Statistical Science
1996, Vol. 11, No. 1, 20–34

R. A. Fisher and Multivariate Analysis[1]

T. W. Anderson

Abstract. This paper reviews R. A. Fisher's many fundamental contributions to multivariate statistical analysis—from the derivation of the distribution of the sample correlation coefficient to discriminant analysis. The emphasis here is on the conceptual and mathematical development. All of his papers on multivariate analysis will be included in this survey.

Key words and phrases: Correlation coefficient, multiple correlation, geometric method, discriminant analysis.

1. INTRODUCTION

R. A. Fisher (1890–1962) was a pioneer in almost every aspect of statistical theory and practice, not least in multivariate analysis. He initiated the rigorous derivation of the distributions of many basic statistics, such as the product-moment, partial and multiple correlation coefficients. His geometrical view encouraged intuition as well as provided a powerful method of finding these distributions. His insight into the analysis that multiple measurements afforded led him to discriminant analysis. This paper reviews Fisher's many contributions to multivariate statistical analysis—from one of his first papers to one of his last. The emphasis here is on the work of this one important figure rather than a history of the development of multivariate analysis. For other views of Fisher's role in statistics, see Savage (1976) and Rao (1992), among others.

When I was a graduate student at Princeton University in 1941 ready to start my dissertation research, S. S. Wilks suggested that I read two papers on discriminant analysis by Fisher, published in 1936 and 1938 (summarized below). The first paper was fairly straightforward, but I had great difficulty understanding the second. Finally I gave up and worked out on my own what I thought were Fisher's ideas. That activity got me started on my lifelong interest in multivariate statistics, including research and writing (and revising) a fairly comprehensive book. In my work I have had many occasions to study Fisher's contributions. This R. A. Fisher lecture provides me the opportunity to review the entire sequence of his contributions to the field.

2. THE DISTRIBUTION OF THE PRODUCT-MOMENT CORRELATION COEFFICIENT

Multivariate statistical analysis can be considered as starting with the work of Francis Galton (1822–1911), summarized in *Natural Inheritance* (Galton, 1889), in which the ideas of regression were developed in terms of observations that have the characteristics of samples from a bivariate normal distribution. Galton noted the elliptical contours of equal density. He estimated the lines relating medians of one variable to the values in intervals of the other by plotting these points and determining the slopes by eye. Francis Ysidro Edgeworth (1845–1926) explored other methods for estimating the slopes of what we now know as "regression lines." He arrived at the estimator $S(xy)/S(x^2)$, where S denotes summation over the sample and x and y are standardized observations. However, it is not clear whether the standardization (subtraction of the means and division by the standard deviations) was done by sample quantities or population quantities assumed known. Edgeworth also studied the mathematical question of how to obtain the coefficients of the quadratic form in the density of a multivariate normal distribution from the variances and correlations of the variables.

Karl Pearson (1857–1936) is usually credited with defining the product-moment correlation coefficient

$$(1) \qquad r = r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

[1]The R. A. Fisher Memorial Lecture presented to the Annual Meeting of the American Statistical Association, the Institute of Mathematical Statistics and the Eastern North American Region of the International Biometric Society, August 7, 1985.

T. W. Anderson is Professor Emeritus of Statistics and Economics at Stanford University.

as the "best value" estimating the correlation coefficient of the bivariate normal distribution (Pearson, 1896). (His notation made for some ambiguity between sample and population quantities.) In treating the regression of one standardized variable x on two other standardized variables y and z he expressed the regression coefficients in terms of the product-moment correlations; these "coefficients of double regression" are proportional to partial correlation coefficients. George Udny Yule (1871–1951), starting as an assistant to Pearson, clarified the partial correlation coefficient, say $r_{xy\cdot z}$, as the correlation between the residuals of x and y regressed on z and the multiple correlation coefficient, say $R_{x\cdot yz}$, as the correlation between x and its regression on y and z (Yule, 1897). Pearson, Yule and others developed asymptotic sampling theory for these measures of association. At the end of the nineteenth century the basic parameters and statistics of multivariate normal analysis had been defined and studied. For more detail on this history, see Chapters 8, 9, and 10 of Stigler (1986), Chapter V of Walker (1929) and Chapters 8 and 9 of Porter (1986).

In his first paper on multivariate analysis, Fisher (1915) found the exact distribution of the product-moment correlation coefficient in a sample of n observations (x_i, y_i), $i = 1, \ldots, n$, from

$$N\left[\begin{pmatrix}\mu_x\\ \mu_y\end{pmatrix}, \begin{pmatrix}\sigma_x^2 & \sigma_x\sigma_y\rho\\ \sigma_x\sigma_y\rho & \sigma_y^2\end{pmatrix}\right]. \tag{2}$$

This paper was pathbreaking in that it derived in a mathematically rigorous fashion the "small-sample" distribution of a fairly complicated nature and it introduced a basic geometric approach to such a distribution problem. It was a substantial achievement, especially for a young man of 25 years of age. The geometric method was consistent with Fisher's choice of spherical trigonometry as an optional subject in school and may have been stimulated by his experience of being tutored in mathematics at a time when his reading was restricted because of eye trouble. (See Box, 1978, pages 11–15.)

Fisher was drawn to the problem by an article on the distribution by Soper (1913), who had been stimulated by two papers of Student: one (Student, 1908a) on what we now know as the t-distribution and one (Student, 1908b) on the correlation coefficient. At that time the significance of a sample mean was determined by referring the ratio of the mean to its standard error to the standard normal distribution. Student realized that for small samples the variability in the sample standard error would augment the variability of the ratio.

Fisher's paper on the correlation coefficient included a derivation of the distribution of the sample standard deviation, which had earlier been surmised by Student (1908a). Student defined the standard deviation as

$$s = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}} \tag{3}$$

and calculated the first four moments of s^2 on the assumption that the n observations are independently normally distributed with common variance. He observed that these four moments agreed with those of a Pearson type III distribution and then asserted that this type III distribution was indeed the distribution of s^2. This assertion is equivalent to the statement that s^2 is proportional to a chi-square variable with $n - 1$ degrees of freedom. It was customary at that time to use one of the seven Pearson families of frequency curves to approximate a given density. Next Student argued that $\bar{x}$ and s^2 were uncorrelated and showed by an algebraic computation that $\bar{x}^2$ and $(s^2)^2$ were uncorrelated. These facts led Student to write the density of $\bar{x}$ and s^2 as the product of the normal density of $\bar{x}$ and the type III density of s^2. He then made the usual change of variables and integrated out s^2 to obtain what we now know as the Student t-density.

Student and other British statisticians seemed to have been unaware that the German astronomer Helmert (1876a, b) had derived the distribution of s^2 much earlier. From the density of n independent normal variables with the same mean and variance, Helmert found

$$\frac{h^{n-1}}{\Gamma((n-1)/2)}\sigma^{(n-3)/2}\exp(-h^2\sigma) \tag{4}$$

as the density of σ, where σ is the sum of squared deviations of the observations from the sample mean and $2h^2$ is the reciprocal of the population variance (Helmert's notation). He used the linear transformation that is now known as the Helmert transformation. Note that the exponent of σ corresponds to $n - 1$ degrees of freedom. It is curious that Karl Pearson, who had studied at Heidelberg and Berlin, was unaware of Helmert's work. (Pearson was so affected by his experience in Germany that he changed the spelling of his name from Carl to Karl.) Pearson (1900) showed that the quadratic form in the exponent of a multivariate normal distribution has a distribution now called chi-square with number of degrees of freedom equal to the dimensionality, but Student did not refer to this paper nor did Pearson in an editorial comment.

(For more history of the chi-square distribution, see Lancaster, 1966, and Kruskal, 1946, 1968, for example.)

Fisher was stimulated even more by the second paper of Student (1908b). In this paper Student tackled the problem of the exact distribution of the correlation coefficient. In a sample of size 2 the correlation coefficient can take on only the values of 1 and -1. Student deduced that

$$\cos[\pi \Pr\{r = 1\}] = \rho \tag{5}$$

by a simple geometric argument. To investigate larger sample sizes he did a Monte Carlo study for $n = 4$ with $\rho = 0$ and $\rho = 0.66$, $n = 8$ with $\rho = 0$ and $\rho = 0.66$ and $n = 30$ with $\rho = 0.66$. These days we would consider his simulation rather crude; he considered the number of replications to be 750. He then calculated the first four moments of the empirical distributions and attempted to fit Pearson curves. In the case of $\rho = 0$ he was successful, coming up with the uniform distribution for $n = 4$ and $\text{const}(1 - r^2)^2$ for $n = 8$. From these two cases he generalized by a productive insight to the suggestion that the general formula for the density of the correlation coefficient in the case of independence was

$$\text{const}(1 - r^2)^{(n-4)/2}, \tag{6}$$

that is, the distribution of r^2 is a beta distribution with parameters $1/2$ and $(n - 2)/2$. For $\rho = 0.66$ he was unsuccessful in fitting a Pearson distribution and could not suggest a form.

Soper (1913) had fitted the function

$$\text{const}(1 - r)^{m_1}(1 + r)^{m_2} \tag{7}$$

by use of approximate first and second moments of the correlation coefficient, using the fact that the range of r is $[-1, 1]$. Fisher wrote "To Mr. Soper's laborious and intricate paper I cannot hope to do justice." Neither can I. Incidentally, Fisher made a practice of reading *Biometrika* during his lunch hour.

To explain Fisher's derivation of the density of r, first let me describe his attack on the problem of the distribution of $ns^2 = ns_x^2$, because this shows his geometric approach in a simple case. We start with finding the probability of the set

$$S_x = \{\mathbf{x} | \bar{x}^* \le \bar{x} \le \bar{x}^* + d\bar{x}^*,\ s_x^* \le s_x \le s_x^* + ds_x^*\} \tag{8}$$

in the n-dimensional space of $\mathbf{x} = (x_1, \ldots, x_n)'$, where $\bar{x}^*$ and s_x^* have specified values and $d\bar{x}^*$ and ds_x^* are small quantities. The density of $x_1, \ldots, x_n$ is

$$\begin{aligned} &\text{const} \exp\left\{-\frac{1}{2\sigma_x^2}\sum_{i=1}^{n}(x_i - \mu_x)^2\right\} \\ &= \text{const} \exp\left\{-\frac{1}{2\sigma_x^2}[n(\bar{x} - \mu_x)^2 + ns_x^2]\right\}. \end{aligned} \tag{9}$$

The probability of $\mathbf{x}$ falling in S_x is the integral of the density (9) over the set S_x, which is approximately the product of the density and the volume, namely,

$$\begin{aligned} &\Pr\{S_x\} \\ &= \text{const} \exp\left[-\frac{1}{2\sigma_x^2}n(\bar{x}^* - \mu_x)^2 - \frac{1}{2\sigma_x^2}ns_x^{*2}\right] \\ &\quad \cdot \text{Vol}(S_x) + o(d\bar{x}^*\, ds_x^*), \end{aligned} \tag{10}$$

where $o(u)$ means $o(u)/u \to 0$ as $u \to 0$. The volume $\text{Vol}(S_x)$ is found by geometry. The first pair of inequalities in (8),

$$n\bar{x}^* \le \sum_{\alpha=1}^{n} x_\alpha \le n(\bar{x}^* + d\bar{x}^*), \tag{11}$$

defines a slab—the space between two parallel hyperplanes orthogonal to the equiangular line—with thickness $n\, d\bar{x}^*$. See Figure 1. The second pair of inequalities,

$$ns_x^{*2} \le \sum_{\alpha=1}^{n}(x_\alpha - \bar{x})^2 \le n(s_x^* + ds_x^*)^2, \tag{12}$$

defines a cylindrical shell—the space between two concentric spherical cylinders with the equiangular line as axis—with thickness $\sqrt{n}\, ds_x^*$. The radius of the inner spherical cylinder is $\sqrt{n}s_x^*$ and of the outer is $\sqrt{n}(s_x^* + ds_x^*)$. The set S_x is the intersection of these two sets:

$$S_x = \text{slab} \cap \text{spherical cylindrical shell}. \tag{13}$$

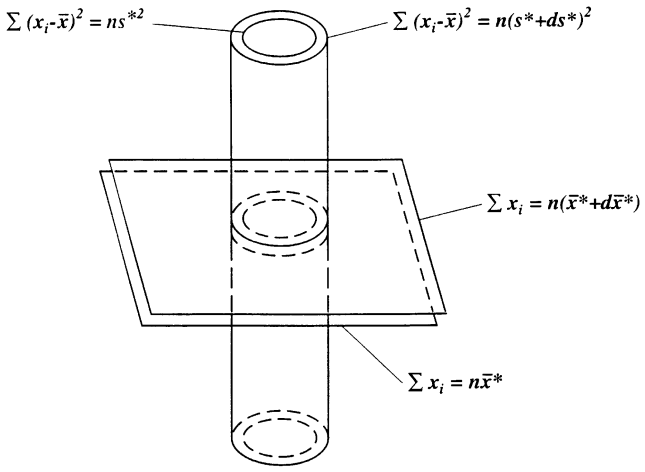

FIG. 1.

Since the radius of the inner spherical cylinder is $\sqrt{n}\ s_x^*$ and the intersection with a hyperplane (orthogonal to the axis) is a sphere in $n-1$ dimensions, the volume of S_x is

$$\text{Vol}(S_x) = \text{const}\, s_x^{*n-2}\, d\bar{x}^*\, ds_x^* + o(d\bar{x}^*\, ds_x^*). \tag{14}$$

The joint density of $\bar{x}$ and s_x at $\bar{x} = \bar{x}^*$ and $s_x = s_x^*$, obtained by dividing (10) by $d\bar{x}^*\, ds_x^*$ and letting $d\bar{x}^* \to 0$ and $ds_x^* \to 0$, is

$$\begin{aligned} &\text{const}\exp\left[-\frac{1}{2\sigma_x^2}\, n(\bar{x}^* - \mu_x)^2\right] \\ &\quad \cdot \text{const}(s_x^*)^{n-2}\exp\left[-\frac{1}{2\sigma_x^2} n s_x^{*2}\right]. \end{aligned} \tag{15}$$

This density factors into the density of $\bar{x}$ and the density of s_x. Thus s_x is proportional to the square root of a chi-square variable with $n-1$ degrees of freedom.

Actually Fisher included in his paper only the calculation of the volume of S_x, but he had apparently communicated the derivation of the distribution of s_x to Student some three years earlier. Student wrote to Karl Pearson September 12, 1912 (E. S. Pearson, 1968):

> Dear Pearson,
>
> I am enclosing a letter which gives a proof of my formulae for the frequency distribution of z $(= x/s)$, where x is the distance of the mean of n observations from the general mean and s is the S.D. of the n observations. Would you mind looking at it for me; I don't feel at home in more than three dimensions even if I could understand it otherwise.
>
> The question arose because this man's tutor is a Caius man whom I have met when I visit my agricultural friends at Cambridge and as he is an astronomer he has applied what you may call Airy to their statistics and I have fallen upon him for being out of date. Well, this chap Fisher produced a paper giving 'A new criterion of probability' or something of the sort. A neat but as far as I could understand it, quite unpractical and unserviceable way of looking at things. (I understood it when I read it but its gone out of my head and as you shall hear, I have lost it.) By means of this he thought he proved that the proper formula for the S.D. is
>
> $$\sqrt{\frac{\Sigma(x-m)^2}{n}} \quad \text{vice} \quad \sqrt{\frac{\Sigma(x-m)^2}{n-1}}.$$
>
> This, Stratton, the tutor, made him send me and with some exertion I mastered it, spotted the fallacy (as I believe) and wrote him a letter showing, I hope, an intelligent interest in the matter and incidentally making a blunder. To this he replied with two foolscap pages covered with mathematics of the deepest dye in which he proved, by using n dimensions that the formula was, after all
>
> $$\sqrt{\frac{\Sigma(x-m)^2}{n-1}}$$
>
> and of course exposed my mistake. I couldn't understand his stuff and wrote and said I was going to study it when I had time. I actually took it up to the Lakes with me—and lost it!
>
> Now he sends this to me. It seemed to me that if it's all right perhaps you might like to put the proof in a note. It's so nice and mathematical that it might appeal to some people. In any case I should be glad of your opinion of it.

Student wrote later in the month

> Dear Pearson,
>
> Since I wrote to you Fisher's first letter has turned up. It is nearly as incomprehensible to me as the other, but shows signs of supplying the missing links in the argument. I am sending it to you in case it should interest you.

The paper Student referred to was Fisher (1912) recommending what we now know as maximum likelihood estimation. (A psychoanalyst might make something of Student misplacing Fisher's communication.)

It is worth noting that (9) shows that the density of $x_1, \ldots, x_n$ depends on only $\bar{x}$ and s_x. We now recognize that $\bar{x}$ and s_x are sufficient statistics for μ_x and σ_x, but Fisher did not introduce the general idea of sufficient statistics until much later (Fisher, 1920).

To treat the correlation coefficient Fisher evaluated the probability of a set in the $2n$-dimensional space:

$$\begin{aligned} S = \{(\mathbf{x}, \mathbf{y}) | &\bar{x}^* \le \bar{x} \le \bar{x}^* + d\bar{x}^*, \\ &s_x^* \le s_x \le s_x^* + ds_x^*,\ \bar{y}^* \le \bar{y} \le \bar{y}^* + d\bar{y}^*, \\ &s_y^* \le s_y \le s_y^* + ds_y^*,\ r^* \le r \le r^* + dr^*\}. \end{aligned} \tag{16}$$

The density of $\mathbf{x}=(x_1,\ldots,x_n)'$ and $\mathbf{y}=(y_1,\ldots,y_n)'$ is, with C representing a generic constant,

$$
\begin{aligned}
f(\mathbf{x},\mathbf{y}) &= C\exp\Bigg\{-\frac{1}{2(1-\rho^2)}\sum_{\alpha=1}^{n}\Bigg[\frac{(x_\alpha-\mu_x)^2}{\sigma_x^2} \\
&\qquad -2\rho\frac{(x_\alpha-\mu_x)(y_\alpha-\mu_y)}{\sigma_x\sigma_y} \\
&\qquad +\frac{(y_\alpha-\mu_y)^2}{\sigma_y^2}\Bigg]\Bigg\} \\
&= C\exp\Bigg\{-\frac{n}{2(1-\rho^2)}\Bigg[\frac{(\bar{x}-\mu_x)^2}{\sigma_x^2} \\
&\qquad -2\rho\frac{(\bar{x}-\mu_x)(\bar{y}-\mu_y)}{\sigma_x\sigma_y}+\frac{(\bar{y}-\mu_y)^2}{\sigma_y^2}\Bigg]\Bigg\} \\
&\qquad \cdot\exp\Bigg\{-\frac{n}{2(1-\rho^2)}\Bigg[\frac{s_x^2}{\sigma_x^2}-2\rho r\frac{s_xs_y}{\sigma_x\sigma_y}+\frac{s_y^2}{\sigma_y^2}\Bigg]\Bigg\}.
\end{aligned}
\tag{17}
$$

(Note that $\bar{x}$, $\bar{y}$, s_x, s_y and r are sufficient statistics for μ_x, μ_y, σ_x, σ_y and ρ.) The first two pairs of inequalities in (16) define the set S_x in the n-dimensional space of $\mathbf{x}$ [which is a cylinder set in the $2n$-dimensional space of $(\mathbf{x},\mathbf{y})$]. Now consider the intersection of S with the hyperplane defined by an $\mathbf{x}$ in S_x. That intersection in the y-space satisfies the third, fourth and fifth pairs of inequalities in (16). The third pair defines a slab of thickness $n\,d\bar{y}^*$. The region defined by the fourth and fifth pairs is illustrated in Figure 2. The fourth pair defines a spherical cylindrical shell in the $n-1$ dimensions of the y-space orthogonal to the equiangular line. The radius of the inner cylinder is $\sqrt{n}s_y^*$ and the thickness of the shell is $\sqrt{n}\,ds_y^*$. The correlation r can be interpreted as $\cos\phi$, where ϕ is the angle between the vectors $(x_1-\bar{x},\ldots,x_n-\bar{x})$ and $(y_1-\bar{y},\ldots,y_n-\bar{y})$. The fifth pair of inequalities states that this angle is between $\phi^*=\cos^{-1}r^*$ and $\cos^{-1}(r^*+dr^*)$, which is approximately $\phi^*+d\phi^*$, where $d\phi^*=dr^*/\sin\phi^*$ is the increment in angle. This pair of inequalities defines the set between two concentric cones with axes along the vector $(x_1-\bar{x},\ldots,x_n-\bar{x})$. The intersection of the inner sphere and the inner cone is an $(n-2)$-dimensional sphere with radius $\sin\phi^*\sqrt{n}\,s_y^*$. The last two pairs of inequalities define the shaded region in Figure 2. Hence the two-dimensional area of the shaded cross section is $(\sqrt{n}\,ds_y^*)(\sqrt{n}\,s_y^*\,d\phi^*)$. The volume in n-space is proportional to

$$
\begin{aligned}
&\left(\sin\phi^*\sqrt{n}\,s_y^*\right)^{n-3}\left(\sqrt{n}\,s_y^*\,d\phi^*\right)\left(\sqrt{n}\,ds_y^*\right) \\
&\quad=\left(\sqrt{1-r^{*2}}\sqrt{n}\,s_y^*\right)^{n-3} \\
&\qquad\cdot\left(\sqrt{n}\,s_y^*\frac{dr^*}{\sqrt{1-r^{*2}}}\right)\left(\sqrt{n}\,ds_y^*\right) \\
&\quad=(1-r^{*2})^{(n-4)/2}n^{(n-1)/2}(s_y^*)^{n-2}\,dr^*\,ds_y^*.
\end{aligned}
\tag{18}
$$

Note that this volume does not depend on the fixed vector $\mathbf{x}$ and is in the n-dimensional space orthogonal to the x-space. Hence the probability of S is approximately

$$
\begin{aligned}
&C\exp\Bigg\{-\frac{n}{2(1-\rho^2)}\Bigg[\frac{(\bar{x}^*-\mu_x)^2}{\sigma_x^2} \\
&\qquad -2\rho\frac{(\bar{x}^*-\mu_x)(\bar{y}^*-\mu_y)}{\sigma_x\sigma_y}+\frac{(\bar{y}^*-\mu_y)^2}{\sigma_y^2}\Bigg]\Bigg\} \\
&\qquad\cdot\exp\Bigg\{-\frac{n}{2(1-\rho^2)}\Bigg[\frac{s_x^{*2}}{\sigma_x^2}-2\rho r^*\frac{s_x^*s_y^*}{\sigma_x\sigma_y}+\frac{s_y^{*2}}{\sigma_y^2}\Bigg]\Bigg\} \\
&\cdot s_x^{*n-2}s_y^{*n-2}(1-r^{*2})^{(n-4)/2}\,d\bar{x}^*\,d\bar{y}^*\,ds_x^*\,ds_y^*\,dr^*.
\end{aligned}
\tag{19}
$$

The joint density of $(\bar{x},\bar{y})$ and (s_x,s_y,r) factors into the density of $(\bar{x},\bar{y})$ and the density of (s_x,s_y,r). To integrate out s_x and s_y (dropping *), Fisher made the change of variables

$$
e^z=\frac{s_x/\sigma_x}{s_y/\sigma_y},\qquad w=\frac{s_xs_y}{\sigma_x\sigma_y}
\tag{20}
$$

(with constant Jacobian). The integration with respect to w leaves the density of r as

$$
\begin{aligned}
&\text{const}(1-\rho^2)^{(n-1)/2}(1-r^2)^{(n-4)/2} \\
&\qquad\cdot\int_0^\infty\frac{dz}{(\cosh z-\rho r)^{n-1}},
\end{aligned}
\tag{21}
$$

where $\cosh z=(e^z+e^{-z})/2\geq 1$. Let

$$
-\rho r=\cos\theta.
\tag{22}
$$

Then for $n=2$ the integral is

$$
\int_0^\infty\frac{dz}{\cosh z+\cos\theta}=\frac{\theta}{\sin\theta}.
\tag{23}
$$

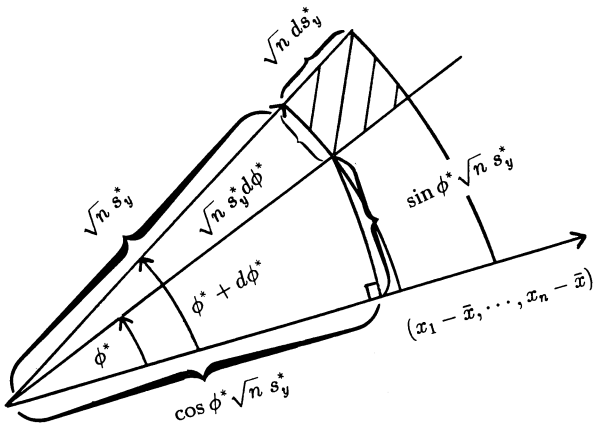

FIG. 2.

Fisher writes that "Professor Pearson has shown this last result" and indicates Pearson's proof. (See Problem 24 of Chapter 4 of Anderson, 1984.) By differentiating (23) under the integral sign with respect to θ we obtain

$$\int_0^\infty \frac{dz}{(\cosh z + \cos\theta)^{n-1}} = \frac{1}{(n-2)!}\left(\frac{\partial}{\sin\theta\,\partial\theta}\right)^{n-2}\frac{\theta}{\sin\theta}. \tag{24}$$

The density of r can be expressed as

$$\text{const}(1-\rho^2)^{(n-1)/2}(1-r^2)^{(n-4)/2}\frac{\partial^{n-2}}{\partial r^{n-2}}\frac{\theta}{\sin\theta}. \tag{25}$$

Fisher worked out some cases of small n, calculated some moments and gave a table of the functions of the first four moments for "inspection rather than for reference." He pointed out the use of Legendre functions in some of these integrations (showing his knowledge of classical mathematics). If $\rho = 0$, $\cos\theta = 0$ and (21) reduces to $\text{const}(1-r^2)^{(n-4)/2}$, which is exactly Student's suggestion.

He considered the transformation

$$t = \frac{r}{\sqrt{1-r^2}} \tag{26}$$

and noted that when $\rho = 0$ the distribution of t is Student's distribution. In passing he suggested the useful transformation

$$z = \frac{1}{2}\log\frac{1+r}{1-r} = \tanh^{-1} r, \tag{27}$$

but did not explore its use in this paper. It turned out that this transformation was eminently practical since the distribution of z is close to normal. He proposed that to estimate the population correlation coefficient one should maximize (25) with respect to ρ, referring to his paper (Fisher, 1912). This amounts to maximizing the likelihood of ρ.

Immediately following Fisher's paper is an Editorial, unsigned but obviously by the editor, Karl Pearson (1915), commenting on the distribution of the sample standard deviation. He was particularly interested in results for large samples, which he thought necessary. (In the second letter quoted above Student wrote "If I'm the only person you've come across that works with too small samples you are very singular.") In a footnote Pearson writes "Of course the form reached above shows that for normal distributions there is no correlation between deviations in the means and in the standard deviations of samples, a familiar fact." Though the basis for this remark is the factoring of the density of the two quantities, he did not write that they were independent.

Fisher's paper was actually submitted in September, 1914, but Pearson asked for additions such as graphs of some of the curves "and tracing as n increased the change of the frequency form towards a normal distribution." He further wrote "I don't think, myself, that values of r for n less than 10 ought ever to be considered, but tables of the distribution of r for $n = 10$ to 25, say, and for $\rho = 0.1$ to 1.0 by tenths would be of special value." Fisher did not carry out all of Pearson's suggestions, but made some revisions and the paper appeared in the May 1915 issue of *Biometrika*.

Before Fisher's paper appeared, Pearson and his associates began an intense study of the distribution and extensive computation of the density function; the density could be integrated numerically to obtain the cumulative distribution and hence significance levels. The work was ready by May, 1916, but publication was delayed about a year because *Biometrika* was having financial difficulties (Soper, Young, Cave, Lee and Pearson, 1917).

Pearson and Fisher had frequent and friendly exchanges of letters on their respective computations, but the published paper contained a section "On the determination of the 'most likely' value of the correlation in the sampled population." Pearson took Fisher's recommendation of using as an estimate the value of ρ that maximizes the density for an observed r as an application of Bayes' theorem on the assumption that a priori all values of ρ are equally likely to occur. The authors devote about 8 pages to discussing alternatives to a uniform prior for ρ, strongly criticizing the use of the uniform.

Fisher understandably was disturbed by this misrepresentation of his position, though he was perhaps partly to blame by using terms such as "most likely value" and "inverse probability". He incorporated a reply in his next paper on correlation (described below) and submitted it to *Biometrika*. Pearson responded August 21, 1920, as follows:

> Dear Mr. Fisher,
>
> Only in passing through Town today did I find your communication of August 3rd. I am very sorry for the delay in answering it, but it is not my fault. During my holiday no journals or pamphlets are forwarded to me and your paper being enclosed in an envelope with a large 'Lawes Agricultural Trust' heading had been taken by the laboratory steward for an offprint and not forwarded.
>
> As there has been a delay of three weeks already, and as I fear if I could give full attention to your paper, which I

> cannot at the present time, I should be unlikely to publish it in its present form, or without a reply to your criticisms which would involve also a criticism of your work of 1912—I would prefer you published elsewhere. Under present printing and financial conditions, I am regretfully compelled to exclude all that I think erroneous on my own judgment, because I cannot afford controversy.

Note the gentle, but firm "I would prefer you published elsewhere."

This was not the first paper of Fisher that Pearson rejected. In 1916 Fisher had sent Pearson a note commenting on a paper in *Biometrika* that used minimum chi-square as a criterion. Pearson's letter on this matter ends

> if I were to publish your note, it would have to be followed by another note saying that it missed the point, and that would be a quarrel among contributors.

The other note would, of course, be by the editor. Two years later Fisher wrote another communication on this subject, and Pearson replied in part

> Also I fear that I do not agree with your criticism of Dr. Kirstine Smith's paper and under present pressure of circumstances must keep the little space I have in *Biometrika* from controversy which can only waste what power I have for publishing original work.

After three rejections Fisher decided that *Biometrika* was not a vehicle for the publication of his papers.

3. DISTRIBUTIONS OF OTHER CORRELATION COEFFICIENTS

His next paper on correlation was published in *Metron* (Fisher, 1921), a new journal in Italy. Here he treated the estimation of ρ in (2) when $\mu_x = \mu_y$ and $\sigma_x = \sigma_y$. The statistic is the intraclass correlation coefficient given by

$$
\begin{aligned}
r &= \frac{\sum_{i=1}^n (x_i - \hat{\mu})(y_i - \hat{\mu})}{\frac{1}{2}[\sum_{i=1}^n (x_i - \hat{\mu})^2 + \sum_{i=1}^n (y_i - \hat{\mu})^2]} \\
&= \frac{2\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) - \frac{1}{2}n(\bar{x} - \bar{y})^2}{\sum_{i=1}^n (x_i - \bar{x})^2 + \sum_{i=1}^n (y_i - \bar{y})^2 + \frac{1}{2}n(\bar{x} - \bar{y})^2},
\end{aligned} \tag{28}
$$

where $\hat{\mu} = (\bar{x} + \bar{y})/2$. By another geometric argument Fisher showed that the density of r is

$$
\begin{aligned}
&\text{const}(1+\rho)^{n/2-1}(1-\rho)^{(n-3)/2}(1+r)^{(n-3)/2} \\
&\quad \cdot (1-r)^{n/2-1}(1-\rho r)^{-(n-1/2)},
\end{aligned} \tag{29}
$$

a simpler result than the earlier case. Now we know that (28) has the distribution of

$$
\frac{((1+\rho)/(1-\rho))((n-1)/n)F_{n-1,n} - 1}{((1+\rho)/(1-\rho))((n-1)/n)F_{n-1,n} + 1}, \tag{30}
$$

where $F_{n-1,n}$ has the F-distribution with $n-1$ and n degrees of freedom. However, at that time Fisher had not yet derived the F-distribution.

In this paper Fisher also responded to the criticism of Pearson and his associates, asserting that he did not use Bayes' theorem or "equal distribution of ignorance." He replaced the term "most likely value" by "optimum value" and clarified that what he meant was simply the value of ρ that maximizes the density given the observed r. Taking an example from the "Cooperative Study" of an observed $r = 0.6$ he obtained the (maximum likelihood) estimate of 0.5918, but in the cooperative study the estimate of 0.462 was obtained on the basis of a prior distribution of ρ (normal with mean 0.46 and standard deviation 0.02). Fisher wrote that "they enforce [the prior] with such rigour that a sample which expresses the value 0.6000 has its message so modified in transmission that it is finally reported as 0.462 at a distance of 0.002 only above that value which is assumed a priori to be most probable." In passing Fisher objects to a uniform prior on the grounds that the result depends on what function of ρ is assigned the uniform prior.

Fisher (1922) derived "The exact distribution of χ^2 when σ is determined from the data," which is equivalent to the F-distribution. In the regression problem he considered there are several observations of the dependent variable for each value of the independent variable. The χ^2 of the section title refers to the sum of squares of deviations of the cell means about their regression on the independent variables divided by an estimate of the variance obtained from the deviations of the dependent variables about their cell means. Except for a constant multiplier this ratio is an F-statistic. (The notation F was suggested later by Snedecor in honor of its inventor.) In this paper Fisher also finds that the ratio of an estimated regression coefficient to its estimated standard error has Student's distribution.

Fisher (1924a) showed by means of an orthogonal transformation that the partial correlation coefficient $r_{xy\cdot z}$ has the distribution of r_{xy} with one less degree of freedom.

About this time Fisher was impressed that the χ^2, t and F-distributions kept coming up in various contexts; one of several papers on this subject was Fisher (1925). In Section 3 "Proof of the exactitude of Student's distribution for normal samples,"

Fisher again derives the joint density of $\bar{x}$ and s_x as the product of the normal density of $\bar{x}$ and the chi-density of s_x. He remarks that "the two distributions must be wholly independent" and "from which the distribution of t has already been derived." The main difference between this paper and the 1915 paper is that the earlier paper gave the joint density of two means, two standard deviations and the correlation coefficient, while this paper gives the joint density of only one mean and one standard deviation.

There is a more interesting part of this paper. "It is perhaps worthwhile to give, at length, an algebraic method of proof, since analogous cases have hitherto been demonstrated only geometrically, by means of a construction in Euclidean hyperspace, and the validity of such methods of proof may not be universally admitted." Suppose $y_1, \ldots, y_n$ are independently normally distributed with mean 0 and standard deviation 1. Then $z_1, \ldots, z_n$ resulting from an orthogonal transformation are also independently normally distributed with mean 0 and standard deviation 1. Now suppose that the first h of these new variables have been defined; then, Fisher argues, one can find $n - h$ linear combinations of the y_i's so that the entire set of new variables constitutes an orthogonal transformation of the original y_i's. Because of orthogonality, the difference between the sum of squares of the y_i's and the sum of squares of the first h new variables is equal to the sum of squares of the last $n - h$ new variables and hence the difference and the last sum of squares (of new variables) are independent. In the standard simple linear regression problem $[y_i - \alpha - \beta(x_i - \bar{x})]/\sigma$ are normally and independently distributed with mean 0 and standard deviation 1. Here $\bar{y}$ and the sample regression coefficient properly normalized are two orthogonal linear combinations of the original y_i's. Because the estimates are least squares, Fisher writes, the sum of squared deviations around the sample regression is equal to the sum of squared deviations around the population regression minus a normalized square of $\bar{y}$ and minus the normalized square of the sample regression coefficient. Fisher concludes this discussion with pointing out that the fact that the ratio of the sample regression coefficient to its estimated standard error has Student's distribution does not depend on the distribution of the x_i's.

About a decade later in a paper presented to a meeting of the Royal Statistical Society, John Wishart (1934) referred to a published paper of Irwin (1934) and an unpublished paper of Wilks proving that various sums of squares in the analysis of variance were independent. In the ensuing discussion of the paper Fisher expressed his anger at what he thought was Wishart ignoring the 1925 paper. Fisher claimed that the proof of independence outlined in the previous paragraph was all that was needed. As a result of the interchange, Wishart revised his paper; in the printed version he refers to Fisher (1925), mentioning Irwin and Wilks only in a footnote.

It happens that I heard about this matter from another point of view. Wilks, having received his Ph.D. at the University of Iowa under H. L. Rietz in 1931, spent 1931–1932 at Columbia University with Harold Hotelling, the first half of the next academic year at University College, London, where Karl Pearson directed the Department of Statistics, and the second half of 1932–1933 at Cambridge University with Wishart, M. S. Bartlett and W. G. Cochran. When I was a graduate student during the early 1940s, Wilks told me that in the early 1930s he had thought that the independence of sums of squares in the analysis of variance had not been justified with sufficient rigor and generality. He developed a general proof and submitted it to the Royal Statistical Society. In the second of two letters to Wilks questioning the value of this paper Fisher wrote "I should judge that I should be doing both the Society and yourself an ill service if I were to recommend your paper, as it now stands, for publication. If you are in any doubt it might be wise to ask Professor Hotelling or some other American friend of standing" (Bennett, 1990, pages 303–304). Wilks subsequently received notice that his paper was rejected.

Shortly thereafter, Irwin (1934) published a paper giving a general proof of the independence of quadratic forms [although earlier Irwin (1931) had referred to Fisher (1925) "where a general proof is given of a theorem of which this and all similar theorems are particular cases"]. Since Irwin was an associate of Fisher at Rothamstead and Irwin's paper was published with Fisher's approval in the journal to which Wilks had submitted his paper, Wilks felt his contribution was treated unfairly and was understandably upset. He made no further attempt to publish his paper.

As we have seen in the earlier discussion, Fisher often omitted steps in proofs. Fisher's theorem in *Metron* was not a justification for all of the designs in the analysis of variance. There was a need for a general proof, which Wilks gave. Cochran at the suggestion of Wishart studied the problem of independence of quadratic forms and published his famous theorem in the *Proceedings of the Cambridge Philosophical Society* (Cochran, 1934).

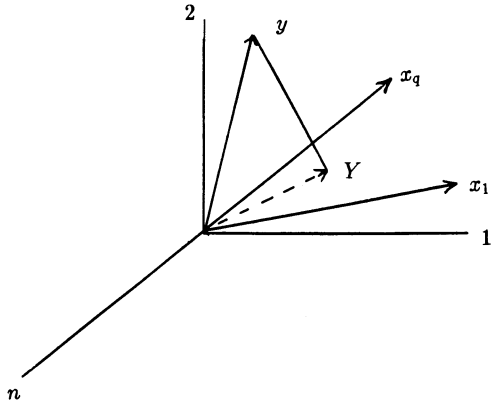

FIG. 3.

Fisher (1924b) applied his geometric approach to the distribution of the multiple correlation coefficient when the population correlation coefficient is zero. For simplicity I have illustrated the problem in Figure 3 without taking account of the effect of subtracting the means of the variables. The vector $\mathbf{y}$ represents the dependent variates and $\mathbf{x}_1, \ldots, \mathbf{x}_q$ the independent variates, and $\mathbf{Y}$ is the regression of $\mathbf{y}$ on $\mathbf{x}_1, \ldots, \mathbf{x}_q$. Fisher writes as follows:

> The multiple correlation, R, is then the cosine of the angle which the line OP makes with the space of q dimensions. If the variate y is unrelated to the variates $x_1, \ldots, x_q$, the line OP may be regarded as drawn at random through O, in the space of n dimensions; using this fact it may be shown without difficulty that the chance that R^2 falls into the elementary range $d(R^2)$ is

$$df = \frac{((n-3)/2)!}{((n-q-3)/2)!((q-2)/2)!} \cdot (R^2)^{(q-2)/2}(1-R^2)^{(n-q-3)/2}d(R^2). \tag{31}$$

This is the complete derivation! (The argument really should be in the $n-1$ dimensions orthogonal to the equiangular line; Fisher did not take this into account in his words, but he did in the result.)

The multiple correlation R is the correlation between the dependent vector $\mathbf{y}$ and its projection on the q-dimensional x-space. Let $\|\mathbf{y}\|$ denote the norm (Euclidean length) of $\mathbf{y}$. If the variance of a component of $\mathbf{y}$ is σ^2, then $\|\mathbf{Y}\|^2/\sigma^2 = \cos^2\phi\|\mathbf{y}\|^2/\sigma^2$ and $\|\mathbf{y}-\mathbf{Y}\|^2/\sigma^2 = \sin^2\phi\|\mathbf{y}\|^2/\sigma^2$ are distributed as independent χ^2 variables with q and $n-q-1$ degrees of freedom, respectively. The joint density of the two χ^2's can be transformed to the joint density of $R^2 = \cos^2\phi$ and $\|\mathbf{y}\|^2/\sigma^2$; then integration with respect to the latter variable leads to (31) as the probability element [density $d(R^2)$] of R^2. Equivalently, this method leads to

$$\frac{n-q-1}{q}\frac{R^2}{1-R^2} =_{\mathrm{d}} F_{q,n-q-1}. \tag{32}$$

To put this in geometric terms similar to that of the Pearson correlation coefficient consider the set

$$\sqrt{n}s_y^* \le \|\mathbf{y}\| \le \sqrt{n}(s_y^* + dy^*), \tag{33}$$

$$R^{*2} \le \cos^2\phi \le R^{*2} + dR^{*2}. \tag{34}$$

This set can be approximated by the set defined by each of $\|\mathbf{Y}\|$ and $\|\mathbf{y}-\mathbf{Y}\|$ satisfying a pair of inequalities. Each pair defines a spherical cylindrical shell in q and $n-q-1$ dimensions, respectively.

Fisher's approach was to consider $\mathbf{y}/\|\mathbf{y}\|$ $(= OP)$, which has the uniform distribution on the unit sphere when $\mathbf{y}$ has the distribution $N(\mathbf{0}, \sigma^2\mathbf{I})$, and carried out the corresponding analysis, projecting $\mathbf{Y}$ and $\mathbf{y}-\mathbf{Y}$ on the unit sphere.

Distributions of the multiple correlation coefficient when the dependent variable y and the independent variables $x_1, \ldots, x_q$ are not independent were obtained in Fisher (1928). If the joint distribution of the dependent and independent variables is normal, Fisher argued that the multiple correlation is invariant with respect to a linear transformation of the independent variables; that is, R and $\bar{R}$, the population multiple correlation coefficient, are invariant. Hence, for the sake of obtaining the distribution of R we can suppose that the parent normal distribution has covariance matrix

$$\begin{pmatrix} \sigma_y^2 & \sigma_y\sigma_1\bar{R} & 0 & \cdots & 0 \\ \sigma_y\sigma_1\bar{R} & \sigma_1^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & \sigma_q^2 \end{pmatrix}. \tag{35}$$

Let $\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_q$ be the vectors of observations on $y, x_1, \ldots, x_q$ in a sample, in which $\bar{y}, \bar{x}_1, \ldots, \bar{x}_q$ are the means, $s_y^2, s_1^2, \ldots, s_q^2$ are the variances and r_{y1} is the correlation between y and x_1. The density of this sample can be written as the product of the density of $\bar{y}, \bar{x}_1, \ldots, \bar{x}_q$ and

$$\text{const}\exp\left\{-\frac{n}{2(1-\bar{R}^2)}\left[\frac{s_y^2}{\sigma_y^2} - 2\bar{R}r_{y1}\frac{s_y s_1}{\sigma_y\sigma_1} + \frac{s_1^2}{\sigma_1^2}\right]\right\} \cdot \prod_{j=2}^{q}\exp\left\{-\frac{1}{2}\frac{s_j^2}{\sigma_j^2}\right\}. \tag{36}$$

The first exponential in (36) is similar to the second exponential on the right-hand side of (17). The sufficient statistics are $\bar{y}$, $\bar{x}_1, \ldots, \bar{x}_q$, s_y^2, $s_1^2, \ldots, s_q^2$ and r_{1y}. Fisher writes that it is evident that the density of R in this case is the density in the null case times

$$\frac{[\frac{1}{2}(n-2)]!}{[\frac{1}{2}(n-3)]!\sqrt{\pi}}(1-\bar{R}^2)^{(n-1)/2} \cdot \int_{-\infty}^{\infty} \frac{dz}{(\cosh z - \bar{R}r_{y1})^{n-1}}. \tag{37}$$

Now we relate r_{y1} to R by considering the correlation of $\mathbf{y}$ and $\mathbf{Y} = b_1\mathbf{x}_1 + \cdots + b_q\mathbf{x}_q$, the regression of $\mathbf{y}$ on $\mathbf{x}_1, \ldots, \mathbf{x}_q$, which is $R = r_{yY}$, the correlation between $\mathbf{y}$ and $\mathbf{Y}$. The sample partial correlation between $\mathbf{y}$ and $\mathbf{x}_1$, given $\mathbf{Y}$, is

$$r_{y1\cdot Y} = \frac{r_{y1} - r_{yY}r_{1Y}}{\sqrt{1-r_{yY}^2}\sqrt{1-r_{1Y}^2}} = 0 \tag{38}$$

(because $\mathbf{y} - \mathbf{Y}$ is orthogonal to the space spanned by $\mathbf{x}_1, \ldots, \mathbf{x}_q$); here r_{1Y} is the sample correlation between $\mathbf{x}_1$ and $\mathbf{Y}$. Hence, $r_{y1} = r_{yY}r_{1Y}$, and r_{y1} in (37) can be replaced by $r_{yY}r_{1Y}$. Fisher refers to r_{1Y} as $\cos\psi$ and $\sqrt{1-r_{1Y}^2}$ as $\sin\psi$ because r_{1Y} is the cosine of the angle between $\mathbf{x}_1$ and $\mathbf{Y}$ (in q dimensions). He writes that given the value of R "the frequency with which ψ falls in the range $d\psi$ is evidently

$$\frac{[\frac{1}{2}(q-2)]!}{[\frac{1}{2}(q-3)]!\sqrt{\pi}}\sin^{q-2}\psi\, d\psi." \tag{39}$$

The argument (which Fisher does not give) is that leading to the volume element in the case of the sample correlation coefficient; the details will be omitted here. This leads to the density for R of

$$\text{const}(1-\bar{R}^2)^{(n-1)/2}(R^2)^{q/2-1}(1-R^2)^{(n-q-3)/2} \cdot \int_{-\infty}^{\infty}\int_0^{\pi} \frac{\sin^{q-2}\psi\, d\psi\, dz}{(\cosh z - \bar{R}R\cos\psi)^{n-1}}. \tag{40}$$

Expansion of the integrand in a series and integration term by term yields the density of the form of

$$\frac{\Gamma[\frac{1}{2}(n-1)]}{\Gamma(\frac{1}{2}q)\Gamma[\frac{1}{2}(n-q-1)]} \cdot (1-\bar{R}^2)^{(n-1)/2}(R^2)^{(q-1)/2}(1-R^2)^{(n-q-3)/2} \cdot F\left[\frac{1}{2}(n-1), \frac{1}{2}(n-1), \frac{1}{2}q, \bar{R}^2R^2\right], \tag{41}$$

where $F(a, b, c; x)$ is a hypergeometic distribution. (See Anderson, 1984, page 113, for example.) Fisher termed (41) the A distribution. A generalization of the derivative expression (25) was also given.

Let $nR^2 = B^2$ and $n\bar{R}^2 = \beta^2$. Then the limiting density of B^2 as $n \to \infty$ is found from (41), termed by Fisher the B distribution and now known as the noncentral χ^2 distribution with noncentrality parameter β^2. When the x's are fixed, the distribution of (32) is the noncentral F.

4. WORK OF OTHERS

A fundamental contribution to multivariate analysis was made by Wishart (1928) in deriving the joint distribution of the components of the sample covariance matrix; this distribution is a generalization of the distribution of two sample standard deviations and the correlation coefficient obtained by Fisher in 1915 (involving a change of variables). Wishart, who wrote this paper during his four years at Rothamstead, used the geometrical approach. Let $\mathbf{x}_1, \ldots, \mathbf{x}_n$ be independent vector observations from $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, and define

$$(a_{ij}) = \mathbf{A} = \sum_{i=1}^{n}(\mathbf{x}_i - \bar{\mathbf{x}})(\mathbf{x}_i - \bar{\mathbf{x}})'. \tag{42}$$

Then $a_{ij} = (n-1)r_{ij}s_is_j$, $i, j = 1, \ldots, p$, where s_i is the sample standard deviation of the ith component of the observation vectors and r_{ij} is the sample correlation between the ith and jth components. Wishart first derived the density of the $\bar{x}_i$'s, s_i's and r_{ij}'s by consideration of the angles between the vectors with coordinates $x_{i1} - \bar{x}_i, \ldots, x_{in} - \bar{x}_i$, $i = 1, 2, 3$. His exposition is easier to understand than that of Fisher (1915) because he gives more details. Then Wishart goes on to the general case in which the density of $a_{11}, \ldots, a_{pp}, a_{12}, \ldots, a_{p-1,p}$ is

$$\frac{|\mathbf{A}|^{(n-p-2)/2}\exp\left(-\frac{1}{2}\operatorname{tr}\boldsymbol{\Sigma}^{-1}\mathbf{A}\right)}{2^{pn/2}\pi^{p(p-1)/4}|\boldsymbol{\Sigma}|^{(N-1)/2}\prod_{i=1}^{p}\Gamma[(N-i)/2]}. \tag{43}$$

It is clear that Fisher influenced this research; Wishart writes "My thanks are due to Dr. R. A. Fisher in whose laboratory this paper was written, and without whose critical help it would have been difficult to generalise the geometrical methods employed by him."

Subsequently, the density (43) has been derived in other ways. See Wishart (1948) for a review of some of these. Many of these methods amount to analytic copies of Wishart's geometry (including Section 7.2 of Anderson, 1984).

Harold Hotelling (1931), who had been with Fisher at Rothamstead in 1929–1930, developed the distribution of the generalization of Student's t,

$$T^2 = n\bar{\mathbf{x}}'\mathbf{S}^{-1}\bar{\mathbf{x}}, \tag{44}$$

where

$$\mathbf{S} = \frac{1}{n-1}\mathbf{A}, \tag{45}$$

by means of a geometrical argument. In the n-dimensional space he makes an orthogonal transformation so that $\sqrt{n}\bar{\mathbf{x}} = \mathbf{y}$ and $\mathbf{A} = \sum_{\alpha=1}^{n-1} \mathbf{z}_\alpha \mathbf{z}_\alpha'$, where $\mathbf{z}_1, \ldots, \mathbf{z}_{n-1}, \mathbf{y}$ are independently normally distributed with covariance matrix $\mathbf{\Sigma}$ and means $\mathbf{0}$ except for $\mathbf{y}$ having mean $\sqrt{n}\,\boldsymbol{\mu}$. Because T is invariant with respect to transformations $\mathbf{x} = \mathbf{B}\mathbf{w}$, the distribution can be derived with $\mathbf{\Sigma} = \mathbf{I}$. Let $(z_{i1}, \ldots, z_{i,n-1}, y_i)'$ be a vector $\mathbf{v}_i$ in the n-dimensional space, $i = 1, \ldots, p$. Then $T^2/(n-1)$ is the cotangent squared of the angle between the nth coordinate axis and the p-dimensional hyperplane spanned by the p vectors $\mathbf{v}_1, \ldots, \mathbf{v}_p$. Because $\mathbf{\Sigma}$ is taken to be $\mathbf{I}$, the p vectors are independently distributed and if $\boldsymbol{\mu} = \mathbf{0}$, each has a spherically symmetric distribution. The distribution of the angle between a fixed vector (in particular, the nth coordinate axis) and the p-dimensional hyperplane is the same as the distribution of an angle between a fixed hyperplane and a random vector with a spherically symmetric distribution. Thus

$$\frac{T^2}{n-1} = \frac{R^2}{1-R^2}, \tag{46}$$

where R refers to the multiple correlation for $q = p$. The distribution is that given by Fisher (1924b). However, Hotelling carries out a transformation to n polar coordinates, which includes the angle between the random vector and the fixed hyperplane, and integrates out the $n-1$ irrelevant angles to obtain the density. Hotelling, who was not given to writing tersely (in contrast to Fisher), remarks in a footnote "The omitted steps in Fisher's argument may be supplied with the help of generalized polar coordinates as in the text." Incidentally, Hotelling did not notice that his argument can be carried out for the general case of $\boldsymbol{\mu} \neq \mathbf{0}$ to obtain the distribution C of Fisher (1928) though he cites the last. (A transformation to polar coordinates is sketched in Problems 2, 3, and 4 of Chapter 5 in Anderson, 1984.)

5. DISCRIMINANT ANALYSIS

Fisher began a new direction in multivariate analysis with his 1936 paper in the *Annals of Eugenics*, the journal whose editorship he inherited from Karl Pearson when he became the Galton Professor of Eugenics at University College in 1933. Consider two samples, $\mathbf{x}_1^{(1)}, \ldots, \mathbf{x}_{n_1}^{(1)}$ and $\mathbf{x}_1^{(2)}, \ldots, \mathbf{x}_{n_2}^{(2)}$ from multivariate distributions with common covariance matrix. Let the difference in the sample means be

$$\mathbf{d} = \bar{\mathbf{x}}^{(1)} - \bar{\mathbf{x}}^{(2)} \tag{47}$$

and let the "within sum of squares" be

$$\mathbf{A} = \sum_{k=1}^{2} \sum_{\alpha=1}^{n_k} (\mathbf{x}_\alpha^{(k)} - \bar{\mathbf{x}}^{(k)})(\mathbf{x}_\alpha^{(k)} - \bar{\mathbf{x}}^{(k)})', \tag{48}$$

leading to the sample covariance matrix

$$\mathbf{S} = \frac{1}{n_1 + n_2 - 2}\mathbf{A}. \tag{49}$$

Consider an arbitrary linear combination of the observed variables

$$X = \mathbf{b}'\mathbf{x}. \tag{50}$$

Then the difference in the sample means of this linear combination is

$$\bar{X}^{(1)} - \bar{X}^{(2)} = \mathbf{b}'\mathbf{d} \tag{51}$$

and the sample variance of X is

$$\mathrm{Var}(X) = \mathbf{b}'\mathbf{S}\mathbf{b}. \tag{52}$$

Fisher asked what linear combination has the greatest difference of sample means relative to its sample standard deviation; this linear combination discriminates best between the two samples. The algebraic problem is to maximize

$$\frac{(\bar{X}^{(1)} - \bar{X}^{(2)})^2}{\mathrm{Var}(X)} = \frac{(\mathbf{b}'\mathbf{d})^2}{\mathbf{b}'\mathbf{S}\mathbf{b}} \tag{53}$$

with respect to $\mathbf{b}$. A solution is

$$\mathbf{b} = \mathbf{S}^{-1}\mathbf{d}. \tag{54}$$

The linear function $\mathbf{b}'\mathbf{x}$ is the linear discriminant function. For this $\mathbf{b}$ the difference in sample means is

$$\bar{X}^{(1)} - \bar{X}^{(2)} = \mathbf{x}'\mathbf{d} = \mathbf{d}'\mathbf{S}^{-1}\mathbf{d}, \tag{55}$$

and hence the maximum of (53) is

$$\frac{(\mathbf{b}'\mathbf{d})^2}{\mathbf{b}'\mathbf{S}\mathbf{b}} = \mathbf{d}'\mathbf{S}^{-1}\mathbf{d}. \tag{56}$$

The linear discriminant function $\mathbf{b}'\mathbf{x}$ can be used to classify a future observation $\mathbf{x}$ as coming from the first population if $\mathbf{b}'\mathbf{x} > \frac{1}{2}(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)})$ and from the second if $\mathbf{b}'\mathbf{x} < \frac{1}{2}(\bar{\mathbf{x}}^{(1)} + \bar{\mathbf{x}}^{(2)})$. This procedure can be derived from decision theory; see Anderson (1951b), for example. Fisher applied this technique to the famous iris data that he had obtained from botanist Edgar Anderson. He gave an analysis of variance table of the linear combination. The total number of degrees of freedom is $n_1 + n_2 - 1$. The number of degrees of freedom within species is the total number minus 1 for the mean and minus $p-1$ because

there are $p-1$ ratios of coefficients determined in $\mathbf{b}$; that is, n_1+n_2-p-1. That leaves p degrees of freedom for between species.

Another approach is to set up the dummy variables

$$y_\alpha^{(1)} = \frac{n_2}{n_1+n_2}, \qquad y_\alpha^{(2)} = -\frac{n_1}{n_1+n_2}. \tag{57}$$

The assigned values are chosen so that

$$\sum_{k,\alpha} y_\alpha^{(k)} = 0, \qquad \sum_{k,\alpha} (y_\alpha^{(k)})^2 = \frac{n_1 n_2}{n_1+n_2}. \tag{58}$$

Then regress the $y_\alpha^{(k)}$ on $\mathbf{x}_\alpha^{(k)}$, $\alpha = 1,\dots,n_k$, $k = 1,2$. The regression function is proportional to the discriminant function. Fisher gave an analysis of variance table for the dummy variable; the entries are proportional to those of the previous table.

In his next paper, Fisher (1938) reviewed the work of other contributors and related it to his discriminant function. He showed that

$$\frac{(\mathbf{b}'\mathbf{d})^2}{\mathbf{b}'\mathbf{S}\mathbf{b}} = \frac{p}{n_1+n_2-p-1} T^2, \tag{59}$$

where T^2 is Hotelling's T^2 for the two-sample problem. In this case T^2 has the distribution of $(n_1+n_2-2)F_{p,\,n_1+n_1-p-1}$. Mahalanobis (1930) had defined

$$\Delta^2 = \frac{1}{p}(\boldsymbol{\mu}^{(1)}-\boldsymbol{\mu}^{(2)})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}^{(1)}-\boldsymbol{\mu}^{(2)}) \tag{60}$$

as a measure of the difference between two populations; $p\Delta^2$ is known as the "Mahalanobis squared distance." Fisher visited the Indian Statistical Institute in Calcutta in the winter of 1937–1938 and learned of the research of Indian statisticians. The D^2-statistic is (60) with $\boldsymbol{\mu}^{(1)}$ and $\boldsymbol{\mu}^{(2)}$ replaced by $\bar{\mathbf{x}}^{(1)}$ and $\bar{\mathbf{x}}^{(2)}$, respectively, and $n_1^{-1}+n_2^{-1}$ subtracted; the studentized D^2-statistic is

$$D^2 = \frac{1}{p}\mathbf{d}'\mathbf{S}^{-1}\mathbf{d} = \text{const}\, T^2. \tag{61}$$

Bose (1936) had found the distribution of D^2; Fisher pointed out that it was equivalent to his B distribution (Fisher, 1928). He also wrote "In a very brilliant research R. C. Bose & S. N. Roy have demonstrated that the distribution of D^2, so defined, takes a form derivable from distribution (C) of my 1928 paper." This work was published as Bose and Roy (1938).

Bose told me that he and Roy wrote up notes on Fisher's lectures during that visit. Finding many explanations lacking detail they filled in the gaps and sent the draft to Fisher for his approval. When it came back, Fisher had eliminated all of their additions.

Now comes "V. Extension of discriminant analysis." Consider q samples with means $\bar{\mathbf{x}}^{(1)},\dots,\bar{\mathbf{x}}^{(q)}$. Define the "within sum of squares"

$$\mathbf{A} = \sum_{k=1}^{q}\sum_{\alpha=1}^{n_k}(\mathbf{x}_\alpha^{(k)}-\bar{\mathbf{x}}^{(k)})(\mathbf{x}_\alpha^{(k)}-\bar{\mathbf{x}}^{(k)})'. \tag{62}$$

Let

$$\bar{\mathbf{X}} = (\bar{\mathbf{x}}^{(1)},\dots,\bar{\mathbf{x}}^{(q)}), \qquad \mathbf{S} = \frac{1}{n-q}\mathbf{A}, \tag{63}$$

where $n = \sum_{k=1}^{q} n_k$. For an arbitrary q-component vector $\boldsymbol{\lambda}$ define

$$\mathbf{d} = \bar{\mathbf{X}}\boldsymbol{\lambda} = \sum_{k=1}^{q}\bar{\mathbf{x}}^{(k)}\lambda_k. \tag{64}$$

Thus $\mathbf{d}$ is an arbitrary linear combination of the q sample means. In the 1936 paper, $q=2$, $\mathbf{d} = \bar{\mathbf{x}}^{(1)}-\bar{\mathbf{x}}^{(2)}$ and $\boldsymbol{\lambda} = (1,-1)'$. In general,

$$\max_{\mathbf{b}} \frac{(\mathbf{b}'\mathbf{d})^2}{\mathbf{b}'\mathbf{S}\mathbf{b}} \tag{65}$$

is attained by

$$\mathbf{b} = \mathbf{S}^{-1}\mathbf{d} = \mathbf{S}^{-1}\bar{\mathbf{X}}\boldsymbol{\lambda}, \tag{66}$$

and this maximum is

$$\frac{(\mathbf{b}'\mathbf{d})^2}{\mathbf{b}'\mathbf{S}\mathbf{b}} = \mathbf{d}'\mathbf{S}^{-1}\mathbf{d} = \boldsymbol{\lambda}'\bar{\mathbf{X}}'\mathbf{S}^{-1}\bar{\mathbf{X}}\boldsymbol{\lambda}. \tag{67}$$

Let

$$\mathbf{N} = \begin{pmatrix} n_1 & 0 & \cdots & 0 \\ 0 & n_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & n_q \end{pmatrix}, \tag{68}$$

$$\mathbf{e}' = (1,1,\dots,1). \tag{69}$$

The vector $\boldsymbol{\lambda}$ is normalized by

$$\mathbf{e}'\boldsymbol{\lambda} = 0, \tag{70}$$

$$\boldsymbol{\lambda}'\mathbf{N}^{-1}\boldsymbol{\lambda} = 1. \tag{71}$$

Then $\boldsymbol{\lambda}'\mathbf{x}$ represents a contrast. We can consider a vector $\boldsymbol{\lambda}$ that maximizes (67) subject to (70) and (71). This paper was one of the papers reproduced (photographically) in Fisher's *Contributions to Mathematical Statistics* (1950), with comments and editing by the author of the papers. He had crossed out the last three pages of this paper. No wonder I had trouble understanding this paper as a graduate student; it was wrong!

In the material that was crossed out Fisher indicated in the case of $q=3$ and $p=2$ that the maximum of (67) is the largest root of

$$|\bar{\mathbf{X}}'\mathbf{S}^{-1}\bar{\mathbf{X}} - \theta\mathbf{N}^{-1}| = 0. \tag{72}$$

Some algebraic manipulation shows that the roots of (72) are the roots of

$$|\mathbf{H} - \theta\mathbf{S}| = 0, \tag{73}$$

where

$$\mathbf{H} = \sum_{i=1}^{q} n_i(\bar{\mathbf{x}}^{(i)} - \bar{\mathbf{x}})(\bar{\mathbf{x}}^{(i)} - \bar{\mathbf{x}})' \tag{74}$$

is the "between sum of squares."

Fisher raised the question whether the population means $\boldsymbol{\mu}^{(1)}, \ldots, \boldsymbol{\mu}^{(q)}$ lie on a line and suggested as a criterion

$$\operatorname{tr}\mathbf{H}\mathbf{S}^{-1} - \theta_1 = \sum_{i=1}^{p} \theta_i - \theta_1 = \sum_{i=2}^{p} \theta_i, \tag{75}$$

where $\theta_1 \geq \theta_2 \geq \cdots \geq \theta_p$ are the ordered roots of (73). The hypothesis of collinearity is rejected if (75) is too large. Hsu (1941) stated the problem of rank more generally and gave a large-sample distribution of the criterion. In my dissertation (Anderson, 1945) I showed that the likelihood ratio criterion for collinearity is $\prod_{i=2}^{q}(1+\theta_i)^{-n/2}$; see Anderson (1951a).

A natural question follows from the suggestion of using the roots of (73) for testing collinearity, coplanarity and so forth, namely, what is the distribution of the roots? If

$$\boldsymbol{\mu}^{(1)} = \cdots = \boldsymbol{\mu}^{(q)}, \tag{76}$$

$\mathbf{H}$ and $\mathbf{A}$ have Wishart distributions with $q-1$ and $n-q$ degrees of freedom, respectively, and are independent. The density of the roots in this case was found independently and almost simultaneously by at least five statisticians. Fisher (1939) derived the density for $p = 2$. "A more formal demonstration" in the general case was given by Hsu (1939), apparently at the suggestion of Fisher. Girshick (1939) had hoped to make his work part of his doctoral dissertation, but could not do so when the results were already published. Roy (1939) published in *Sankyhā* at about the same time; because of the war, that issue arrived in the United States several years later. Mood had been working on the problem as his Ph.D. research. Wilks told me "You should have seen Alex's face drop when he saw that issue of the *Annals of Eugenics*." Mood gave up that area and turned in another direction. Later as editor of the *Annals of Mathematical Statistics*, I persuaded Mood to publish a version since his proof was different (Mood, 1951).

To describe this result, we let $\phi_i = \theta_i/(n-q)$, $i = 1, \ldots, p$, which are the roots of $|\boldsymbol{H} - \phi\boldsymbol{A}| = 0$. If $q > p$, the density of $\phi_1, \ldots, \phi_p$ is

$$C\prod_{i=1}^{p}\phi_i^{(q-p-2)/2}\prod_{i=1}^{p}(1+\phi_i)^{-(n-1)/2}\prod_{i<j}(\phi_i - \phi_j) \tag{77}$$

for $\phi_1 \geq \phi_2 \geq \cdots \geq \phi_p \geq 0$ and 0 otherwise; here

$$C = \pi^{p/2}\prod_{i=1}^{p}\Gamma\left[\tfrac{1}{2}(n-i)\right]\left\{\Gamma\left[\tfrac{1}{2}(p+1-i)\right]\right. \cdot \left.\Gamma\left[\tfrac{1}{2}(q-i)\right]\Gamma\left[\tfrac{1}{2}(n-q+1-i)\right]\right\}^{-1}. \tag{78}$$

The hard part of the derivation is to find the Jacobian of the transformation from $\mathbf{H}$ and $\mathbf{A}$ to the roots and p^2 other variables and to obtain the constant C.

In the 1938 paper Fisher also considered testing the null hypothesis that the discriminant function for two populations is a given linear function $\boldsymbol{\delta}'\mathbf{x}$. His approach was to ask whether the two population means and a fictitious third were collinear and let the size of the third sample approach infinity. By an analysis of variance argument he reduced the test statistic to a T^2. This analysis was incorrect, however, and was crossed out in the reprint and replaced by a sketch of the argument (Fisher, 1940), which is proper. The idea is that if the specified linear function is indeed the discriminant function, there is no more information in the samples to distinguish them. Consider the conditional distribution of the variables conditioned on the value of the specified linear function. If the hypothetical discriminant function is the true one, then the vector of intercepts in the conditional distribution is the $\mathbf{0}$ vector. Hence, a T^2-test can be constructed; a multiple of it has the F-distribution with $p-1$ and n_1+n_2-p-1 degrees of freedom under the null hypothesis. Bartlett (1939) had already published the correct solution.

The last section of Fisher's 1940 paper is an application of this algebra to contingency tables. The analysis amounts to finding canonical correlations and variates among two sets of dummy variables. This was a forerunner of "correspondence analysis." (See Benzécri, 1973.)

Fisher's last paper in multivariate analysis was in 1962 and was on the simultaneous distribution of correlation coefficients. Let $a_{ij} = (n-1)r_{ij}s_is_j$, $i, j = 1, \ldots, p$. From (43) one finds the density of $s_1, \ldots, s_p, r_{12}, \ldots, r_{p-1,\,p}$ as

$$\text{const}\prod_{i=1}^{p} s_i^{n-2}\,|r_{ij}|^{(n-p-2)/2} \cdot \exp\left\{-\frac{n-1}{2}\sum_{i,\,j=1}^{p}\frac{\rho^{ij}s_is_jr_{ij}}{\sigma_i\sigma_j}\right\}, \tag{79}$$

where ρ^{ij} is an element of the inverse of the correlation matrix (ρ_{gh}). At this point Fisher makes the

change of variables

$$(80) \qquad u_i = \frac{s_i}{\sigma_i}\sqrt{(N-1)\rho^{ii}}, \qquad \gamma_{ij} = -\frac{\rho^{ij} r_{ij}}{\sqrt{\rho^{ii}\rho^{jj}}}.$$

Then integration of the density with respect to u_i from 0 to ∞, $i = 1, \ldots, p$, gives the density of the r_{ij}'s. However, Fisher does not suggest a way of carrying this out.

ACKNOWLEDGMENTS

The author is indebted to Yasuo Amemiya, William Kruskal and Ingram Olkin for helpful suggestions for the exposition in this paper.

REFERENCES

ANDERSON, T. W. (1945). The noncentral Wishart distribution and its application to problems in multivariate statistics. Dissertation, Princeton Univ.

ANDERSON, T. W. (1984). *An Introduction to Multivariate Statistical Analysis*. Wiley, New York.

ANDERSON, T. W. (1951a). Estimating linear restrictions on regression coefficients for multivariate distributions. *Ann. Math. Statist.* **22** 327–351.

ANDERSON, T. W. (1951b). Classification by multivariate analysis. *Psychometrika* **16** 31–50.

BARTLETT, M. S. (1939). A note on tests of significance in multivariate analysis. *Proceedings of the Cambridge Philosophical Society* **35** 180–185.

BENNETT, J. H. (1990). *Statistical Inference and Analysis: Selected Correspondence of R. A. Fisher*. Clarendon Press, Oxford.

BENZÉCRI, J.-P. (1973). *L'Analyse des Données*. Dunod, Paris.

BOSE, R. C. (1936). On the exact distribution of D^2-statistic. *Sankhyā* **2** 143–154.

BOSE, R. C. and ROY, S. N. (1938). The distribution of the studentized D^2-statistic. *Sankhyā* **4** 19–38.

BOX, J. F. (1978). *R. A. Fisher: The Life of a Scientist*. Wiley, New York.

COCHRAN, W. G. (1934). The distribution of quadratic forms in a normal system, with applications to the analysis of variance. *Proceedings of the Cambridge Philosophical Society* **30** 178–191.

FISHER, R. A. (1912). On an absolute criterion for fitting frequency curves. *Messenger of Mathematics* **41** 155–160.

FISHER, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika* **10** 507–521.

FISHER, R. A. (1920). A mathematical examination of the methods of determining the accuracy of an observation by the mean error, and the mean square error. *Monthly Notices of the Royal Astronomical Society* **80** 758–770.

FISHER, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron* **1** 3–32.

FISHER, R. A. (1922). The goodness of fit of regression formulae, and the distribution of regression coefficients. *J. Roy. Statist. Soc.* **85** 597–612.

FISHER, R. A. (1924a). The distribution of the partial correlation coefficient. *Metron* **3** 329–332.

FISHER, R. A. (1924b). The influence of rainfall on the yield of wheat at Rothamsted. *Philos. Trans. Roy. Soc. London Ser. B* **213** 89–142.

FISHER, R. A. (1925). Application of "Student's" distribution. *Metron* **5** 90–104.

FISHER, R. A. (1928). The general sampling distribution of the multiple correlation coefficient. *Proc. Roy. Soc. London Ser. A* **121** 654–673.

FISHER, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7** 179–188.

FISHER, R. A. (1938). The statistical utilization of multiple measurements. *Annals of Eugenics* **8** 376–386.

FISHER, R. A. (1939). The sampling distribution of some statistics obtained from nonlinear equations. *Annals of Eugenics* **9** 238–249.

FISHER, R. A. (1940). The precision of discriminant functions. *Annals of Eugenics* **10** 422-429.

FISHER, R. A. (1950). *Contributions to Mathematical Statistics* (J. W. Tukey, ed.). Wiley, New York.

FISHER, R. A. (1962). The simultaneous distribution of correlation coefficients. *Sankhyā Ser. A.* **24** 1–8.

GALTON, F. (1889). *Natural Inheritance*. Macmillan, London.

GIRSHICK, M. A. (1939). On the sampling theory of roots of determinantal equations. *Ann. Math. Statist.* **10** 203–224.

HELMERT, F. R. (1876a). Die Genauigkeit der Formel von Peters zur Berechnung des wahrscheinlichen Beobachtungsfehlers directer Beobachtungen gleicher Genauigkeit. *Astronom. Nachr.* **88** 113–132.

HELMERT, F. R. (1876b). Über die Wahrscheinlichkeit der Potenzsummen der Beobachtungsfehler und über einige damit im Zusammenhange stehende Fragen. *Zeitschrift für Mathematik und Physik* **21** 192–218.

HOTELLING, H. (1931). The generalization of Student's ratio. *Ann. Math. Statist.* **2** 360–378.

HSU, P. L. (1939). On the distribution roots of certain determinantal equations. *Annals of Eugenics* **9** 250–258.

HSU, P. L. (1941). On the problem of rank and the limiting distribution of Fisher's test function. *Annals of Eugenics* **11** 39–41.

IRWIN, J. O. (1931). Mathematical theorems involved in the analysis of variance. *J. Roy. Statist. Soc.* **94** 284–300.

IRWIN, J. O. (1934). On the independence of the constituent items in the analysis of variance. *J. Roy. Statist. Soc. Suppl.* **1** 236–251.

KRUSKAL, W. H. (1946). Helmert's distribution. *Amer. Math. Monthly* **53** 435-438.

KRUSKAL, W. H. (1968). Review of "Forerunners of the Pearson χ^2" by H. O. Lancaster. *Math. Rev.* **36** 916.

LANCASTER, H. O. (1966). Forerunners of the Pearson's χ^2. *Austral. J. Statist.* **8** 117–126.

MAHALANOBIS, P. C. (1930). On tests and measures of group divergence. Part I. Theoretical formulae. *Journal and Proceedings of the Asiatic Society of Bengal* **26** 541–588.

MOOD, A. M. (1951). On the distribution of characteristic roots of normal second-moment matrices. *Ann. Math. Statist.* **22** 266–273.

PEARSON, E. S. (1968). Studies in the history of probability and statistics. XX. Some early correspondence between W. S. Gossett, R. A. Fisher and Karl Pearson, with notes and comments. *Biometrika* **55** 445–467.

PEARSON, K. (1896). Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia. *Philos. Trans. Roy. Soc. London Ser. A* **187** 253–318.

PEARSON, K. (1900). On the criterion that a given system of derivations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science, Series 5* **50** 157–175.

PEARSON, K. (1915). On the distribution of the standard deviations of small samples. Appendix I to papers by "Student" and R. A. Fisher. *Biometrika* **10** 522–529.

PORTER, T. M. (1986). *The Rise of Statistical Thinking, 1820–1900*. Princeton Univ. Press.

RAO, C. R. (1992). R. A. Fisher: the founder of modern statistics. *Statist. Sci.* **7** 34–48.

ROY, S. N. (1939). *p*-statistics or some generalisations in analysis of variance appropriate to multivariate problems. *Sankhyā* **4** 381–396.

SAVAGE, L. J. (1976). On rereading R. A. Fisher. *Ann. Statist.* **4** 441–500.

SOPER, H. E. (1913). On the probable error of the correlation coefficient to a second approximation. *Biometrika* **9** 91–115.

SOPER, H. E., YOUNG, A. W., CAVE, B. M., LEE, A. and PEARSON, K. (1917). On the distribution of the correlation coefficient in small samples. Appendix II to the papers of "Student" and R. A. Fisher. A co-operative study. *Biometrika* **11** 328–413.

STIGLER, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty before 1900*. Harvard Univ. Press.

STUDENT (1908a). The probable error of a mean. *Biometrika* **6** 1–25.

STUDENT (1908b). Probable error of a correlation coefficient. *Biometrika* **6** 302–310.

WALKER, H. M. (1929). *Studies in the History of Statistical Method*. Williams and Wilkins, Baltimore.

WISHART, J. (1928). The generalized product moment distribution in samples from a normal multivariate population. *Biometrika* **20A** 32–52.

WISHART, J. (1934). Statistics in agricultural research. *J. Roy. Statist. Soc. Suppl.* **1** 26–51.

WISHART, J. (1948). Proofs of the distribution law of the second order moment statistics. *Biometrika* **35** 55–57.

YULE, G. U. (1897). On the theory of correlation. *J. Roy. Statist. Soc.* **60** 812–854.