©Journal of Business & Economic Statistics, Vol. 1, No. 4, October 1983

A Nonstochastic Interpretation of Reported Significance Levels

David Freedman
Department of Statistics, University of California, Berkeley, CA 94720

David Lane
Department of Statistics, University of Minnesota, Minneapolis, MN 55455

Tests of significance are often made in situations where the standard assumptions underlying the probability calculations do not hold. As a result, the reported significance levels become difficult to interpret. This article sketches an alternative interpretation of a reported significance level, valid in considerable generality. This level locates the given data set within the spectrum of other data sets derived from the given one by an appropriate class of transformations. If the null hypothesis being tested holds, the derived data sets should be equivalent to the original one. Thus, a small reported significance level indicates an unusual data set. This development parallels that of randomization tests, but there is a crucial technical difference: our approach involves permuting observed residuals; the classical randomization approach involves permuting unobservable, or perhaps nonexistent, stochastic disturbance terms.

KEY WORDS: Significance levels; P values; Hypothesis testing; Multiple regression.

1. INTRODUCTION

Tests of significance are often made in situations where the standard assumptions underlying the probability calculations do not hold. As a result, the reported significance levels become difficult to interpret; they are often computed by applying some standard formula to the data, without concern for assumptions. This article sketches an alternative interpretation of a reported significance level, valid in considerable generality. In brief, such a level is shown to locate the given data set within the spectrum of other data sets derived from the given one by an appropriate class of transformations.

It is not always possible, or even desirable, to view the derived data sets as alternative possible realizations of some random experiment that generated the observed data. The key property of these derived data sets, loosely put, is as follows: If the null hypothesis being tested holds, the derived data sets should appear equivalent to the original one. Thus, a small reported significance level indicates a data set that is unusual.

Our idea is quite widely applicable. It is presented here for the conventional t tests and F tests of the null hypothesis that certain coefficients in a regression equation should vanish. In this case, the transformations amount to permuting the residuals from a restricted equation, where the coefficients in question are set equal to zero.

Section 2 presents the idea for simple regression of a dependent variable y on an explanatory variable z. The null hypothesis is that the coefficient of z should be 0. In this situation, the ideas of permutation testing are applicable and lead to the same derived data sets as our approach. This is not the case in Section 3, which extends the idea to a multiple regression equation, with the null hypothesis that the coefficients of certain variables should vanish identically. A permutation interpretation of the P value from the conventional F test is developed. This interpretation is believed to be new, as is the theorem that justifies it.

This article should not be construed as an attack on significance testing or as a defense. We are deliberately avoiding all such issues. Investigators often do fit linear models and make F tests without paying much attention to stochastic assumptions on the disturbance terms. Our only concern is to interpret the resulting P values. For more discussion of the issues and our framework, and other examples, see Freedman and Lane (1983).

Our interpretation is related to but distinct from the ideas of randomization tests and permutation tests. Such tests have been extensively considered (see, e.g., Fisher 1935); they use permutation distributions for the calculation of significance probabilities. Indeed, the mathematical theory of permutation distributions arose in this context (see, e.g., Hoeffding 1951 and Wald and Wolfowitz 1944). But such tests differ from the ones discussed in this article.

For both randomization and permutation tests, significance probabilities are interpreted as long-run frequencies, in repetitions of the stochastic mechanism that generates the data. So both theories require stochastic models for the generating mechanism; the reference sets to which the observed data are compared, and on which the permutation distributions are defined, consist of other possible realizations under repetition of the underlying experiment. By contrast, our reference sets are derived by permuting residuals, and our significance level is a descriptive statistic rather than a probability. In short, we permute observed residuals rather than unobservable stochastic disturbance terms. Again, the objective is to provide an interpretation of P values, for use when the investigator does not really have a stochastic model for the data.

For a lucid discussion of randomization tests and a comparison with permutation tests, see Kempthorne (1955, 1966, 1972, 1975). The primary justification for randomization tests is a phycial act of randomization. Our interpretation is useful even when no such act has occurred. Too, randomization theory has a hard time with analysis of covariance, let alone multiple regression. See Cox (1956) for more discussion. Our approach is presented in the context of a multiple regression equation.

For the underlying theory of permutation tests, see Pitman (1937, 1938). This theory rests on the assumption of an exchangeable error distribution. Conditional on its order statistics, the vector of stochastic disturbance terms is uniformly distributed over all its permutations. Pitman's study of the sample mean can be interpreted this way, and the approach extends to the regression context (see Box and Watson 1962). Still, the idea is to permute unobservable errors, not observable residuals.

Another set of papers in which a reference set is created from the data is Efron (1979) and its successors. These papers are related to ours in technique, but the focus is different. Efron takes the standard statistical models for granted; his concern is to approximate significance probabilities by resampling the data. A similar approach is offered by Geisser (1975).

We should also point out that the idea of interpreting significance probabilities in a nonstochastic way has been around for a long time in the social science literature. See, for example, Morrison and Henkel (1970).

2. A SPECIAL CASE

We begin by presenting a theorem due to Wald and Wolfowitz (1944); for a more recent treatment, see Hajek (1961) or Ho and Chen (1978). The theorem justifies the permutation interpretation of P values from Fisher's t test. To state the theorem, suppose y and z are n vectors, $(y_1, \ldots, y_n)$ and $(z_1, \ldots, z_n)$ respectively. Let $\bar{y}$ be the mean of y, and let $s(y)$ be its standard deviation:

$$\bar{y} = \frac{1}{n} \sum_{m=1}^{n} y_m \quad \text{and} \quad s(y)^2 = \frac{1}{n} \sum_{m=1}^{n} (y_m - \bar{y})^2.$$

Likewise for z. Let $r(y, z)$ be the correlation between y and z:

$$s(y)s(z)r(y, z) = \frac{1}{n} \sum_{m=1}^{n} (y_m - \bar{y})(z_m - \bar{z}).$$

Let π be a permutation of $I_n = (1, 2, \ldots, n)$, that is, a 1–1 mapping of I_n onto itself. Thus, π moves $i \in I_n$ to $\pi i \in I_n$. Let y_π be the permuted vector $(y_{\pi 1}, y_{\pi 2}, \ldots, y_{\pi n})$. Assign equal weight $1/n!$ to each of the $n!$ permutations π.

Theorem 1. Suppose K is a finite, positive constant and for all $m = 1, \ldots, n$,

$$|y_m - \bar{y}| < Ks(y) \quad \text{and} \quad |z_m - \bar{z}| < Ks(z).$$

Let $R(\pi) = \sqrt{n}\, r(y_\pi, z)$. As $n \to \infty$, the permutation distribution of R converges to normal, with mean 0 and variance 1. (The convergence is uniform in y and z, with for example the Levy metric for weak convergence.)

We now use this theorem to make the permutation interpretation of the P value from Fisher's t test. We have a data set with two variables, (y_i, z_i) for subjects $i = 1, \ldots, n$. Consider the regression of y on z,

$$y_i = a + bz_i + u_i, \tag{1}$$

where a and b minimize the error sum of squares

$$\sum_{1}^{n} (y_i - a - bz_i)^2. \tag{2}$$

The null hypothesis to be tested is that the coefficient of z in (1) should be 0, that is, the difference between b and 0 is accidental. The usual procedure is to compute the statistics $r = r(y, z)$ and

$$t = \sqrt{n-2}\, r/\sqrt{1-r^2}.$$

(This is the usual t statistic for the coefficient b.) The observed significance level P is then computed as the area beyond $|t|$ under Student's curve with $n - 2$ degrees of freedom.

If very specific assumptions are made about the stochastic mechanism that generated the data (y, z), this P value can be interpreted as a probability. However, there is a permutation interpretation that makes no assumptions about the mechanism for generating the data. We now turn to this interpretation.

Suppose the informal null hypothesis that y is unrelated to z. Then the connection between y_1 and z_1 is accidental. Likewise for the connection between y_2 and z_2. So the permuted data set (y_2, z_1), (y_1, z_2), (y_3, z_3), ... should be indistinguishable from the original data set. Indeed, the original data set should be a typical element of the $n!$ permuted data sets (y_π, z). The P value from Fisher's t test locates the original data set (y, z) among these $n!$ permuted data sets. Indeed, suppose n is large, and there are no outliers in the data. Theorem 1 demonstrates that to a good approximation, P is the fraction of permutations π for which $|r(y_\pi, z)| > |r(y, x)|$. This is the permutation interpretation of P.

The informal null hypothesis was not a statement about the parameters of a conventional linear model. Indeed, there was no chance model for the data. Nothing was said about the stochastic behavior of the u_i in (1). In fact, these u_i were treated as data, not as random variables.

In this example, it was satisfactory to permute y; the residuals were not involved. In the next section, however, it will be necessary to permute residuals; Theorem 1 will be restated in terms of permuting residuals.

The boundedness condition in the theorem can be relaxed to a condition of the Lindeberg type. The uniformity will be helpful in the next section. To be more explicit, let $d(U, V)$ be the Levy distance between the laws of U and V, namely, the least $\delta > 0$ such that

$$P(U \le x) \le P(V \le x + \delta) + \delta$$

and

$$P(V \le x) \le P(U \le x + \delta) + \delta.$$

For any $\epsilon > 0$, there is an $N(\epsilon, K)$ such that for any $n > N$ and any y, z satisfying the boundedness condition of the theorem, $d(R, S) < \epsilon$, where S is normal with mean 0 and variance 1.

3. THE GENERAL CASE

We now turn to multiple regression. The data consist of an $n \times 1$ column vector $y = (y_1, \ldots, y_n)$, an $n \times p$ matrix X, and an $n \times q$ matrix Z. Each column of X is considered as n observations on a variable, and likewise for Z. The jth column of X is denoted by X^j and likewise for Z. So we have one dependent variable y, and $p + q$ independent variables $X^1 \ldots X^p$ and $Z^1 \ldots Z^q$. By assumption, X^1 is identically 1, so the equation has an intercept. This will simplify the mathematics appreciably. Informally, the explanatory variables $X^1 \ldots X^p$ are "known" to be in the regression; it is desired to test whether $Z^1 \ldots Z^q$ belong in the equation too. Ordinarily, X would be used to denote the design matrix whose first p columns are $X^1 \ldots X^p$ and whose last q columns are $Z^1 \ldots Z^q$. The present notation is more convenient for our purposes, and it is also convenient to write X_i for the ith row of X, and likewise for Z. Write X_{ij} for the ijth element of X; this number may also be viewed as the ith element of X^j or the jth element of X_i. Likewise for Z.

We fit the regression

$$y_i = X_i \cdot a + Z_i \cdot b + u_i, \tag{3}$$

where a is a $p \times 1$ column vector and b is a $q \times 1$ column vector, chosen to minimize the sum of squares

$$\sum_{i=1}^{n} (y_i - X_i \cdot a - Z_i \cdot b)^2. \tag{4}$$

The null hypothesis to be tested is that the coefficient b of Z in (3) should be 0, that is, the difference between b and 0 is accidental. Let $F(y, X, Z)$ be the conventional F statistic used to test this hypothesis; F will be defined algebraically in Equation (15). The degrees of freedom for F are q in the numerator and $n - p - q$ in the denominator. Let P be the observed significance level of the F test, that is, the probability that an F variable with the specified degrees of freedom exceeds $F(y, X, Z)$.

We present a theorem concerning the permutation distribution of $F(y, X, Z)$ and then show how to interpret P. To state the theorem, consider the regression of y on X alone:

$$y_i = X_i \cdot a' + v_i, \tag{5}$$

where the $p \times 1$ column vector a' is chosen to minimize

$$\sum_{i=1}^{n} (y_i - X_i \cdot a')^2. \tag{6}$$

Recall that π is a permutation of $1, \ldots, n$. Let

$$y_i^\pi = X_i \cdot a' + v_{\pi i}. \tag{7}$$

Thus, y_i^π is obtained from the regression (5) of y on X, by permuting the residuals v_i.

We orthogonalize the columns $Z^1, \ldots, Z^q$ of the data matrix, as follows:

$\tilde{Z}^1$ is the part of Z^1 orthogonal to $X^1, \ldots, X^p$,

$\tilde{Z}^2$ is the part of Z^2 orthogonal to $X^1, \ldots, X^p, Z^1$,

$\vdots$

$\tilde{Z}^q$ is the part of Z^q orthogonal to $X^1, \ldots, X^p, Z^1, \ldots, Z^{q-1}$.

Note that $\tilde{Z}^q$ has mean 0, being orthogonal to X^1.

The following theorem is believed to be new; its proof is deferred to Section 5. The notation s for standard deviation was introduced in Section 2.

Theorem 2. Suppose X^1 is identically 1. Let K be a finite, positive constant. Suppose that for all $m = 1, \ldots, n$ and $j = 1, \ldots, q$,

$$|\tilde{Z}_{mj}| < Ks(\tilde{Z}^j).$$

Suppose too that for all $m = 1, \ldots, n$,

$$|v_m| < Ks(v),$$

where v was defined by (5). Let $Q(\pi) = qF(y^\pi, X, Z)$. The conclusion is that as $n \to \infty$, the permutation distribution of Q converges to chi squared with q degrees of freedom. (The convergence is uniform in y, X, and Z, with for example the Levy metric for weak convergence.)

The boundedness condition in the theorem is assummed to simplify the proof; it can be relaxed to a uniform integrability condition of the Lindeberg type.

We now return to the F test of the null hypothesis that the coefficient of Z in the regression (3) should be 0. The observed significance level of this F test was called P. Theorem 2 gives a permutation interpretation of P. To motivate the interpretation, suppose the residuals u_i in the regression (3) are permutable against X and Z. In other terms, (X, Z, u_π) should look about the same as (X, Z, u), for most π; if not, the regression may not be appropriate. This can be investigated by examination of the usual residual plots.

On the null hypothesis that b ought to be 0, the residuals v_i in the restricted regression (5) should be about the same as the u_i, and hence are permutable against X and Z. That is, if the null hypothesis is right, and the regression is appropriate, the v_is can be permuted against X and Z without losing anything essential in the data. In particular, the original data set (y, X, Z) should be a typical element of the permuted data sets (y^π, X, Z), where y^π is defined by (7). Now P does scale (y, X, Z) against (y^π, X, Z). Indeed, suppose n is large, there are no outliers in the data set, and collinearity is not excessive. (These conditions will be discussed shortly.) Then Theorem 2 shows that to a good approximation, P is just the fraction of permutations π for which $F(y^\pi, X, Z) > F(y, X, Z)$, that is, the F statistic from the permuted data set exceeds the F statistic from the original data set. This is the permutation interpretation of P.

The construction of y^π in this section is more complicated than in the previous section. The reason is as follows: On the null hypothesis, y need not be permutable against X and Z, nor should Z be permutable against y and X. Indeed, substantial correlations may well exist among all the variables. The null hypothesis is only that given X, there is no further relationship between y and Z, that is, y is only related to Z through X. It is this idea that is captured by permuting the residuals from the regression of y on X, but leaving Z fixed. Permuting Z rather than the residuals would deny any relationship between X and Z, and this would seldom be reasonable.

The conditions of Theorem 2 may seem hard to check, for they involve the orthogonalized Z's and the residuals from the regression (5). The following conditions are easier to check and are sufficient; indeed, they imply the stated conditions of the theorem with $K = L/\sqrt{\epsilon}$. Fix ϵ with $0 < \epsilon < 1$ and L with $0 < L < \infty$; in practice, ϵ should be small and L moderate, for example, $\epsilon = 1/10$ and $L = 3$. Suppose that for all n,

$$\left| Z_{mj} - \frac{1}{n}\sum_{i=1}^{n} Z_{ij} \right| < Ls(Z^j)$$

$$\text{for} \quad m = 1, \ldots, n \quad \text{and} \quad j = 1, \ldots, q, \tag{8}$$

$$r_j^2 < 1 - \epsilon \quad \text{for} \quad j = 1, \ldots, q, \tag{9}$$

$$\left| y_m - \frac{1}{n}\sum_{i=1}^{n} y_i \right| < Ls(y) \quad \text{for} \quad m = 1, \ldots, n \tag{10}$$

$$r^2 < 1 - \epsilon. \tag{11}$$

In (9), r_j is the multiple correlation between Z^j and $\{X^1, \ldots, X^p, Z^1, \ldots, Z^{j-1}\}$. In (11), r is the multiple correlation between y and $\{X^1, \ldots, X^p\}$. Thus conditions (8) and (10) prevent outliers; conditions (9) and (11) prevent collinearity. Conditions (8) and (10) may be weakened appreciably to conditions of the Lindeberg type: that the data points far from the mean make only a small contribution to the variance, uniformly in n. We omit the details.

We venture to remind the reader that Theorem 2 covers the two-tailed t test as the special case $q = 1$. Indeed, the square of the t statistic equals the F statistic, and the conventional degrees of freedom match up. The P value from a one-tailed t test can be interpreted in a similar way, as the argument for Theorem 2 shows. Again, we omit the details.

The present discussion is asymptotic, so we do not differentiate between t and normal, or χ^2 and F. Theorem 1 is a special case of Theorem 2, when $p = 1$ and $q = 1$. Thus X consists of a column of 1's. In (5), $a' = \bar{y}$ is the mean of y, and $v_i = y_i - \bar{y}$ is the deviation from the mean. In particular, $y_i^\pi = \bar{y} + v_{\pi i} = y_{\pi i}$, so in this case (and this one only) permuting the residuals from the restricted regression is tantamount to permuting y.

4. AN EXAMPLE

By way of example, take the stock market weekly closing quotes from 1981, as reported in the Business and Finance section of the *New York Times.* Index the weeks by $i = 1, \ldots, 52$. Consider explaining the price of Control Data Corporation common stock (CDC_i) by a regression on the Dow Jones Industrial Average (DJ_i) and the price of International Business Machines common stock (IBM_i). To avoid serial correlation in the errors, an equation in first differences is presented:

$$\begin{aligned}\log(\mathrm{CDC}_i/\mathrm{CDC}_{i-1}) &\\ &= a_0 + a_1 \log(\mathrm{DJ}_i/\mathrm{DJ}_{i-1}) \\ &\quad + b \log(\mathrm{IBM}_i/\mathrm{IBM}_{i-1}) + u_i.\end{aligned} \tag{12}$$

This equation was run, with the results shown in Table 1. The usual residual plots showed no trend or pattern. In the notation of the previous section, $n = 52$, $p = 2$, $q = 1$. The dependent variable is $y_i = \log(\mathrm{CDC}_i/\mathrm{CDC}_{i-1})$. The independent variables are $X_{i1} \equiv 1$, $X_{i2} = \log(\mathrm{DJ}_i/\mathrm{DJ}_{i-1})$, $Z_{i1} = \log(\mathrm{IBM}_i/\mathrm{IBM}_{i-1})$.

Does IBM belong in the equation? A two-sided t test

Table 1. Fitting Equation (12) to 1981 Data

Parameter	Estimate	Standard Error	t Statistic	Two-Sided P Value
a_0	.0024	.0045	.52	.60
a_1	1.08	.34	3.15	.0028
b	.11	.29	.45	.65
	$R^2 = .33$	$F = 12$	$P = 6 \times 10^{-5}$	

Table 2. Fitting the Restricted Equation (13) to 1981 Data

Parameter	*Estimate*	*Standard Error*	*t Statistic*	*Two-Sided P Value*
a'_0	.0022	.0045	.49	.63
a'_0	1.18	.24	4.93	.0001
	$R^2 = .33$	$F = 24$	$P = 10^{-5}$	

on b gives a P value of .65. To obtain the permutation interpretation, consider the restricted equation

$$\log(\mathrm{CDC}_i/\mathrm{CDC}_{i-1}) = a'_0 + a'_1\log(\mathrm{DJ}_i/\mathrm{DJ}_{i-1}) + v_i. \quad (13)$$

This equation was run, with the results shown in Table 2. Again, the residuals showed no trend or pattern.

Now subject the residuals v in (13) to a random permutation π, obtaining a permuted data set, in which the explanatory variables DJ and IBM are as before, but the dependent variable is

$$y_i^\pi = a'_0 + a'_1\log(\mathrm{DJ}_i/\mathrm{DJ}_{i-1}) + v_{\pi i}. \quad (14)$$

For each π, run a regression of y_i^π on $\log(\mathrm{DJ}_i/\mathrm{DJ}_{i-1})$ and $\log(\mathrm{IBM}_i/\mathrm{IBM}_{i-1})$. Let t_π be the t statistic for the coefficient of the latter variable. The P value of .65 is interpreted to mean that 65% of all permutations π will have $|t_\pi| > .45$, the value of the t statistic for b reported in Table 1.

In a simulation, 100 pseudorandom π's were selected, and 67 had $|t_\pi| > .45$. This is quite respectable. As further confirmation of the theory, a plot of relative frequencies against normal probabilities is shown in Figure 1. With 49 degrees of freedom for error, there seemed to be no point in distinguishing the t distribution from the normal. The plot should be close to a straight line, and it is. A two-sided Kolmogorov-Smirnov test gives $P \approx 44\%$; see Smirnov (1948). Informally, a comparison plot of relative frequencies against normal probabilities is shown in Figure 2 for 100 independent pseudorandom standard normal variates.

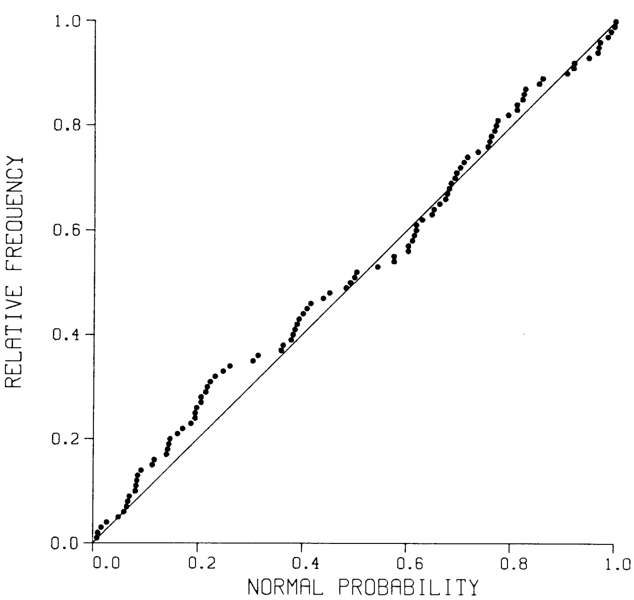

Figure 1. PROBABILITY PLOT. For each real x, the horizontal axis shows $\Phi(x)$, where Φ is the standard normal distribution function. The vertical axis shows the relative frequency with which $t_\pi \leq x$, among 100 pseudorandom π's.

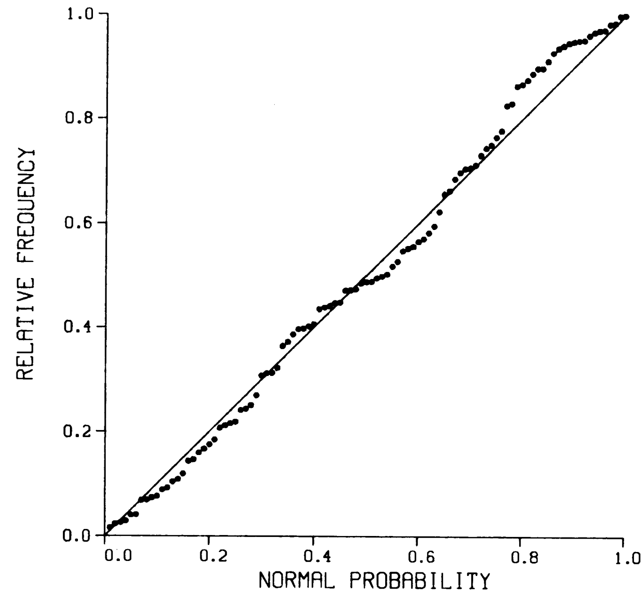

Figure 2. COMPARISON PLOT. For each real x, the horizontal axis shows $\Phi(x)$, where Φ is the standard normal distribution function. The vertical axis shows the relative frequency with which $\zeta_i \leq x$, where $\zeta_1, \ldots, \zeta_{100}$ are independent standard normal variates.

All calculations were done in SAS on an IBM 4341 at the University of California, Berkeley. The regressions were run using GLM, and the pseudorandom permutations were generated by PLAN.

5. THE PROOF OF THEOREM 2

We begin with a fact whose easy proof is omitted. Recall from Section 2 that s is standard deviation.

Lemma 1. Let $x = (x_1, \ldots, x_n)$ and $v = (v_1, \ldots, v_n)$ be n vectors. Suppose $\bar{v} = 0$. Let π be a random permutation of $1, \ldots, n$ and $S_\pi = (1/n)\sum_{m=1}^n x_m v_{\pi m}$. Then

$$\frac{1}{n!}\sum_\pi S_\pi = 0 \quad \text{and} \quad \frac{1}{n!}\sum_\pi S_\pi^2 = \frac{1}{n-1}s(x)^2 s(v)^2.$$

It is enough to prove Theorem 2 under the side conditions that $\bar{y} = 0$ and $s(y) = 1$, while X has full rank p, and $s(\tilde{Z}^j) = 1$ for $j = 1, \ldots, q$. Note that $\tilde{Z}^j$ has mean 0 by construction, and $s(v) > 0$ by assumption: the residuals v were defined in (5).

Abbreviate $r_j(\pi)$ for the correlation coefficient between

y^π and $\tilde{Z}^j$, and $r(y^\pi, X)$ for the multiple correlation coefficient between y^π and the columns of X. Now

$$qF(y^\pi, X, Z) = \frac{\sum_{j=1}^q (n-p-q) r_j(\pi)^2}{1 - r(y^\pi, X)^2 - \sum_{j=1}^q r_j(\pi)^2}. \quad (15)$$

There are three ideas in the proof. First, for most π,

$$1 - r(y^\pi, X)^2 \doteq s(v)^2 \quad (16)$$

by Lemma 1. Second, for most π,

$$\sum_{j=1}^q r_j(\pi)^2 \ll s(v)^2, \quad (17)$$

again by Lemma 1. Third, $r_1(\cdot), \dots, r_q(\cdot)$ are close to being independent and normally distributed, with common mean 0 and common asymptotic variance $s(v)^2/n$, by Theorem 1. So $\sum_{j=1}^q r_j(\cdot)^2$ is about $\chi_q^2 s(v)^2/n$. Thus, (15) has the requisite limiting distribution.

To flesh this out, note that $\bar{v}_\pi = \bar{v} = 0$ and $s(v_\pi) = s(v)$. Abbreviate x_i for the term $X_i \cdot a'$ in (5), so $y_i^\pi = x_i + v_{\pi i}$. As a vector, x is in the column space of X. Let $\delta^\pi = \text{proj}_X v_\pi$, the projection of v_π into the column space of X. A remark on the notation: the vector v_π has ith coordinate $v_{\pi i}$; the vectors y^π and δ^π are not permutations in that way, hence the use of superscripts.

We must now make some preliminary estimates. The vectors y, x, v and v_π, y^π, and δ^π all have mean 0; for such vectors, $s(\ )$ is the Euclidean norm times the scaling factor $1/\sqrt{n}$.

The first estimate is just the triangle inequality:

$$|s(v_\pi - \delta^\pi) - s(v)| \le s(\delta^\pi). \quad (18)$$

Next, we claim

$$|s(y^\pi) - 1| \le 4s(\delta^\pi). \quad (19)$$

Indeed, by orthogonality,

$$s(y^\pi)^2 = s(x + \delta^\pi)^2 + s(v_\pi - \delta^\pi)^2,$$

and

$$1 = s(y)^2 = s(x)^2 + s(v)^2 = s(x)^2 + s(v_\pi)^2.$$

So

$$s(y^\pi)^2 - 1 = (2/n)(x - v_\pi) \cdot \delta^\pi. \quad (20)$$

Now

$$\| x - v_\pi \| \le \| x \| + \| v_\pi \| = \| x \| + \| v \| \le 2 \| y \| \le 2n.$$

The Schwarz inequality completes the proof of (19).

We now show that $s(\delta^\pi)$ is negligible by comparison with $s(v)$, for most π's. First,

$$\frac{1}{n!} \sum_\pi s(\delta^\pi)^2 = \frac{(p-1)s(v)^2}{n-1}. \quad (21)$$

This is easy to check, by choosing an orthonormal basis in the column space of X and appealing to lemma 1 for each basis vector. Fix any large positive number L. By Chebychev's inequality, except for a set of π's of small total probability $(p-1)/L$,

$$s(\delta^\pi) \le \sqrt{L/(n-1)}\, s(v) \le \sqrt{L/(n-1)}. \quad (22)$$

For π's satisfying (22), relation (19) shows

$$1 - 4\sqrt{L/(n-1)} \le s(y^\pi) \le 1 + 4\sqrt{L/(n-1)}. \quad (23)$$

We now estimate $1 - r(y^\pi, X)^2$ in the denominator of the F statistic (15), completing the first step (16) in the proof. Clearly,

$$1 - r(y^\pi, X)^2 = s(v_\pi - \delta^\pi)^2 / s(y^\pi)^2.$$

For π's satisfying (22), relation (18) shows

$$s(v)^2[1 - \sqrt{L/(n-1)}]^2 \le s(v_\pi - \delta^\pi)^2 \le s(v)^2[1 + \sqrt{L/(n-1)}]^2.$$

This and (23) prove (16).

We now complete the second step (17) in the proof. Indeed, $y^\pi = x + v_\pi$ and $\tilde{Z}^j$ is orthogonal to X, so

$$r_j(\pi) = r(y^\pi, \tilde{Z}^j) = \frac{s(v)}{s(y^\pi)} r(v_\pi, \tilde{Z}_j). \quad (24)$$

But

$$\frac{1}{n!} \sum_\pi r(v_\pi, \tilde{Z}^j)^2 = \frac{1}{n-1}$$

by Lemma 1. Remember that L is a large positive number. By Chebychev's inequality, except for a set of π's of total probability q/L, we have $r(v_\pi, \tilde{Z}^j)^2 < L/(n-1)$ for all $j = 1, \dots, q$. Except for another set of π's of probability $(p-1)/L$, we have from (23) that

$$s(y^\pi) \ge 1 - 4\sqrt{L/(n-1)}.$$

So, except for a set of π's of small total probability $(p + q - 1)/L$,

$$\sum_{j=1}^q r_j(\pi)^2 < \frac{1}{1 - 4\sqrt{L/(n-1)}} \frac{1}{n-1} s(v)^2.$$

We turn now to the third step in the proof: computing the asymptotic distribution of the $r_j(\pi)$ in the numerator of the F statistic (15). We use the representation (24). Fix arbitrary constants $c_1, \dots, c_q$ with $\sum_j c_j^2 = 1$. Check that

$$\sqrt{n} \frac{s(y^\pi)}{s(v)} \sum_{j=1}^q c_j r_j(\pi) = \sqrt{n} r \left[v^\pi, \sum_{j=1}^q c_j \tilde{Z}^j \right] \quad (25)$$

because the $\tilde{Z}^j$ are orthonormal, by our side conditions. The permutation distribution of the right side of (25) is nearly $N(0, 1)$, by Theorem 1. And $s(y^\pi)$ is nearly 1, by (23). This completes the proof of Theorem 2.

6. THE CONNECTION WITH FINCH

A referee states, "The basic idea is nearly *all* in Finch." With all due respect to the referee and to Finch, we differ. Finch (1979) presents an interesting theory of how to judge explanations for observed rankings. For the leading special case, let S denote the set of subjects in a study. These are ranked on some outcome variable, with no ties. Let ω stand for the ranking. An "explanation" is a binary relation ρ, that is, a subset of $S \times S$. The idea is that if subject x is related to y by ρ, that is, $(x, y) \in \rho$, then x tends to rank below y according to ω. For simplicity, suppose x is never related to itself by ρ.

The adequacy of ρ as an explanation of ω is judged by the number of mistakes, that is, the number of pairs (x, y) of subjects such that either

- x is related to y by ρ, but x does not rank below y according to ω,

or

- x ranks below y according to ω, but is not related to y by ρ.

Finch introduces the "descriptive power" of ρ at ω as the proportion of competing binary relations ρ' that do worse than ρ in explaining ω; while the "characterizing power" of ρ at ω is the proportion of alternative rankings ω' that are worse described by ρ than is ω. He gives asymptotic formulas for computing these powers, in terms of the normal curve.

Broadly speaking, Finch seems to have the same objective we do, namely, to develop statistical procedures for use in nonstochastic situations. He too offers a combinatorial solution. However, in detail, he formulates the technical question very differently. For instance, he does not consider multiple regression. There is little overlap between his paper and ours. For a critical discussion of Finch's work, see Mallows (1983).

ACKNOWLEDGMENTS

We would like to thank Arnold Zellner for editorial support. We would like to thank two very helpful, if somewhat irascible, referees. Research was partially supported by National Science Foundation Grant MCS-80-02535.

[Received September 1982. Revised May 1983.]

REFERENCES

BOX, G. E. P., and WATSON, G. S (1962), "Robustness to Non-Normality of Regression Tests," *Biometrika*, 49, 93–106.

COX, D. R. (1956), "A Note on Weighted Randomization," *The Annals of Mathematical Statistics*, 27, 1144–1151.

EFRON, B (1979), "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, 7, 1–26.

FINCH, P. D. (1979), "Description and Analogy in the Practice of Statistics," *Biometrika*, 68, 195–208.

FISHER, R. (1935), *Design of Experiments* (1st ed.), Edinburgh: Oliver and Boyd; reprinted in 1971 (9th ed.), New York: Hafner.

FREEDMAN, D., and LANE, D (1983), "Significance Testing in a Nonstochastic Setting," in *Lehmann Festschrift*, ed. P. Bickel, K. Doksum, and J. L. Hodges, Jr., Belmont, Calif.: Wadsworth.

GEISSER S. (1975), "The Predictive Sample Reuse Method With Applications," *Journal of the American Statistical Association*, 70, 320–328.

HAJEK, J. (1961), "Some Extensions of the Wald-Wolfowitz-Noether Theorem," *The Annals of Mathematical Statistics*, 32, 506–523.

HO, S. T., and CHEN, L. (1978), "An L_p Bound for the Remainder in a Combinatorial Central Limit Theorem," *The Annals of Probability*, 6, 231–239.

HOEFFDING, W. (1951), "A Combinatorial Central Limit Theorem," *The Annals of Mathematical Statistics*, 22, 556–558.

KEMPTHORNE, O. (1955), "The Randomization Theory of Experimental Inference," *Journal of the American Statistical Association*, 50, 946–967.

——— (1966), "Some Aspects of Experimental Inference," *Journal of the American Statistical Association*, 61, 11–34.

——— (1972), "Theories of Inference and Data Analysis," *Statistical Papers in Honor of G. W. Snedecor*, Ames, Iowa: Iowa State University Press.

——— (1975), "Inferences From Experiments and Randomization," *A Survey of Statistical Design and Linear Models*, Amsterdam: North-Holland.

KEMPTHORNE, O., and DORFLER, T. E. (1969), "The Behavior of Some Significance Tests Under Experimental Randomization," *Biometrika*, 56, 231–248.

MALLOWS, C. L. (1983), "Data Descriptions," in *Scientific Inference, Data Analysis, and Robustness*, ed. T. Leonard and C. F. Wu.

MORRISON, D. and HENKEL, R. (1970), *The Significance Test Controversy: A Reader*, New York: Academic Press.

PITMAN, E. (1937), "Significance Tests Which May Be Applied to Samples From Any Population I and II," *Journal of the Royal Statistical Society*, Supp. 4, 119–130, 225–232.

——— (1938), "Significance Tests Which May Be Applied to Samples From Any Populations: III. The Analysis of Variance Tests," *Biometrika*, 29, 322–335.

SMIRNOV, H. (1948), "Table for Estimating Goodness of Fit," *The Annals of Mathematical Statistics*, 19, 279–281.

WALD, A., and WOLFOWITZ, J. (1944), "Statistical Tests Based on Permutations of the Observations," *The Annals of Mathematical Statistics*, 15, 358–372.