*The Economic and Social Review, Vol. 24, No. 3, April, 1993, pp. 275-295*

# Geary on Inference in Multiple Regression and on Closeness and the Taxi Problem

JOHN E. SPENCER
*The Queen's University, Belfast*

ANN LARGEY
*The Queen's University, Belfast*

*Abstract*: This paper deals with some minor aspects of Roy Geary's work. Two areas are selected for discussion — (a) his work with Leser on "paradoxical" situations in multiple regression and (b) his work on estimation of the unknown upper bound, N, of a uniform distribution, based on a sample of n values from that distribution. This work is explained, expanded and evaluated. The concept of "paradoxes" in multiple regression is slightly extended and applied to the case of estimating means in a multinormal situation with a known covariance matrix. Geary's estimator of N is compared with several other estimators, on the basis, *inter alia*, of mean squared errors, in both the cases of a continuous distribution and a discrete distribution sampling without replacement. In the latter case, a "large N minimum mean squared error" estimator is derived and assessed.

## I INTRODUCTION

Roy Geary's achievements in the field of mathematical statistics have been described in Spencer (1983 a,b,c) as, in the main, falling under three broad headings — (a) sampling problems involving ratios, (b) testing for normality and issues concerning robustness and (c) the estimation of relationships where the variables are subject to errors of measurement. These are all areas involving difficult theoretical problems and, crucially, all involving issues of great practical importance.

The object of this paper is not to give another account of his contributions in these areas nor to discuss his work as a whole, but to focus on some minor

aspects of his work, especially on analysing and extending two papers which lay outside the three main threads of his theoretical work. In a letter to one of the authors in 1976, Geary spoke of having the impression of his technical papers being "all over the place" and an inability to remember the content of any of them. He did, however, note that they were often motivated by something in Fisher and that he had no difficulty recalling points in many of them. In fact, of course, his papers can be seen with hindsight to form a superbly focused body of work — though with attractive asides, many of which were of considerable importance. Several of these are described in Spencer (1976, and 1983b) and include, as two examples, the propositions that maximum likelihood minimises the generalised variance and that independence of sample mean and variance implies underlying normality under quite general conditions.

In this article we concentrate on two papers, Geary (1944) and Geary and Leser (1968). The former involves an estimation problem known to have intrigued him and is discussed in Section III, while the Geary and Leser paper, discussed in Section II, deals with the possibilities for seemingly paradoxical inference in multiple regression. Geary never believed that individual coefficients in multiple regression had much importance (e.g. Geary, 1963) and in the paper written with Leser, he analysed the relationship between the individual t-ratios and the overall F test in a relationship involving the constant term.

## II "PARADOXICAL" SITUATIONS IN INFERENCE

The paradoxical situations studied in Geary and Leser (1968) are in particular:

*PS.1* All individual coefficients insignificant and the regression as a whole significant.

*PS.2* All individual coefficients significant and the regression as a whole insignificant.

Taking the general model for multiple normal regression, in standard notation, where $b^0$ is the (k+1)x1 vector of coefficients including a constant term, ($b_0$), and X is non stochastic.

$$Y = Xb^0 + \varepsilon \qquad \text{Model 1}$$

the relevant hypotheses to be tested for PS.1 and PS.2 above are:

$$\text{(i) } H_0: \begin{pmatrix} b_1 \\ b_2 \\ \cdot \\ \cdot \\ \cdot \\ b_k \end{pmatrix} = 0 \quad \text{(ii) } H_0: b_i = 0,\ i = 1 \ldots k$$

i.e. $b = 0$.[1]

PS.1 arises when each hypothesis in (ii) is accepted while (i) is rejected, and PS.2 when (i) is accepted while each hypothesis in (ii) is rejected.

Note $b_0 = 0$ does not form part of the joint hypothesis in (i) nor is it tested as a single hypothesis in (ii).

In order to be able to write the tests for these hypotheses as expressions involving correlations between the independent variables two transformations of model 1 are performed.

Assuming that the X's have, where necessary, been multiplied by $-1$ in order to obtain positive values for all estimated coefficients, the first transformation takes deviations, giving

$$y = x\beta + u \qquad \text{Model 2}$$

where $\beta' = [\beta_1 \ldots \beta_k]$ is the coefficient matrix with the constant term omitted and with the property that $|\hat{b}_i| = \hat{\beta}_i$, i=1...k.

Secondly, using any k dimensional invertible matrix W, Model 2 can be expressed as:

$$y = xWW^{-1}\beta + u \qquad \text{Model 3}$$
$$= Z\gamma + u$$

where $Z = xW$, $\gamma = W^{-1}\beta$, $u \sim N(O, \sigma^2 I)$.

1. Savin (1984) deals with the case of induced tests arising from a hypothesis such as (1) $\beta=0$. In this framework he illustrates not only the possibility of conflict between tests of type (1) and the resulting induced tests, but also computes the probability of such occuring.

For $\beta$ a 2x1 vector he tabulates the probability of agreement between a chi square test of $\beta=0$ and the induced Bonferroni tests, showing that for any given $\alpha$, as the correlation between the variables increases the probability of agreement between the two tests decreases, but not greatly, e.g. for $\alpha=0.1$ with r=0, prob (agreement) = 0.965 and r=0.9999, prob (agreement) = 0.934. The discrepancy in probability of agreement for the cases r=0 and r=0.9999 falls as the value of $\alpha$ is reduced.

Choosing W as the diagonal matrix with $w_{ii} = \frac{\sqrt{n}}{s_i}$,

where $s_i = \left( \sum_{i=1}^{n} (X_i - \overline{X})^2 \right)^{\frac{1}{2}}$ we have that $Z'Z = n\,R$ where R is the $k \times k$ correlation matrix for our original (Model 1) independent variables $X_1 \ldots X_k$ (having been transformed where necessary to ensure positive estimated regression coefficients).

$$r_{ij} = \frac{\sum_{l=1}^{n} x_{il} x_{jl}}{s_i s_j} = \frac{\sum_{l=1}^{n} (X_{il} - \overline{X}_i)(X_{jl} - \overline{X}_j)}{s_i s_j}$$

Hypotheses similar to (i) and (ii) are now based on Model 3 i.e.

(i') $$\gamma = 0 \text{ i.e. } W^{-1}\beta = \frac{1}{\sqrt{n}} \begin{bmatrix} s_1\beta_1 \\ . \\ . \\ . \\ s_k\beta_k \end{bmatrix} = 0$$

is equivalent to testing $\beta = 0$ (recalling that X is non-stochastic).

(ii') $$\gamma_i = 0 \text{ i.e. } \frac{1}{\sqrt{n}} s_i\beta_i = 0$$

is equivalent to testing $\beta_i = 0$, $i = 1 \ldots k$.

Model 3 is estimated as $\hat{y} = Z\hat{\gamma}$, $\hat{\gamma} = (Z'Z)^{-1}Z'y$.

The F statistic for testing (i') is derived from the independent statistics

$$\hat{\gamma}'[Z'Z]\hat{\gamma} / \sigma^2 \sim \chi^2_k$$

and

$$\frac{e'e}{\sigma^2} \sim \chi^2_{n-k}, \quad \text{where } e = y - \hat{y}. \tag{2.1}$$

$$F = \frac{n\hat{\gamma}'R\hat{\gamma} / k}{e'e / (n-k)} \sim F_{k,n-k}$$

i.e.
$$F = \frac{n\left(\sum_{i=1}^{k}\hat{\gamma}_i^2 + \sum_{i=1}^{k}\sum_{j\neq i}\hat{\gamma}_i\hat{\gamma}_j r_{ij}\right)/k}{\sum_{i=1}^{n} e_i^2/(n-k)} \sim F_{k,n-k}. \qquad (2.2)$$

Substituting for $\hat{\gamma} = W^{-1}\hat{\beta}$ and $Z = xW$ in Equation (2.1) the F statistic reduces to

$$\frac{\hat{\beta}'(W^{-1})'(W'x'xW)W^{-1}\hat{\beta}/k}{e'e/(n-k)}$$

$$= \frac{\hat{\beta}'(x'x)\hat{\beta}/k}{e'e/(n-k)}$$

— the test statistic for hypothesis (i).

The t statistic for testing (ii') is given by

$$t_i = \frac{\hat{\gamma}_i - 0}{s\sqrt{[Z'Z]^{-1}_{ii}}} \quad \text{where } s^2 = \frac{e'e}{n-k} \quad i = 1..k$$

Since $\hat{\beta}_i$ is ensured positive, $\hat{\gamma}_i = \frac{1}{\sqrt{n}} s_i\hat{\beta}_i$ is also positive and so too then is each $t_i$ statistic.

Using (a) $|\lambda M| = \lambda^n |M|$

(b) $[M]^{-1}_{ij} = \frac{|\overline{M}_{ij}|}{|M|}$, (where $|\overline{M}_{ij}|$ is the cofactor of element ji)

for any invertible matrix M of dimension $n \times n$, then the t statistic above can be rewritten as:

$$t_i = \hat{\gamma}_i / s\sqrt{\frac{n^{k-1}|C_{ii}|}{n^k|R|}} = \frac{\hat{\gamma}_i}{s}\sqrt{\frac{n|R|}{|C_{ii}|}}$$

where $|C_{ii}|$ is the cofactor of the $ii^{th}$ element in the correlation matrix R.

Rearranging we obtain, $\hat{\gamma}_i = s\sqrt{\frac{|C_{ii}|}{n|R|}}t_i$ and $(\hat{\gamma}_i)^2 = \frac{s^2|C_{ii}|}{n|R|}t_i^2$, $i = 1..k$.

We can now obtain the Geary and Leser relationship between the F and t statistics, by substituting these results in Equation (2.2).

$$F = \frac{\sum_{i=1}^{k} |C_{ii}| t_i^2 + \sum_{i=1}^{k} \sum_{j \neq 1} \sqrt{|C_{ii}||C_{jj}|}\, r_{ij} t_i t_j}{|R| k} \qquad (2.3)$$

From (2.3), Geary and Leser determine that PS.1 could arise when all or most of the variables are highly positively correlated, in which case $|R|$ is small and F becomes large relative to the $t_i^2$. Strong positive correlation is not however a necessary condition for PS.1. If all variables are uncorrelated (2.3) reduces to $F = \Sigma t_i^2/k$. Given that the critical F value is lower than the critical t value for more than 3 degrees of freedom, it would still be possible for all t values to be approximately equal and all non-significant while F is significant.

PS.2 may arise when variables are predominantly negatively correlated, but not so strongly correlated that $|R|$ will become small relative to the $|C_{ii}|$ and F become large. In the case of k=2, for example, $X_1$ and $X_2$ could fluctuate in different directions so that their contribution to Y roughly cancels out leaving Y determined as if by a constant and a random term.

Although referred to as "paradoxical" by Geary and Leser, Cramer (1972) stressed that these situations did not embody contradictory results. The existence of situations where tested singly coefficients are insignificantly different from their hypothesised values, while jointly tested they appear to be all significantly different, or vice versa does not imply a logical contradiction. A t test on $\beta_i$, say, is a test between two linear models, one including $X_i$ with k–1 other variables, and the other with $X_i$ excluded.

Now, even if all t values for a regression are insignificant, we cannot conclude that all $\beta$ values are simultaneously insignificantly different from zero. What this result implies is simply that focussing on any particular $\beta_j$, a model which excludes $X_j$ but includes the other k–1 variables would explain the data as well as the model with all k variables included. On the basis of this result we could not omit more than one variable from our regression.

On the other hand the F test considers whether all $\beta_i$ values are simultaneously zero. It compares a model where one or more of the $\beta$'s are non-zero to a model where they are all constrained to zero.

Largey and Spencer (1992) show there is in fact much potential for occurrence of these so-called "paradoxical events". The paper addressed the multiple regression problem in an alternative way focusing in particular on analysing hypotheses of types (i) and (ii) as applied to two regression variables. It concluded that hypothesised values of the $\beta$'s could be found such

that at least one of PS.1 and PS.2 is always possible, and both could possibly occur for any sample size, depending on the correlation between the estimated coefficients of the variables in the regression.

In fact, the existence of seemingly paradoxical situations and the techniques for analysing when such may occur are not limited to the multiple regression problem. The same approach can be extended to a much wider range of hypothesis testing problems.

Using the techniques of the Largey and Spencer paper, a simple example illustrating the point may be set up. Let X be a vector of normal random variables with known covariance matrix, and $\overline{X}$ the vector of sample means of n values from each population, such that:

$$X \sim N(\mu, \Sigma).$$

Then, $\overline{X} \sim N(\mu, \frac{1}{n}\Sigma)$.

Suppose we wish to test the hypotheses:

(i) $H_0: \mu = \mu^0$ (ii) $H_0^1: \mu_1 = \mu_1^0$

$H_0^2: \mu_2 = \mu_2^0$

.

.

.

$H_0^k: \mu_k = \mu_k^0$

i.e.
$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ . \\ . \\ . \\ \mu_k \end{pmatrix} = \begin{pmatrix} \mu_1^0 \\ \mu_2^0 \\ . \\ . \\ . \\ \mu_k^0 \end{pmatrix}$$

Paradoxical situations analogous to PS.1 and PS.2 are

P1: $H_0$ (i) rejected and each hypothesis in (ii) accepted.
P2: $H_0$ (i) accepted and each hypothesis in (ii) rejected.

Since $\Sigma$ is assumed known we test hypothesis (i) using the result $n(\overline{X} - \mu^0)'\Sigma^{-1}(\overline{X} - \mu^0) \sim X_k^2$. $H_0$ is accepted if $n(\overline{X} - \mu^0)'\Sigma^{-1}(\overline{X} - \mu^0) \leq \chi_k^{2*}$, the relevant $\chi^2$ value representing a significance level $\alpha$. Each hypothesis in (ii) is tested using $\overline{X}_i \sim N\left(\mu, \frac{1}{n}\sigma_i^2\right)$ where $\sigma_i^2$ is the $ii^{th}$ element of the matrix $\Sigma$. Thus $H_0^i$ is accepted if $\mu_i^0$ lies within the confidence interval $[\overline{X}_i - Z^* s_i, \overline{X}_i + Z^* s_i]$ where $s_i = \sigma_i / \sqrt{n}$. $Z^*$ is the normal distribution critical value allowing a confidence level $(1-\alpha)$ for the test.

Setting k=2 in the sets of hypotheses above, or singling out two means to

test, the confidence regions for both tests can be shown diagrammatically and the paradoxical situations labelled. The confidence region for (i) forms a two dimensional ellipse while that for both hypotheses in (ii) forms a rectangle.

P1 occurs when the hypothesised values for the means fall within the box (abcd in Figure 1), but outside the ellipse in regions marked 1, while P2 occurs when hypothesised $\mu$ values fall within the ellipse but outside the box, in regions marked 2.

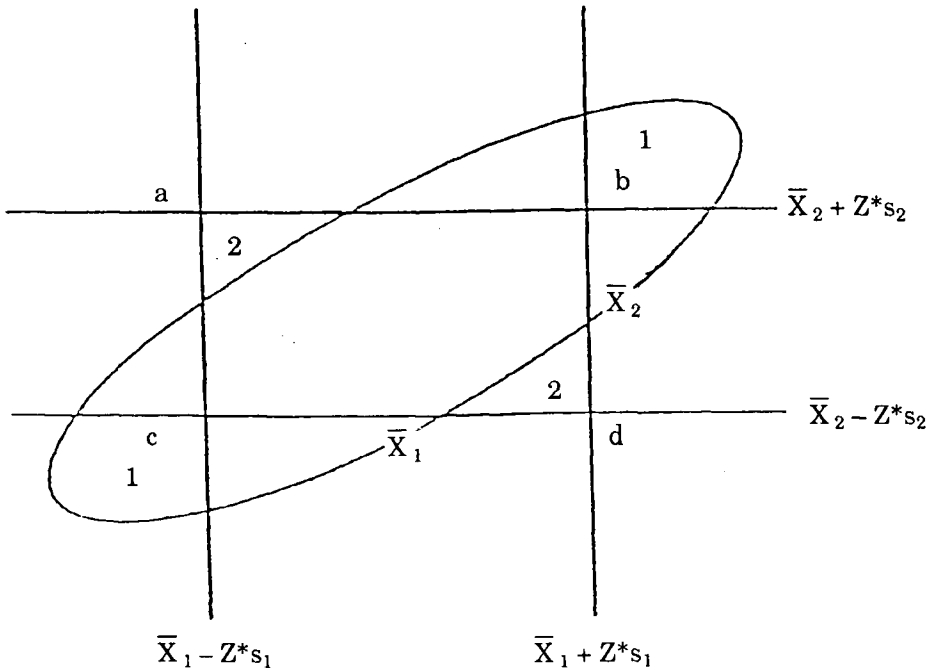


Figure 1: *P1 and P2 Characterised Diagrammatically*

As in the multiple regression case analysed in Largey and Spencer (op. cit.), three classes of situation may arise. Class A occurs when of the two paradoxical situations only P2 is possible which implies in the two dimensional case that abcd is completely enclosed by the ellipse. Class B, where only P1 is possible, occurs when all the corners of the rectangle abcd lie outside the ellipse. Class C which allows both P1 and P2 as possible events occurs when only two opposite corners of abcd are contained within the ellipse. (See Figure 2.)

We can relate the existence of these 3 classes to the correlation between the means.

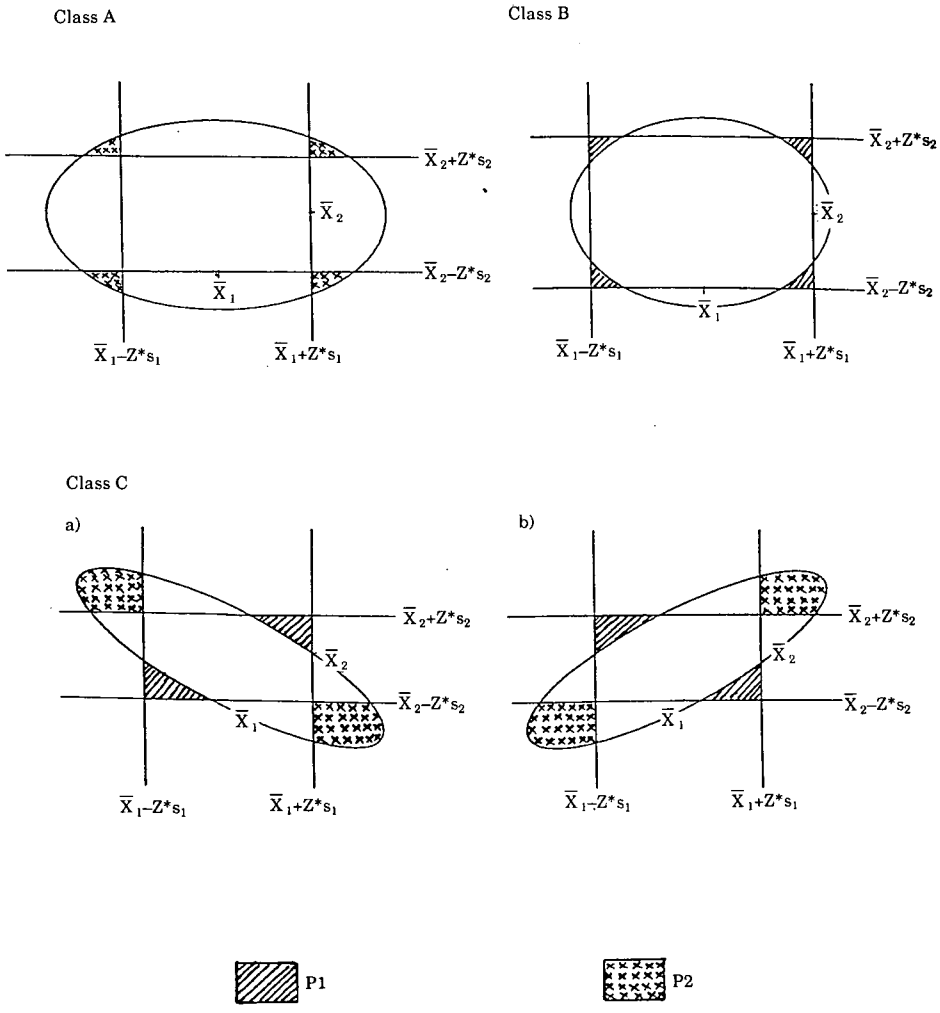


Figure 2: *Three Possible Classes of Situation*

Let $\rho_{12}$ be the correlation between $\bar{X}_1$ and $\bar{X}_2$.

$$\rho_{12} = \frac{\sigma_{12} / n}{\sqrt{\left(\sigma_1^2 / n\right)\left(\sigma_2^2 / n\right)}} = \frac{\sigma_{12}}{\sigma_1 \sigma_2}$$

where $\sigma_{12}/n$ is the covariance of $\bar{X}_1$ with $\bar{X}_2$ and $\sqrt{\sigma_i^2 / n}$ is the s.e. for $\bar{X}_i$. Note that $\rho_{12}$ reduces simply to the correlation between the two variables $X_1$ and $X_2$.

Thus this example provides the interesting result that analysis of the paradoxical situations using the correlation of the estimated parameters (the means of the variables) relates back immediately to the correlation between the variables themselves. (This was not the case with regard to the same exercise in multiple regression where there the classes were described in terms of the correlation between the estimated regression coefficients, not the correlation between regression variables themselves.)

The 3 classes can be characterised using identical formulae to those derived in the Largey and Spencer paper in the case of a multiple regression problem with a known error covariance matrix. Hence setting $\theta^* = \frac{\chi_2^{2*}}{2Z^{*2}}$ we have occurrence of the various classes limited to the following situations:

Class A: All corners of R in E

$$\frac{1-\theta^*}{\theta^*} \leq \rho_{12} \leq \frac{\theta^*-1}{\theta^*}$$

Class B: No corners of R in E

$$\frac{\theta^*-1}{\theta^*} < \rho_{12} < \frac{1-\theta^*}{\theta^*}$$

Class C: Two opposite corners in E

either (a) $\rho_{12} \leq \frac{\theta^*-1}{\theta^*}, \quad \rho_{12} < \frac{1-\theta^*}{\theta^*}$

or (b) $\rho_{12} \geq \frac{1-\theta^*}{\theta^*}, \quad \rho_{12} > \frac{\theta^*-1}{\theta^*}.$

The value of $\theta^*$ depends only on the value of $\alpha$, not on the sample size, so the bounds on the correlation coefficient for each of these classes depend solely on the strictness of the hypothesis tests. Figure 3 illustrates the bounds for each class.

For $\alpha > 0.2152$ only Classes A and C could occur. Class B will not. This implies P1 can only occur if P2 is also possible (through Class C), whereas it is possible for P2 to exist alone (through Class A). The lower bound on $|\rho_{12}|$ for which P1 is possible increases as $\alpha$ rises.

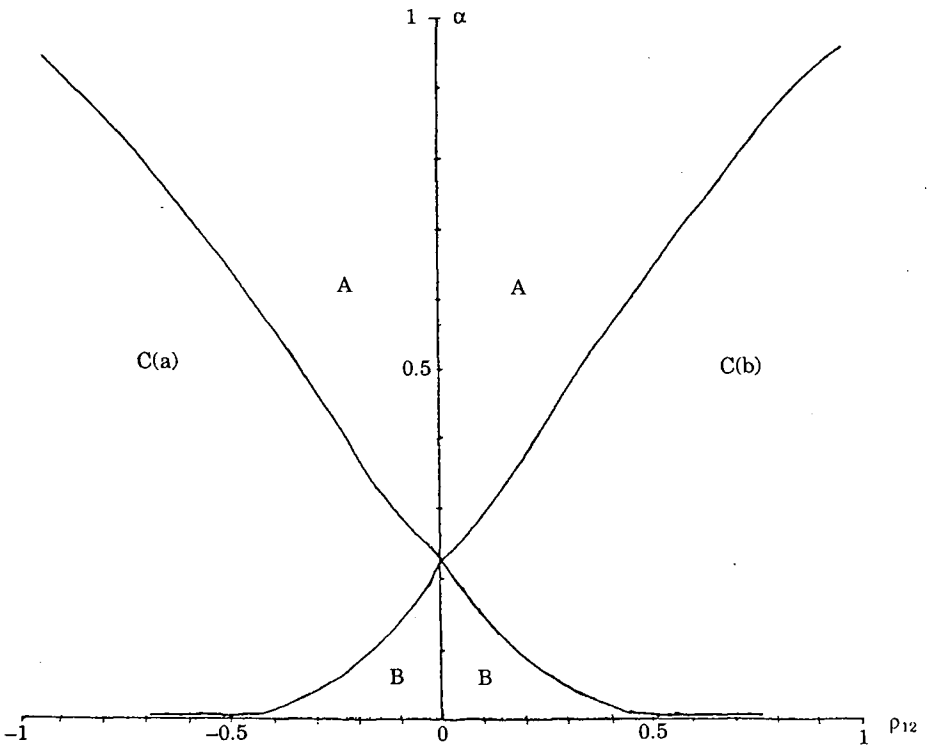


Figure 3: *Bounds on $\rho_{12}$ for Classes A, B and C as $\alpha$ Varies*

For more reasonable sized tests, $\alpha \leq 0.2152$, the opposite is the case. Here Classes B and C can occur, but A cannot. P2 cannot occur without the possibility of P1 also, and the bounds on $|\rho_{12}|$ for which P2 is possible increase as the test becomes stricter, tending towards 1 in the limit as $\alpha$ tends to zero.

It is clear therefore that of the two paradoxical situations analogous to those set out by Geary and Leser, at least one is always possible, in the sense that suitable hypothesised values of the parameters exist, and both could possibly result provided the value of $\rho_{12}$ is such as to allow the existence of Class C. With commonly used significance levels ($\alpha = 0.05$, $\alpha = 0.025$), P1 is possible for any $\rho_{12}$ value, while $|\rho_{12}|$ must be larger for P2 to be also possible.

These results reflect those of the 2 variable multiple regression case, where for commonly used $\alpha$, the multicollinearity case (PS.1) is possible for any value of the coefficient of correlation between the estimated parameters, while PS.2 is possible also for higher absolute values of the correlation coefficient.

Note that in the multiple regression problem with variance — covariance

matrix $\sigma^2 I$, $\sigma^2$ unknown, the same formulae characterising each of the three classes would still apply if we redefine $\theta^*$ as $\theta^* = \frac{F^*}{t^{*2}}$. This implies a second determining factor for the value of $\theta^*$ in addition to the significance level of the test, the number of degrees of freedom for the test. The basic shape of Figure 3 will hold, for more than three degrees of freedom and Figure 3 represents the limit of the changing graph as degrees of freedom tend to infinity. This more realistic case is dealt with in Largey and Spencer (op. cit.).

## III THE TAXI OR SCHRÖDINGER[2] PROBLEM

The second paper, Geary (1944), contains a problem which greatly intrigued him and which he called the Schrödinger problem, following its proposal by Schrödinger at a meeting in the early 1940s of the Dublin University Mathematical Society. It can be stated as follows: A spectator at a street corner is observing cars passing. The N cars in the town are numbered 1, 2, ..., N, N unknown. The spectator wishes to estimate N after observing that n cars numbered $x_1, \ldots, x_n$ have passed. Each of the N cars is assumed equally likely to pass.[3] This problem is now extremely well known and has various names including the "taxi problem". It is mentioned in many books including Feller (1957, pp. 211-212), Whittle (1970, pp. 63, 87-88) and Edwards (1972, pp. 165-167).

The problem arose in the context of Geary's analysis of Pitman's notion of "closeness" (Pitman, 1937). If X and Y are estimators of $\theta$, X is closer than Y if:

$$\Pr[|X - \theta| < |Y - \theta|] > \tfrac{1}{2}.$$

2. Erwin Schrödinger (1887-1961), the Austrian theoretical physicist, received the Nobel Prize in 1933 for his work on wave mechanics. After difficulties with the Nazis in 1938 he left Austria and was enticed by de Valera to Dublin where he arrived in October 1939. He was to remain there, working in the newly established Institute for Advanced Studies, until 1956. In 1943/1944 he was working on statistics, one of his earliest loves (Moore, 1989, pp. 415).

3. In fact, the problem is a good deal older than the early 1940s. John Aldrich has pointed out to us that it appears in Jeffreys (1939) and in an essay of C.S. Peirce written in 1911, reprinted in Eisele (1976), pp. 157-210. Jeffreys writes that the problem, with a "tramcar" setting, was suggested to him by M.H.A. Newman "several years ago", i.e., in the early to middle 1930s. He provides rough analysis for n=1 and says this is easily extendable. Earlier still, Peirce (1839-1914), the American founder of pragmatism and a pioneer in the development of formal logic, discusses the problem in a long letter of June 22, 1911 to a Mr Kehler (see Eisele, ibid, especially p. 188). In a section critical of Laplace (1749-1827) Peirce ascribes to Laplace an attempt to estimate N from a random drawing from an urn containing balls numbered 1,2,3,...,N but asserts that "no deductive conclusion on the subject can be drawn from those premisses correctly". Thus, the problem goes back at least to Laplace's famous 1812-1814 writings on probability.

X is a closest statistic if it is closer than any other. Pitman analysed closeness mainly in situations where sufficient statistics exist and he found the closest estimates of parameters of certain distributions including the rectangular distribution, $a-\frac{1}{2}c \leq X \leq a+\frac{1}{2}c$. The mid-point is a and the closest estimate of this, whether or not c is known, is shown to be (v+w)/2 where v is the minimum value in the sample and w the maximum.

Geary (1944) compared efficiency (relative variances) and closeness and showed, for example, that if X and Y are unbiased estimators of $\theta$ with a joint normal distribution and arbitrary correlation, X will be closer than Y if and only if var X < var Y. The ordering of estimators by closeness will not in general follow the ordering by efficiency, of course. Indeed, the former ordering is not necessarily transitive in that X could be closer than Y, Y than Z and Z closer than X, a fact noted by Pitman himself. Geary, however, calculated that closeness and efficiency would rank unbiased estimators similarly at least in large samples.

As an illustration he applied Pitman's theory to the Schrödinger problem. Formally the situation is that of estimating N in a rectangular (or uniform distribution) where N is the unknown upper bound and the lower bound is known. Geary assumed N large in order to justify treating the distribution as that of a continuous random variable X, with $f(x) = 1/N$, all x, $0 \leq x \leq N$.

His argument was as follows: Let W be the largest value in a random sample of n independent drawings.

$$\begin{aligned} \Pr(W \leq w) &= \Pr(X_1 \leq w, X_2 \leq w, \ldots, X_n \leq w) \\ &= (w/N)^n \end{aligned}$$

since $\Pr(X \leq w) = w/N$.

Thus, the median of W is the value of w such that $(w/N)^n = \frac{1}{2}$ i.e., $w = (\frac{1}{2})^{1/n}N$. Consequently, $w2^{1/n}$ is the value of N for which w is the median and this, from a theorem of Pitman, is the closest estimate.

We refer to $2^{1/n}W$ as the Geary estimator, an estimator which he showed to be sufficient.

Since the likelihood function is $(1/N)^n$, the maximum likelihood (ML) estimator is the smallest possible value of N consistent with the obvious consideration that N has to be at least as large as the maximum observed value. Thus W is the ML estimator. [4]

4. See Carlton (1946) for some discussions of ML and other matters regarding the estimation of the parameters of $\left[a-\frac{1}{2}c,\ a+\frac{1}{2}c\right]$. A survey of the rectangular distribution is provided by Johnson and Kotz (1970, Chapter 25). It is clear that Geary (1944) is one of the earliest references to the estimation problems of [O, N] and that $W\,2^{1/n}$ is new. Note that the likelihood function is zero for $N < \max x_i$ and so has a discontinuity at $\max x_i$.

The density of W is easy to derive. Since $F(w) = (w/N)^n$,

$$f(w) = n(w/N)^{n-1}/N$$
$$= nw^{n-1}/N^n.$$

Hence EW is the integral of $nw^n/N^n$ from O to N i.e. $nN/(n+1)$.

Similarly var $W = N^2n/(n+1)^2(n+2)$.

It follows at once that an unbiased estimator of N is given by $W(n+1)/n$ (see, e.g., Davis (1951,p. 48)).

This estimator is indeed the minimum variance unbiased (MVU) estimator as shown by Davis (ibid), Tate (1959) and Kendall and Stuart (1973, p. 36). There are, of course, many other unbiased estimators including $2\bar{X}$ and 2 (median sample value). The variances of these estimators are $N^2/3n$ and $N^2/(n+2)$ and hence they are totally inefficient relative to estimators of the form $\lambda W$, $\lambda$ constant. Apparently, Schrödinger criticised Geary's estimator on the grounds that it, unlike $\bar{X}$, did not use all the data. Geary convinced his critic by pointing out that the variance of $\bar{X}$ was $0(1/n)$ while that of W was $0(1/n^2)$ i.e., of a different and superior order of magnitude. (See also Carlton, 1946 Section 4 for the same point.)

Johnson (1950) discussed the comparison of estimators and advocated mean square error (MSE) as a criterion for judging them. For the Schrödinger problem, he computed the MSE (= variance + bias squared) of $\lambda W$, $\lambda$ constant and showed this was minimised when $\lambda = (n+2)/(n+1)$, a value which is independent of the unknown N.

We now have four estimators worth discussing. These are set out in Table 1.

Table 1: *Estimators of N and their Characteristics*

| | *Estimator* | *Expected Value* | *Variance* | *MSE* |
|---|---|---|---|---|
| Geary | $2^{1/n}W$ | $2^{1/n}nN/(n+1)$ | $\frac{2^{2/n}nN^2}{(n+1)^2(n+2)}$ | $\frac{N^2}{(n+1)^2}\left\{\frac{2^{2/n}n}{n+2}+\left[2^{1/n}n-(n+1)^2\right]\right\}$ |
| ML | W | $nN/(n+1)$ | $\frac{nN^2}{(n+1)^2(n+2)}$ | $\frac{2N^2}{(n+1)(n+2)}$ |
| MVU (Davis) | $W(n+1)/n$ | N | $\frac{N^2}{n(n+2)}$ | $\frac{N^2}{n(n+2)}$ |
| Min MSE (Johnson) | $W(n+2)/(n+1)$ | $n(n+2)N/(n+1)^2$ | $\frac{n(n+2)N^2}{(n+1)^4}$ | $\frac{N^2}{(n+1)^2}$ |

The next table compares MSE for the first three estimators relative to the min MSE estimator. Similar calculations appear in Johnson (ibid).

Table 2: *MSE Relative to Min MSE*

| | *n=1* | *2* | *5* | *20* | *40* | *∞* |
|---|---|---|---|---|---|---|
| Geary | 1.333 | 1.029 | 1.008 | 1.061 | 1.0765 | 1.094 |
| ML | 1.333 | 1.50 | 1.714 | 1.909 | 1.952 | 2. |
| MVU | 1.333 | 1.125 | 1.029 | 1.022 | 1.001 | 1. |

The Geary estimator performs excellently, especially for moderate n and remembering that it is absolutely the closest. (Johnson makes some calculations on the closeness and also remarks on the poor performance of the ML estimator.)

Regarding a confidence interval, Geary argues (in five lines) as follows. We look for an upper limit on N which should not be so great as to render too unlikely the occurrence of the largest number actually observed, say 50. Now if N was indefinitely large, this probability would be negligible and getting a sample maximum as low as 50 could effectively be ruled out.

Write $$\alpha = \Pr(W \leq w) = (w/N)^n.$$

If N was 100 and n was 5, $\alpha$ would be $(50/100)^5 = .031$ and in about 3 samples in 100 would as low a sample maximum as 50 be observed.

$\alpha = .05$ would be generated if N were to satisfy $(50/N)^5 = .05$

i.e. $$N = 50/(.05)^{1/5}$$
$$= 91.03$$

yielding a 95 per cent confidence interval of

$$50 \leq N \leq 91.03$$

or, generally, a $1-\alpha$ confidence interval for N

$$w \leq N \leq w/\alpha^{1/n}.$$

Geary takes $\alpha = .05$, $w = 247$ and $n = 30$ and writes that N "will be less than 273 unless in taking the particular sample an event, the probability of which was 1/20, occurred". His point estimate, of course, is 253.

This method of finding a probabilistic upper bound appears in Kendall and

Stuart (op. cit., pp. 138-139) and in Patel, Kapadia and Owen (1976, p. 191).

Returning to the discrete rectangular distribution case, i.e. that formally implied by the Schrödinger problem, there are two distinct situations — with and without replacement. It is not clear which situation Geary or Schrödinger had in mind but that without replacement may be easier to analyse and we consider it next.

Several estimators have been proposed (see Noether, 1971 and Mosteller, 1965).

(a) The ML estimator W.

(b) The estimator implied by adding the average gap between the observed numbers to the largest observed number. For example, if the observed numbers were 5, 12, 34, 48 the gaps would be 4, 6, 21, 13 with average gap 11 and estimate 48 + 11 = 59. This estimator is W [(n+1)/n] – 1. (As regards the minus one, note that the lower bound on X is 1 unlike the continuous case above, where it was zero). Analysing the Schrödinger problem by paying explicit attention to these gaps or spaces is advocated in an interesting paper by Rao (1981). Using regression analysis, treating the gaps as observations, yields this estimator as BLUE.

(c) Geary's $2^{1/n}$ W. He regarded it as likely that his solution would hold in the discrete case for any N.

(d) Since the median of the distribution lies halfway between 1 and N, i.e. (N+1)/2, N equals one less than twice the median of the distribution. Thus N might be estimated by one less than twice the median of the observed numbers — or, replacing the median by the sample mean, $2\overline{X}-1$.

A useful account of the distribution is in Tenenbein (1971).

$$\Pr(X_1 \le a, X_2 \le a, \ldots X_n \le a)$$

$$= \frac{a}{N}\frac{a-1}{N-1} \cdots \frac{a-n+1}{N-n+1}$$

$$= \binom{a}{n} / \binom{N}{n}.$$

Hence, with $W = \max X_i$ , $i=1 \ldots n$

$$\Pr(W \le w) = \binom{w}{n} / \binom{N}{n}.$$

The density of W is:

$$P(W \leq w) - P(W \leq w-1)$$

$$= \binom{w-1}{n-1} / \binom{N}{n}, \quad w = n,\ldots,N.$$

Obviously $w \geq n$ when there is no replacement.

The mean and variance of W are easily calculated (see Tenenbein, ibid):

$$E\,W = \left(\frac{n}{n+1}\right)(N+1)$$

$$\mathrm{Var}W = \frac{n(N+1)(N-n)}{(n+1)^2(n+2)}$$

Tenenbein shows that W is sufficient for N and that the average gap estimator is the unique minimum variance unbiased estimator.

Since it is of the form $\lambda W-1$, it is natural to look for the minimum MSE estimator of this form, $\lambda$ constant.

$$\lambda W-1 \text{ has MSE}$$
$$\lambda^2 \mathrm{var}W + [\lambda EW-1-N]^2.$$

Writing this as $F(\lambda)$ and differentiating, we find

$$F'(\lambda) = 0 \text{ if } \lambda = \frac{(n+2)(N+1)}{n+N(n+1)}$$

This will minimise F but is unusable as it depends on the unknown N.

However, as N gets large, the solution for $\lambda$ tends to $(n+2)/(n+1+n/N)$ so we have the large N min MSE estimator as

$$w\left(\frac{n+2}{n+1}\right) - 1, \text{ provided n is not too large relative to N.}$$

From the expressions for EW and var W, it is straightforward to calculate the mean and variance of all these estimators apart from that involving the mean and median. It can easily be shown that:

$$E\,\bar{X} = (N + 1)/2$$

$$\text{var}\,\bar{X} = \left(\frac{N - n}{N - 1}\right)(\text{var } X)/n$$

and less easily (see Wilks, 1962, p. 251)

$$E\ \text{Median} = (N+1)/2$$

$$\text{var Median} = (N-n)(N+1)/4(n+2).$$

As in the continuous case, the estimators based on mean or median are hopelessly inefficient and are not discussed further.

The authors have examined the cases N=10, N=100 and N=1000 in some detail for the four leading estimators. In the tables, an asterisk denotes the lowest MSE for the particular sample size, n.

It is apparent from the theoretical results and from the tables that Geary's estimator stands up excellently in the discrete case without replacement. In the three cases considered, it does best in the N = 100 case where it dominates for small n in MSE terms and does well for moderate n.

It should be pointed out that informal information can be useful in estimating N (Mosteller, op. cit.). Rosenberg and Deely (1976) provide an analysis using a Bayesian approach.

Table 3: *Comparison of Estimators of N when N = 10*

| | *n* | *1* | *2* | *4* | *6* | *8* |
|---|---|---|---|---|---|---|
| W | E | 5.50 | 7.33 | 8.80 | 9.43 | 9.78 |
| | Var | 8.25 | 4.89 | 1.76 | 0.67 | 0.22 |
| | MSE | 28.50 | 12.00 | 3.20 | 1.00 | 0.27* |
| $2^{1/n}W$ | E | 11.00 | 10.37 | 10.47 | 10.58 | 10.66 |
| | Var | 33.00 | 9.78 | 2.49 | 0.85 | 0.26 |
| | MSE | 34.00 | 9.92* | 2.71* | 1.19 | 0.70 |
| $W\left(\frac{n+1}{n}\right)-1$ | E | 10.00 | 10.00 | 10.00 | 10.00 | 10.00 |
| | Var | 33.00 | 11.00 | 2.75 | 0.92 | 0.28 |
| | MSE | 33.00 | 11.00 | 2.75 | 0.92* | 0.28 |
| $W\left(\frac{n+2}{n+1}\right)-1$ | E | 7.25 | 8.78 | 9.56 | 9.78 | 9.86 |
| | Var | 18.56 | 8.69 | 2.53 | 0.88 | 0.27 |
| | MSE | 26.13* | 10.19 | 2.73 | 0.93 | 0.29 |

Table 4: *Comparison of Estimators of N when N = 100*

| | *n* | *5* | *10* | *20* | *40* | *80* |
|---|---|---|---|---|---|---|
| W | E | 84.17 | 91.82 | 96.19 | 98.54 | 99.75 |
| | Var | 190.38 | 62.60 | 16.66 | 3.43 | 0.30 |
| | MSE | 441.07 | 129.55 | 31.17 | 5.57 | 0.36 |
| $2^{1/n}W$ | E | 96.68 | 98.41 | 99.58 | 100.26 | 100.62 |
| | Var | 251.20 | 71.91 | 17.85 | 3.55 | 0.31 |
| | MSE | 262.21* | 74.45* | 18.03* | 3.62 | 0.69 |
| $W\left(\frac{n+1}{n}\right)-1$ | E | 100.00 | 100.00 | 100.00 | 100.00 | 100.00 |
| | Var | 274.14 | 75.75 | 18.36 | 3.61 | 0.31 |
| | MSE | 274.14 | 75.75 | 18.36 | 3.61 | 0.31* |
| $W\left(\frac{n+2}{n+1}\right)-1$ | E | 97.19 | 99.17 | 99.77 | 99.94 | 99.98 |
| | Var | 259.12 | 74.50 | 18.28 | 3.60 | 0.31 |
| | MSE | 267.00 | 75.20 | 18.33 | 3.61* | 0.31 |

Table 5: *Comparison of Estimators of N when N = 1,000*

| | *n* | *5* | *20* | *80* | *200* | *500* |
|---|---|---|---|---|---|---|
| W | E | 834.17 | 953.33 | 988.64 | 996.02 | 999.00 |
| | Var | 19,761.81 | 2,022.22 | 136.94 | 19.63 | 1.99 |
| | MSE | 47,262.50 | 4,200.00 | 265.94 | 35.47 | 2.98 |
| $2^{1/n}W$ | E | 958.21 | 986.95 | 997.25 | 999.48 | 1,000.39 |
| | Var | 26,075.86 | 2,167.36 | 139.33 | 19.76 | 1.99 |
| | MSE | 27,822.61 | 2,337.60 | 146.92 | 20.03 | 2.14 |
| $W\left(\frac{n+1}{n}\right)-1$ | E | 1,000.00 | 1,000.00 | 1,000.00 | 1,000.00 | 1,000.00 |
| | Var | 28,457.00 | 2,229.50 | 140.38 | 19.82 | 1.99 |
| | MSE | 28,457.00 | 2,229.50 | 140.38 | 19.82 | 1.99* |
| $W\left(\frac{n+2}{n+1}\right)-1$ | E | 972.19 | 997.73 | 999.85 | 999.98 | 1,000.00 |
| | Var | 26,898.01 | 2,219.40 | 140.34 | 19.82 | 1.99 |
| | MSE | 27,671.16* | 2,224.55* | 140.36* | 19.82* | 1.99 |

We have a final footnote to this analysis of the Schrödinger problem. According to Noether (op. cit.), analysis of the Schrödinger problem was most useful during World War II , when German tanks took the place of "taxi cabs". He writes that statistical estimates of German tank production were much more accurate than estimates based on more orthodox intelligence sources. Since Geary's paper appeared in 1944 it would likely have been sent to *Biometrika* by 1942 or 1943. The editor of *Biometrika* in that period was

E.S. Pearson (1895-1980) who would therefore have known of Geary's work on the problem and who was head of a group of statisticians working on weapons assessment with the Ordnance Board (see the article on E.S. Pearson in Kotz and Johnson, 1985, p. 652). It seems therefore very likely that Geary's analysis played at least a major part in the success referred to by Noether and therefore in contributing to the war effort. One wonders if he knew this. Given that his great mathematical powers were always targeted at application to real problems, how satisfying this would have been, had he known it.

## *REFERENCES*

CARLTON, A.G., 1946. "Estimating the Parameters of a Rectangular Distribution", *Annals of Mathematical Statistics*, Vol. 17, pp. 355-358.

CRAMER, E.M., 1972. "Significance Tests and Tests of Models in Multiple Regression", *The American Statistician*, Vol. 26, No. 4, pp .26-30.

DAVIS, R.C., 1951. "On Minimum Variance in Non-regular Estimation", *Annals of Mathematical Statistics*, Vol. 22, pp. 43-57.

EDWARDS, A.W.F., 1972. *Likelihood*, Cambridge: Cambridge University Press.

EISELE, C. (ed.), 1976. "The New Elements of Mathematics" by C.S. Peirce, Vol. III/1, *Mathematical Miscellanea*, Atlantic Highlands, New Jersey: Humanities Press.

FELLER, W., 1957. *An Introduction to Probability Theory and Its Applications*, Vol.1, 2nd Edition, New York: Wiley.

GEARY, R.C., 1944, "Comparison of the Concepts of Efficiency and Closeness for Consistent Estimates of a Parameter", *Biometrika*, Vol. 33, pp. 123-128.

GEARY, R.C., 1963. "Some Remarks about Relations between Stochastic Variables: A Discussion Document", *Review of International Statistical Institute*, Vol. 31, No. 2, pp. 163-181.

GEARY, R.C. and LESER, C.E.V., 1968. "Significance Tests in Multiple Regression", *The American Statistician*, Vol. 22, No. 1, pp. 20-21.

JEFFREYS, H., 1939. *Theory of Probability*, Oxford: Clarendon Press.

JOHNSON, N.L., 1950, "On the Comparison of Estimators", *Biometrika*, Vol. 37, pp. 281-287.

JOHNSON, N.L. and KOTZ, S., 1970. *Continuous Univariate Distributions — 2*, Boston: Houghton Mifflin.

KENDALL, M.G. and STUART, A., 1973. *The Advanced Theory of Statistics*, Vol. 2, 3rd Edition, London: Griffin.

KOTZ, S. and JOHNSON, N.L., 1985. *Encyclopedia of Statistical Sciences*, Vol. 6, New York: Wiley.

LARGEY, A. and SPENCER, J.E., 1992, *F and t Tests in Multiple Regression*, Queen's University, Belfast, Department of Economics, Working Paper No. 37.

MOORE, W., 1989. *Schrödinger: Life and Thought*, Cambridge: Cambridge University Press.

MOSTELLER, F., 1965. *Fifty Challenging Problems in Probability*, London: Addison Wesley.

NOETHER, G., 1971. *Introduction to Statistics: A Fresh Approach*, Boston: Houghton Mifflin.

PATEL, J.K., KAPADIA, C.H. and OWEN, D.B., 1976. *Handbook of Statistical Distributions*, New York: Dekker.

PITMAN, E.J.G., 1937, "The 'Closest' Estimates of Statistical Parameters", *Proceedings Cambridge Philosophy Society*, Vol. 33, pp. 212-222.

RAO, J.S., 1981. "Estimation Problems for Rectangular Distributions (or the Taxi Problem Revisited)", *Metrika*, Vol. 28, pp. 257-262.

ROSENBERG, W.J. and DEELY, J.J., 1976. "The Horse-Racing Problem — A Bayesian Approach", *The American Statistician*, Vol. 30, No. 1, pp. 26-29.

SAVIN, N.E., 1984. "Multiple Hypothesis Testing", Chapter 14 in Z. Griliches and M.D. Intriligator (eds.), *Handbook of Econometrics*, Vol. 2, Amsterdam, North-Holland, pp. 827-879.

SPENCER, J.E., 1976. "The Scientific Work of Robert Charles Geary", *The Economic and Social Review*, Vol. 7, No. 3, pp. 233-241.

SPENCER, J.E., 1983a. "Robert Charles Geary — An Appreciation", *The Economic and Social Review*, Vol. 14, No. 3, pp. 161-164.

SPENCER, J.E., 1983b. "Robert Charles Geary, Mathematicial Statistician", *Irish Mathematical Society Newsletter*, No. 8, pp. 12-20.

SPENCER, J.E., 1983c. "Robert Charles Geary 1896-1983", *Econometrica*, Vol. 51, No. 5, pp. 1,599-1,601.

TATE, R.F., 1959. "Unbiased Estimation: Functions of Location and Scale Parameters", *Annals of Mathematical Statistics*, Vol. 30, pp. 341-366.

TENENBEIN, A., 1971. "The Racing Car Problem", *The American Statistician*, Vol. 25, No. 1, pp. 38-40.

WHITTLE, P., 1970. *Probability*, Harmondsworth: Penguin.

WILKS, S.S., 1962. *Mathematical Statistics*, New York: Wiley.