	5.0	6.2	9.3
	10	3.3	4.6	6.9	8.9
	15	3.5	4.8	6.3	8.0
	20	2.8	4.8	6.7	7.4
	25	2.8	5.2	6.4	7.6
10%	2	10.3	9.4	9.9	12.3
	3	9.3	9.6	9.9	12.5
	5	8.1	9.5	10.1	13.3
	7	7.7	8.9	10.1	13.3
	10	6.9	8.1	10.9	12.3
	15	6.8	8.1	9.7	10.5
	20	5.8	8.2	9.6	9.5
	25	5.3	8.2	9.3	9.2

$m = 3$

α	k \ $\bar{\gamma}$	0.1	0.2	0.3	0.5
1%	2	0.9	0.9	1.1	1.9
	3	0.8	1.2	1.2	2.4
	5	0.7	1.1	1.0	2.6
	7	0.8	1.3	1.6	3.1
	10	0.7	1.0	1.8	3.5
	15	0.8	1.2	1.8	3.5
	20	0.6	1.2	2.3	3.7
	25	0.5	1.6	2.7	3.9
5%	2	5.1	4.9	5.0	6.3
	3	4.7	5.2	5.0	6.9
	5	4.1	4.9	5.3	7.2
	7	4.5	5.4	5.4	8.1
	10	3.6	4.9	5.6	8.0
	15	3.6	4.2	5.8	7.0
	20	3.5	4.7	6.4	6.6
	25	2.8	4.8	6.3	6.5
10%	2	10.6	9.8	9.9	11.2
	3	9.7	9.7	10.2	12.3
	5	9.1	9.6	10.2	12.6
	7	8.5	9.6	10.2	12.9
	10	7.9	8.9	10.0	12.0
	15	7.6	8.2	9.2	10.0
	20	6.9	8.3	10.2	9.5
	25	6.3	8.6	10.0	8.6

$m = 5$

α	k \ $\bar{\gamma}$	0.1	0.2	0.3	0.5
1%	2	0.9	1.0	1.0	1.6
	3	0.8	1.0	1.0	2.0
	5	0.8	1.1	1.2	2.2
	7	0.7	1.2	1.5	2.7
	10	0.7	1.0	1.7	3.1
	15	0.8	0.9	1.4	2.9
	20	0.6	0.9	2.2	3.2
	25	0.6	1.5	2.3	3.2
5%	2	5.2	4.8	4.9	6.2
	3	4.7	5.2	5.1	6.3
	5	4.5	4.6	5.4	7.0
	7	4.5	5.2	5.7	7.9
	10	4.1	4.8	5.6	8.0
	15	4.0	4.3	5.5	7.0
	20	3.9	4.6	6.2	6.9
	25	3.4	4.8	6.1	6.5
10%	2	10.3	9.5	9.6	11.4
	3	9.9	10.1	10.1	12.6
	5	9.0	9.7	10.8	12.5
	7	9.0	9.6	11.1	13.2
	10	8.5	9.8	10.1	12.9
	15	8.5	9.1	9.8	10.7
	20	8.2	8.2	10.1	10.9
	25	7.4	8.7	10.2	9.3

$m = 10$

α	k \ $\bar{\gamma}$	0.1	0.2	0.3	0.5
1%	2	0.9	1.0	1.1	1.1
	3	0.7	1.0	1.1	1.1
	5	0.8	0.8	1.0	1.3
	7	0.9	1.0	1.2	1.7
	10	0.7	1.0	1.1	2.6
	15	0.8	0.8	1.3	2.3
	20	0.7	0.9	1.6	3.0
	25	0.7	1.1	1.8	3.0
5%	2	5.2	4.9	5.2	5.5
	3	5.1	5.3	5.6	6.0
	5	4.5	5.4	5.7	6.7
	7	4.7	5.6	5.9	8.1
	10	4.3	5.4	5.7	8.9
	15	4.5	5.3	5.5	8.4
	20	4.4	5.6	6.3	9.1
	25	4.4	5.3	6.7	8.7
10%	2	10.7	9.8	9.9	11.0
	3	10.2	10.6	11.1	11.7
	5	9.7	10.6	10.9	13.5
	7	9.5	10.8	11.4	14.7
	10	9.1	11.0	11.5	15.8
	15	8.8	10.0	11.1	15.2
	20	8.9	10.7	11.8	15.0
	25	9.2	10.6	12.6	14.8

Table 5. Analysis of variance table for the relative differences between the exact and nominal levels on the z score.

Factor	Degrees of freedom	% of total sum of squares	Ratio of mean square to residual mean square
k	7	0.6	9
m	3	3.0	98
$\bar{\gamma}$	3	58.1	1880
Levels	2	4.7	228
Var(γ)	1	0.7	73
$k \times m$	21	1.3	6
$k \times \bar{\gamma}$	21	8.2	38
$m \times \bar{\gamma}$	9	0.6	7
$k \times$ Levels	14	3.6	25
$m \times$ Levels	6	1.8	29
$\bar{\gamma} \times$ Levels	6	1.4	23
$k \times$ Var(γ)	7	1.5	21
$m \times$ Var(γ)	3	0.9	28
$\bar{\gamma} \times$ Var(γ)	3	1.1	36
Levels $\times$ Var(γ)	2	4.6	226
$k \times m \times \bar{\gamma}$	63	1.2	2
$m \times \bar{\gamma} \times$ Levels	18	0.4	2
$k \times \bar{\gamma} \times$ Levels	42	0.6	1
$k \times m \times$ Levels	42	0.4	1
Residual	494	5.3	1
Total	767		

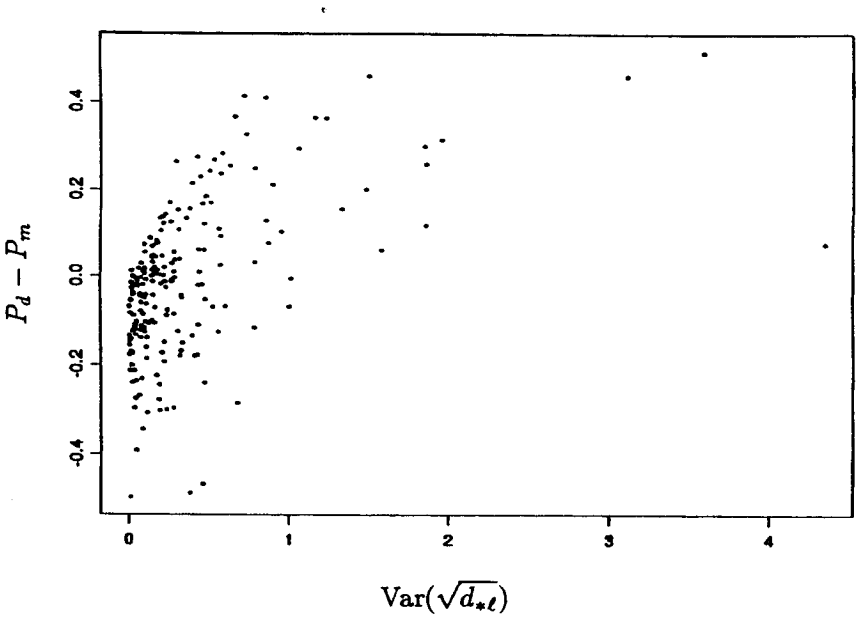

Figure 1a. Plot of $P_d - P_m$ vs $\mathrm{Var}(\sqrt{d_{*\ell}})$ for $k = 5$.

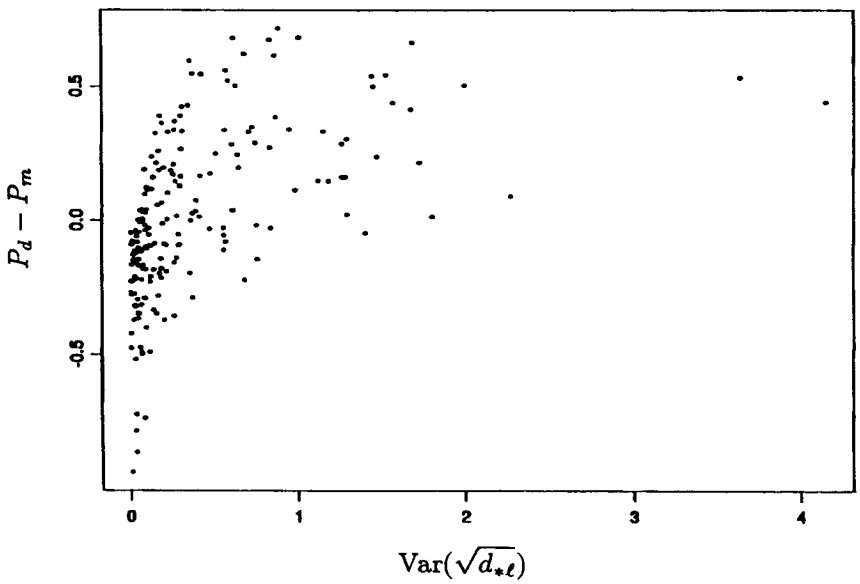

Figure 1b. Plot of $P_d - P_m$ vs $\mathrm{Var}(\sqrt{d_{*\ell}})$ for $k = 20$.

References

Aspin, A. A. (1948). An examination and further development of a formula arising in the problem of comparing two mean values. *Biometrika* **35**, 88-96.

Cochran, W. G. (1964). Approximate significance levels of the Behrens-Fisher test. *Biometrics* **20**, 191-195.

Heitjan, D. F. and Little, R. J. A. (1991). Multiple imputation for the fatal accident reporting system. To appear in *Applied Statistics*.

Heitjan, D. F. and Rubin, D. B. (1990). Inference from coarse data via multiple imputation with application to age heaping. *J. Amer. Statist. Assoc.* **85**, 304-314.

Jeffreys, H. (1940). Note on the Behrens-Fisher formula. *Ann. Eugen.* **10**, 48-51.

Johnson, Palmer O. and Neyman, J. (1936). Tests of certain linear hypotheses and their application to some educational problems. *Statistical Research Memoirs* **1**, 57-93.

Li, K. H. (1985). Hypothesis testing in multiple imputation – with emphasis on mixed-up frequencies in contingency tables. Ph.D. Thesis, Department of Statistics, University of Chicago.

Li, K. H., Raghunathan, T. E. and Rubin, D. B. (1990). Large sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution. Research Report, Department of Statistics, Harvard University.

Meng, X. L. (1988). Significance levels from repeated significance levels in multiple imputation. Ph.D. Qualifying paper, Department of Statistics, Harvard University.

Meng, X. L. and Rubin, D. B. (1990). Likelihood ratio tests with multiply-imputed data sets. *Proceeding of the Statistical Computing Section of the American Statistical Association.* To appear.

Patnaik, P. B. (1949). The non-central χ^2- and F-distributions and their applications. *Biometrika* **36**, 202-232.

Raghunathan, T. E. (1987). Large sample significance levels from multiply-imputed data. Ph.D. Thesis, Department of Statistics, Harvard University.

Robinson, G. K. (1976). Properties of Student's t and of the Behrens-Fisher solution to the two means problem. *Ann. Statist.* **4**, 963-971.

Rubin, D. B. (1978). Multiple imputations in sample surveys – a phenomenological Bayesian approach to nonresponse. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 20-34. Also in *Imputation and Editing of Faulty or Missing Survey Data.* U.S. Dept. of commerce, Bureau of the Census, 1-23.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* John Wiley, New York.

Rubin, D. B. and Schenker, N. (1987). Interval estimation from multiply-imputed data: a case study using census agriculture industry codes. *J. Official Statist.* **3**, 375-387.

Schenker, N., Treiman, D. J. and Weidman, L. (1988). Evaluation of multiply-imputed public-use tapes. *Proceedings of the American Statistical Association Survey Research Methods Section Annual Meetings*, 85-92.

Schenker, N. and Welsh, A. H. (1988). Asymptotic results for multiple imputation. *Ann. Statist.* **16**, 1550-1566.

Smith, H. F. (1936). The problem of comparing the results of two experiments with unequal errors. Council of Scientific and Industrial Research (Australia), *Journal* **9**, 211-212.

Treiman, D. J., Bielby, W. and Cheng, M. (1988). Evaluating a multiple-imputation method for recalibrating 1970 U.S. Census detailed industry codes to the 1980 standard. *Sociological*

Methodology **18**, 309-345.

Wallace, D. L. (1978). The Behrens-Fisher and Feiller-Creasy Problems. *An Appreciation of Fisher* (Edited by S. F. Fienberg and D. V. Hinkley), 119-147.

Welch, B. L. (1937). The significance of the difference between two means when the population variances are unequal. *Biometrika* **29**, 350-362.

Welch, B. L. (1947). The generalization of "Student's" problem when several different population variances are involved. *Biometrika* **34**, 28-35.

Weld, L. H. (1987). Significance levels for public-use data with multiply-imputed industry codes. Ph.D. dissertation, Department of Statistics, Harvard University.

Department of Statistics, Chinese University of Hong Kong, New Territories, Shatin, Hong Kong.

Department of Statistics, Harvard University, Cambridge, MA 02138, U.S.A.

Department of Biostatistics, University of Washington, Seattle, WA 98195, U.S.A.

Department of Statistics, Harvard University, Cambridge, MA 02138, U.S.A.

(Received October 1989; accepted June 1990)