AD 678 505

PREDICTION INTERVALS FOR SUMMED TOTALS

J. A. Dei Rossi

Rand Corporation
Santa Monica, California

October 1968

MEMORANDUM
RM-5806-PR
OCTOBER 1968

PREDICTION INTERVALS FOR SUMMED TOTALS

J. A. Dei Rossi

This research is supported by the United States Air Force under Project RAND—Contract No. F44620-67-C-0045—monitored by the Directorate of Operational Requirements and Development Plans, Deputy Chief of Staff, Research and Development, Hq USAF. Views or conclusions contained in this study should not be interpreted as representing the official opinion or policy of the United States Air Force.

DISTRIBUTION STATEMENT
This document has been approved for public release and sale; its distribution is unlimited.

The RAND Corporation
1700 MAIN ST. • SANTA MONICA • CALIFORNIA

Published by The RAND Corporation

PREFACE

The study presented in this Memorandum is part of the RAND Cost Analysis Department's continuing development of cost analysis methodology. Specifically, the study was undertaken to examine the problems and issues involved in calculating prediction intervals for estimates that are sums of individually derived estimates and to explore the practical implications of these issues for cost analysis.

SUMMARY

This Memorandum sets forth methods for calculating prediction intervals for total estimates that are sums of individually derived estimates. Special attention is given to the problem encountered when the variances of each of the individually derived estimates cannot be assumed to be equal. This problem is essentially identical to the well-known Behren-Fisher problem except that here the context is one of deriving a "t-ratio" for summed means rather than for the difference between means. As a consequence, the prediction interval for the case with unequal variances is based on a statistic with an approximate t-distribution, and the interval itself must be viewed as an approximation. However, for purposes of practical application, the approximate nature of these intervals should cause no difficulty and they can be viewed as reasonably accurate representations of the true intervals.

Section I contains the introduction. Section II describes the standard techniques for calculating prediction intervals for individual estimates, and Section III shows the development of the formula for calculating the prediction interval for summed estimates when the variances of the individual estimates can be assumed equal. Section IV addresses the problem of deriving prediction intervals when the assumption of equal variance is inappropriate. Section V contains examples and discussion for each of the cases considered in the preceding sections.

ACKNOWLEDGMENTS

The author acknowledges his indebtedness to Carl Morris of the RAND Mathematics Department for his aid in clarifying the basic theoretical problems raised by the nonequality of variances and for his useful criticisms and suggestions.

CONTENTS

I. INTRODUCTION

With the increased application of system and program analysis to decision problems, the development and use of statistically derived cost estimating relationships (CERs) in these analyses is also becoming more common. One advantage of using statistical techniques in developing CERs is that, in the process of deriving the relationships, useful information about the reliability of the CER can also be generated. The coefficient of determination (or correlation coefficient), the standard error, the "t" statistics for each of the coefficients and other similar information help the user determine how much confidence to place in a given CER as an estimating tool.

In addition, when it is necessary to use CERs to estimate costs for a specific system with given characteristics, it is also possible to make probabilistic statements about the potential magnitude of error associated with that estimate. Such statements are usually made in the form of prediction intervals, delineated by lower and upper bounds on either side of the estimate, with an average probability that the actual system cost will lie within that interval. There are, of course, many sources of uncertainty associated with such estimates that are not taken into consideration in prediction interval calculations. An example of one is the uncertainty associated with extrapolation, and the judgment that must be made as to whether the relationship is valid outside the range of the sample. Nevertheless, prediction interval statements, when properly understood, are useful tools in systems analysis.

In practice, although what is ultimately sought are estimates of total cost, it is often desirable or necessary to develop the CERs for separate subcomponents or subelements which are then summed. This need can arise for various reasons: analytically, it is useful to examine the different life cycle phases separately, and therefore to develop separate estimates for R&D, investment and operation costs; empirically, it becomes necessary to disaggregate when different subcomponents of cost, such as R&D, engineering, and production materials, relate to different variables or react to the same variables differently.

II. PREDICTION INTERVALS FOR SINGLE ESTIMATES

Assume the following array of observations on cost and a corresponding set of characteristics:

$$y_i \; x_{i1} \; x_{i2} \; . \; . \; . \; x_{ip} \quad (i=1, \; . \; . \; ., \; n) \tag{1}$$

where y_i = the ith observation on the cost variable,

x_{ij} = the corresponding ith observation on each of the p characteristics,

n = number of observations,

and the following hypothesis

$$y_i = b_o + b_1 x_{i1} + \; . \; . \; . + b_p x_{ip} + z_i \tag{2}$$

where x_{ij} = observable nonrandom variables (j=1, . . ., p),

b_j = unknown coefficients (j=0, . . ., p), and

z_i = unobservable independent normal random variables with zero expected value and a constant and unknown variance; i.e., $z_i \sim N(0,\sigma^2)$.

With these assumptions

$$E(y_i|x_i) = b_o + b_i x_{i1} + \; . \; . \; . + b_p x_{ip} \tag{3}$$

where x_i = the vector $(x_{i1}, x_{i2}, \; . \; . \; ., x_{ip})$, and

b_j = the unknown coefficients of Eq. (2).

For the sake of brevity, let

$$E(y_i|x_i) \equiv \mu_i \tag{4}$$

so that the CER derived by the method of least squares may be denoted as

$$\hat{\mu}_i = \hat{b}_o + \hat{b}_1 x_{ij} + . . . + \hat{b}_p x_{ip} \tag{5}$$

where $E(\hat{\mu}_i) = \mu_i$ by hypothesis and

$\hat{b}_j$ = the estimated parameters of Eq. (3) chosen so as to minimize $\sum_{i=1}^{n} (y_i - \hat{\mu}_i)^2$.

From the hypothesis and the use of the method of least squares it follows that

$$\frac{r\hat{\sigma}^2}{\sigma^2} \sim X_r^2 \tag{6}$$

i.e., Eq. (6) has a chi-square distribution with r degrees of freedom,*

where $r = n-(p+1)$ and

$$\sigma^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\mu}_i)^2}{r}$$

It also follows that

$$\frac{y_i - \hat{\mu}_i}{a\sigma} \sim N(0,1) \tag{7}$$

since $E(y_i - \hat{\mu}_i) = 0$ and $a^2\sigma^2 = var(y_i - \hat{\mu}_i)$.**

---

*Kendall, M. G. and A. Stuart, The Advanced Theory of Statistics, Fol. 2, Harper Publishing Co., New York, p. 83.

**Ibid., p. 363, where $a^2 = [1 + \frac{1}{n} + x'A^{-1}x]$, A equals (n-1) times the covariance matrix of the variables x_{ij} (i=1, . . . n) and (j=1, . . ., p), and x' equals $[(x_{i1} - \bar{x}_1), . . . (x_{ip} - \bar{x}_p)]$.

Dividing Eq. (6) by its degrees of freedom, taking the square root of this variable and dividing it into Eq. (7) results in a variable with a t distribution with r degrees of freedom:

$$t_r = \frac{(y_i - \hat{\mu}_i)}{\hat{\sigma}a} \quad . \tag{8}$$

The probability statement for the prediction interval based on this t statistic is

$$P(-t_{\alpha/2} < t_r < t_{\alpha/2}) = 1 - \alpha \tag{9}$$

where α is the level of significance or average probability that t_r will lie outside the interval $[-t_{\alpha/2}, t_{\alpha/2}]$.

Rewriting Eq. (9) by substituting Eq. (8) gives

$$P(-t_{\alpha/2} < \frac{y_i - \hat{\mu}_i}{\hat{\sigma}a} < t_{\alpha/2}) = 1 - \alpha \tag{10}$$

so that

$$P(\hat{\mu}_i - \hat{\sigma}a t_{\alpha/2} < y_i < \hat{\mu}_i + \hat{\sigma}a t_{\alpha/2}) = 1 - \alpha \tag{11}$$

and the prediction interval for $\hat{\mu}_i$ at the α level of significance is

$$\hat{\mu}_i \pm t_{\alpha/2}\, \hat{\sigma}a \quad . \tag{12}$$

III. SUMMED TOTALS WITH EQUAL VARIANCES

Now we will consider the case where a prediction interval is desired for a total cost estimate which is the sum of m subelements and the variances for each of these m elements are equal. Let

$$\hat{M}_i = \sum_{i=1}^{m} \hat{\mu}_{ik} \tag{13}$$

and

$$Y_i = \sum_{i=1}^{m} y_{ik} \tag{14}$$

where the subscript k (k=1, . . ., m) has been added to indicate that the $\hat{\mu}_{ik}$ and y_{ik} are values for the kth subelement of the total and that there are now m sets of hypotheses such as that for Eq. (2) and m conditional means estimated from m distinct CERs, such as that shown in Eq. (5).

Assuming that the distributions of the conditional means, μ_{ik}, are independent, the following is a normally distributed random variable with zero mean and unit variance.*

---

*The assumption that the distributions are independent will be maintained throughout the development of prediction interval formulas. While there is evidence that in practice there will be a tendency toward negative correlation in the types of problems for which these prediction interval formulas are likely to be used, this will be treated as essentially an <u>ad hoc</u> problem in this Memorandum. Since the observed sample residuals for each of the m subelement CERs are hypothesized to be drawn from a normally distributed population, it is possible to test the assumption of independence by testing for correlation between the sets of residuals (see p. 14). Thus, it is possible to assess the applicability of the assumption in each case. Further, in instances where the negative correlation is pronounced, its effect should be adverse enough on the results of the regression analysis to discourage the use of the related CERs.

$$\frac{Y_i - \hat{M}_i}{\sqrt{\sum_k \sigma_k^2 a_k^2}} \tag{15}$$

Adding the subscript "k" to Eq. (6) to indicate that it corresponds to the k^{th} component of the summed total gives

$$\frac{r_k \hat{\sigma}_k^2}{\sigma_k^2} \sim \chi^2_{r_k} \quad . \tag{16}$$

Summing the expression in Eq. (16) over the m subelements gives

$$\sum_{k=1}^{m} \left(\frac{r_k \hat{\sigma}_k^2}{\sigma_k^2} \right) . \tag{17}$$

Since by assumption the variances σ_k are equal, Eq. (17) becomes

$$\frac{1}{\sigma^2} \sum_{k=1}^{m} (r_k \hat{\sigma}_k^2) \tag{18}$$

which because of the assumption of independence has a chi-square distribution with $r(=\sum_k r_k)$ degrees of freedom. Similarly, Eq. (15) becomes

$$\frac{Y_i - \hat{M}_i}{\sigma \sqrt{\sum_k a_k^2}} \tag{19}$$

Dividing Eq. (18) by its degrees of freedom, taking the square root and dividing into Eq. (19) gives

$$t_r = \frac{(Y_i - \hat{M}_i)\ (r)^{\frac{1}{2}}}{(\Sigma a_k^2)^{\frac{1}{2}}\ (\Sigma r_k\ \hat{\sigma}_k^2)^{\frac{1}{2}}} \tag{20}$$

which has a "t" distribution with r degrees of freedom.

The corresponding prediction interval is:

$$\hat{M}_i \pm t_{\alpha/2} \left[\frac{\Sigma a_k^2\ \Sigma r_k\ \hat{\sigma}_k^2}{\Sigma r_k} \right]^{\frac{1}{2}} \tag{21}$$

IV. SUMMED TOTAL WITH UNEQUAL VARIANCES

If the unknown variances in the denominator of Eq. (17) are not equal, then the sum is a linear combination of chi-squares rather than the sum of multiples of a chi-square, and therefore, no longer a chi-square. Consequently, even though Eq. (15) still has a normal distribution with zero mean and unit variance, the ratio shown in Eq. (20) will no longer have a "t" distribution and standard techniques for deriving prediction intervals are no longer valid. It is possible, however, to develop approximate prediction interval formulas based on a statistic which has the approximate distribution required. The basic procedure is to define a variable which has the same first moment as the chi-square divided by its degrees of freedom and to estimate the degrees of freedom for which this variable has approximately a chi-square distribution by matching the second moments.* This variable then can be used in developing a variable which, in turn, has approximately a "t" distribution.

Since the variable in Eq. (15) has a normal distribution with zero mean and unit variance, it follows that

$$U = (Y_i - \hat{M}_i) \sim N\left(0, \sum_k \sigma_k^2 a_k^2\right) \tag{22}$$

---

*This technique was worked out for the two-variance case by B. L. Welch ("The Significance of the Difference Between Two Means When the Population Variances are Unequal," Biometrika, No. 29, 1938, p. 350) and is described by M. G. Kendall and A. Stuart in The Advanced Theory of Statistics, Vol. 2, Harper Publishing Co., New York, 1961, pp. 146-147.

Define a new variable θ^2 and its estimated value $\hat{\theta}^2$ as follows:

$$\theta^2 = \sum_k \sigma_k^2 a_k^2 \tag{23}$$

and

$$\hat{\theta}^2 = \sum_k \hat{\sigma}_k^2 a_k^2 \tag{24}$$

then

$$\frac{U}{\hat{\theta}} \sim \frac{\theta N(0,1)}{\sqrt{\frac{\hat{\theta}^2}{\theta^2}\,\theta^2}} \sim \frac{N(0,1)}{\sqrt{\frac{\hat{\theta}^2}{\theta^2}}} \; . \tag{25}$$

Since $E(\hat{\sigma}_k^2) = \sigma_k^2$, it follows that $E(\hat{\theta}) = \theta$;

so that

$$E\left(\frac{\hat{\theta}^2}{\theta^2}\right) = E\left(\frac{\chi_r^2}{r}\right) = 1 \; . \tag{26}$$

Matching the second moments:

$$\mathrm{var}\left(\frac{\hat{\theta}^2}{\theta^2}\right) = \mathrm{var}\left(\frac{\chi_r^2}{r}\right) = \frac{2}{r} \tag{27}$$

and solving for the degrees of freedom gives

$$r = \frac{\theta^4}{\sum_k a_k^4 \sigma_k^4 / r_k} \; . \tag{28}$$

Using $\hat{\theta}$ and $\hat{\sigma}_k$ to estimate the degrees of freedom:

$$\hat{r} = \frac{\hat{\theta}^4}{\Sigma \hat{\sigma}_k^4 a_k^4 / r_k} \quad . \tag{29}$$

Since the first two moments of $\frac{\hat{\theta}^2}{\theta^2}$ with r degrees of freedom, as defined in Eq. (28), match those of $\frac{\chi_r^2}{r}$ with r degrees of freedom as defined in Eq. (18), the variable $\frac{U}{\hat{\theta}}$ as defined in Eq. (25) has approximately a t distribution with $\hat{r}$ degrees of freedom. Thus, the approximate prediction intervals for $\hat{M}_i$ may be expressed as

$$\hat{M}_i \pm t_{\alpha/2} \left[\Sigma \hat{\sigma}_k^2 a_k^2 \right]^{\frac{1}{2}} \tag{30}$$

where $t_{\alpha/2}$ has $\hat{r} = \frac{\hat{\theta}^4}{\sum_k \hat{\sigma}_k^4 a_k^4 / r_k}$ degrees of freedom.

V. EXAMPLE

This section contains an example of the use of formulas presented in the preceding sections, highlighting the significant features of the prediction interval approximation described in Section IV. Modified data on direct base maintenance personnel, taken from RM-4748-PR, are used in the example.* The RM presents CERs for Organizational Maintenance Manhours (ODM), Field Maintenance Manhours (FDM), the sum of ODM and FDM (O + FDM), Communications-Armament-Electronics Maintenance Manhours (CAE) and Total Maintenance Manhours (TDM), the sum of ODM, FDM, and CAE.

For the example, these same CERs were recalculated, dropping the one classified data point. The data required for the calculation of the prediction intervals, based on these results, are shown in Table 1. Lines 1, 2, and 4 represent subcomponents which would normally be added together in an analysis of total costs and comprise the type of summed total of central interest in this Memorandum. However, CERs for the subtotal of ODM and FDM as well as the total sum of ODM, FDM, and CAE were also calculated in the referenced Memorandum, as indicated on lines 3 and 5, respectively, of Table 1. As a consequence, it is possible not only to compare the difference between the results for the summed totals, with and without the assumption of constant variance, but also to compare the difference between intervals for the summed individual estimates and the direct estimate of the summed total.

---

*Dienemann, P. F., and G. C. Sumner, Estimating Aircraft Base Maintenance Personnel, RM-4748-PR, The RAND Corporation, October 1965. The data used are on maintenance manhours per flying hour for fighter aircraft as shown on pp. 24-33.

Table 1

DATA FOR PREDICTION INTERVAL CALCULATION

Item	Sample Mean ($\bar{Y}_k$)	Coefficient of Determination (R^2)	Standard Error (σ_k)	Number of Observations (n_k)	Degrees of Freedom (r_k)
1. ODM	13.556	.923	.922	9	6
2. FDM	12.333	.958	.735	9	6
3. (O+F)DM	25.889	.986	.890	9	5
4. CAE	7.111	.808	2.345	9	7
5. Total	33.000	.962	2.488	9	5

To test the assumption of the independence of the distributions of the conditional means, simple regressions were run on the observed residuals of each of the pairwise combinations: the residuals of the ODM regression were run against the residuals of the FDM regression, etc. The resulting simple correlations are tabulated below:

Residuals on Variables	Simple Correlations
ODM and FDM	-.043
ODM and CAE	.149
FDM and CAE	.321

Since the residual values are assumed to be normally distributed, the hypothesis that distributions are independent can be tested with the hypothesis that the correlations are zero. The critical value for this test at the five-percent level of significance is .798 for 7 degrees of freedom.* None of the sample correlations are greater than this value, so the hypothesis of independence cannot be rejected.

In deciding whether or not the assumption of equal variance is appropriate one may either: (1) test for equality of variance and use the approximation only if the hypothesis of equal variance must be rejected; or (2) make no attempt to bring in the assumption of equality, and use approximation of Section III regardless of the values of the sample variances. As the example will illustrate, the difference in the results will be slight when the sample variances are nearly equal,

---

*Dixon, W. J., and F. J. Massey, Introduction to Statistical Analysis, McGraw-Hill Book Company, Inc., New York, 1959, Table 30-a, p. 468.

regardless of which technique for calculating the intervals is used. However, because of the basic weakness of variance tests with small samples, the first approach is not recommended for small sample cases.

For purposes of illustration, however, tests will be performed on the variances in our example regardless of the small sample size and prediction interval results using both techniques will be presented. Considering the case for the sum of ODM and FDM first, the null hypothesis is var(ODM) = var(FDM) and the test statistic is:

$$\frac{\hat{\sigma}^2_{ODM}}{\hat{\sigma}^2_{FDM}} = \frac{.850}{.540} = 1.57 \ .$$

The critical F value at the 5-percent level of significance is 4.39. Thus, the null hypothesis is not rejected.

Consider next the sum of ODM, FDM, and CAE. In a case with more than two subelements, the hypothesis of equal variance may be tested by testing for equality between the largest sample variance and the smallest. In our example, the null hypothesis for the assumption that var(ODM) = var(FDM) = var(CAE) is simply var(CAE) = var(FDM) and the test statistic is:

$$\frac{\hat{\sigma}^2 CAE}{\hat{\sigma}^2 FDM} = \frac{5.499}{0.540} = 10.2$$

so that the null hypothesis is rejected.

On the basis of these tests, it appears that the assumption of equality of variance is warranted for the sum of ODM and FDM but not for the sum of ODM, FDM, and CAE. However, Table 2 shows the calculated

Table 2

PREDICTION INTERVAL CALCULATIONS FOR INDIVIDUAL ESTIMATES AND SUMMED TOTALS

(α = .05)

Item	Direct Estimates	Prediction Intervals		Summed Total	Prediction Intervals: Equal Variance Assumed		Prediction Intervals: Unequal Variance Assumed	
Item	$\hat{\mu}_i$	$\pm t_{\alpha/2}(\hat{\sigma}^2 a^2)^{\frac{1}{2}}$	Degrees of Freedom	$\Sigma \hat{\mu}_{ik} = \hat{M}_i$	$t_{\alpha/2}\left[\frac{\Sigma a_k^2 \Sigma r_k \hat{\sigma}_k^2}{\Sigma r_k}\right]^{\frac{1}{2}}$	Degrees of Freedom	$t_{\alpha/2}\left[\Sigma \hat{\sigma}_k^2 a_k^2\right]^{\frac{1}{2}}$	Degrees of Freedom
1. ODM	13.556	± 2.378	6					
2. FDM	12.333	± 1.894	6					
3. (O+F)DM	25.889	± 2.412	5	25.889	± 2.711	12	± 2.807	9
4. CAE	7.111	± 5.844	7					
5. Total	33.000	± 6.744	5	33.000	± 5.996	19	± 6.256	9
6. ODM + CAE				20.667	± 5.895	13	± 6.279	6
7. FDM + CAE				19.444	± 5.767	13	± 6.120	7
8. (O+F)DM + CAE	33.000	± 6.744	5	33.000	± 6.116	12	± 6.253	7

prediction intervals for each sum under both assumptions. To facilitate comparisons, all "predicted" values were calculated using the sample mean value of each of the independent variables. Because of the properties of the least-squares estimating relationships, the predicted values are, therefore, equal to the corresponding mean value of each of the dependent variables. In addition to calculating the five cases for which CERs exist, prediction intervals were calculated for three additional combinations. These are shown in lines 6, 7, and 8 of Table 2.

Perhaps the most dominant feature of the intervals for the summed totals in Table 2 is that the approximation is always somewhat more conservative but that the results are not strikingly different. However, several additional general observations can be made. When the assumption of equality appears warranted because of the near equality of the sample variances, the interval for the summed total is, approximately, $t_{\alpha/2}$ times the square root of the number of subelements squared times their weighted average variance. When the assumption of equality is not warranted, the worst case for the approximation occurs if there is significant disparity in both the sample variances and the number of degrees of freedom in each sample.

As hinted in the results of Table 2, the number of degrees of freedom in the smallest sample sets the lower limit for the degrees of freedom for the approximation. The limits of the degrees of freedom for the approximation formula may be expressed as $\min r_k \leq \hat{r} \leq \sum_k r_k$. As the disparity in the sample variances increases the lower limit is approached. Thus, in line 3, where the difference in the sample variances is about 57 percent, there is a reduction of degrees of freedom of 25 percent, from 12 to 9. In line 6, however, the difference is over 600 percent and the reduction in the degrees of freedom is from 13 to the lower limit, 6, which in this case is about a 50 percent reduction.

In comparing the results of the intervals for the summed total with the intervals for the direct estimates, it can be seen that the summed total intervals may be either larger (line 3) or smaller (line 5). It is difficult to establish definitive generalizations as to which total will have the smallest prediction interval, but one interesting fact is that the existence of intervals of unacceptable size on a subelement will not necessarily produce unacceptable results for the summed total. The intervals for CAE (line 3) are almost as large as the estimate for CAE. However, the intervals for the summed total (line 5) are only about 20 percent on either side of the estimate. This difference is due to the fact that CAE is a small percentage of the total and that the intervals (and sample variances) for the other elements are relatively small. This demonstrates that it is possible to develop statistically acceptable estimates of total cost, as measured by the prediction intervals, by subcomponent, even when it is not possible to derive strong estimating relationships for every subelement in the total. Thus, it is not always necessary or desirable to discard useful information about subelement costs by aggregating to higher levels when difficulty is encountered in individual sublement CERs.

DOCUMENT CONTROL DATA

1 ORIGINATING ACTIVITY	2a. REPORT SECURITY CLASSIFICATION
THE RAND CORPORATION	UNCLASSIFIED
	2b. GROUP

3. REPORT TITLE

PREDICTION INTERVALS FOR SUMMED TOTALS

4. AUTHOR(S) (Last name, first name, initial)

Dei Rossi, J. A.

5. REPORT DATE	6a. TOTAL No. OF PAGES	6b. No. OF REFS.
October 1968	27	---

7. CONTRACT OR GRANT No.	8. ORIGINATOR'S REPORT No.
F44620-C-0045	RM-5806-PR

9a. AVAILABILITY/LIMITATION NOTICES	9b. SPONSORING AGENCY
DDC 1	United States Air Force, Project RAND

10. ABSTRACT

Methods for calculating prediction intervals for total estimates that are sums of individually derived estimates. Special attention is given to the problem encountered when the variances of each of the individually derived estimates cannot be assumed to be equal. This problem is essentially identical to the well-known Behren-Fisher problem except that here the context is one of deriving a "t-ratio" for summed means, ~~rather~~ than for the difference between ~~means.~~ Thus, for the case of unequal variances, the prediction interval is based on a statistic with an approximate t-distribution and the interval itself must be viewed as an approximation. This should cause no difficulty, however, since such intervals can be viewed as reasonably accurate representations of the true intervals. Examples are given for each of the cases considered.

11. KEY WORDS

Cost estimating relationships
Statistical methods and processes
Cost analysis