VARIANCE COMPONENT ANALYSIS OF CENSUS DATA AND THE USE OF THE HARMONIC MEAN

Ibtihaj S. Arafat, City College of City University of New York

Like every other statistical quantity, averages are used for comparative purposes. Specifically, averages' are used to compare the locations of distinct groups or distributions on the same scale. However, it should be obvious that the different types of averages are not comparable. The ancient Greeks were interested in the phenomenon of the progress of numbers from low to high in regular but different ways. In pursuing this interest, they noticed in particular that the interval between adjacent values may be constant as in the progression 1, 2, 3, 4, 5; that the ratio of adjacent values may be uniform as in the progression 2, 4, 8, 16, 32; that for any three consecutive numbers, A, B, and C, the percentage drop from A to B may be equal to the percentage jump from C to B, as in the progression 1, 1/2, 1/3, 1/4, 1/5. Today, we designate these progressions as arithmetic, geometric, and harmonic respectively (Mueller, et. al., 1970: 146-150). The middle term of each of the series could be calculated by specific formulas. We sum values and divide by N to get the arithmetic mean. Consider the aforementioned geometric series: 2, 4, 8, 16, 32. The middle term is 8, but the arithmetic mean of this progression is 12.4, a value which misrepresents the "central tendency" of the series. To get the middle term, we find the product of all terms and take the Nth root of that product: $\sqrt[5]{(2x4x8x16x32)} = 8$. We refer to a result obtained in this manner as the geometric mean. Thus instead of summing and dividing by N, as in the case of the arithmetic mean, we multiply and find the Nth root.

Consider the harmonic series as given above: 1, 1/2, 1/3, 1/4, 1/5. Neither the arithmetic nor the geometric mean correspond to the middle term of this series. However, the middle term may be obtained by taking the reciprocal of the arithmetic mean of the reciprocals:

Mean of the Reciprocals = $\frac{1+2+3+4+5}{5} = 3$

Reciprocal of Mean of Reciprocals = 1/3.

Although the geometric mean and the harmonic mean occur with some regularity in advanced statistical analysis, they are not likely to be encountered in simple descriptive statistics (Mueller, et.al., 1970: 149). Occasionally, the geometric mean is used to find the size of a population at the midpoint of an interval of time, on the assumption that its growth during that period has been at a constant rate. To illustrate, if the population of a particular community is 5,000 at the time of the first observation and 10,000 at the time of the second observation, the population at the midpoint of the interval would have been approximately 7,000 provided growth was at the same rate throughout the period.

Then what is the harmonic mean, and how could we manipulate it in analyzing census data? According to Marks (1971: 78), "The harmonic mean 'H' is the middle term in a harmonic progression. The harmonic mean of n numbers is 1/nth of the sum of their reciprocals." Spiegel (1961: 49) explained it as follows: "The harmonic mean H of a set of N numbers $X_1, X_2, X_3, \dots X_N$ is the reciprocal of the arithmetic mean of the reciprocals of the numbers:

$$H = \frac{1}{\frac{1}{N}\sum_{j=1}^{N}\frac{1}{X_j}} = \frac{N}{\sum\frac{1}{X}}$$

In practice it may be easier to remember that

$$\frac{1}{H} = \frac{\sum\frac{1}{X}}{N} = \frac{1}{N}\sum\frac{1}{X}."$$

As a computed average, the harmonic mean gives more weight to the smaller values (this is just the opposite of the arithmetic mean), it is capable of algebraic manipulation, and it yields consistent results in most cases (Chou, 1969: 67,82). Chou also mentions that for values which are not all the same and do not have a value of zero, the harmonic mean is always smaller than both the arithmetic mean and the geometric mean. Spiegel (1961: 49) illustrates the relative value of the three means as follows: $H \leq G \leq \bar{X}$ (the equality signs hold only if all the numbers $X_1, X_2, \dots, X_N$ are identical).

In population data, it is not always possible to have an equal number of observations on all treatment combinations. When disproportionality occurs, the analysis of the data becomes difficult. The objective of the researcher in this study is to show how the harmonic mean "H" could be utilized in analyzing census data, which consist of unequal numbers of observations on all subclass combinations. Since the level of measurement of the samples in general includes nominal, ordinal, and ratio measures, the devices used by most of the users of the census data were general statistical descriptive techniques with some ratio and graphic devices. According to Steel and Torrie (1960: 15), the main use of the harmonic mean is in averaging ratios and rates. Steel and Torrie (p.274) also states that in a test of significance for interaction, where unequal subclass numbers are used, the method of weighted squares of means may be used to estimate mean squares for main effects. An alternative procedure is to use the harmonic mean of all subclass n_{ij}'s. They add that calculations are less involved in this case. While the general methods are still applicable, special computing procedures should be used, which depend upon the presence or absence of interaction, although the initial steps are the same. In this case, an analysis of variance could be used with the estimation of variance components rather than F test.

To illustrate the procedure, let us study the different components of the total variation of age distribution by sex, color, and ethnicity. The data are taken from the one-in-a-thousand sample of the 1960 census of the United States of America. (See Table I)

An analysis of variance of the cell means of sex, color, and ethnicity combinations could be performed as illustrated in Table II. The means were considered in the analysis because the number of observations was unequal in the different cells, as shown in Table I. The error degrees of freedom was obtained as the pooled degrees of freedom within cells. The error sum of squares was obtained in the same manner. It was weighted for the inequality of the number of observations in cells by multiplying by the inverse of the harmonic mean of the number of observations within cells

$$\frac{\sum_{i=1}^{K} \frac{1}{N_i}}{K}$$

where K is the number of cells. The effects due to color, ethnicity, and sex are fixed, and hence the effects due to their different interactions are fixed. For this reason θ^2 in the column of the expected mean square (EMS) denotes the sum of the squares of the true effects of the factor or interaction indicated by the subscript of θ^2, divided by the corresponding degrees of freedom. On the other hand, the error was assumed to be random, and its EMS is σ_ε^2, designating the error variance. The mean squares were equated with their corresponding EMS's, and the resulting equations were solved to obtain estimates of the error variance and the variation due to different factors and interactions. The variation due to different factors and interactions was obtained in percent of the total variation. The completed analysis is given in Table II.

In interpreting the results, it could be seen clearly that age distribution in this example was mainly influenced by ethnicity, as illustrated about 50 percent of the total variation was due to ethnicity. The next highest variation, 37.98 percent, was due to color. The estimated θ_s^2 was a small negative quantity and hence could be considered zero, which implies that there is no significant variation between males and females with respect to the average age. The same could be noted about the estimate of color-ethnicity-sex interaction which was a small negative quantity. And so is the case with the estimates of variation due to color-sex and ethnicity-sex interaction. This indicated that the behavior of color under the two sexes was the same, and the behavior of ethnicity was also the same under the two sexes. However, color-ethnicity interaction showed higher variation, 10.74 of the total percent variation. This implied that the behavior of color was somewhat different under different ethnic groups. These observations could be seen clearly by looking on Table II. From the analysis obtained by the use of the harmonic mean it was easy to forecast the magnitude of variation in age distribution due to sex, color, and ethnicity.

The method of the harmonic mean which was used and which was illustrated above could be applied on any data which uses actual numbers such as age or income with unequal number observations on all subclass combinations.

In conclusion, the harmonic mean deserves more attention than it receives in most of the behavioral statistics textbooks. However, the proper use of the harmonic mean depends upon two main considerations. First of all, it must be remembered that an average refers to some class of units that must be appropriate to the use that the average is to serve. And next, it is specially adapted to a situation where the observations are expressed inversely to what is required in the average (Chou, 1969: 68-69).

References

Chou, Ya-Lun.
1969 Statistical Analysis. New York Holt,Rinehart, and Winston, pp. 67-69.

Marks, Robert.
1971 New Mathematics Dictionary and Handbook. New York: Bantam Books, pp. 78-79.

Mueller, John, et.al.
1970 Statistical Reasoning in Sociology. Boston: Houghton Mifflin Co., pp. 146-150.

Spiegel, Murray.
1961 Schaum's Outline Series: Theory and Problems of Statistics. New York: Schaum Publishing Co., p.49.

Steel, Robert and James Torrie.
1960 Principles and Procedures of Statistics. New York: McGraw-Hill, pp. 15, 252-257, 274-276.

TABLE I

AGE DISTRIBUTION BY SEX, COLOR, AND ETHNICITY

Age Interval in Years	White: 3rd Generation Sex M No	%	F No	%	2nd Generation Sex M No	%	F No	%	1st Generation Sex M No	%	F No	%
Less than 1	1803	1.43	1628	1.29	96	0.40	80	0.34	3	0.03	2	0.02
1- 4	6559	5.18	6452	5.10	357	1.51	322	1.36	35	0.39	42	0.47
5- 9	7841	6.19	7433	5.87	416	1.75	399	1.68	60	0.67	79	0.88
10-14	5913	5.46	6928	5.47	444	1.87	371	1.57	114	1.27	86	0.96
15-19	5382	4.25	5311	4.19	441	1.86	414	1.75	94	1.05	81	0.90
20-24	4156	3.28	4209	3.32	503	2.12	492	2.08	104	1.16	147	1.64
25-29	3758	2.97	3836	3.03	693	2.92	681	2.87	163	1.81	172	1.92
30-34	4022	3.18	4113	3.25	936	3.95	1063	4.48	176	1.96	237	2.64
35-39	4134	3.27	4269	3.37	1239	5.23	1293	5.45	244	2.72	306	3.41
40-44	3606	2.85	3760	2.97	1334	5.63	1349	5.69	210	2.34	207	2.30
45-49	3303	2.61	3389	2.68	1255	5.29	1273	5.37	284	3.16	327	3.64
50-54	2951	2.33	2990	2.36	1078	4.55	1074	4.53	369	4.11	427	4.75
55-59	2357	1.86	2588	2.04	848	3.58	871	3.67	482	5.37	495	5.51
60-64	1881	1.49	2112	1.67	614	2.59	780	3.29	500	5.57	545	6.07
65-69	1569	1.24	1830	1.45	545	2.30	672	2.83	564	6.28	516	5.75
70-74	1146	0.91	1415	1.12	366	1.54	503	2.12	469	5.22	450	5.01
75 or more	1221	0.96	1739	1.37	360	1.52	543	2.29	478	5.32	513	5.71
Total	62602	49.45	64002	50.55	11525	48.62	12180	51.38	4349	48.42	4632	51.58
Total	126604				23705				8981			
Color %	79.48				14.88				5.64			
% in Grand Total	70.61				13.22				5.01			

Age Interval in Years	Non-White: 3rd Generation Sex M No	%	F No	%	2nd Generation Sex M No	%	F No	%	1st Generation Sex M No	%	F No	%
Less than 1	315	1.65	289	1.51	8	1.58	4	0.79	0	0.0	0	0.0
1- 4	1059	5.54	1062	5.55	21	4.16	19	3.76	4	1.00	2	0.50
5- 9	1216	6.36	1183	6.18	17	3.37	14	2.77	6	1.50	4	1.00
10-14	1050	5.49	1075	5.62	25	4.95	26	5.15	5	1.25	7	1.75
15-19	767	4.01	826	4.32	22	4.36	22	4.36	8	2.01	6	1.50
20-24	576	3.01	669	3.50	16	3.17	14	2.77	19	4.76	14	3.51
25-29	547	2.86	596	3.12	24	4.75	20	3.96	20	5.01	24	6.02
30-34	567	2.96	674	3.52	21	4.16	36	7.13	21	5.26	18	4.51
35-39	575	3.01	600	3.14	37	7.33	30	5.94	16	4.01	15	3.76
40-44	525	2.74	592	3.10	22	4.36	27	5.35	13	3.26	9	2.26
45-49	504	2.64	525	2.74	17	3.37	18	3.56	14	3.51	17	4.26
50-54	417	2.18	458	2.39	7	1.39	7	1.39	22	5.51	5	1.25
55-59	342	1.79	401	2.10	7	1.39	8	1.58	31	7.77	15	3.76
60-64	255	1.33	287	1.50	2	0.40	2	0.40	15	3.76	16	4.01
65-69	238	1.24	253	1.32	3	0.59	3	0.59	9	2.26	11	2.76
70-74	139	0.73	184	0.96	3	0.59	1	0.20	17	4.26	3	0.75
75 or more	168	0.88	193	1.01	1	0.20	1	0.20	10	2.51	3	0.75
Total	9260	48.41	9867	51.59	253	50.10	252	49.90	230	57.64	169	42.36
Total	19127				505				399			
Color %	95.49				2.52				1.99			
% in Grand Total	10.67				0.28				0.22			

Grand Total = 179321 = 100.0% Sample Total
White = 159290 = 88.83% of Sample Total
Non-White = 20031 = 11.17% of Sample Total
*Percentages for each generation equal 100.0

TABLE II

ANALYSIS OF VARIANCE OF AGE DISTRIBUTION BY SEX, COLOR, AND ETHNICITY

Source	Degrees of Freedom	Sum of Squares	Mean Square	Expected Mean Square
Total	19	2162.2746		
Color (C)	1	642.8056	642.8056	$\sigma_e^2 + 10\ \theta_C^2$
* Ethnicity (E)	4	1345.8717	336.4679	$\sigma_e^2 + 4\ \theta_E^2$
Sex (S)	1	0.5139	0.5139	$\sigma_e^2 + 10\ \theta_S^2$
Col. x Ethn. (CE)	4	155.4215	38.8554	$\sigma_e^2 + 2\ \theta_{CE}^2$
Col. x Sex (CS)	1	4.2421	4.2421	$\sigma_e^2 + 5\ \theta_{CS}^2$
Ethn. x Sex (ES)	4	10.7637	2.6909	$\sigma_e^2 + 2\ \theta_{ES}^2$
Col. x Ethn. x Sex (CES)	4	2.6561	0.6640	$\sigma_e^2 + \theta_{CES}^2$
+ Error	179,301	473,489.9080	2.6408	σ_e^2

TOTAL VARIATION OF PERCENTAGES

Parameters	Estimates of Parameters	Estimate in percent of total
θ_C^2	64.0165	37.98
θ_E^2	83.4568	49.51
θ_S^2	0.0	0.0
θ_{CE}^2	18.1073	10.74
θ_{CS}^2	0.3203	0.19
θ_{ES}^2	0.0251	0.01
θ_{CES}^2	0.0	0.0
σ_e^2	2.6408	1.57
Total	168.5668	100.00

+ The error was calculated as within cells, the error sum of squares was calculated as within cell sum of squares, divided by the harmonic mean of the number of observations within cells.

σ_e^2 = the variance of the error

θ^2 = sum of the squares of the true effect of the factor or the interaction shown by the subscript, divided by the corresponding degrees of freedom.

* In the statistical analysis, Ethnicity was divided into five categories: (1) Native-born with native-born parents, (2) Native-born with foreign-born father and native-born mother, (3) Native-born with native-born father and foreign born mother, (4) Native-born, with both parents, foreign born, and (5) Foreign-born. (Categories 2, 3, and 4 combined make the 2nd generation American group).