A GENERALIZED ANALYSIS OF VARIANCE

By Franklin E. Satterthwaite

University of Iowa and Aetna Life Insurance Company

The analysis of variance is a statistical technique whose fields of application are only beginning to be explored. A few simple standard designs appear in the literature and a great deal has been done with them. However, if the applied statistician limits himself to such standard designs, he soon finds that many of his problems are receiving inadequate or inappropriate treatment. The writer has found this particularly true in his own field where most of the raw data are in the nature of frequencies or averages which lack homogeneity of variance. Also the nature of the problem usually indicates the use of weighted averages rather than simple averages and sometimes part of the data are missing.

The purpose of this study is to examine the fundamental principles underlying analysis of variance designs and to show how designs may be constructed and applied to practically any data which can be assumed to be normally distributed.

1. **Test of independence.** In the analysis of variance we calculate two or more statistics of the types,

$$\chi^2 = \Sigma(x_i - m_i)^2,$$
$$\chi^2 = \Sigma\theta_i^2 .$$

The x_i's are considered to be independent variables from a normal population. The m_i's and the θ_i's are homogeneous linear functions of the x_i's. Heretofore the demonstration of the independence of the χ^2's used has only been made for certain special θ_i's and m_i's. To make our analysis general we shall let our θ_i's be general homogeneous linear functions of the x_i's and we shall define our m_i's through certain linear homogeneous restrictions.

Let us define Chi-square as

$$\chi^2 = \Sigma(x_i - m_i)$$

where the x_i's are independent normally distributed variables with mean zero and unit variance. We also define certain linear functions of the x_i's,[1]

$$\theta_j = a_{ji}x_i , \qquad j = 1, 2, \cdots s, \tag{1}$$

which we shall assume to have been orthogonalized.[2] To define the m_i's we make use of the linear restrictions

[1] A repeated lower case subscript will always indicate summation with respect to that subscript. All subscripts range from 1 to n unless otherwise specified. The Kronecker Delta, δ_{ij}, equals one or zero depending on whether i equals or does not equal j.

[2] The θ_j's are orthogonal if $a_{ik}a_{jk} = \delta_{ij}$. Any algebraically independent set may be

$$a_{ji}(x_i - m_i) = 0, \qquad j = 1, 2, \cdots s, \tag{2}$$

or

$$a_{ji}m_i = a_{ji}x_i = \theta_j .$$

This system has an $(n - s)$-infinitude of solutions and we should not expect all of these to be suitable for our purposes. For reasons which will appear later we shall choose the single solution,

$$m_k = a_{jk}\theta_j = a_{jk}a_{ji}x_i , \qquad j = 1, \cdots s. \tag{3}$$

This is the solution which follows if we complete the system (2) with $n - s$ additional linear restrictions on the m_i's which are homogeneous and which form an orthogonal set with (2). Thus

$$a_{ji}m_i = \theta_j , \qquad j = 1, \cdots s,$$

$$a_{ji}m_i = 0, \qquad j = s + 1, \cdots n.$$

This is consistent with standard analysis of variance designs. For the usual one way analysis, we have

$$\sum_{i=1}^{r} \frac{1}{r^{1/2}} m_{ji} = \sum \frac{1}{r^{1/2}} x_{ji} \qquad j = 1, \cdots s, \tag{4}$$

which yield a solution according to (3),

$$m_{ji} = \frac{1}{r^{1/2}} \frac{1}{r^{1/2}} \Sigma_i x_{ji} , \qquad j = 1, \cdots s.$$

The additional homogeneous restrictions in this case might have been taken as

$$m_{j1} = m_{j2} = \cdots = m_{jr} , \qquad j = 1, \cdots s,$$

which are orthogonal to (4) and may be easily orthogonalized among themselves.

Substituting the values of the m_i's obtained in (3) into Chi-square, we obtain,

$$\begin{aligned} \chi^2 &= (x_i - m_i)(x_i - m_i) \\ &= (\delta_{ik} - a_{ji}a_{jk})x_k(\delta_{il} - a_{mi}a_{ml})x_l , \qquad j, m = 1, \cdots s, \\ &= (\delta_{kl} - a_{mk}a_{ml} - a_{jl}a_{jk} + \delta_{jm}a_{jk}a_{ml})x_k x_l \\ &= (\delta_{kl} - a_{jk}a_{jl})x_k x_l . \end{aligned}$$

---

replaced by an equivalent orthogonal set. Thus, if θ_2 is not orthogonal to θ_1, it may be replaced by $\theta_2^1 = \theta_2 + k\theta_1$, where k is determined by

$$a_{1j}(a_{2j} + ka_{1j}) = 0$$

or $k = -a_{1j}a_{2j}/a_{1j}a_{1j}$.
The condition $\Sigma a_{1j}^2 = \Sigma a_{2j}^2 = 1$ can always be met by simple division.

The sum of the squares of the θ_j's is

$$\Sigma\theta_j^2 = \theta_j\theta_j, \qquad j = 1, 2, \cdots s,$$
$$= a_{jk}a_{jl}x_kx_l .$$

Therefore we have the relation,

$$\chi^2 + \Sigma\theta_j^2 = \delta_{kl}x_kx_l = \Sigma x_k^2 . \tag{5}$$

The rank, R_j, of each θ_j^2 is obviously equal to unity since it is the square of a linear form. The rank, R_0, of χ^2 is at least equal to $n - s$ since the rank of the right hand side of (5) is n. Also, R_0 can not be greater than $n - s$ since,

$$a_{jk}(\delta_{kl} - a_{ik}a_{jk}) = a_{jl} - \delta_{ji}a_{il}, \qquad i, j = 1, \cdots s,$$
$$= 0$$

gives s independent relations between the rows of its coefficient matrix. Therefore we have the relation,

$$R_0 + R_1 + \cdots + R_s = n. \tag{6}$$

The two conditions, (5) and (6), are sufficient[3] conditions for χ^2 and the θ_j^2's each to be independent of the others and each to be distributed as is Chi-square with the number of degrees of freedom equal to its rank.

2. **Adjustment of data.** The above development is not general enough for many practical problems. We do not always have given data, y_i, which are normally distributed about a mean zero with unit (or homogeneous) variance. Of course if the means, $\hat{m}_i$, and variances, σ_i^2, are known, we may make the transformation,

$$x_i = \frac{y_i - \hat{m}_i}{\sigma_i} \tag{7}$$

and apply our theory in a straight forward manner. We shall now check the effect on our analysis if the $\hat{m}_i$'s and σ_i's are determined, in part at least, from our data, the y_i's.

Let us assume that the x_i's of (7) are normally and independently distributed variables about a mean zero and with unit variance. Let us also define certain linear orthogonal functions of the first r of the θ_j's by

$$\phi_k = b_{kj}\theta_j = b_{kj}a_{ji}x_i \qquad k = 1, 2, \cdots q,$$
$$= b_{kj}a_{ji}\left(\frac{y_i - \hat{m}_i}{\sigma_i}\right) \qquad j = 1, 2, \cdots r. \tag{8}$$

We next form the characteristic function of the joint distribution of χ^2, of $\Sigma_1^r\theta_j^2$, of $\theta_{r+1}^2, \cdots \theta_s^2$, and of $\phi_1, \cdots \phi_q$. This is

[3] See A. T. Craig, "On the independence of certain estimates of variance," *Annals of Math. Stat.*, Vol. 9(1938), pp. 46-55.

$$\Phi(t, u, v_{r+1}, \cdots, v_s, w_1, \cdots w_q)$$

$$= K \int \cdots \int \exp [it\chi^2 + iu\Sigma_1^r \theta_j^2 + i\Sigma_{r+1}^s v_j \theta_j^2 + i\Sigma_1^q w_j \phi_j - \tfrac{1}{2}\Sigma_1^n x_j^2] \, dx_n \cdots dx_1.$$

The conditions (5) and (6) are sufficient[4] for there to exist an orthogonal transformation of the x_i's which will convert

$$\begin{aligned} &\theta_j \quad \text{to} \quad \theta_j, && j = 1, \cdots s, \\ &\chi^2 \quad \text{to} \quad \Sigma_{s+1}^n \theta_j^2, \\ &\Sigma_1^n x_j^2 \quad \text{to} \quad \Sigma_1^n \theta_j^2, \\ &dV = \Pi \, dx_j \quad \text{to} \quad \Pi \, d\theta_j. \end{aligned}$$

The characteristic function then takes the form,

$$\Phi = K \left\{ \Pi_1^r \int \exp \left[-\tfrac{1}{2}(1 - 2iu) \left(\theta_j - \frac{iw_k b_{k,j}}{1 - 2iu} \right)^2 \right] d\theta_j \right\}$$

$$\{\Pi_1^q \exp [w_k^2/2(1 - 2iu)]\}$$

$$\left\{ \Pi_{r+1}^s \int \exp [-\tfrac{1}{2}(1 - 2iv)\theta_j^2] \, d\theta_j \right\}$$

$$\left\{ \Pi_{s+1}^n \int \exp [-\tfrac{1}{2}(1 - 2it)\theta_j^2] \, d\theta_j \right\},$$

where

$$\Sigma(w_k b_{kj})^2 = \Sigma w_k^2,$$

since the b_{kj}'s are orthogonal.

At the beginning of this section we stated that we wished in some way to use our data, the y_i's, to estimate the $\dot{m}_i$'s and the σ_i's. A suitable method is to restrict the ϕ functions, (8), to zero.

Our problem thus reduces to finding the distribution of the "array" in our joint distribution for which

$$\phi_1 = \phi_2 = \cdots = \phi_q = 0.$$

Except for perhaps a constant factor, the characteristic function of the distribution of such an array is obtained from Φ by integrating out the w_k's.[5] Thus, on performing the integrations, we have,

---

[4] See A. T. Craig, *ibid.*

[5] This is easily seen since if one passes from the characteristic function to the joint distribution, equates the ϕ_k's to zero, and then passes back to the characteristic function, all the integrations except the above appear in pairs of the form

$$\frac{1}{2\pi} \int e^{itz} \int e^{-itz} \Phi \, dt \, dx,$$

which leave Φ unchanged.

$$\Phi'(t, u, v_{r+1}, \cdots, v_s) = K\{(1 - 2iu)^{-(r-q)/2}\}\{\Pi_{r+1}^{s}(1 - 2iv_j)^{-1/2}\}$$
$$\{(1 - 2it)^{-(n-s)1/2}\}.$$
$$= \Phi_0(u)\{\Pi_{r+1}^{s}\Phi_j(v_j)\}\Phi_n(t),$$

which shows that $\Sigma_1^r \theta_j^2$, θ_{r+1}^2, $\cdots$ θ_s^2, and χ^2 are each independent of the others and that each is distributed according to the Chi-square distribution with $r - q$, 1, $\cdots$ 1, and $n - s$ degrees of freedom respectively.

3. **Numerical application.** The developments of the preceding sections have been abbreviated to cover technical points alone. We shall now take a definite practical problem and see how we may work out its solution with the aid of the above techniques.

In Table I are given the losses, the exposures (in car years) and the indicated pure premiums from the Massachusetts Statutory Liability automobile insurance experience for four towns and for three different classes of cars. (To illustrate the effect of missing items, the data for town D, class W, and for town C, class Y, have been omitted.) Our problem is to determine if there is a significant variation in the indicated pure premium between the different towns and between the different classes of cars.

Our first problem is to set up a normally distributed variable about a mean zero and with homogeneous variance. The true mean, $\hat{m}_i$, of the distribution of the indicated pure premiums, P_i, is unknown. Under the hypothesis that the different towns and classes of cars are homogeneous with each other, we may assume that the $\hat{m}_i$'s are all equal. We may estimate their value by using the combined indicated P for the whole territory, which is \$32.44. By a preliminary argument, which need not concern us here, we show that the variance, σ_i^2, of an indicated pure premium is inversely proportional to the exposure, E_i, on which it is based but the constant of proportionality is unknown. If we now make the assumption that the indicated pure premiums are normally distributed, we may convert them to the form

$$x_i = \frac{P_i - 32.44}{1/E_i^{1/2}}$$

which will be normally distributed about a mean zero with homogeneous variance. We have calculated these statistics and entered them in the table. Because the expected value of P_i, \$32.44, was estimated from our data, the x_i's are subject to the single homogeneous linear restriction,

$$\begin{aligned} 0 &= \Sigma(L_i - \bar{P}E_i) \\ &= \Sigma E_i^{1/2}\frac{P_i - \bar{P}}{1/E_i^{1/2}} \\ &= \Sigma E_i^{1/2}x_i. \end{aligned} \tag{9}$$

TABLE I

i	Class	Town	Exposure E_i	Losses L_i	Premium P_i	$\sqrt{E_i}$	$\sqrt{E}(P-\bar{P})=x_i$	Restrictions, (j)								
								1	2	3	4	5′	6′	5	6	7
1	W	A	4,800	154,500	32.19	69.2	−17.3	69.2				37.1	3.4	69.2		
2		B	3,041	102,900	33.85	55.1	77.6		55.1			30.5	1.6	55.1		
3		C	4,012	147,500	36.77	63.4	274.5			63.4		32.2	−4.9	63.4		
5		T	11,853	404,900	34.16	108.9	(187.2)									
4	X	A	4,188	131,800	31.46	64.7	−63.4	64.7				−30.0	13.2		64.7	
5		B	2,995	83,900	28.02	54.8	−242.2		54.8			−24.4	10.1		54.8	
6		C	4,141	131,700	31.79	64.4	−41.9			64.4		−31.7	4.9		64.4	
7		D	1,004	29,700	29.60	31.6	−89.7				31.6		9.2		31.6	
6		T	12,328	377,100	30.59	111.0	(−205.4)									
8	Y	A	1,365	57,500	42.14	37.0	358.9	37.0				−17.1	−29.4			37.0
9		B	778	16,300	20.93	27.9	−321.1		27.9			−12.4	−22.8			27.9
10		D	406	11,200	27.67	20.2	−96.4				20.2		−14.3			20.2
7		T	2,549	85,000	33.35	50.5	(46.0)									
1	T	A	10,353	343,800	33.21	101.8	(77.6)									
2		B	6,814	203,100	29.81	82.5	(−217.4)									
3		C	8,153	279,200	34.25	90.3	(163.0)									
4		D	1,410	40,900	29.01	37.5	(−128.9)									
Total			26,730	867,000	32.44		(0)									
1′	Σa_{ij}^2						395,360	10,340	6,810	8,170	1,410	6,290	2,010			
2′	$k_{5j}=\Sigma a_{i5}a_{ij}/\Sigma a_{ij}^2$							.464	.446	.492	0					
3′	$k_{6j}=\Sigma a_{i6}a_{ij}/\Sigma a_{ij}^2$							.404	.440	.507	.708	−.846				
4′	Σa_{ij}							170.9	137.8	127.8	51.8	−15.8	−29.0	187.7	215.5	85.1
5′	$\theta_j^2=x_i^2$ or $=(\Sigma\, a_{ij}x_i)^2/\Sigma\, a_{ij}^2$						(173,051)	6,022	47,263	26,568	16,615	49,006	27,576	(35,044)	(42,189)	(2,116)

The next step is to express the indicated pure premiums for each town and for each class of car as θ_j's as defined in equation (1). For town A we have an indicated pure premium of \$33.21 when all classes of cars are combined. This breaks down as follows:

$$\begin{aligned} 33.21 &= \Sigma E_i P_i / \Sigma E_i , && i = 1, 4, 8, \\ &= [\Sigma E_i (x_i / E_i^{1/2} + 32.44)] / \Sigma E_i \\ &= \Sigma (E_i^{1/2} / \Sigma E_i) x_i + 32.44. \end{aligned}$$

Dividing this by the square root of the sum of the squares of the coefficients, we obtain,

$$\begin{aligned} \theta_1 &= (\Sigma E_i)^{1/2} (33.21 - 32.44), && i = 1, 4, 8, \\ &= \Sigma (E_i^{1/2} / (\Sigma E_i)^{1/2}) x_i , \end{aligned} \tag{10}$$

which is of the form of (1). We have entered the coefficients of θ_1 (except for the common denominator, $(\Sigma E_i)^{1/2}$, whose square is entered on line (1')) under Restriction (1) in the table. Similarly, we have entered the values for the other towns under Restrictions (2), (3), and (4). The values for the classes of cars are entered under (5), (6), and (7).

The next step is to orthogonalize the θ_j's. The first four have no common elements so they are orthogonal by inspection. To make θ_5 orthogonal to θ_1 we must add to θ_5,

$$k_{51} = -\Sigma a_{i5} a_{i1} / \Sigma a_{i1}^2$$

times θ_1. This and similar coefficients for making θ_5 orthogonal to θ_2, θ_3, and θ_4 are entered on line (2'). We may now replace θ_5 by the equivalent $\theta_{5'}$ by the formula

$$a_{i5'} = a_{i5} + k_{51} a_{i1} + k_{52} a_{i2} + k_{53} a_{i3} + k_{54} a_{i4} . \tag{11}$$

Similar k's for θ_6 are entered on line (3') and θ_6 is replaced by $\theta_{6'}$. θ_7 should be ignored since it is algebraically dependent on the other θ_j's:

$$\theta_7 \equiv \theta_1 + \theta_2 + \theta_3 + \theta_4 - \theta_5 - \theta_6 .$$

Note that on line (4') we have entered Σa_{ij} for checking the calculation (11).

We next calculate the θ_j^2's according to the formula,

$$\theta_j^2 = [\Sigma a_{ij} x_j]^2 / \Sigma a_{ij}^2 .$$

Note that for this particular design all the θ_j's except $\theta_{5'}$ and $\theta_{6'}$ are numerically equal to the corresponding x_i's (enclosed in parentheses).

Returning to equation (9), we see that it is equivalent to either of the following restrictions on the θ_j's:

$$E_1^{1/2} \theta_1 + E_2^{1/2} \theta_2 + E_3^{1/2} \theta_3 + E_4^{1/2} \theta_4 = 0$$

or

$$E_5^{1/2}\theta_5 + E_6^{1/2}\theta_6 + E_7^{1/2}\theta_7 = 0.$$

Therefore we may conclude that

$$S_1^2/\sigma_x^2 = (\theta_1^2 + \theta_2^2 + \theta_3^2 + \theta_4^2)/\sigma_x^2 = 96{,}469/\sigma_x^2$$

is distributed as is Chi-square with three degrees of freedom. Also we may conclude that

$$S_2^2/\sigma_x^2 = (\theta_5^2 + \theta_6^2 + \theta_7^2)/\sigma_x^2 = 79{,}349/\sigma_x^2,$$

is distributed as is Chi-square with two degrees of freedom. Note that we have not proved, and indeed it is not so, that S_1^2 and S_2^2 are independent.

We have yet to obtain our interaction sum of squares. Equation (5) is of assistance, here giving,

$$S_3^2/\sigma_x^2 = [\Sigma x_i^2 - (\theta_1^2 + \theta_2^2 + \theta_3^2 + \theta_4^2 + \theta_{5'}^2 + \theta_{6'}^2)]/\sigma_x^2$$

$$= \frac{395{,}360 - 173{,}051}{\sigma_x^2} = \frac{222{,}309}{\sigma_x^2}.$$

This is distributed as is Chi-square with $10 - 6 = 4$ degrees of freedom. Also it is independent of S_1^2 and of S_2^2.

Lastly we form the variance ratios

$$F_1 = \frac{96.469/3}{222{,}309/4} = 0.58,$$

$$F_2 = \frac{79{,}349/2}{222{,}309/4} = 0.71,$$

which are not significant.

We therefore conclude that as far as the present data and analysis show, we have no reason to believe that these three classes of cars and these four towns are not all subject to the same true premium rate.