Approximate Error Degrees of Freedom with Randomly Missing Components of Multivariate Vectors

by

M. J. Symons and Y. C. Yuan

Department of Biostatistics

University of North Carolina at Chapel Hill

Presented at

1981 Joint Statistical Meetings

Detroit, Michigan

August 13, 1981

Revised: March, 1982

1.0 Introduction

Missing data items is a practical reality of most statistical investigations. The reasons for an item being missing are numerous and frequently out of the control of the investigator. Experimental animals may die during the course of the experiment; a field plot may be ravaged by pests; subjects may move unexpectedly; records may be lost or otherwise unavailable for some subjects. These are but a few of the uncontrollable reasons for missing data. Additional losses may result from lack of clarity and thoroughness of the protocol, or poor planning and care in data management. However, those losses related to study management are under the control of the investigator and efforts to minimize them are demanding but prudent.

For whatever reasons, every item of missing information introduces problems for the data analyst. In fact, as the amount of missing data increases, the credibility of the statistical inferences is decreased and consequently the validity of the conclusions is weakened. It should be noted that the distribution of missing data items is presumed to be random as regards the treatments being compared. If there is more missing data for one treatment, for example, the reason should be sought out, recorded as a response, and analyzed accordingly.

In the presence of missing data that are randomly dispersed through the data set, iterative strategies are available to estimate the missing values. The analysis often proceeds as if these imputed values were never missing. In particular, the degrees of freedom for error in a likelihood ratio F-test comparing two treatments are specified as if no data were missing. Simulation results suggest that this strategy yields a level of significance that is smaller than the true value, the discrepancy being greater with more missing data. This point was presented by Symons, et al [1978] and replicated in that and the two other situations presented in Section 3.0.

Other situations where only an asymptotic chi-square distribution is available to test the hypothesis of interest will not be addressed.

In this paper we report the simulation results from approximating a reduced number of error degrees of freedom for situations where the likelihood ratio has an F distribution, so that the true level of significance will be more closely represented. Simulation was utilized to determine the 'effective' degrees of freedom for error with a likelihood ratio test of no treatment effects when comparing two treatments.

2.0 General Situation and Summary

Consider the multivariate general linear model situation, specifically

$$\underset{nxp}{\underset{\sim}{Y}} = \underset{nxm}{\underset{\sim}{X}} \; \underset{mxp}{\underset{\sim}{\beta}} + \underset{nxp}{\underset{\sim}{\varepsilon}} \qquad (1)$$

where p responses on n subjects are modelled by the m column design matrix $\underset{\sim}{X}$ and corresponding mxp matrix of parameters, $\underset{\sim}{\beta}$. It is presumed that the i^{th} response vector, $\underset{\sim}{Y}_i$, is approximately p-variate normally distributed with mean $\underset{\sim}{X}_i\underset{\sim}{\beta}$ and covariance matrix $\underset{\sim}{\Sigma}$.

A test of the hypothesis H_0: $\underset{\sim\sim\sim}{C\beta U} = \underset{\sim}{0}$ is desired, where s is the rank of C with $s \leq m$ and u is the rank of U with $u \leq p$. The matrix $\underset{\sim}{C}$ selects/constrains the rows of the matrix $\underset{\sim}{\beta}$, which corresponds to columns of the design matrix, while the matrix $\underset{\sim}{U}$ selects/constrains the columns of $\underset{\sim}{\beta}$ corresponding to the p responses. With

$$\underset{uxu}{\underset{\sim}{S}_E} = \underset{\sim}{U}'[\underset{\sim}{Y}'\underset{\sim}{Y} - \underset{\sim}{Y}'\underset{\sim\sim}{X\hat{\beta}}]\underset{\sim}{U} \quad , \qquad (2)$$

where

$$\underset{m\tilde{x}p}{\hat{\underset{\sim}{\beta}}} = (\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{X}'\underset{\sim}{Y} \qquad (3)$$

and

$$\underset{uxu}{\underset{\sim}{S}_H} = (\underset{\sim\sim\sim}{C\beta U})'[\underset{\sim}{C}(\underset{\sim}{X}'\underset{\sim}{X})^{-1}\underset{\sim}{C}'](\underset{\sim\sim\sim}{C\hat{\beta}U}) \quad , \qquad (4)$$

the largest root criterion of S.N. Roy is

$$F_s = \frac{n - m - u + 1}{|s - u| + 1} \lambda_1 , \qquad (5)$$

where λ_1 is the largest root of the matrix $S_{\sim H} S_{\sim E}^{-1}$. When the minimum of s and u is one, the likelihood ratio criterion and trace criterion of Lawley and Hotelling are also equivalent to F_s and further F_s is then distributed as F with $|s - u| + 1$ and $n - m - u + 1$ degrees of freedom. Of particular interest is a test of two treatments (s=1) compared over each of the u=p responses. Then

$$F_s \doteq F_{p;n-m-p+1} , \qquad (6)$$

under a null hypothesis of no treatment effects and with approximate p-variate normality. For more details, see Anderson [1958], Morrison [1967] and Puri and Sen [1971].

When the sample vectors are not complete and the missing components are distributed at random, a natural way to proceed is to estimate those values by the Missing Information Principle of Orchard and Woodbury [1972], or EM algorithm of Dempster, et al [1976], and then analyze according to the above sketch of standard linear model theory. However, there are not n complete response vectors and consequently the error degrees of freedom are too large. The following tabulation of percentiles for three simulated sample F distributions for testing for a difference in two drugs with 10% missing data. This is a simplified version of Example #1 in Section 3.1. The three sample F distributions corresponds to analyzing the complete data, analyzing the data remaining after deleting any vector with a missing component, and analyzing the data after imputing the missing values. As can be seen the sample F with the missing estimated is too large and appears inferior to the sample F based upon the deletion strategy for this case with a moderate amount

of missing data.

Percentile	Complete Data	Deletion	Missing Estimated
50%	0.688	0.717	0.838
75%	1.326	1.358	1.661
90%	2.253	2.299	2.790
95%	2.971	2.975	3.722
99%	4.389	4.822	5.241

If in fact the missing data are at random, the simulation results described in the next section suggest that the sample F for the strategy of analysis based upon estimating missing values should be adjusted as follows:

$$F_s' = \frac{n'-m-p+1}{n-m-p+1} F_s \quad , \tag{7}$$

where n' is the effective number of complete vectors with fraction f of the n by p data matrix $\underset{\sim}{Y}$ randomly missing. On the basis of simulation results, when m=2 then n' approximately satisfies

$$(1-f)^{p/2}n \leq n' \leq \frac{1}{2}[1 + (1-f)^p]n \quad . \tag{8}$$

That is, the effective sample size with randomly missing data lies between the geometric mean and the arithmetic mean of the total sample size, n, and the number of complete vectors expected presuming independence for each of the p responses with the fraction f of missing data. Therefore, the significance of the statistic (7) should be compared with percentiles of an F with p and n'-m-p+1 degrees of freedom. When there are three or more independent variables, it appears that the effective sample size decreases quickly to number of complete data vectors.

3.0 Simulation Rationale and Results

Simulations were conducted for three experimental situations described below. For each a full data set was generated according to the design matrix;

response components were randomly deleted to acheive a particular proportion of missing data; values were imputed for those missing; and the test statistic (5) was calculated. Under the null hypotheses of no treatment effects, the p-variate normal vectors were simulated by first generating one variate and subsequently generating the rest of the variates conditional on the previously generated component(s). The scheme is described by Kennedy and Gentle [1980]. The IMSL routine GGNML, a random normal generator, was utilized. Due to the different time requirements for the three examples, the simulation specifications for each could not be identical for practical reasons. Therefore, the additional details of simulation are described separately for each example.

However, the rationale leading to the adjusted sample F in (7) and the approximation (8) to the number of effectively complete vectors is common to each example and is sketched next. The F statistic (5) computed for a data set with missing values estimated, was viewed as the ratio of a p degree of freedom chi-square divided by p over an independent n'-m-p+1 degree of freedom chi-square divided by n'-m-p+1. But standard analysis packages applied to the complete data set will, report n-m-p+1 degrees of freedom in the denominator, where n is the number of completed observations. The number of effectively complete observations for the data matrix with missing values is n'. Consequently, the F statistic (5) needs adjustment when calculated for a data set with imputed values for those missing. The rationale for the adjustment indicated by (7) can be seen by writing F_s in (5) as

$$F_s = \frac{n-m-p+1}{n'-m-p+1} F_s' \quad , \tag{9}$$

where F_s' is distributed as an F with p and n'-m-p+1 degrees of freedom under a hypothesis of no treatment effect.

Now n' must be determined. The average over simulated values of F_s

computed for a data set with imputed values will approximately be

$$\overline{F}_s(\text{est}) = \frac{n-m-p+1}{n'-m-p-1} \quad . \tag{10}$$

This can be seen from (9) and recalling that the mean of F_s' is $(n'-m-p+1)/(n'-m-p-1)$. The notation $\overline{F}_s(\text{est})$ is for the average of simulated values of F_s when the fraction f of missing items in the n by p data matrix are estimated. Solving expression (10) for the effective number of complete vectors, we have

$$n' = m+p+1 + \frac{n-m-p+1}{\overline{F}_s(\text{est})} \quad . \tag{11}$$

This gives n' as a function of the number of linearly independent columns m in the design matrix $\underset{\sim}{X}$, the number of responses p, results from simulation $\overline{F}_s(\text{est})$, and the total number observations n. Examining the values of n' by (11) as they relate to the fraction f of missing data led to the conjecture in (8).

3.1 Example #1: Bivariate Response for Two Treatments

A simplified clinical trial setting forms the basis for the first example. The bivariate response corresponds to two follow-up visit measurements on each participant with a baseline evaluation level subtracted off. Two treatments, the experimental with 40 subjects randomly assigned and the standard with 30 subjects, were compared. The n=70 observations were generated for each simulation presuming the covariance matrix

$$\underset{\sim}{\Sigma} = 0.64 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} , \tag{12}$$

with ρ variously specified at -0.7, -0.4, 0.0, 0.1, 0.4, 0.7 and 0.95.

Patterns of missing data with 10%, 20% and 30% were simulated by randomly deleting 4, 8 and 12 elements, respectively, from each bivariate response on the 40 experimental treatment subjects and deleting 3, 6 and 9 elements from each bivariate response on the 30 standard treatment subjects. Averages of 1000 simulated test statistics were computed for the complete vectors, $\bar{F}_s(com)$, vectors with estimated missing values, $\bar{F}_s(est)$, and those vectors that had no missing components, $\bar{F}_s(del)$. The results with $\bar{F}_s(com)$ and $\bar{F}_s(del)$ are reported in the Appendix: Simulation Checks. Reasonable agreement between the simulation averages and F distribution theory was observed. Tables 1, 2 and 3 contain the $\bar{F}_s(est)$ results for the various values of correlation and the calculations of n' for missing data levels 100f = 10%, 20% and 30%, respectively. The results do not appear to depend upon the strength of correlation between the bivariate responses, just as the null distribution of $\bar{F}_s$ in (5) does not.

The determination of the effective number of complete observations, n', in the 70 by 2 data matrix as a function of the percentage 100f of missing data is summarized for all three examples in Section 3.4.

3.2 Example #2: Bivariate Response for Two Treatments, One Covariate and Four Investigators

The features of a clinical trial is the essence of the second example. The bivariate response corresponds to two follow-up visit measurements on each participant. A baseline evaluation is included as a covariate rather than the change from baseline being the bivariate response as in Example #1. The n=128 observations were divided among the four investigators:

39, 24, 33 and 32. Patterns of missing data by investigator for each variable corresponding to roughly 5%, 10% and 15% missing were as follows:

i) Investigator #1: 2, 4 and 6 of 39 participants

ii) Investigator #2: 1, 2 and 4 of 24 participants

iii) Investigator #3: 2, 3 and 5 of 33 participants

iv) Investigator #4: 2, 3 and 5 of 32 participants

Totals: 7(5.47%), 12(9.38%) and 20(15.63%) of 128 (100%)

The data were simulated as a trivariate normal with covariance matrix

$$= 0.64\begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix} \qquad (13)$$

with $\rho = 0.70$. The bivariate response, formed as the conditional distribution of the two follow-up visits conditional on the baseline value, has a correlation of approximately $\rho_{12.0} = 0.57$. The design matrix included $m = 6$ columns: an intercept, the baseline evaluation, the treatment designation and three indicator variables to separate the investigators. No investigator effects were included in the simulation and, of course, there was no treatment effect.

The checks on the simulation were acceptable and are contained in the Appendix: Simulation Checks. Table 4 contains the average of the $\overline{F}_s$(est) results for missing data levels 100f = 5%, 10% and 15%. The larger design matrix slowed the iterative estimation of the missing values considerably, consequently the reduction in the percentage missing for this example. The relation between n' and f is examined in Section 3.4.

3.3 Example #3: Two Correlated Differences from Daily Baseline for Four Days in Comparing Two Treatments

Effects of two pollutants on pulmonary function performance were compared in the third example. Daily changes from baseline performances at two and four hours after exposure were taken as correlated within day. Daily performances when adjusted for baseline determinations were taken as uncorrelated. The p=8 measurements were presumed to have the covariance matrix ($\rho = 0.7$)

$$\underset{\sim}{\Sigma} = 0.16 \begin{bmatrix} 1 & \rho & & & & & & \\ \rho & 1 & & & & & & \\ & & 1 & \rho & & & & \\ & & \rho & 1 & & & & \\ & & & & 1 & \rho & & \\ & & & & \rho & 1 & & \\ & & & & & & 1 & \rho \\ & & & & & & \rho & 1 \end{bmatrix} \qquad (14)$$

0's

The m=2 column design matrix contained a constant and an indicator for type of pollutant exposure. The n=40 subject's response simulated were subjected to random deletion of 2, 4 and 6 pairs of observations on each of the four days.

The checks on the simulation are given in the appendix. Table 5 contains the simulation results for calculating n' at the levels of missing data of 5%, 10% and 15%. With p=8, the proportion missing was reduced to keep simulations within reasonable bounds. The relation between n' and f is examined in the next section for all three examples.

3.4 Relationship of the Effective Sample Size (n'), Fraction Missing (f) and Number of Independent Variables (m)

The relationship of the effective number of complete vectors n' in the n by p response matrix with fraction f of randomly missing elements was not generally quantified. With a simple design matrix of a constant term and one treatment indicator, i.e., m=2, it appears that n' lies between the geometric mean and the arithmetic mean of the total sample size, n, and the number of vectors without any missing components n^0. This is shown in Figure 1 for a bivariate response, i.e., p=2.

However, as Example #2 illustrates also for p=2 but m=6, n' was approximately equal to $n^0 = (1-f)^p n$. This was not anticipated and two additional simulations were conducted based upon Example #1, but with more complex design matrices. One was with m=3 that includes a constant, treatment indicator and covariate. The other was with m=4 allowing two investigators rather than the single covariate of the m=3 cases. These confirmed the result of this one simulation and suggests that n' tends to n^0 rather abruptly in more complex designs.

The results of Example #3 are for p=8, although the missing data were deleted in pairs corresponding to the two measurements for one day both being missing or available. This attempt to reflect practicality may have disturbed the simulation, however. The results are shown in Figure 1, but the number of variables may not be p=8 but something less, and greater than p=4, since the missing values were deleted in pairs.

In summary, it appears for simple designs and with two or three responses that n' could be conservatively approximated by the geometric mean of n and n^0. For more complex designs or experiments with several independent variables, n' should be conservatively specified as n^0.

4.0 Discussion

Pilot simulations were performed to examine the effect on test statistics for multivariate data sets with missing values imputed by the Missing Information Principle or EM algorithm. The findings can be summarized as follows:

(a) The sample F for a test of difference in comparing two treatments that is obtained for usual multivariate general linear model theory, when missing values have been estimated to complete the response vectors, is biased. Uncorrected, the sample F will tend to give a p-value that is too small. A correction is suggested, namely to multiply the F statistic so obtained by $(n'-m-p+1)/(n-m-p+1)$. Recall that n is the total sample size of the experiment, m is the number of columns in the design matrix and p is the number of response variables.

(b) The number of effectively complete response vectors with the missing items is n'. Based upon these few simulation points n' appears to be approximated by the geometric mean or arithmetic mean of the total sample size n and the number of response vectors with no missing items n^0 when the experiment is a comparison of two groups. Conservatism would suggest the geometric mean, $\sqrt{n\ n^0}$

(c) For more complex designs or the inclusion of covariates, n' probably should be conservatively set equal to n^0, the number of response vectors with no missing components. These few simulations suggests that n' approaches n^0 quickly with an increasing fraction of missing components.

The theoretical premise utilized in these simulations should be reexamined. The rationale for the corrective factor and subsequent attempt to approximate n' as a function of f, p and m given n was based on the presumption that the multivariate statistic (5) can be satisfactorily expressed by the definition of an F variate, namely

$$F_s = \frac{\chi^2_{p/p}}{\chi^2_{n'-m-p+1/(n'-m-p+1)}}$$

This may not be generally appropriate. However, apparently it was approximate for the simplest situations considered in these simulation.

Several comments and suggestions for further work are noteworthy.

(i) The error degrees of freedom for F_s in (5) are n-m-p+1 when no components are missing. The detail of the design matrix, represented by the number of linearly independent columns m, and the number of responses p should not be allowed to overwhelm the total sample size n. Probably five or more degrees of freedom for error are a minimum if statistical power of any magnitude is desired. With missing data this consideration becomes even more critical, as the losses to missing data reduce n, and consequently, the error degrees of freedom for the test(s) of interest.

(ii) The patterns of missing data considered in these simulations were random over the full set of p responses. This is not likely to represent actual practice, especially where the p responses correspond to follow-up measurements. Greater losses are probably experienced later in the follow-up period. Since treatment differences may be greatest at these later

times, patterns of missing data with the fraction missing increased at later follow-up times may be very important to consider.

5.0 Relevance to Statistical Practice

Various methods to filling in missing values provide quite satisfactory estimates. The use of selected sub-group means of simple linear regression schemes may provide close estimates to those available from the sophisticated iterative approach of the Missing Information Principle of EM algorithm. The subsequent use of standard packages generates a test statistic that is anti-conservative, unless a price is paid in lost error degrees of freedom. Unfortunately, these simulations permitted only partial examination of the required adjustments.

The alternative is to use the Missing Information Principle or EM algorithm to obtain estimates of the components of β and associated variances. Theoretically sound inferences will be the result. The articles by Rubin [1974], and Beale and Little [1975], John and Prescott [1975] and Jarrett [1978] are useful in this regard.

Acknowledgments

Encouragement and interest in follow-up work on the 1978 presentation in San Diego by the Biometry Division of the Environmental Protection Agency, Research Triangle Park, North Carolina, made this work possible. The interest of Victor Hasselblad, Dennis House and Bill Nelson is appreciated. Special thanks go to John Creason for arranging the Intergovernmental Personnel Act appointment for MJS and the graduate assistantship support for YCY and for his professional encouragment throughout this project.

SELECTED REFERENCES

1. Anderson, T. W. 1958. An Introduction to Multivariate Statistical Analysis. Wiley, New York.

2. Afifi, A. A. and Elashoff, R. M. 1966. Missing values in multivariate statistics--I. Review of the literature. Journal of the American Statistical Association, 61. 595-604.

3. Beale, E. M. L. and Little, R. J. A. 1975. Missing values in multivariate analysis. Journal of the Royal Statistical Society, Series B, 37: 129-145.

4. Dempster, A. P., Rubin, D. B. and Hughes N. 1976. Maximum likelihood from incomplete data via the EM algorithm. Harvard University Research Report 5-38, NS-320.

5. GGNML. IMSL Random Normal Deviate Generator.

6. Jarrett, R. G. 1978. The analysis of designed experiments with missing observations. Applied Statistics, 27. 38-46.

7. John, J. A. and Prescott, P. 1975. Estimating missing values in experiments. Applied Statistics, 24:190-192.

8. Kennedy, W. J. and Gentle, J. E. 1980. Statistical Computing. Marcel Dekker, New York.

9. Morrison, D. F. 1967. Multivariate Statistical Methods. McGraw-Hill, New York.

10. Orchard, T. and Woodbury, M. A. 1972. A missing information principle: theory and applications. Proceedings of the Sixth Berkeley Symposium, I. 697-715.

11. Puri, M. L. and Sen, P. K. 1971. Nonparametric methods in multivariate analysis. Wiley, New York.

12. Rubin, D. B. 1974. Characterizing the estimation of parameters in incomplete-data problems. Journal of the American Statistical Association, 69. 467-474.

13. Symons, M. J., Gillings, D. B., and Donelan, M. A. 1978. A practical comparison of some multivariate analysis strategies for clinical trials with missing data. Presented at Joint Statistical Meetings, San Diego, California.

Table 3. Average* F_s Statistic for Simulations of Size 1000 Corresponding to a Test of Equality of Two Treatments with 30% Missing Data; n = 70; p = 2; m = 2. Seven Values of Correlation.

Simulation (ρ) (1000 per)	$\bar{F}_s$ (est)	$n' = 5 + \frac{67}{\bar{F}_s \text{ (est)}}$
1. ρ = -0.70	1.4348	51.70
2. ρ = -0.40	1.4414	51.48
3. ρ = 0.00	1.4733	50.48
4. ρ = 0.10	1.4367	51.64
5. ρ = 0.40	1.4521	51.14
6. ρ = 0.70	1.4953	49.81
7. ρ = 0.95	1.4098	52.52
Average	1.4491	51.25
Std. Dev.	0.0280	0.88

* The standard deviation of the sample F statistics in the seven simulations of size 1000 each ranged from 0.0449 to 0.0522.

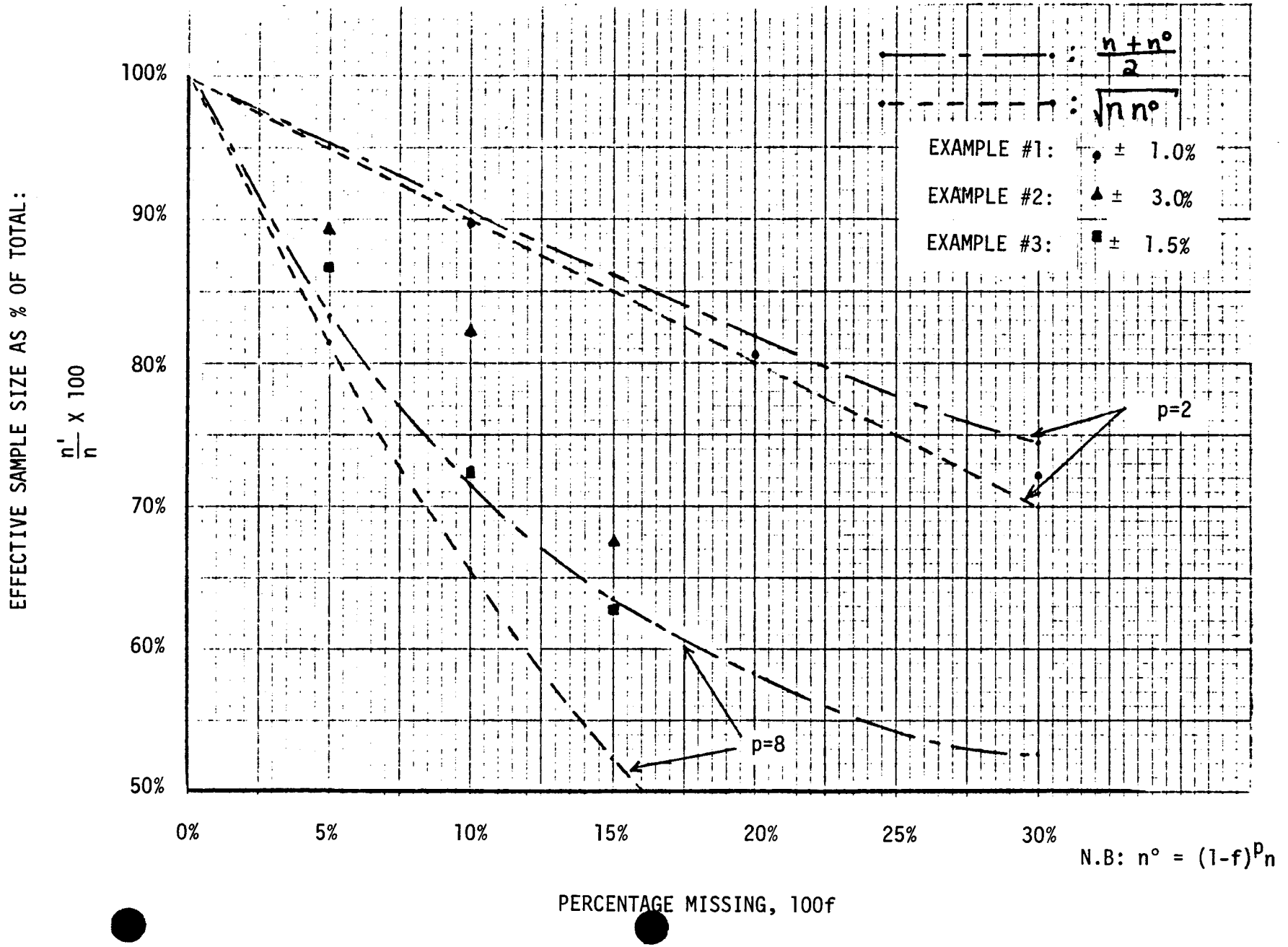
Table 4. Average F_s Statistic for Simulations of Size 1000 Corresponding to a Test of Equality of Two Treatments in Example #2: n = 128; p = 2; m = 6; ρ = 0.70; 100f = 5%, 10% and 15%

Simulation	F_s(est) $\pm$ (std. dev.)	$n' = 9 + \frac{121}{\bar{F}_s(est)}$
1. 5% Missing	1.1428 $\pm$ (0.0405)	114.88
2. 10% Missing	1.2406 $\pm$ (0.0390)	106.53
3. 15% Missing	1.5196 $\pm$ (0.0526)	88.63

Table 5. Average F_s Statistic for Simulations of Size 1000 Corresponding to a Test of Equality of Two Treatments in Example #3; n = 40; p = 8; m =2; 100f = 5%, 10% and 15%

Simulation	F_s(est) $\pm$ (std. dev.)	$n' = 11 + \frac{31}{\bar{F}_s(est)}$
1. 5% Missing	1.2422 $\pm$ 0.0223	35.96
2. 10% Missing	1.5145 $\pm$ 0.0301	31.47
3. 15% Missing	1.7739 $\pm$ 0.0363	28.48

FIGURE 1. EFFECTIVE SAMPLE SIZE, n', AS FUNCTION OF THE PERCENTAGE MISSING, 100f.



Appendix: Simulation Checks

Several internal and external checks were considered to assure the validity of the simulations. Briefly these are identified as follows:

(1) For each simulation replication a complete data set was generated and from this a pattern of missing data was then randomly imposed. The sample F for the complete data was calculated, which should be distributed under a hypothesis of no treatment effects as an F-variate with p and $n-m-p+1$ degrees of freedom. The average over each 1000 simulated data sets of this test statistic, $F_s(com)$, should be approximately $(n-m-p+1)/(n-m-p-1)$.

(2) Analagous to the test statistic with complete data, a sample F was calculated for the data set composed of only complete vectors after items were randomly deleted. Under a hypothesis of no treatment effects, this test statistic based upon the data set deleting observations with any missing components should be distributed as an F-variate with p and $n^o-m-p+1$ degrees of freedom. Of course n^o could vary from simulation to simulation even with a fixed fraction f of missing items from the data matrix. On average, however, one would expect

$$n^o \doteq (1-f)^p n \quad , \tag{A1}$$

with data being missing for independent reasons from each of the p responses at fraction f, i.e., with <u>randomly</u> missing items. Also, the average over each 1000 simulated data sets

of the corresponding test statistic, $\overline{F}_s(del)$, should be approximately $(n^o-m-p+1)/(n^o-m-p-1)$, the mean of the corresponding F distribution.

(3) The adjusted average of $\overline{F}_s(est)$ over each 1000 simulated data sets should be between $\overline{F}_s(com)$ and $\overline{F}_s(del)$. The adjustment is that proposed in (7), specifically as applied to $\overline{F}_s(est)$,

$$\overline{F}'_s(est) = \frac{n'-m-p+1}{n-m-p+1}\,\overline{F}_s(est) \quad . \tag{A2}$$

(4) The average correlation in the simulated data sets should match that specified for the simulation.

(5) Since the simulated data were generated using a normal distribution generator, the sample F statistics computed for the data sets before items were randomly deleted and for the complete vectors remaining with randomly missing items should be distributed as F with p and n-m-p+1 and as F with p and n^o-m-p+1 degrees of freedom, respectively.

The simulation results all agreed very well with those based upon the F distribution and the random generation of the missing values in the n by p response matrix. Tables A1, A2 and A3 contain these comparisons for Example #1, Table A4 for Example #2 and Table A5 for Example #3.

Table A1. Checks on Simulation for Example #1, 10% Missing Data

Simulation (ρ) (1000 per)		$\bar{r}$	$\bar{F}_s$(com)	$\bar{F'}_s$ (est)	$\bar{F}_s$(del)	$\bar{n^o}$
1.	ρ = -0.70	-0.7001	1.0851	1.0361	1.0864	56.68
2.	ρ = -0.40	-0.3989	1.0412	1.0347	1.0593	56.69
3.	ρ = 0.00	0.0038	1.0265	1.0337	1.0322	56.70
4.	ρ = 0.10	0.1021	1.0514	1.0348	1.0479	56.67
5.	ρ = 0.40	0.3982	1.0973	1.0363	1.0854	56.73
6.	ρ = 0.70	0.7001	1.0123	1.0332	1.0036	56.70
7.	ρ = 0.95	0.9494	0.9958	1.0337	1.0208	56.73
Simulation:		Average	1.0442	1.0346	1.0479	56.70
		Standard Deviation	0.0370	0.0371	0.0315	0.02
Theoretical*:		Average	1.0308	1.0340	1.0387	56.70
		Standard Deviation	0.0336	0.0339	0.0341	---

*Based upon $E(F_{k_1,k_2}) = k_2/(k_2-2)$ and $V(F_{k_1,k_2}) = \dfrac{2k_2^2\,(k_1+k_2-2)}{k_1\,(k_2-2)^2\,(k_2-4)}$; $n^o = (1-f)^p\, n$.

Table A2. Checks on Simulation for Example #1, 20% Missing Data

Simulation (ρ) (1000 per)		$\bar{r}$	$\bar{F}_s$(com)	$\bar{F'}_s$ (est)	$\bar{F}_s$(del)	$\bar{n^o}$
1. ρ = -0.70		-0.6962	1.0679	1.0399	1.0946	44.76
2. ρ = -0.40		-0.3892	0.9559	1.0368	1.0402	44.85
3. ρ = 0.00		-0.0018	1.0237	1.0379	1.0764	44.79
4. ρ = 0.10		0.1040	1.0737	1.0406	1.1069	44.81
5. ρ = 0.40		0.4004	1.0719	1.0396	1.0638	44.80
6. ρ = 0.70		0.6988	1.0068	1.0360	0.9791	44.78
7. ρ = 0.95		0.9492	0.9979	1.0379	1.0455	44.81
Simulation:	Average		1.0283	1.0384	1.0581	44.80
	Standard Deviation		0.0451	0.0475	0.0424	0.03
Theoretical*:	Average		1.0308	1.0377	1.0503	44.80
	Standard Deviation		0.0336	0.0341	0.0349	---

*Based upon $E(F_{k_1,k_2}) = k_2/(k_2-2)$ and $V(F_{k_1,k_2}) = \dfrac{2k_2^2\,(k_1+k_2-2)}{k_1\,(k_2-2)^2\,(k_2-4)}$; $n^o = (1-f)^p\, n$.

Table A3. Checks on Simulation for Example #1, 30% Missing Data

Simulation (ρ) (1000 per)	$\bar{r}$	$\bar{F}_s$(com)	$\bar{F'}_s$ (est)	$\bar{F}_s$(del)	$\bar{n^o}$
1. ρ = -0.70	-0.6957	1.0203	1.0429	1.0646	34.27
2. ρ = -0.40	-0.4001	1.0720	1.0430	1.0817	34.42
3. ρ = 0.00	0.0030	1.0448	1.0441	1.0678	34.31
4. ρ = 0.10	0.0979	1.0522	1.0430	1.0587	34.41
5. ρ = 0.40	0.3863	1.0222	1.0433	1.0866	34.31
6. ρ = 0.70	0.6955	1.0040	1.0447	1.0657	34.28
7. ρ = 0.95	0.9488	0.9894	1.0420	1.0672	34.24
Simulation:	Average	1.0293	1.0433	1.0703	34.32
	Standard Deviation	0.0287	0.0214	0.0100	0.07
Theoretical*:	Average	1.0308	1.0432	1.0683	34.30
	Standard Deviation	0.0336	0.0344	0.0362	---

*Based upon $E(F_{k_1,k_2}) = k_2/(k_2-2)$ and $V(F_{k_1,k_2}) = \dfrac{2k_2^2\,(k_1+k_2-2)}{k_1\,(k_2-2)^2\,(k_2-4)}$; $n^o = (1-f)^p\, n$.

Table A4. Checks on Simulation for Example #2

Simulation (1000 per)		$\bar{F}_s$(com)	$\bar{F}'_s$(est)	$\bar{F}_s$(del)	$\bar{n}_o$
1. 5% Missing					
Simulation:	Average	0.9925	1.0189	1.0029	114.42
	Std. Dev.	0.0334	0.0361	0.0351	--
Theoretical*:	Average	1.0168	1.0189	1.0190	114.38
	Std. Dev.	0.0327	0.0328	0.0328	--
2. 10% Missing					
Simulation:	Average	0.9975	1.0205	1.0202	105.14
	Std. Dev.	0.0315	0.0321	0.0317	--
Theoretical*:	Average	1.0168	1.0205	1.0208	105.13
	Std. Dev.	0.0327	0.0329	0.0330	--
3. 15% Missing					
Simulation:	Average	1.0010	1.0252	1.0165	91.18
	Std. Dev.	0.0325	0.0355	0.0333	--
Theoretical*:	Average	1.0168	1.0252	1.0244	91.13
	Std. Dev.	0.0327	0.0332	0.0332	--

*Based upon $E(F_{k_1,k_2}) = k_2/(k_2-2)$ and $V(F_{k_1,k_2}) = \dfrac{2k_2^2\,(k_1+k_2-2)}{k_1\,(k_2-2)^2\,(k_2-4)}$; $n^o = (1-f)^p\, n$.

Table A5. Checks on Simulation for Example #3

Simulation (1000 per)	$\bar{F}_s$(com)	$\bar{F}'_s$(est)	$\bar{F}_s$(del)	$\overline{n^o}$
1. 5% Missing				
Simulation: Average	1.0636	1.0803	1.0730	32.60
Std. Dev.	0.0193	0.0194	0.0196	---
Theoretical*: Average	1.0690	1.0801	1.0927	32.58
Std. Dev.	0.0198	0.0205	0.0212	---
2. 10% Missing				
Simulation: Average	1.0954	1.0978	1.1637	26.24
Std. Dev.	0.019	0.0218	0.0248	---
Theoretical*: Average	1.0970	1.1147	1.1827	20.77
Std. Dev.	0.0198	0.0215	0.0237	---
3. 15% Missing				
Simulation: Average	1.0970	1.1147	1.1827	20.77
Std. Dev.	0.0200	0.0228	0.0288	---
Theoretical*: Average	1.0690	1.1144	1.2024	20.88
Std. Dev.	0.0198	0.0226	0.0286	---

*Based upon $E(F_{k_1,k_2}) = k_2/(k_2-2)$ and $V(F_{k_1,k_2}) = \dfrac{2k_2^2\,(k_1+k_2-2)}{k_1\,(k_2-2)^2\,(k_2-4)}$; $n^o = (1-f)^p\, n$.