100cias@uned, Revista de la Facultad de Ciencias.
(2000), **3**, 51-55

R. A. FISHER: EL INICIO DEL ANALISIS MULTIVARIANTE

Miguel A. Gómez Villegas
Dpto. de Estadística e I. O.
Fac. de C. Matemáticas
Universidad Complutense

---

1 Algunos aspectos biográficos de R. A. Fisher

Sir Ronald Aylmer Fisher (1890-1962) puede ser considerado sin duda el creador del análisis multivariante, interesa recordar la cita de L. J. Savage, que decía que era mas fácil señalar las partes de la Estadística a las que no había contribuido, que referirse a las que si lo había hecho.

Fisher nace en East Finchley (Londres). Era el mas joven de ocho hermanos; tuvo otro gemelo que no sobrevivió; acudió a la escuela en Stanmore y posteriormente estudió en Harrow. En su juventud tuvo prohibido leer con luz eléctrica y se le recomendó no fijar la vista demasiado, se ha especulado sobre si sus problemas de visión ayudaron a desarrollar su capacidad para lograr resultados sin necesidad de realizar todos los pasos y reforzar su intuición geométrica. Gracias a una beca, estudió en el Casius College en Cambridge, donde se graduó entre 1909 y 1912; en 1913 es lector de física matemática, dedicándose al estudio de biometría y genética. Entre 1913 y 1915 trabaja en una compañía de inversiones, pero pronto descubre que no es ésta su vocación.

En 1916 escribe un artículo en el que demuestra que las teorías de Mendel no se ven rechazadas por los datos; este artículo fué referenciado por Karl Pearson como estadístico y por Punnet como genetista, al no ser aceptada su publicación en la versión que Fisher deseaba, va a dar lugar a una de las muchas polémicas que Fisher mantuvo a lo largo de su vida y que le llevó a afirmar que su artículo había sido referenciado por un estadístico que no sabía genética y por un genetista que no sabía estadística, lo que le creó una fuerte enemistad con Karl Pearson.

En 1917 se casó con Ruth E. Guinnes, con la que tuvo dos hijos y seis hijas. Fisher era una persona muy partidaria de su familia pero mantenía la teoría de que a partir de una determinada edad, los hijos debían vivir fuera de la unidad familiar, teoría que procuró mantener.

En 1919 se une a la estación experimental de Rothamsted, este fué su particular esfuerzo a la contribución de Inglaterra en la primera guerra mundial, ya que aunque quiso alistarse, por su mala vista, no fué admitido. En esta estación experimental, desarrolló dos de sus principales contribuciones a la ciencia estadística: *el análisis de la varianza* , alrededor del año 1991 y *los principios del diseño de experimentos*, entre 1923 y 1924.

En 1929 es elegido miembro de la Royal Society por sus contribuciones a la estadística, por cierto que opinaba que era un error que la citada sociedad estuviera integrada por un alto porcentaje de personas por encima de 50 años. Al año siguiente, publica su libro *The Genetical Theory of Natural Selection*, donde apoya la teoría de Darwin de la evolución de las especies y modifica la inmutabilidad del concepto de dominancia.

Acepta, a la retirada de Karl Pearson, la cátedra de Eugenesia en el University College de Londres, quedando Egon Pearson, el hijo de Karl Pearson, como catedrático de Estadística. Durante este tiempo se dedica a la investigación en genética, Fisher ha sido de los pocos científicos que han destacado en dos campos distintos del conocimiento.

En 1938 viaja a la India invitado por Mahalanobis y en 1943, a Estados Unidos, como profesor visitante en la Universidad de Carolina del Norte. La etapa entre 1938 y 1962 es en la que se dedica a desarrollar sus trabajos en el campo de la inferencia estadística. Durante el bienio 1953-1954 es presidente de la Royal Statistical Society y dedica sus intervenciones a glosar las contribuciones de los primeros estadísticos.

En 1956, con 66 años de edad, publica el libro *Statistical Methods and Scientific Inference*, que da la impresión de ser un manual para principiantes más que un libro de texto, pero en eso radicó su éxito. A lo largo del libro, anima a trabajar con ejemplos, discute problemas prácticos y términos teóricos, todo a partir de ejemplos numéricos. En él se separa de los matemáticos diciendo que en estadística hay que hacer razonamiento inductivo, en lugar de razonamiento deductivo, para lo cual es necesaria una gran formación matemática, para aplicarla a obtener conclusiones de los datos con que se trabaja.

Se retira en 1957 y se marcha a Australia donde trabaja como investigador en el CSIRO (Commonwealth Scientific and Industrial Research Organitation). En 1962 muere de cancer de boca en Adelaida (Australia), a la edad de 72 años.

Un amplio estudio biográfico sobre Fisher puede verse en el libro de J. B. Box (1978).

2 Contribuciones a la Inferencia Estadística

Fisher publicó 140 artículos sobre genética, 129 sobre estadística y 16 sobre otros temas. Si uno tuviera que quedarse con las contribuciones que mas impacto han producido, éstas posiblemente serían: la calibración del nivel de significación, la diferencia entre muestra y población, el método de la máxima verosimilitud para la

construcción de estimadores, el análisis de la varianza y el diseño de experimentos. A continuación se van a desarrollar brevemente cada uno de estos aspectos.

2.1 La calibración del nivel de significación

En su libro titulado *Statistical Methods and Scientific Inference*, publicado por vez primera en 1956 (su última edición es de 1973), Fisher valora la evidencia suministrada por el p-valor del siguiente modo:

- si el p-valor $\in [0, 0.01]$, existe evidencia decisiva contra H_0.
- si el p-valor $\in (0.01, 0.05]$, existe evidencia fuerte contra H_0.
- si el p-valor $\in (0.05, 0.1]$, existe evidencia sustancial contra H_0.
- si el p-valor $\in (0.1, 1]$, existe evidencia a favor de H_0.

Sin duda esta asignación de valores, junto con el peso de su autoridad, han contribuido a la gran difusión de los tests de hipótesis mediante esta aproximación.

2.2 La diferencia entre muestra y población

También se debe a Fisher la distinción nítida entre muestra y población, hasta entonces no quedaba muy claro el entorno en el que se estaba trabajando. Se podía estar manejando una colección de observaciones y encontrar características de la misma o bién se trataban esas características como los parámetros desconocidos de una población.

2.3 El método de construcción de estimadores de la máxima verosimilitud

Es claramente el método más importante y con mejores propiedades de obtención de estimadores que se conoce. Ha de tenerse en cuenta que hasta entonces el método más empleado era el de los momentos, ampliamente utilizado por Karl Pearson para aproximar su célebre familia de curvas.

Es conocido, que dada la muestra $(x_1, \ldots, x_n)$, el método de la máxima verosimilitud para estimar el valor θ, consiste en utilizar el valor de $\widehat{\theta}$ tal que

$$max_\theta f(x_1, \ldots, x_n \mid \theta) = f(x_1, \ldots, x_n \mid \hat{\theta})$$

o equivalentemente, si se da regularidad suficiente, el valor que sea solución del sistema

$$\left.\begin{array}{r} \frac{\partial}{\partial\theta_i} lnf(x_1,\ldots,x_n \mid \theta) = 0 \\ i = 1,\ldots,k \end{array}\right\}$$

La idea de considerar el modelo como función de θ en lugar de como función de la muestra es una genialidad, aunque perfectamente razonable si uno está interesado en la estimación del parámetro; y precisamente el que $\widehat{\theta}$ cumpla que las derivadas parciales se anulen en él, dota al estimador de máxima verosimilitud de *buenas* propiedades.

2.4 El análisis de la varianza

El análisis de la varianza fué desarrollado en la estación experimental de Rohamsted alrededor de 1921. Siguiendo el método recomendado por Fisher se puede introducir a través de un ejemplo: Conociendo que el trabajador que usa una máquina influye en el rendimiento de ésta, se trata de controlar estadísticamente la influencia de la máquina y del trabajador en el rendimiento. Para ello supuestas I máquinas y J operarios y considerando la tabla de rendimientos en la forma

	1	2	...	J
1	y_{11}	y_{12}	$\cdots$	y_{1J}
2	y_{21}	y_{22}	$\cdots$	y_{2J}
$\vdots$	$\vdots$	$\vdots$		$\vdots$
I	y_{I1}	y_{I2}	$\cdots$	y_{IJ}

se postula el modelo:

$$y_{ij} = \mu + \alpha_i + \beta_j + u_{ij}, \quad i = 1,\ldots,I \quad j = 1,\ldots,J$$

donde

- μ es el efecto global,
- α_i es el efecto máquina,
- β_j es el efecto debido al operario,
- u_{ij} es el efecto aleatorio.

Fisher recomienda que la asignación de operario a máquina sea aleatoria. Lo que se pretende contrastar es la hipótesis nula $H_0 : \alpha_i = 0 \quad \forall i, \quad \beta_j = 0 \quad \forall j$. Así se construye la conocida Tabla ADEVA (del Análisis de la Varianza)

Fuente	Suma de cuadrados	S.c. medios	Esperanza
e. máquina	$J\Sigma(\overline{y}_{i\cdot} - \overline{y}_{\cdot\cdot})^2 = J\sum \widehat{\alpha}_i^2$	$J\Sigma\widehat{\alpha}_i^2/(I-1)$	$\sigma^2 + J\frac{\Sigma\alpha_i^2}{I-1}$
e. operario	$I\Sigma(\overline{y}_{\cdot j} - \overline{y}_{\cdot\cdot})^2 = I\sum \widehat{\beta}_j^2$	$I\Sigma\widehat{\beta}_j^2/(J-1)$	$\sigma^2 + I\frac{\Sigma\beta_j^2}{J-1}$
e. no explicado	$\Sigma\Sigma(y_{ij} - \widehat{\mu} - \widehat{\alpha}_i - \widehat{\beta}_j)^2$	$\frac{\Sigma\Sigma(y_{ij}-\widehat{\mu}-\widehat{\alpha}_i-\widehat{\beta}_j)^2}{(I-1)(J-1)}$	σ^2

De manera que el efecto máquina nulo se contrasta mediante la hipótesis nula $H_{01} : \alpha_i = 0 \quad i = 1, \ldots, I$, por lo que la región crítica del test viene dada mediante

$$RC = \{\frac{J\Sigma\widehat{\alpha}_i^2/(I-1)}{\frac{\Sigma\Sigma(y_{ij}-\widehat{\mu}-\widehat{\alpha}_i-\widehat{\beta}_j)^2}{(I-1)(J-1)}} \geq f_{I-1,(I-1)(J-1);\alpha}\}$$

Análogamente, el efecto operario nulo, se contrasta mediante la hipótesis nula $H_{02} : \beta_j = 0 \quad j = 1, \ldots, J$, por lo que la región crítica del test viene dada mediante

$$RC = \{\frac{I\Sigma\widehat{\beta}_j^2/(J-1)}{\frac{\Sigma\Sigma(y_{ij}-\widehat{\mu}-\widehat{\alpha}_i-\widehat{\beta}_j)^2}{(I-1)(J-1)}} \geq f_{J-1,(I-1)(J-1);\alpha}\}$$

2.5 El diseño de experimentos

Fisher fué el creador del diseño de experimentos. La importancia de esta aportación la pone de manifiesto el hecho de que su libro sobre este tema, conoció nueve ediciones entre 1935 y 1966. Sólo se recogerá aquí su cita:

> "...Un examen cuidadoso del proceso de recogida de datos, o diseño experimental, puede incrementar la precisión de los resultados, diez o doce veces. Consultar a un estadístico depués de que se haya concluido un experimento es, muy a menudo, pedirle que realice un examen postmortem. Quizás le pueda decir de qué murió el experimento."
>
> Fisher (1935)

3 Contribuciones al Análisis Multivariante

En esta sección se realiza un comentario, breve y no técnico, de todos los artículos de Fisher que tratan sobre el Análisis Multivariante; serán citados únicamente por el año y se corresponden con los que están en la bibliografía al final del artículo.

(1915)- En él se obtiene la distribución en el muestreo del *coeficiente de correlación lineal*, es decir de

$$r = \frac{\sum(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum(x_i - \overline{x})^2 \sum(y_i - \overline{y})^2}},$$

cuando la población tiene distribución normal.

Está inspirado en un trabajo de Student en el que éste prueba que si la población $X \epsilon N(\mu, \sigma)$ entonces la media muestral y la cuasi varianza son variables aleatorias independientes, y además $\overline{X} \epsilon N(\mu, \frac{\sigma}{\sqrt{n}})$ y $(n-1)\frac{s^2}{\sigma^2} \epsilon \chi^2_{n-1}$.

El artículo fué publicado con retraso por Karl Pearson y fué la causa de la enemistad entre ambos y el motivo por el cual Fisher no volvió a publicar en la revista Biometrika, liderada por Karl Pearson.

(1921)- Este artículo fué publicado en la revista italiana *Metron* y en él estima el coeficiente de correlación poblacional, cuando ésta tiene una distribución normal bivariante cuyas medias son iguales y la matriz de covarianzas es

$$\Sigma = \begin{pmatrix} \sigma_x^2 & \sigma_x y \\ \sigma_{xy} & \sigma_x^2 \end{pmatrix}.$$

Introduce como estimador el valor del coeficiente de correlación muestral

$$r = \frac{\sum_{i=1}^{n}(x_i - \hat{\mu})(y_i - \hat{\mu})}{.5\left[\sum_{i=1}^{n}(x_i - \hat{\mu})^2 + \sum_{i=1}^{n}(y_i - \hat{\mu})^2\right]}$$

donde $\widehat{\mu}$ es el estimador de máxima verosimilitud para el parámetro μ, dado por

$$\hat{\mu} = 0.5(\overline{x} + \overline{y}).$$

También obtiene en este artículo la distribución en el muestreo del coeficiente de correlación muestral, sin emplear la distribución de Snedecor, así cómo la distribución *fiducial* del coeficiente de correlación poblacional.

(1922)- En este trabajo vuelve a obtener la distribución de la F de Snedecor, al tratar de encontrar la distribución de estadísticos asociados a la regresión lineal.

Demuestra que el cociente entre el coeficiente de regresión estimado y su error estándar estimado tiene distribución de Student.

(1924)- Contiene la distribución en el muestreo del coeficiente de correlación parcial $r_{xy \cdot z}$ cuando se quita la dependencia lineal de una tercera variable y obtiene que la distribución coincide con la del coeficiente de correlación lineal r_{xy} pero con un grado de libertad menos.

(1925)- Pone de manifiesto la importancia de las distribuciones χ^2, t y F en los contextos del análisis de la varianza, para poblaciones normales.

En este artículo incluye el argumento de que si una variable n dimensional $\overrightarrow{X} \epsilon N(0, \sigma I_n)$ una transformación ortogonal hace que la nueva variable $\overrightarrow{Y} \epsilon N(0, \sigma I_n)$

y si además las primeras k variables aleatorias de $Y_1, \ldots, Y_n$ se construyen convenientemente y se escogen las restantes $Y_{k+1}, \ldots, Y_n$ hasta completar las n, entonces cómo

$$\sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{k} Y_i^2 = \sum_{i=k+1}^{n} Y_i^2,$$

se sigue que $\sum_{i=k+1}^{n} Y_i^2$ tiene una distribución $\sigma^2 \chi_{n-k}^2$ e independiente de $\sum_{i=1}^{k} Y_i^2$. Es decir, generaliza el razonamiento que había utilizado para demostrar que en una población normal la media muestral es independiente de la cuasi varianza muestral.

(1928)- Obtiene la distribución del coeficiente de correlación múltiple muestral, cuando las variables x e y no son independientes. Además introduce la distribución de la variable χ^2 no centrada.

(1936)- Este artículo está publicado en los *Anales de Eugenesia*, la revista que había fundado Karl Pearson y que Fisher dirigía, desde su nombramiento cómo profesor de Eugenesia en el University College de Londres.

En él introduce el *análisis discriminante* como el mecanismo que dadas dos muestras $x^{(1)} = (x_1^{(1)}, \ldots, x_n^{(1)})$ y $x^{(2)} = (x_1^{(2)}, \ldots, x_n^{(2)})$ obtiene la combinación lineal de la diferencia $b(x^(1) - x^(2))$ que maximice la expresión $(x^{(1)} - x^{(2)})^2 / V[X]$ con lo que construye la siguiente regla de decisión cuando se obtiene la nueva observación x:

si $b'x > 0.5(x^{(1)} + x^{(2)}) \Longrightarrow x$ pertenece a la primera población
si $b'x < 0.5(x^{(1)} + x^{(2)}) \Longrightarrow x$ pertenece a la segunda población

(1938)- Incluye las contribuciones de otros autores relacionadas con el análisis discriminante y, en particular, incluye aspectos tratados por Mood, el coautor de un celebrado libro clásico sobre inferencia, por Wilks, el principal causante del alto nivel alcanzado por los *Annals of Statistics* y por Hsu, el estadístico que aproximó el número de grados de libertad de la distribución de la χ^2 en el problema de Berhens-Fisher.

(1940)- Trata problemas de tests de hipótesis relacionados con el análisis discriminante, estudia las tablas de contingencia y anticipa el análisis de correspondencias.

(1962)- En este artículo, que apareció el mismo año de su muerte, aborda la distribución de distintos coeficientes de correlación tratados de forma conjunta.

Desearía terminar este estudio con lo que G.E.P. Box decía de Fisher:

¿Era un estadístico aplicado? ¿era un estadístico matemático? ¿era un analista de datos? ¿era un diseñador de experimentos? Seguramente, él era todas estas cosas y mucho más que la suma de éstas. El constituye el ejemplo que nosotros deberíamos seguir.

Box(1978)

AGRADECIMIENTOS

Este trabajo se ha subvencionado, en parte, con la ayuda de la *Dirección General de Investigación Científica y Técnica* (DGICYT) correspondiente al proyecto número PB98–0797.

REFERENCIAS

Anderson, T. W. (1996) R. A. Fisher and multivariate analysis. *Statist. Science*, **11**,1, 20–34.

Bennett, J. H. (1990) *Statistical Method, Experimental Design and Scientific Inference*. Oxford. Oxford University Press.

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

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

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. (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. Stat. Soc.*, **85**, 597-612.

Fisher, R. A. (1925, 1970) *Statistical Methods for Research Workers*. Edimburgo. Oliver and Boyd. (Hay edición en español).

Fisher, R. A. (1924) The distribution of the partial correlation coefficient. *Metron,* **3**, 329-332.

Fisher, R. A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **8**, 376–386.

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. (1935,...,1966) *The Design of Experiments*. Edimburgo. Oliver and Boyd. (Hay edición en español).

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

Fisher, R. A. (1956, 1959) *Statistical Methods and Scientific Inference*. Edimburgo. Oliver and Boyd. (1973) New York. Hafner.

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

Fisher, R. A. (1990) *Statistical Methods, Experimental Design and Scientific Inference*. Edited by Bennett, J. M. with a foreword by Yates, F. Oxford. Oxford University Press.

Girón, F. J. y Gómez Villegas, M. A. (1998). R. A. Fisher: su contribución a la Ciencia Estadística. En *Historia de la Matemática en el siglo XX*. Ed. Real Acad. Cien. Exac. Fis. Nat., pp. 43–61.