U.S. AIR FORCE RTD-TDR-63-4020

AD ~~429435~~
$0.50

A STATISTICAL METHOD OF REDUCING RANDOM ERRORS IN DATA

A GOVERNMENT RESEARCH REPORT

U.S. DEPARTMENT OF COMMERCE

OFFICE OF TECHNICAL SERVICES

distributes this and thousands of similar reports in the interest of science, industry, and the public—for which research and new products mean better health, better living, and a stronger economy.

HOW TO GET OTHER REPORTS

The Office of Technical Services is the Nation's clearinghouse for reports of research supported by the Army, Navy, Air Force, Atomic Energy Commission, and other Government agencies.

Abstracts of new reports available are published twice a month in U. S. GOVERNMENT RESEARCH REPORTS ($15 a year domestic).

Selected Reports of particular interest to small business are described monthly in TECHNICAL REPORTS NEWSLETTER ($1 a year domestic).

Translations of foreign technical material are also available from the Office of Technical Services and other sources. These are listed or abstracted semimonthly in TECHNICAL TRANSLATIONS ($12 a year domestic).

The above periodicals may be ordered from Superintendent of Documents, U. S. Government Printing Office, Washington, D. C., 20402, or through a U. S. Department of Commerce Field Office.

Inquiries about the availability of reports and translations on any particular subject may be directed to Office of Technical Services, U. S. Department of Commerce, Washington, D.C., 20230, or to any Commerce field office.

Reports and translations are published by the Office of Technical Services for use by the public. Thus, you may use the know-how or reprint the information therein except that where patent questions appear to be involved the usual preliminary search is advised, and where copyrighted material is used permission should be obtained for its further publication.

These documents are reprinted by OTS from the best available copy.

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]
[illegible]
[illegible] COMMAND
[illegible] BASE, OHIO

NOTICES

When Government drawings, specifications, or other data are used for any purpose other than in connection with a definitely related Government procurement operation, the United States Government thereby incurs no responsibility nor any obligation whatsoever; and the fact that the Government may have formulated, furnished, or in any way supplied the said drawings, specifications, or other data, is not to be regarded by implication or otherwise as in any manner licensing the holder or any other person or corporation, or conveying any rights or permission to manufacture, use, or sell any patented invention that may in any way be related thereto.

[illegible]

[illegible]

Copies of this report should not be returned to the Research and Development Division unless return is required by security considerations, contractual obligations, or notice on a specific document.

[illegible] 1953 [illegible]

FOREWORD

This report was prepared by the Fibrous Materials Branch and was initiated under Project No. 7342, "Fundamental Research on Macromolecular Materials and Lubrication Phenomena," Task No. 734202, "Investigations of Structure-Property Relationships of Polymeric Materials." This program was administered under the direction of the Nonmetallic Materials Division, Air Force Materials Laboratory, with Dr. Pierce M. Williamson acting as Project Engineer.

This report covers work from January 1963 to May 1963.

ABSTRACT

A method of modifying observed or calculated data to reduce random errors has been devised. It is based on applying standard statistical calculations to sets of first order differences, the sets varying from each other by the length of intervals used. It is suitable for unrepeatable observations, rather than for laboratory experimental results, and for timewise non-cyclical numerical indexes.

This technical documentary report has been reviewed and is approved.

C. A. Willis

C. A. WILLIS
Chief, Fibrous Materials Branch
AF Materials Laboratory

INTRODUCTION

The concern here is only with series of associated variables for which (1) no theoretical relationships may be posited a priori, and (2) which cannot be repeated. Timewise economic data are of this nature, and so are non-cyclical historical data in general. The concern is with observations and composite effects that cannot be repeated by design either through experimental control of conditions or through assumption of cyclical conditions. An example of the use of cycles is the recording and averaging of lowest temperature at a given place during each February 22. Historical data may span many years, or only the short life of an errant missile.

The problem is to reduce the random errors in a series of data distributed along the coordinate of a variable, such as time. Since no theoretical relationship involving the data can be stated, they are thus treated as unrelated to their surroundings, that is, as constituting a closed system. Hence, only the statistical properties of the data are left to provide means for reducing random errors.

Other methods have been devised for reducing random errors by operating on sets of higher order differences, i.e., differences between differences, differences between differences between differences, etc. Another kind of method is herein proposed, to wit, a method to utilize only sets of first order differences.

Smaller average differences are assumed to exist in a series of data between all data and their nearer predecessors, or successors, than between all data and their farther predecessors, or successors. In principle, the supposition is made that a number of combined, quantitative effects of unknown causes may be observed at points distributed along the coordinate of a continuous variable, and that the lower limit of interval distances between successive points of possible observations is zero. It is then assumed that the average differences between magnitudes observed at both ends of all intervals of a chosen length approaches zero as the chosen length approaches zero. This line of reasoning leads to taking the average difference, $\overline{D}$, and the standard deviation, σ as the definitive criteria of each set of first order differences.

WEIGHTING OF MOVING MEANS

Table 1 presents a series of observations made at successive, equal, time intervals. Now the arithmetic mean is the standard device for indicating central tendency, that is, for reducing random errors which obscure the tendency. Thus, the central tendency of V_{31}, V_{32}, and V_{33} appears to be 104.8 from averaging, respectively, 104.1, 101.0, and 109.4. However, this example of an unweighted mean has the defect of treating the values 104.1, 101.0, and 109.4 as if they were repeated observations of the same thing, which they are not. Each V has its own unknown information content plus or minus a random error. Each observed known is the sum or difference of two unknowns. Furthermore, it is assumed that the true parts of successive V's fluctuate and exhibit non-theoretical trends.

Manuscript released by the author September 1963 for publication as an ASD Technical Documentary Report.

A set of differences, designated as D_1's, between successive V's may be created by taking the absolute difference between each V and its immediate predecessor. Each D_1 is a composite of true information plus or minus a random error. The D_1's are, obviously, not completely random but only partly so.

The true information in each D_1 is always a difference between the true information in two V's, whereas the random error in each D_1 may be either the difference or the sum of the random errors in two V's, depending upon their signs. Since the D_1's are smaller than the V's and contain the sums and differences of original random errors, it seems, intuitively, that the content of randomness in the D_1's is greater than in the V's. Treating the D_1's as random, the average, $\overline{D}_1$, is found and deviations, d_1's, from D_1 are found and squared.

Table 2 shows the results of these computations, including squared standard deviations. Attributing to the d_1's a Gaussian distribution on both sides of $\overline{D}_1$, the well known equation,

$$P_1 = \frac{e^{-\frac{d_1^2}{2\sigma_1^2}}}{\sqrt{2\pi}\ \sigma_1^2}$$

is invoked to state the probability of the occurrence of any d_1.

Again using the data of Table 1, another set of differences is created, designated as D_2's, by taking the absolute difference between each V and its second predecessor. The properties of the D_2's are shown in Table 2 with those of D_1's, D_3's, and D_4's. Likewise,

$$P_2 = \frac{e^{-\frac{d_2^2}{2\sigma_2^2}}}{\sqrt{2\pi}\ \sigma_2^2}$$

states the probability of the occurrence of any d_2.

Table 2 presents four sets of differences. It shows that $\overline{D}_4 > \overline{D}_3 > \overline{D}_2 > \overline{D}_1$, and that $\sigma_4 > \sigma_3 > \sigma_2 > \sigma_1$. (Compare the presence of a trend in magnitude of D's with the absence of a trend in the rolling of a cubical die. Complete randomness in the die rolling results in: $\overline{D}_1 \to \overline{D}_2 \to \overline{D}_3 \to \overline{D}_4 \to \overline{D}_n \to 1.944$ --- with increasing number of rolls.)

Let W_m be the weight assigned to the V corresponding to the median, t_m, of an odd number of successive t's. Then $(W_{m+4} = W_{m-4})$, $(W_{m+3} = W_{m-3})$, $(W_{m+2} = W_{m-2})$, $(W_{m+1} = W_{m-1})$ to be consistent with $\sigma_4 > \sigma_3 > \sigma_2 > \sigma_1$. If W_m is assigned a value of 1, then $W_{m\pm1}$, $W_{m\pm2}$, $W_{m\pm3}$, and $W_{m\pm4}$ should be fractions of decreasing size. A systematic way of assigning decreasing fractional values to $W_{m\pm1}$, $W_{m\pm2}$, $W_{m\pm3}$, $W_{m\pm4}$ is to equate each respectively with P_1, P_2, P_3, and P_4, and to calculate the latter with characteristic and corresponding values of d_1, d_2, d_3 and d_4, and σ_1, σ_2, σ_3, and σ_4.

Now the comparable influences of σ_1, σ_2, σ_3, and σ_4 could be shown by calculating P_1, P_2, P_3, and P_4 for the same value of d, i. e., take $d_1 = d_2 = d_3 = d_4$. But such a demonstration would not utilize the influences of $\overline{D}_1$, $\overline{D}_2$, $\overline{D}_3$ and $\overline{D}_4$. To utilize them, simply take $d_1 = \overline{D}_1$, $d_2 = \overline{D}_2$, $d_3 = \overline{D}_3$ and $d_4 = \overline{D}_4$. It is interesting to note that the resultant P_1 is the probability that $\bar{D}_1 = 0$, i.e., two immediately successive V's are equal. Likewise, P_2 is the probability that two secondly successive V's are equal. Table 2 contains values of P_1, P_2, P_3 and P_4 which are, of course, fractional and progressively smaller. These are now used as values for $W_{m\pm1}$, $W_{m\pm2}$, $W_{m\pm3}$, and $W_{m\pm4}$ in the calculation of weighted means.

Table 2 presents values for $W_{m\pm1}$, $W_{m\pm2}$, $W_{m\pm3}$, and $W_{m\pm4}$. As an example, the weighted mean of V_{31}, V_{32}, and V_{33} is found as follows:

$$\bar{V}_{(31,\,32,\,33)} = \frac{101.0 + 0.0960 \times 104.1 + 0.0960 \times 109.4}{1.0000 + 0.0960 + 0.0960}$$

$$= \frac{101.0 + 20.5}{1.1920}$$

$$= 101.9$$

Likewise, $$\bar{V}_{(30,\,31,\,32,\,33,\,34)} = \frac{101.0 + 20.5 + 0.0586 \times 103.6 + 0.0586 \times 111.0}{1.1920 + 0.0586 + 0.0586}$$

$$= \frac{101.0 + 20.5 + 12.6}{1.3092}$$

$$= 102.4$$

OPTIMUM BASE FOR MOVING MEANS

Obviously, a method of assigning weights does not answer the question of how many weighted data to include in each moving mean. A collection of moving means based on error laden data is smoother than the original data because the means are smoothed by the cancelling tendency of positive and negative random errors. But as the number of data included in each mean is increased, the number of means is decreased. The process could be carried to the ultimate absurdity of converting the whole collection into only one mean. To study the effect of using different numbers of data in weighted moving means, four columns of means are presented in Table 3. The weights used are, of course, P_1, P_2, P_3, and P_4 from Table 2. The root mean square (not standard deviation) of the differences, Δ's, between consecutive V's in each column is calculated. The root mean square derived from the data of each column is a measure of the total dispersion of those data. Hence, decreases in root mean square represent smoothing of data which, in turn, represents reduction of random error. Table 4 shows that broadening the base of averaging achieves only progressively smaller improvements in smoothing.

Returning to the data of Table 3, each value of V_{iii} is subtracted from the corresponding value of V_i on the assumption that each V_{iii} contains less random error than the corresponding V_i. If weighted averaging had accomplished only random error reduction, the sum of all remainders would be zero, since the total of positive random errors should equal the total of negative random errors. However, the total is -0.6, which indicates that V_{iii} values are slightly too high. The operation is repeated with values of V_v, V_{iii}, and V_{ix}, and the results are shown below:

$$(V_i - V_{iii}) = -0.6$$

$$(V_i - V_v) = -0.6$$

$$(V_i - V_{vii}) = -1.6$$

$$(V_i - V_{ix}) = -2.7$$

If V_{iii} and V_v values are slightly too high, the values of V_{iii} are more than slightly too high and those of V_{ix} are even higher. Now recalculated data values that are too high represent loss of information. Therefore, broadening the base of the weighted moving mean increases loss of information. As just shown, the method provides a relative loss of information indicator. Data may be smoothed as far as one wishes to pay the price.

TABLE I

TIME	OBSERVATION	TIME	OBSERVATION
t	V	t	V
1	86.8	21	71.0
2	88.5	22	70.4
3	90.2	23	71.5
4	87.8	24	78.8
5	85.4	25	81.9
6	86.2	26	81.1
7	87.0	27	80.3
8	86.5	28	77.8
9	86.0	29	93.0
10	85.0	30	103.6
11	83.9	31	104.1
12	84.2	32	101.0
13	77.1	33	109.4
14	74.2	34	111.0
15	71.0	35	109.0
16	72.1	36	110.0
17	72.9	37	106.9
18	72.0	38	107.6
19	73.6	39	110.3
20	73.0	40	113.4
		41	112.5
		42	113.1
		43	114.0

TABLE 2

Span of D in Time Intervals	N	$\sum D$	$\bar{D}$	$\sum d^2$	σ^2	σ	P	
1	42	102.2	2.4	379.96	9.2673	3.0	0.0960	$= W_{m\pm1}$
2	41	165.8	4.0	935.78	23.395	4.8	0.0586	$= W_{m\pm2}$
3	40	209.7	5.2	1,319.37	33.830	5.8	0.0460	$= W_{m\pm3}$
4	39	256.5	6.6	1,520.93	40.024	6.3	0.0366	$= W_{m\pm4}$

TABLE 3

Time	Original Data	Weighted Mean of 3 data	Weighted Mean of 5 data	Weighted Mean of 7 data	Weighted Mean of 9 data
t	V_i	V_{iii}	V_v	V_{vii}	V_{ix}
1	86.8				
2	88.5	88.5			
3	90.2	89.8	89.5		
4	87.8	87.8	87.8	87.8	
5	85.4	85.7	85.9	86.1	86.1
6	86.2	86.2	86.3	86.4	86.4
7	87.0	86.9	86.8	86.7	86.7
8	86.5	86.6	86.5	86.4	86.4
9	86.0	86.0	85.9	85.9	85.7
10	85.0	85.1	85.1	84.9	84.6
11	83.9	84.1	83.8	83.6	83.4
12	84.2	83.6	83.3	83.0	82.8
13	77.1	77.4	77.5	77.5	77.6
14	74.2	74.2	74.5	74.8	74.9
15	71.0	71.3	71.6	72.1	72.4
16	72.1	72.1	72.1	72.3	72.6
17	72.9	72.7	72.7	72.8	72.8
18	72.0	72.2	72.3	72.2	72.2
19	73.6	73.4	73.3	73.2	73.0
20	73.0	72.9	72.7	72.7	72.8
21	71.0	71.1	71.3	71.5	71.8
22	70.4	70.6	71.0	71.5	71.8
23	71.5	72.1	72.5	72.8	73.0
24	78.8	78.5	78.3	78.1	78.0
25	81.9	81.6	81.1	80.6	80.7
26	81.1	81.1	80.9	81.0	81.3
27	80.3	80.2	80.8	81.5	81.8
28	77.8	79.2	80.4	81.3	81.7
29	93.0	92.6	92.6	92.4	92.6
30	103.6	102.8	101.6	101.1	100.9

TABLE 3 (Cont'd)

Time	Original Data	Weighted Mean of 3 data	Weighted Mean of 5 data	Weighted Mean of 7 data	Weighted Mean of 9 data
t	V_i	V_{iii}	V_v	V_{vii}	V_{ix}
31	104.1	103.8	103.5	102.9	102.5
32	101.0	101.9	102.4	102.4	101.9
33	109.4	108.9	108.7	108.6	108.2
34	111.0	110.8	110.3	110.0	109.7
35	109.9	110.1	109.9	109.5	109.3
36	110.0	109.8	109.8	109.8	109.7
37	106.9	107.2	107.5	107.8	107.9
38	107.6	107.8	108.1	108.3	108.5
39	110.3	110.3	110.3	110.4	110.5
40	113.4	113.1	112.8	112.6	
41	112.5	112.7	112.7		
42	113.1	113.1			
43	114.0				

TABLE 4

Observations	N	$\sum \Delta^3$	Root Mean Square	Consecutive Decrease of r.m.s.	Total Decrease of r.m.s.
Original, V_i	42	628.60	3.869		
Weighted Av of 3 Data, V_{iii}	40	495.09	3.518	0.351	0.351
Weighted Av of 5 Data, V_v	38	403.05	3.257	0.261	0.612
Weighted Av of 7 Data, V_{vii}	36	346.65	3.103	0.154	0.766
Weighted Av of 9 Data, V_{ix}	34	321.30	3.074	0.029	0.795