Robustness to Non-Normality of Regression Tests

G. E. P. Box; G. S. Watson

Biometrika, Vol. 49, No. 1/2 (Jun., 1962), 93-106.

Stable URL:
http://links.jstor.org/sici?sici=0006-3444%28196206%2949%3A1%2F2%3C93%3ARTNORT%3E2.0.CO%3B2-4

Biometrika is currently published by Biometrika Trust.

Biometrika (1962), **49**, 1 *and* 2, *p.* 93
With 1 *text-figure*
Printed in Great Britain

Robustness to non-normality of regression tests[†]

By G. E. P. BOX and G. S. WATSON

University of Wisconsin and University of Toronto

1. Summary

A number of statistical procedures involve the comparison of a 'regression' mean square with a 'residual' mean square using the normal-theory F distribution for reference. The use of the procedure for the analysis of actual data implies that the distribution of the mean-square ratio is insensitive to moderate non-normality. Many investigators, in particular Pearson (1931), Geary (1947), Gayen (1950), have considered the sensitivity of this distribution to parent non-normality for important special cases and a very general investigation was carried out by David & Johnson (1951*a*, *b*).

The principal object of this paper is to demonstrate the overriding influence which the *numerical values of the regression variables* have in deciding sensitivity to non-normality and to demonstrate the essential nature of this dependency.

We first obtain a simple approximation to the distribution of the regression F statistic in the non-normal case. This shows that it is 'the extent of non-normality' in the regression variables (the x's), which determines sensitivity to non-normality in the observations (the y's). Our results are illustrated for certain familiar special cases. In particular the well-known robustness of the analysis of variance test to compare means of equal-sized groups and the notorious lack of robustness of the test to compare two estimates of variance from independent samples are discussed in this context. We finally show that it is possible to choose the regression variables so that, to the order of approximation we employ, non-normality in the y's is without effect on the distribution of the test statistic. Our results demonstrate the effect which the choice of *experimental design* has in deciding robustness to non-normality.

2. The mean square ratio R

Suppose that the *response* y_u, observed at the uth set of levels $x_{1u}, x_{2u}, \ldots, x_{pu}$ of p *regression variables*, may be represented by the response function

$$y_u = \beta_0 + \beta_1 x_{1u} + \ldots + \beta_i x_{iu} + \ldots + \beta_p x_{pu} + \epsilon_u \quad (u = 1, 2, \ldots, N), \tag{1}$$

where the *error* $\epsilon_u = y_u - E(y_u)$ is a random variable. The regression variables may be quantitative or merely indicator variables denoting presence or absence of a certain quality. We speak of the N values x_{iu} as the elements of the ith *regression vector* x_i and suppose that the model is set up so that the regression vectors are linearly independent, with $\sum_u x_{iu} = 0$. We denote the usual least squares estimates of $\beta_1, \ldots, \beta_p$ by $b_1, \ldots, b_p$ and write

$$c_{ij} = \sum_{n=1}^{N} x_{iu} x_{ju} \quad (i, j = 1, \ldots, p).$$

To shed light on the plausibility of $\beta_1^*, \beta_2^*, \ldots, \beta_p^*$ as possible values of the coefficients, the ratio of mean squares

$$R_\beta^* = \{S_{\beta*}/p\}/\{S_E/(N-p-1)\} \tag{2}$$

[†] This research was supported in part by the United States Navy, through the Office of Naval Research, and by the United States Army through the Office of Ordnance Research.

may be calculated. The appropriateness of this ratio can be seen from the fact that the *regression* sum of squares

$$S_{\beta*} = \sum_{i=1}^{p} \sum_{j=1}^{p} c_{ij}(b_i - \beta_i^*)(b_j - \beta_j^*) \tag{3}$$

is a measure of overall discrepancy between the least-squares estimates b_i and the contemplated values β_i^*, while the *residual* sum of squares

$$S_E = \sum_{u=1}^{N} (y_u - \hat{y}_u)^2 \quad \text{with} \quad \hat{y}_u = b_0 + \sum_{i=1}^{p} b_i x_{iu} \tag{4}$$

measures the internal consistency of the data independently of the choice of β_i^*. Furthermore, provided that the assumed form of response function is adequate, and that the errors have equal variances and are uncorrelated one with another, the expected values of the component sums of squares are

$$E(S_{\beta*}) = \sum_{i=1}^{p} \sum_{j=1}^{p} c_{ij}(\beta_i - \beta_i^*)(\beta_j - \beta_j^*) + p\sigma^2, \tag{5}$$

$$E(S_E) = (N - p - 1)\sigma^2. \tag{6}$$

If in addition the errors could be supposed to follow normal distributions, that is, if the joint distribution of the ϵ_u was a spherical normal distribution, then when $\beta_i^* = \beta_i$ $(i = 1, 2, \ldots, p)$ $R_{\beta*}$ would have an F distribution with $\nu_1 = p$ and $\nu_2 = N - p - 1$ degrees of freedom. Since in the analysis of real data we do not know the precise distribution of the ϵ_u, one would like to know under what circumstances the F distribution still supplied an adequate approximation to the distribution of $R_{\beta*}$. In particular, as has been shown by Fisher (1947) and Scheffé (1953), when the approximation was adequate the statistic $R_{\beta*}$ could be used not only to test hypotheses concerning particular values β^* of the coefficients but also to supply interval estimates for any linear combinations whatever of the β's.

For the 'special' case $\beta_i^* = 0$ $(i = 1, \ldots, p)$ we denote the mean-square ratio by R. Thus

$$R = \frac{S_0/p}{S_E/(N - p - 1)}, \tag{7}$$

where

$$S_0 = \sum_{i=1}^{p} \sum_{j=1}^{p} c_{ij} b_i b_j, \quad S_E = \Sigma\hat{y}^2 - N\bar{y}^2 - S_0. \tag{8}$$

In what follows we consider only this case since the more general situation always can be reduced to it. For example, if we calculate $\dot{R}$ from the constructed observations

$$\dot{y}_u = y_u - \beta_1^* x_{1u} - \beta_2^* x_{2u} - \ldots - \beta_p^* x_{pu},$$

then $\dot{R}$ is identical with $R_{\beta*}$. In what follows the reference of R to the normal-theory F distribution is referred to as the *general regression test.*

3. A familiar comparison

That the answer to the question 'What effect has non-normality on the general regression test?' is profoundly influenced by the nature of the x vectors may be demonstrated by a familiar comparison. By different choices of the x vectors, almost the same regression model can be made to reproduce on the one hand a test to compare means which is little affected by non-normality and on the other a comparison of variances test which is notoriously sensitive to non-normality.

3·1. *Comparison of means*

If in the general regression model of equation (1) we write $N = (p+1)n$, $x_{iu} = p/(p+1)$ for $u = n(i-1)+t$ $(i=1,\ldots,p;\ t=1,2,\ldots,n)$ and $x_{iu} = -1/(p+1)$ otherwise, then we obtain the regression model for the n observations in each of $p+1$ groups, with the elements in each regression vector adjusted to add to zero, so that the appropriate variance mean-square ratio is

$$R_m = \frac{n \sum\limits_{i=1}^{p+1} (\bar{y}_i - \bar{y})^2/p}{\sum\limits_{i=1}^{p+1} \sum\limits_{t=1}^{n} (y_{it} - \bar{y}_i)^2/(N-p-1)}, \tag{9}$$

where $\bar{y}_i$ is the mean for the ith group of n observations and $\bar{y}$ the overall mean.

3·2. *Comparison of variances*

Suppose there are $N-1$ observations and in the general regression model let us temporarily relax the provision that $\sum\limits_{u=1}^{N-1} x_{iu} = 0$ and suppose that the constant term β_0 is known. Now put $x_{iu} = 1$ when $u = i$ $(i=1,2,\ldots,p)$ and $x_{iu} = 0$ otherwise. Then the mean-square ratio is

$$R_v = \frac{\sum\limits_{u=1}^{p} (y_u - \beta_0)^2/p}{\sum\limits_{s=p+1}^{N-1} (y_s - \beta_0)^2/(N-p-1)}. \tag{10}$$

This has the form of the standard test for the comparison of variances (the population means being known to equal β_0) for two independent samples of size $n_1 = p$ and $n_2 = N-p-1$.

The criteria R_m and R_v can each be obtained therefore as particular cases of the general regression criterion. Furthermore, on the usual assumptions, the null distribution of each criterion is the normal-theory F with $\nu_1 = p$ and $\nu_2 = N-p-1$ degrees of freedom. It is well known, however, that whereas R_m has a distribution which is remarkably insensitive to departures from normality in the parent population this is not the case for R_v.

Specifically, using a method which we discuss later in more detail, Box & Andersen (1955) have shown that in the non-normal situation R_m and R_v have distributions which may be approximated by F distributions with modified degrees of freedom. They show that R_m is approximately distributed as F with $\nu_1 = \delta_m p$ and $\nu_2 = \delta_m(N-p-1)$ degrees of freedom and R_v is approximately distributed as F with $\nu_1 = \delta_v p$ and $\nu_2 = \delta_v(N-p-1)$ degrees of freedom, where for moderate non-normality and moderate numbers of observations the δ's are approximately

$$\delta_m^{-1} = 1-(1/N)\,\Gamma_y, \quad \delta_v^{-1} = 1+\tfrac{1}{2}\Gamma'_y. \tag{11}$$

The measures of kurtosis Γ_y and Γ'_y in these expressions are basically similar one to the other and take zero values when the distribution is normal. Explicitly†

$$\Gamma_y = E\{C_y\} = E\left\{\frac{k_4}{k_2^2}\right\}, \quad \Gamma'_y = E\{C'_y\} = E\left\{\frac{N+1}{N-1}\frac{m_4}{m_2^2}\right\} - 3, \tag{12}$$

† Asymptotic expansions for these constants in terms of the standardized cumulants of the parent distributions are given by Box & Andersen (1955).

where k_2 and k_4 are the usual k statistics for the whole sample of N observations and m_q is the qth moment about the mean

$$m_q = \left\{\sum_{u=1}^{n_1}(y_u-\beta_0)^q + \sum_{s=n_1+1}^{n_1+n_2}(y_s-\beta_0)^q\right\}\Big/(N-1)$$

for the sample of $N-1$ observations.

The insensitivity to non-normality shown by R_m arises because the corrective factor is of order N^{-1}, whereas that for R_v is of order 1. From our present point of view the example serves to show that different choices of the x's in the general regression model can change sensitivity to non-normality by a factor as large as $\frac{1}{2}N$.

We shall now investigate the specific nature of this dependence of sensitivity on the nature of the x vectors. We shall be able to show that the corrective factor for the general regression test is, to order N^{-1}

$$\delta^{-1} = 1+\Gamma_y C_X/2N, \tag{13}$$

where C_X is a measure analogous to Γ_y of 'non-normality' in the x's. In the two special examples considered above C_X approaches respectively its smallest possible value of -2 and its largest possible value of N.

4. Permutation moments of R

In matrix notation our model becomes

$$y = \mathbf{1}'\boldsymbol{\beta}_0+\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\epsilon}, \tag{14}$$

where $\mathbf{1}$ is a column of N unities, $\mathbf{X}$ the $N+p$ matrix of the levels of the p regression variables, $\mathbf{y}$ the column of the N observed responses, and $\boldsymbol{\beta}$ the column of coefficients $\beta_1, \ldots, \beta_p$. Then, since we suppose that $\mathbf{X}'\mathbf{1} = \mathbf{0}$, the least-squares estimators b_0 and $\mathbf{b}$ of β_0 and $\boldsymbol{\beta}$ are given by

$$b_0 = \bar{y}, \quad \mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \tag{15}$$

Writing $\mathbf{z}$ for the column of N deviations $z_u = y_u-\bar{y}$ $(u=1, \ldots, N)$ and $\mathbf{M} = \{M_{uv}\}$ for the symmetric idempotent matrix $\mathbf{M} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$, we have $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}$, whence the regression and residual sums of squares defined in (8) and (4) are given by

$$S_0 = \mathbf{z}'\mathbf{M}\mathbf{z}, \quad S_E = \mathbf{z}'(\mathbf{I}-\mathbf{M})\,\mathbf{z}. \tag{16}$$

We are concerned with the distribution of $R = \{(N-p-1)\,S_0\}/\{pS_E\}$ when $\boldsymbol{\beta} = \mathbf{0}$, that is when

$$y = \mathbf{1}'\beta_0+\boldsymbol{\epsilon}.$$

Now if the error vector $\boldsymbol{\epsilon}$ is spherically normally distributed, then R has an F distribution with $\nu_1 = p$ and $\nu_2 = N-p-1$ degrees of freedom. Equivalently

$$W = \frac{S_0}{S_0+S_E} = \frac{S_0}{\mathbf{z}'\mathbf{z}} \tag{17}$$

has a beta distribution with $\nu_1 = p$ and $\nu_2 = N-p-1$ degrees of freedom.

To study the distribution of R on less specific assumptions we suppose only that the errors have a symmetric distribution, that is to say we suppose that the probability density function $p(\epsilon)$ of the vector $\boldsymbol{\epsilon}$ is a symmetric function of the elements of $\boldsymbol{\epsilon}$. As a special case this includes the chief possibility of interest to us here, that $p(\boldsymbol{\epsilon}) = \prod_{u=1}^{N} p(\epsilon_u)$, where $p(\epsilon_u)$ is any distribution whatever. Now suppose a probability density to be associated with a particular $\boldsymbol{\epsilon}$, then this same probability density is associated with every rearrangement of the elements of $\boldsymbol{\epsilon}$. Also, since $z_u = y_u - \bar{y} = \epsilon_u - \bar{\epsilon}$, it follows that whatever be the probability density associated with a particular $\mathbf{z}$ this same probability density is associated with every rearrangement of the elements of $\mathbf{z}$. The distribution obtained by associating a probability $1/N!$ with each of the possible $N!$ values of any function $f(\mathbf{z})$ obtained by rearranging the elements $\boldsymbol{\epsilon}$ is called the permutation distribution of $f(\mathbf{z})$. The mean and variance of this permutation distribution we denote by $E_P f(z)$ and $E_P f(z)$, respectively.

Now suppose we define *different samples* to mean vectors $\mathbf{z}$ which cannot be made identical by rearrangement of the elements, then if E_S denotes the expected value taken over all different samples $\mathbf{z}$, the overall moments $Ef(\mathbf{z})$ and $Vf(\mathbf{z})$ of $f(\mathbf{z})$ are given by

$$\left.\begin{aligned} Ef(\mathbf{z}) &= \underset{s}{E}\{E_P f(\mathbf{z})\}, \\ Vf(\mathbf{z}) &= \underset{s}{E}\{V_P f(\mathbf{z})\}. \end{aligned}\right\} \tag{18}$$

To approximate the distribution of W for any error distribution of form $p(\epsilon) = \prod_{u=1}^{N} p(\epsilon_u)$ it is convenient for our purpose to find $E(W)$ and $V(W)$ by first finding the appropriate permutation moments and then taking their expected values over all samples.

In what follows it is helpful to express our results in terms of the power sums

$$\sum_{u=1}^{N} z_u^r = S_r$$

which, of course, remain constant under permutations of the ϵ_u. We have in particular

$$S_1 = 0, \quad S_2 = \mathbf{z}'\mathbf{z} = S_0 + S_E, \quad W = S_0/S_2.$$

4·1. $E_P(W)$

We have $$E_P(z_u^2) = S_2/N, \quad E_P(z_u z_v) = -S_2/\{N(N-1)\}$$

whence $$E_P(\mathbf{z}\mathbf{z}') = S_2/\{N(N-1)\}\{N\mathbf{I} - \mathbf{1}\mathbf{1}'\}. \tag{19}$$

Now $\mathbf{z}'\mathbf{M}\mathbf{z} = \operatorname{tr}(\mathbf{M}\mathbf{z}\mathbf{z}')$ where $\operatorname{tr}(\mathbf{A})$ denotes the *trace* of the matrix $\mathbf{A}$. Using the linearity of the trace and expectation operators we have

$$\begin{aligned} S_2 E_P(W) = E_P(S_0) = E_P(\mathbf{z}'\mathbf{M}\mathbf{z}) &= E_P \operatorname{tr}(\mathbf{M}\mathbf{z}\mathbf{z}') \qquad (20) \\ &= \operatorname{tr}\{\mathbf{M}E_P(\mathbf{z}\mathbf{z}')\}. \end{aligned}$$

But $\mathbf{M}\mathbf{1} = \mathbf{0}$ and $\operatorname{tr}(\mathbf{M}) = \operatorname{tr}\{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\} = \operatorname{tr}\{\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\} = p$. Hence on substituting (19) in (20) we obtain finally

$$S_2 E_P(W) = pS_2/(N-1),$$

that is $$E_P(W) = p/(N-1), \tag{21}$$

as is obtained on classical regression assumptions. In particular we see that $E_P(W) = E_N(W)$, where $E_N(W)$ is the expected value of W on the usual normal theory.

4·2. $V_P(W)$

In what follows summations and permutation expectations are taken over all combinations for which the subscripts are unequal. Thus $E_P\{z_1^\alpha z_2^\beta z_3^\gamma z_4^\delta\}$ means the average value of $z_t^\alpha z_u^\beta z_v^\gamma z_w^\delta$ taken over all permutations for which t, u, v and w are unequal. Similarly $\Sigma M_{13} M_{23}$

means $\sum_{t} \sum_{u \neq t} \sum_{v \neq u \neq t} M_{tv} M_{uv}$. Also, since the x's and hence the M's can be regarded as remaining fixed whilst the combinations of the z's pass through all possible permutations, we have, for example,

$$E_P \Sigma z_1^\alpha z_2^\beta z_3^\gamma z_4^\delta M_{13} M_{23} = E_P(z_1^\alpha z_2^\beta z_3^\gamma z_4^\delta) \Sigma M_{13} M_{23}.$$

On squaring the expression

$$S_0 = \Sigma z_1^2 M_{11} + \Sigma z_1 z_2 M_{12}$$

and taking expectations, we then have

$$\begin{aligned} E_P(S_0^2) = {} & E_P(z_1^4) \Sigma M_{11}^2 + E_P(z_1^2 z_2^2)(2\Sigma M_{12}^2 + \Sigma M_{11} M_{22}) \\ & + E_P(z_1 z_2^2 z_3)(4\Sigma M_{12} M_{23} + 2\Sigma M_{11} M_{23}) \\ & + E_P(z_1^3 z_2)\, 4\Sigma M_{11} M_{12} \\ & + E_P(z_1 z_2 z_3 z_4) \Sigma M_{12} M_{34}. \end{aligned} \tag{22}$$

Using David & Kendall's tables (1949)

$$\left.\begin{aligned} N E_P(z_1^4) &= S_4, & N^{(2)} E_P(z_1^2 z_2^2) &= S_2^2 - S_4, \\ N^{(3)} E_P(z_1 z_2^2 z_3) &= 2S_4 - S_2^2, & N^{(2)} E_P(z_1^3 z_2) &= -S_4, \\ N^{(4)} E_P(z_1 z_2 z_3 z_4) &= 3S_2^2 - 6S_4, & & \end{aligned}\right\} \tag{23}$$

where
$$N^{(r)} = \prod_{i=0}^{r-1} \{N - i\}.$$

Also, using the relations $\mathbf{M1} = \mathbf{0}$, $\mathbf{M}^2 = \mathbf{M}$, $\operatorname{tr} \mathbf{M} = p$, we find rather remarkably that each of the sums involving the elements of M can be expressed in terms of $m = \sum_{u=1}^{N} M_{uu}^2$. In fact

$$\left.\begin{aligned} \Sigma M_{12} &= p - m, & \Sigma M_{11} M_{32} &= p^2 - m, & \Sigma M_{12} M_{23} &= 2m - p, \\ \Sigma M_{11} M_{23} &= 2m - p^2, & \Sigma M_{11} M_{12} &= -m, & \Sigma M_{12} M_{34} &= -6m + 2p + p^2. \end{aligned}\right\} \tag{24}$$

On substituting (23) and (24) in (22) and writing S_4 and S_2 in terms of Fisher's k statistics

$$(N-1) k_2 = S_2, \quad (N-1)^{(3)} k_4 = N(N-1) S_4 - 3(N-1) S_2^2,$$

we have
$$\begin{aligned} V_P(W) &= E_P(S_0^2)/S_2^2 - \{E_P(W)\}^2 \\ &= \frac{2p(N-p-1)}{(N+1)(N-1)^2} + \frac{k_4/k_2^2}{(N-1)^2} \left\{ m - \frac{p^2}{N} - \frac{2p(N-p-1)}{N(N+1)} \right\}. \end{aligned}$$

We now put $C_y = k_4/k_2^2$ and for reasons which will be apparent later we write

$$C_X = \frac{N(N-1)(N+1)}{p(N-p-1)(N-3)} \left\{ m - \frac{p^2}{N} - \frac{2p(N-p-1)}{N(N+1)} \right\} \tag{25}$$

so that we have finally

$$V_P(W) = V_N(W) \left\{ 1 + \frac{(N-3) C_y C_X}{2N(N-1)} \right\}, \tag{26}$$

where
$$V_N(W) = \frac{2p(N-p-1)}{(N+1)(N-1)^2}$$

is the variance of W on the usual normal theory.

5. C_X AS A MEASURE OF 'NON-NORMALITY' OF THE x'S

We now show that the function C_X of the x's is analogous with the function C_y of the y's and is a measure of 'non-normality' of the x's. We first notice that M is invariant under any non-singular transformation $\mathbf{W} = \mathbf{XT}$. For

$$\mathbf{M} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' = \mathbf{XTT}^{-1}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{T}'^{-1}\mathbf{T}'\mathbf{X}' = \mathbf{W}(\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'.$$

If we now regard $m = \sum_{u=1}^{N} M_{uu}$ as a function of $m\{\mathbf{X}\}$ of the elements x_{iu} of $\mathbf{X}$, we see that its value is unchanged when every element x_{iu} is replaced by the corresponding w_{iu}.

Now let us choose $\mathbf{T}$ so that $\mathbf{W}'\mathbf{W}$ is diagonal with the ith element equal to $\sum_{u=1}^{N} w_{iu}^2$. Then

$$m = \sum_{u=1}^{N} M_{uu}^2 = \sum_{u=1}^{N} \left\{ \sum_{i=1}^{p} \left(w_{iu}^2 \Big/ \sum_{u=1}^{N} w_{iu}^2 \right) \right\}^2,$$

that is

$$m = \sum_{i=1}^{p} \Sigma_4^i/(\Sigma_2^i)^2 + \sum_{i=1}^{p} \sum_{j \neq i=1}^{p} \Sigma_{22}^{ij}/(\Sigma_2^i \Sigma_2^j) \tag{27}$$

where

$$\Sigma_2^i = \sum_{u=1}^{N} w_{iu}^2, \quad \Sigma_4^i = \sum_{u=1}^{N} w_{iu}^4, \quad \Sigma_{22}^{ij} = \sum_{u=1}^{N} w_{iu}^2 w_{ju}^2.$$

Now, because of the invariant property of m, equation (27) is still true if the corresponding power sums for the x's (which we write as $S_2^i, S_4^i, S_{22}^{ij}$) replace the power sums $\Sigma_2^i, \Sigma_4^i, \Sigma_{22}^{ij}$ for the w's.

Defining k statistics for the x's in the usual way by

$$(N-1)\,k_2^i = S_2^i, \quad (N-1)^{(3)}\,k_4^i = N(N-1)\,S_4^i - 3(N-1)\,(S_2^i)^2,$$
$$(N-1)^{(3)}\,k_{22}^{ij} = N(N+1)\,S_{22}^{ij} - (N-1)\,S_2^i\,S_2^j$$

we find after a little reduction that

$$m = \frac{(N-2)(N-3)}{(N+1)N(N-1)}\left[\sum_{i=1}^{p}\left\{\frac{k_4^i}{(k_2^i)^2}\right\} + \sum_{i=1}^{p}\sum_{j \neq i=1}^{p}\left\{\frac{k_{22}^{ij}}{k_2^i\,k_2^j}\right\}\right] + \frac{(N-1)\,p(p+2)}{N(N+1)}. \tag{28}$$

Substituting this expression in (25) we have finally

$$C_X = \frac{N-2}{p(N-p-1)}\left\{\sum_{j=1}^{p}\frac{k_4^i}{(k_2^i)^2} + \sum_{i=1}^{p}\sum_{j \neq i=1}^{p}\frac{k_{22}^{ij}}{k_2^i\,k_2^j}\right\}. \tag{29}$$

We see that just as $C_y = k_4/k_2^2$ is a measure of non-normality (specifically a measure of kurtosis) for the y's so C_X can be regarded as its multi-variable analogue for the x's. In particular:

(*a*) If $p = 1$ then C_X is the same function of the elements of the single vector x_1 as C_y is of the elements of y.

(*b*) The expected value of C_X is zero for samples from a normal population. This is so because the ratio of k statistics which C_X contains are homogeneous functions of degree zero in the x's. All such functions of normal variates are distributed independently of scaling statistics k_2^i $(i = 1, 2, \ldots, p)$ and consequently the expected values of these ratios are the ratios of the expectations. But for a normal distribution,

$$E(k_4^i) = E\{k_{22}^{ij}\} = 0 \quad (i \neq j = 1, 2, \ldots, p).$$

(*c*) C_X is invariant under linear transformations of the x's. Any sensible measure of multi-variate non-normality would clearly need to possess this property.

Upper and lower bounds for C_X

We first show that
$$\frac{p^2}{N} \leqslant m \leqslant \frac{N-1}{N} p.$$

Because of the invariant property of m we may suppose, without loss of generality, that $\mathbf{X}'\mathbf{X} = \mathbf{I}$. Then
$$\sum_{u=1}^{N} M_{uu} = \sum_{u=1}^{N} \sum_{i=1}^{p} x_{iu}^2 = p.$$

But
$$\sum_{u=1}^{N} M_{uu}^2 - \left\{\sum_{u=1}^{N} M_{uu}\right\}^2 \Big/ N \geqslant 0.$$

Hence
$$m = \sum_{u=1}^{N} M_{uu}^2 \geqslant \frac{p^2}{N}.$$

Now suppose $N-p$ further columns are added to $\mathbf{X}$, the first of which is $N^{-\frac{1}{2}}\mathbf{1}$, to form an orthogonal matrix $\mathbf{H}$. Then since the sums of squares of the elements of each row of $\mathbf{H}$ is unity
$$M_{uu} = \sum_{i=1}^{p} x_{iu}^2 \leqslant 1 - 1/N,$$

whence
$$M_{uu}^2 \leqslant M_{uu}(1-1/N).$$

But
$$M_{uu} = p,$$

whence
$$m = \Sigma M_{uu}^2 \leqslant \{(N-1)/N\} \Sigma M_{uu} = p(N-1)/N.$$

Substituting these bounds in (25) we have finally
$$-2 \leqslant \frac{N-3}{N-1} C_X \leqslant N-1. \tag{30}$$

We shall show later that the lower bound is actually obtainable. The upper bound is approached but cannot be attained in finite samples as is clear from the manner of its derivation.

6. Approximate distribution of R

Case 1. *Permutation distribution of R*

Following Pitman (1937) and Welch (1937) we now approximate the permutation distribution of W with a beta distribution with degrees of freedom adjusted so as to have the correct mean and variance. If $\nu_1 = p$ and $\nu_2 = N-p-1$ are the degrees of freedom of the beta distribution appropriate on normal theory, then it is readily shown that the approximating distribution has degrees of freedom $\delta\nu_1$ and $\delta\nu_2$ where
$$\delta_1^{-1} = 1 + \frac{(N+1)\alpha_1}{N-1-2\alpha_1} \quad \text{with} \quad \alpha_1 = \frac{N-3}{2N(N-1)} C_X C_y. \tag{31}$$

Equivalently the permutation distribution of R is approximated by an F distribution with degrees of freedom $\nu_1 = \delta_1 p$ and $\nu_2 = \delta_1(N-p-1)$. In those cases where C_X and C_y are not simultaneously close to upper bounds we have to order N^{-1}
$$\delta_1^{-1} = 1 + C_X C_y / 2N. \tag{32}$$

Case 2. *Distribution of R under general non-normality of y.* $\mathbf{X}$ *fixed*

If we now take expectations of the permutation moments over all samples, then C_y in $V(W)$ is replaced by $E(C_y)$ which we denote by Γ_y. Again using the beta-distribution approximation we have that if $p(\boldsymbol{\epsilon})$ is any symmetric function of the elements of $\boldsymbol{\epsilon}$ and in particular if the ϵ's may be regarded as independent random drawings from any probability distribution whatever, R is distributed approximately as F with $\nu_1 = \delta_2 p$ and $\nu_2 = \delta_2(N-p-1)$ degrees of freedom, where

$$\delta_1^{-1} = 1 + \frac{(N+1)\alpha_2}{N-1-2\alpha_2} \quad \text{and} \quad \alpha_2 = \frac{N-3}{2N(N-1)} C_X \Gamma_y \tag{33}$$

or, to order N^{-1},

$$\delta_2^{-1} = 1 + C_X \Gamma_y / 2N. \tag{34}$$

Case 3. *Distribution of R under general non-normality of y and x*

Finally, we may suppose that the regression variables themselves are random variables distributed independently of the y's in some p-variate distribution and that it is the deviations of these variates from their sample means that is recorded in the matrix X.

Taking expectations of the moments over all realizations of X, C_X in (31) must be replaced by $E\{C_X\}$ which we denote by Γ_X.

Once again using the beta approximation, R will be approximately distributed as F with $\nu_1 = \delta_3 p$ and $\nu_2 = \delta_3(N-p-1)$ degrees of freedom, where

$$\delta_3^{-1} = 1 + \frac{(N+1)\alpha_3}{N-1-2\alpha_3} \quad \text{with} \quad \alpha_3 = \frac{N-3}{2N(N-1)} \Gamma_X \Gamma_y, \tag{35}$$

or, to order N^{-1},

$$\delta_3^{-1} = 1 + \Gamma_X \Gamma_y / 2N. \tag{36}$$

7. Accuracy of the approximation

It should be noticed that the above approximations do not depend for their accuracy on the fitting of an *arbitrary* distribution to the first two moments of W. The distribution of W is known to be given *exactly* by the approximation when in case 2 the observations y are normally distributed and when in case 3 *either* the observations y *or* the regression variables x, or both, are normally distributed. We might expect to be able to represent moderate departures from normality by suitable changes in the mean and variance of a system of curves which are of the right *basic* shape. Evidences that such a hope is justified are:

(*a*) For case 1, Pitman (1937) has shown for analysis of variance tests that except in very unlikely samples the third and fourth moments of W agree fairly closely with those of the approximating beta distribution. We show later that the analysis of variance test for equal groups corresponds to the general regression test when C_X attains its *lower* bound.

(*b*) For case 1, Box & Andersen showed by means of sampling experiments that the permutation distribution appropriate to the comparison of two independent variances was well represented by the approximation for distributions as non-normal as the rectangular and the double exponential. As we shall see, the test they considered can be very nearly reproduced in the general regression framework when C_X approaches its *upper* bound.

(*c*) For case 2, Box & Andersen showed the close agreement between the results of Gayen (1950) and results obtained by this approximation in analysis of variance tests. This confirms in particular the appropriateness of the general regression approximation at the lower bound of C_X.

(*d*) For case 2, Box (1953), using an argument not employing permutation theory, obtained an F approximation to the distribution of R for the comparison of two independent sample variances. In this the degrees of freedom were modified by functions which to order N^{-1} are identical with those given by the approximating F distribution derived via permutation theory. This supports the essential validity of the approximation as C_X approaches its *upper* bound.

8. Special case of the general one-way classification

We can readily check our formulae against those of Welch for the one-way classification analysis of variance with not necessarily equal groups. In the general regression model we arrange that the uth element x_{iu} of the ith x vector is $1-n_i/N$ when y_u falls in the ith group and $-n_i/N$ otherwise. As is well known the general regression test then reduces to the usual 'one way classification' analysis of variance test for the comparison of $p+1$ means. Then

$$S_0 = \mathbf{z}'\mathbf{M}\mathbf{z} = \mathbf{y}'\mathbf{M}\mathbf{y} = \sum_{i=1}^{p+1} n_i \bar{y}_i^2 - N\bar{y}^2,$$

where the overall sample mean is $\bar{y}$ and the sample mean for n_i observations in the ith group is $\bar{y}_i$.

Now if y_u is any one of the n_i observations in the ith group then the corresponding diagonal element M_{uu} of M is $n_i^{-1}-N^{-1}$. Hence

$$m = \sum_{u=1}^{N} M_{uu} = \sum_{i=1}^{p+1} n_i(n_i^{-1}-N^{-1})^2$$

$$= \sum_{i=1}^{p+1} n_i^{-1} - \frac{2p+1}{N}, \tag{37}$$

$$\frac{N-3}{N-1}C_X = 2\left[\frac{N(N+1)}{2p(N-p-1)}\left\{\sum_{i=1}^{p+1} n_i^{-1} - \frac{(p+1)^2}{N}\right\} - 1\right]. \tag{38}$$

Substituting this expression in (21) gives the value obtained by Welch for $V_P(W)$ and the corresponding F approximation given by Box & Andersen.

8·1. *Equal groups*

If the number of observations is the same in every group so that

$$n_i = n = N/(p+1) \quad (i=1,2,\ldots,N),$$

we have using equation (31)

$$m = \frac{p^2}{N} \quad \text{and} \quad \frac{N-3}{N-1}C_X = -2. \tag{39}$$

The lower bound of the inequality (30) is thus *actually attained* for the equal groups analysis of variance test.

Substituting the result in (31) we obtain correctly

$$\delta^{-1} = 1 - \frac{1}{N}\left\{\frac{N+1}{N-1+2C_y/N}\right\}C_y$$

or, to order N^{-1},

$$\delta^{-1} = 1 - C_y/N. \tag{40}$$

The corresponding result with $\Gamma_y = E(C_y)$ replacing C_y provides the appropriate correction factor for case 2 exemplifying the approximate effect of parent non-normality.

8·2. *Very unequal groups*

It is not possible to reproduce *exactly* the test for the comparison of two variances from the present general regression set-up. As we have already mentioned in § 3·2 we can, however, reproduce such a test with a slightly modified model in which the overall mean β_0 is supposed known.

The permutation approximation given by Box & Andersen for the set-up of § 3·2 with $n_1+n_2 = N-1$ observations is

$$\delta^{-1} = 1+\frac{1}{2}\left\{\frac{N+1}{N-2-C'_y}\right\}C'_y$$

or, to order N^0,

$$\delta^{-1} = 1+\tfrac{1}{2}C'_y, \tag{41}$$

where C'_y is defined in equation (12). The corresponding result which gives the effect of non-normality on this comparison of variances test is obtained as before by replacing C'_y with $\Gamma'_y = E(C'_y)$. It is also possible by a different modification in which both the group means are eliminated to reproduce the usual test to compare two variances when their group means are estimated from the samples. With the present unmodified set-up we come closest to reproducing the test for the comparison of two variances by selecting the ith vector to have an element $x_{iu} = (N-1)/N$ when $u = i$ and $x_{iu} = -1/N$ otherwise. This then corresponds to the analysis of variance test discussed above with one observation in each of p groups and the remaining $N-p$ in the remaining groups. After a little manipulation we obtain for this case

$$R = \frac{\left\{\sum\limits_{u=1}^{p}(y_u-\bar{y}')^2+\dfrac{pN}{N-p}(\bar{y}'-\bar{y})^2\right\}\Big/p}{\sum\limits_{s=p+1}^{N}(y_s-\bar{y}'')^2/(N-p-1)}, \tag{42}$$

where $\bar{y}'$ is now used for the sample mean of the first p observations and y'' for the sample mean of the last $N-p$ observations.

The usual test criteria to compare the variances of two samples of p and $N-p$ observations would be

$$R' = \frac{\left\{\sum\limits_{u=1}^{p}(y_u-\bar{y}')^2\right\}\Big/(p-1)}{\sum\limits_{s=p+1}^{N}(y_s-\bar{y}'')^2/(N-p-1)}, \tag{43}$$

which differs from R above only in that the latter contains a single extra comparison which contrasts the mean $\bar{y}'$ of the first p observations with the overall mean $\bar{y}$.

For this extreme case of very unequal groups, which nearly reproduces the comparison of variances test, we have

$$m = \frac{p}{N}\left\{N-1-\frac{(N-p-1)}{N-p}\right\} \quad\text{and}\quad \frac{N-3}{N-1}C_X = N-1-\left(\frac{N+1}{N-p}\right). \tag{44}$$

Provided that the ratio of p to N is small, C_X will approach its upper bound of $N-1$ quite closely.

If we write $\dot{N}$ for $N-1-(N+1)/(N-p)$, the modifying factor appropriate to this case is

$$\delta^{-1} = 1+\frac{1}{2}\frac{N-1}{N}\left\{\frac{\dot{N}}{N-1-\dot{N}C_y/N}\right\}C_y, \tag{45}$$

which is very similar to the corresponding expression in equation (41) and once more, to order N^0, $\delta^{-1} = 1+\frac{1}{2}C'_y$.

8·3. *Group sizes which approximately nullify the effect of non-normality*

We have seen that particular choices of group sizes can be made so that C_X approaches a lower value of -2 and an upper value of N, giving rise to proportionate corrections. Provided the number of observations is not very small, the slight corrective increase in the degrees of freedom at the lower extreme is usually of little concern, but the considerable corrective decrease at the upper extreme is much more serious. Consideration of equations (31), (33) and (35) shows that in general if in cases 1 and 2 we choose X so that $C_X = 0$, or, in case 3, sample from a population in which $\Gamma_X = 0$, the corrective factor supplied by our approximations are all zero irrespective of the y's or their distribution. One way of exploiting this fact in experimental design theory has been noted by Box (1952).

Returning, now, to the one-way classification analysis of variance we see from equation (38) that, if we choose the group sizes n_i so that

$$N\sum_{i=1}^{p+1} n_i^{-1}-(p+1)^2 = \frac{2p(N-p-1)}{(N+1)}, \tag{46}$$

then $C_X = 0$ and the correction term is zero. Rather surprisingly therefore the effect of non-normality as measured by our approximation is *not* smallest for equal group sizes.

As an example, suppose there are just two groups of size $n_1 = rN$ and $n_2 = (1-r)N$. Substituting in (46) we obtain

$$r = \frac{1}{2}\left\{1\pm\sqrt{\frac{N-2}{3N}}\right\} \tag{47}$$

for the optimum ratio of subgroup sizes. If $N = 12$, for example, then approximately the optimal group sizes are 9 and 3. For large N the optimal sizes are approximately in the ratio 4:1.

At first sight the idea that unequal group sizes could produce *less* sensitivity to non-normality seemed sufficiently surprising as to be almost unacceptable. In fact, as we show in the next section, it is not difficult to explain it. The result itself may be confirmed independently by study of the results of Gayen (1950). This author obtained the exact distribution of R for the one-way classification analysis of variance test for a parent population expressed by an Edgeworth series. His method of derivation is, of course, quite different from that used here. His corrective factor for kurtosis is proportional to a quantity which he denotes by

$$(\nu_{22}) = \frac{2\nu_1\nu_2-(k^2-k'^2)(\nu_1+\nu_2+2)}{8(\nu_1+\nu_2+1)(\nu_1+\nu_2+2)}.$$

In our notation $\nu_1 = p$, $\nu_2 = N-p-1$, $k = p+1$, $k' = (p+1)^2-N\sum_{i=1}^{p+1} n_i^{-1}$. Making the necessary substitution we see that when (46) is satisfied Gayen's quantity (ν_{22}) is zero, providing verification of our result.

We do not of course suggest that one would deliberately seek unequal sample sizes to lessen the effect of non-normality in a test to compare means. The reduction in precision with which comparison among the means could be made and the increase in sensitivity to variance inequalities which would result would certainly not be worth the small increase in robustness to non-normality.

9. Robustness determined by 'normality' in the x's

We shall conduct the following discussion in terms of the situation (cases 2 and 3) where the error vector $\boldsymbol{\epsilon}$ is drawn from any symmetric distribution. We have seen that our modifying factor involves a measure of non-normality in the y's multiplying an analogous measure of 'non-normality' in the x's. That such factors would be involved symmetrically is to be expected from geometric considerations. The criterion R is a function of the angle between the observation vector and the plane of the x vectors and, as was first shown by Fisher, will follow its normal theory distribution if the y's *or* the x's *or* both are normally distributed. The multiplicative characteristic shows how non-normality in the y's is magnified or diminished depending on whether the x's are 'normal-looking' or not.

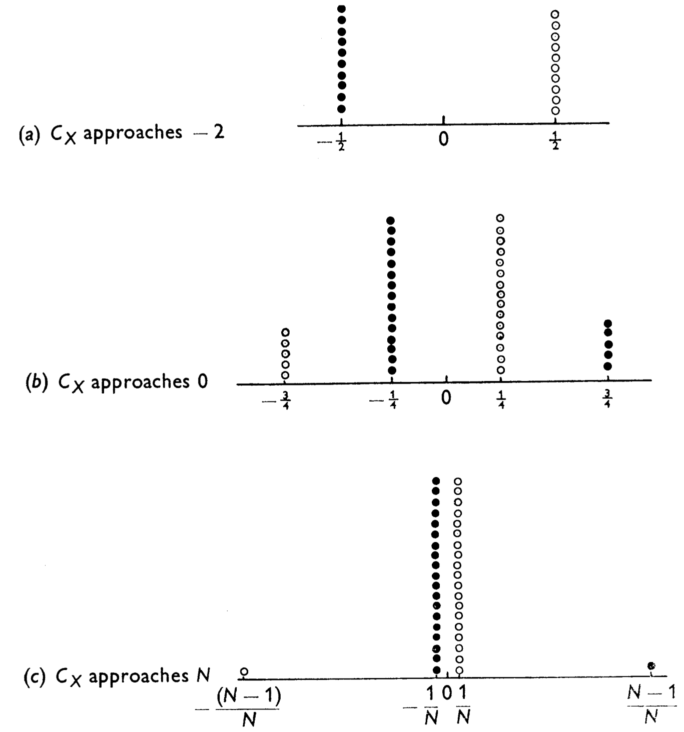

Fig. 1. Some distributions of the elements of x for $N = 20$.

The effects can be understood intuitively by means of particular examples. Consider the analysis of variance test for two equal groups. In this case the single vector x_1 has for its elements $\frac{1}{2}N$ values equal to $-\frac{1}{2}$ and $\frac{1}{2}N$ values equal to $+\frac{1}{2}$. The distribution of individual x's is like that shown in Fig. 1(a) (for $N = 20$) and represents the most 'platykurtic' distribution possible, for which the value C_X tends to its lower limit of -2. The set-up involving very unequal groups, discussed in § 8·2, comes closest to representing the comparison of two independent variances. The x vectors then each have one value equal to $(N-1)/N$ and the remaining $N-1$ values equal to $-1/N$. The distribution for $N = 20$ is that of the full circles shown in Fig. 1 (c). This distribution has the same measure of kurtosis as when its mirror image, shown by open circles, is added. This represents the most leptokurtic distribution possible and the value of C_X tends to its upper limit, N. The full circles in Fig. 1 (*b*) show a distribution in which $\frac{1}{4}$ of the observations are set a distance $\frac{3}{4}$ from the origin and the remaining $\frac{3}{4}$ of the observations are set a distance $-\frac{1}{4}$. This distribution is close to that expected to minimize the effect of non-normality in the y's. Again the mirror image distribution of open circles is added. We see that in this example the distribution of the x's is doing its best to approximate a normal curve which accounts for the resulting insensitivity to non-normality in the y's.

10. Conclusion

Our results may be summarized in the simple statement that sensitivity to non-normality in the y's is determined by the extent of the 'non-normality of the x's'. The small effect in one direction experienced with the equal groups analysis of variance test and the much larger effect in the opposite direction found in the test for the comparison of independent variances provide extremes of sensitivity within which the sensitivity of the general test will be found. In the analysis of data arising from experimental designs such as factorials, a small and usually unimportant degree of sensitivity characteristic of the equal groups analysis of variance may be expected. With data in which the x's themselves are drawn from near normal distributions an even smaller degree of sensitivity is to be expected. Tests which employ x vectors in which one or two elements are very different in magnitude from the remainder may be expected to show much greater sensitivity to non-normality in the y's. In addition to the tests for comparing variances, certain tests concerned with outliers and with missing observations will show this greater sensitivity. One would expect that the usual normal theory *multi-variate* overall regression tests will have analogous corrective factors.

References

Box, G. E. P. (1952). Multifactor designs of the first order. *Biometrika*, **39**, 49–57.

Box, G. E. P. (1953). Non-normality and tests on variances. *Biometrika*, **40**, 318–35.

Box, G. E. P. & Andersen, S. (1955). Permutation theory in derivation of robust criteria and the study of departures from assumption. *J. R. Statist. Soc.* **17**, 1–26.

David, F. N. & Johnson, N. L. (1951*a*). A method of investigating the effect of non-normality and heterogeneity of variance on tests of the general linear hypothesis. *Ann. Math. Statist.* **22**, 382–92.

David, F. N. & Johnson, N. L. (1951*b*). The effect of non-normality on the power function of the *F*-test in the analysis of variance. *Biometrika*, **38**, 43–57.

David, F. N. & Kendall, M. G. (1949). Tables of symmetric functions—Part 1. *Biometrika*, **36**, 431–49.

Fisher, R. A. (1947). *The Design of Experiments*, 4th ed. Edinburgh and London: Oliver and Boyd.

Gayen, A. K. (1950). The distribution of the variance ratio in random samples of any size drawn from non-normal universes. *Biometrika*, **37**, 236–55.

Geary, R. C. (1947). Testing for normality. *Biometrika*, **34**, 209–42.

Pearson, E. S. (1931). Analysis of variance in cases of non-normal variation. *Biometrika*, **23**, 114–33.

Pitman, E. J. G. (1937). Analysis of variance test for samples from any population. *Biometrika*, **29**, 322–35.

Scheffé, H. (1953). A method for judging all contrasts in the analysis of variance. *Biometrika*, **40**, 87–104.

Welch, B. L. (1937). On the *z*-test in randomised blocks and Latin squares. *Biometrika*, **29**, 21–52.