Journal of Mathematical Chemistry Vol. 36, No. 4, August 2004 (© 2004)

Equivalence of the Pecka–Ponec correlation probability and the statistical *F* significance for MLR models

E. Besalú*

Institut de Química Computacional, Facultat de Ciències, Universitat de Girona, Spain
E-mail: emili@iqc.udg.es

J.V. de Julián-Ortiz

Xarxa de Recerca de Malalties Tropicals, Unidad de Investigación de Diseño de Fármacos y Conectividad Molecular, Dep. Química Física, Facultat de Farmàcia, Universitat de València, Spain
E-mail: julian@goya.combios.es

Received 17 June 2004; revised 2 July 2004

In an article of this journal Pecka and Ponec [J. Math. Chem. 27 (2000) 13] have proposed, by means of a probability calculation, a method to evaluate the statistical importance of correlations obtained from multilinear regression equations involving an arbitrary number of experimental points and parameters. Here, it is demonstrated how this probability exactly coincides with a more general concept: the confidence probability of an *F* distribution having the appropriate degrees of freedom.

KEY WORDS: statistical MLR significance, *F*-Snedecor value, Pecka–Ponec probability formula

1. Demonstration

In the article entitled "Simple analytical method for evaluation of statistical importance of correlations in QSAR studies" Pecka and Ponec [1] have elegantly derived, from pure geometrical basis, a probability function which allows to compare different multilinear regression (MLR) equations involving arbitrary numbers of points and independent parameters. Here we demonstrate that this probability function is equivalent to the classical *F* significance method.

Let us consider a MLR equation involving n points and m parameters (not including the independent term) bearing a correlation coefficient equal to r. The probability attached to the left queue in a Snedecor F distribution function is given by the following integral [2]:

$$p(r; n, m) = \frac{\int_0^x t^{\frac{k}{2}-1}(1-t)^{\frac{m}{2}-1}\mathrm{d}t}{\beta\left(\frac{k}{2}, \frac{m}{2}\right)}, \tag{1}$$

*Corresponding author.

being the degrees of freedom equal to m and $k = n - m - 1$. The expression (1) corresponds to the incomplete beta function integral. The denominator is the beta function integral which, in turn, can be expressed in terms of the gamma function or an integral involving trigonometric functions:

$$\beta(z, w) = \frac{\Gamma(z)\,\Gamma(w)}{\Gamma(z+w)} = 2\int_0^{\frac{\pi}{2}} \left(\sin\theta\right)^{2z-1} (\cos\theta)^{2w-1}\,\mathrm{d}\theta. \tag{2}$$

According to Mendenhall and Sincich [3], in (1), the superior limit of integration is $X = \frac{k}{k+mF}$, being $F = \frac{r^2}{1-r^2}\frac{k}{m}$. It is immediate to see that $X = 1 - r^2$. Then, (1) can be re-written according to the following change of variable:

$$t = \sin^2\theta;\ \mathrm{d}t = 2\sin\theta\cos\theta\mathrm{d}\theta, \tag{3}$$

giving

$$p(r; n, m) = \frac{2\int_0^{\arcsin\sqrt{1-r^2}} (\sin\theta)^{k-1}(\cos\theta)^{m-1}\mathrm{d}\theta}{\beta\left(\frac{k}{2}, \frac{m}{2}\right)}. \tag{4}$$

As in the first quadrant $\arccos\alpha = \arcsin\sqrt{1-\alpha^2}$, using (2) one obtains the final expression for the probability p coinciding with Pecka and Ponec's result [1]:

$$p(r; n, m) = \frac{\int_0^{\arccos\sqrt{r}} (\sin\theta)^{n-m-2}(\cos\theta)^{m-1}\mathrm{d}\theta}{\int_0^{\frac{\pi}{2}} (\sin\theta)^{n-m-2}(\cos\theta)^{m-1}\mathrm{d}\theta}.$$

2. Conclusion

It has been demonstrated the equivalence between Pecka and Ponec's formulation and the classic statistical F probability. The F parameter has many applications in statistics and, for the particular case of MLR equations, its attached p value is the probability of obtaining a correlation coefficient lesser than r in a MLR equation involving the mentioned degrees of freedom. It is noteworthy that Pecka and Ponec obtained their formula independently of the Fisher–Snedecor distribution concept. Their derivation additionally provides an elegant geometrical interpretation of the statistical significance.

Acknowledgements

Professor R. Ponec is acknowledged for providing a programming source code implementing the calculation of the integral he promotes. This allowed checking its numerical concordance with standard programs that compute and work with the F statistic. The authors also acknowledge the financial of this research to the grant number BQU2003-07420-C05 of the Ministerio de Ciencia

y Tecnología within the Spanish Plan Nacional I+D. It is also acknowledged the support by the Red Temática de Investigación Cooperativa RICET (Red de Investigación de Centros de Enfermedades Tropicales C03/04) of the Spanish Ministerio de Salud.

References

[1] J. Pecka and R. Ponec, J. Math. Chem. 27 (2000) 13.
[2] M. Abramowitz and I.A. Stegun, *Handbook of Mathematical Functions* (Dover, New York, 1972).
[3] W. Mendenhall and T. Sincich, *Statistics for Engineering and the Sciences* (Prentice Hall, Englewood Cliffs, NJ, 1995).