Estadística Aplicada: "Didáctica de la Estadística y Métodos Estadísticos en Problemas Socioeconómicos"
VII Coloquio Regional de Estadística
XII Seminario de Estadística Aplicada IASI
III Escuela de Verano CEAES
Universidad Nacional de Colombia, Sede Medellín. Medellín, 20-23 de Julio de 2010

Comparación de Intervalos de Confianza para el Coeficiente de Correlación

Juan Carlos Correa[a], *Liliana Vanessa Pacheco*[b]

Email: jccorrea@unal.edu.co

a. *Universidad Nacional-Sede Medellín*

b. *Universidad Nacional-Sede Medellín*

Resumen

La construcción de intervalos de confianza para la estimación de la correlación en la distribución normal bivariable, digamos ρ, es un problema importante en el trabajo estadístico aplicado. Revisamos diferentes procedimientos para su construcción y realizamos un estudio de simulación para analizar el comportamiento de los niveles de confianza reales y compararlos con los teóricos.

Estadística Aplicada: "Didáctica de la Estadística y Métodos Estadísticos en Problemas Socioeconómicos"
VII Coloquio Regional de Estadística
XII Seminario de Estadística Aplicada IASI
III Escuela de Verano CEAES
Universidad Nacional de Colombia, Sede Medellín. Medellín, 20-23 de Julio de 2010

Comparison of Confidence Intervals for the Correlation Coefficient

Juan Carlos Correa[a], *Liliana Vanessa Pacheco*[b]

Email: jccorrea@unal.edu.co

a. *Universidad Nacional-Sede Medellín*

b. *Universidad Nacional-Sede Medellín*

Abstract

Estimation of the correlation coefficient of a bivariate normal distribution using confidence intervals is a common procedure in the statistical practice. We propose a new confidence interval based on relative likelihood and compare it with the Fisher's formula and bootstrap via simulation.

0.1. Introducción

El coeficiente de correlación es una de las medidas estadísticas de más uso dentro del trabajo aplicado. Algunas de sus propiedades fueron estudiadas por Zheng y Matis (1993). Discusión sobre sus interpretaciones puede hallarse en Falk y Well (1997). La estimación del coeficiente de correlación por medio de intervalos es importante, y para ello se disponen de diversos métodos. El problema para el analista es la carencia de reglas sobre cuál fórmula es preferible. Para esto hemos realizado un estudio de simulación que nos permite analizar el comportamiento de los niveles de confianza reales y compararlos con los teóricos de los diversos intervalos disponibles.

Asumamos que $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ es una muestra aleatoria de una normal bivariable con vector de medias μ y matriz de varianzas y covarianzas Σ. El estimador máximo verosímil para ρ es (Graybill, 1976):

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\left(\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2\right)^{1/2}}$$

y el estimador UMVU (uniformly minimum variance and unbiased) de ρ es:

$$\hat{\rho} = R\left(\frac{\Gamma(\frac{n-2}{2})}{\Gamma(\frac{1}{2})\Gamma(\frac{n-3}{2})}\right)\int_0^1 \frac{t^{-\frac{1}{2}}(1-t)^{\frac{(n-2)}{2}}}{\sqrt{1-t(1-R^2)}}\,dt$$

La f.d.p de R es:

$$f_R(r) = \frac{(n-2)(1-\rho^2)^{\frac{(n-2)}{2}}}{\pi}(1-r^2)^{\frac{(n-4)}{2}}\int_0^\infty (\cosh w - \rho r)^{-(n-1)}\,dw$$

donde $-1 < r < 1$ y $-1 < \rho < 1$. El único parámetro de la distribución es ρ.

0.2. Métodos para calcular Intervalos de confianza para el Coeficiente de Correlación

0.2.1. Método I: Basado en la transformación Arco tangente

Este intervalo puede considerarse el intervalo clásico para este parámetro debido a Fisher (Stuart y Ord, 1987).

$$\left(\tanh\left(\operatorname{arctanh}(r) - \frac{z_{\alpha/2}}{\sqrt{n-3}}\right), \tanh\left(\operatorname{arctanh}(r) + \frac{z_{\alpha/2}}{\sqrt{n-3}}\right)\right)$$

0.2.2. Método II: Intervalo de la Razón de Verosimilitud

El siguiente intervalo de confianza no se ha encontrado en la literatura y es un aporte de este documento. Kalbfleish (1985) y Pawitan (2001) presentan la metodología para construir intervalos de verosimilitud. Si $L(\rho)$ es la función de verosimilitud, se define la *función de verosimilitud relativa* como

$$R(\rho) = \frac{L(\rho)}{L(r)}$$

El conjunto de valores de ρ para los cuales $R(\rho) \geq p$ es llamado *intervalo de 100 %p de verosimilitud* para ρ. Los intervalos del 14,7 % y del 3,6 % de verosimilitud corresponden a intervalos de confianza de niveles del 95 % y 99 % aproximadamente. Lo que se debe hacer entonces es hallar las raíces que nos dan los límites del intervalo. Para el caso del parámetro ρ tenemos que un intervalo de confianza del 95 % se halla encontrando el par de raíces tal que

$$R(\rho) = \frac{L(\rho)}{L(r)} = \left(\frac{1-\rho^2}{1-r^2}\right)^{\frac{(n-1)}{2}} \frac{\int_0^\infty (\cosh w - \rho r)^{-(n-1)}\, dw}{\int_0^\infty (\cosh w - r^2)^{-(n-1)}\, dw} \geq K(k, \alpha)$$

Lo anterior se resuelve numéricamente.

0.2.3. Método III: Bootstrap

El método bootstrap proporciona una manera directa y sencilla para hallar el intervalo de confianza para el coeficiente de correlación de la distribución normal

bivariable. La primera aplicación del método bootstrap fue en la determinación del intervalo de confianza del coeficiente de correlación en el artículo seminal de Efron (1979). Polansky (1999) no recomienda esta metodología para tamaños muestrales pequeños, por ejemplo, para la media de una distribución continua, el n que recomienda debe ser mayor de 10 y para estimar la varianza deber ser superior a 20. Para hallarlos se procede así:

1. A partir de la muestra $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ se estiman los parámetros de máxima verosimilitud del vector de medias y de la matriz de varianzas y covarianzas de la distribución normal bivariable.

2. Se generan M muestras de tamaño n de una distribución normal bivariable con parámetros $\hat{\mu}$ y $\hat{\Sigma}$. Y para cada una de estas muestras se estima el parámetro ρ, por ejemplo, para la muestra j el valor del estimador para el coeficiente de correlación es r_j.

3. Para los r_j, $j = 1, \ldots, M$, se construye un histograma y se calculan los percentiles $0{,}025/(k-1)$ y $0{,}975/(k-1)$ los cuales se denotarán: $r_i^{\{0,025\}}$ y $r_i^{\{0,975\}}$.

0.3. Resultados de la Simulación

Para comparar los tres métodos realizamos una simulación en R en la cual se consideraron combinaciones de (ρ, n) con valores de $\rho = 0{,}0, 0{,}1, 0{,}2, \ldots, 0{,}9$ y de $n = 5, 10, 20, 50, 100$. Para cada pareja se realizaron 1000 réplicas y se calcularon las fórmulas previas a un nivel de confianza del 95 % (Este es conocido como el nivel nominal). Para cada método y combinación se calculó la mediana de la longitud de los 1000 intervalos calculados y la proporción de intervalos que cubren el verdadero valor de ρ, esto es lo que llamamos como el nivel de confianza real. Las siguientes tablas presentan los resultados.

	I.C Bootstrap		I.C Transf. ArcTang		I.C L.R	
ρ	Long. 1	Nivel 1	Long. 2	Nivel 2	Long. 3	Nivel 3
			$n = 5$			
0	1.64408	0.901	1.69813	0.95	1.49869	0.931
0.1	1.64079	0.899	1.69203	0.957	1.49227	0.932
0.2	1.59646	0.887	1.66555	0.951	1.46477	0.92
0.3	1.579118	0.898	1.65415	0.951	1.453087	0.928
0.4	1.55363	0.898	1.63683	0.951	1.4355	0.933
0.5	1.42549	0.892	1.55717	0.958	1.35689	0.932
0.6	1.32763	0.9	1.48801	0.951	1.2913	0.93
0.7	1.161843	0.892	1.37531	0.957	1.18831	0.93
0.8	0.832681	0.889	1.10525	0.961	0.95478	0.948
0.9	0.487176	0.901	0.754628	0.956	0.663334	0.935

	I.C Bootstrap		I.C Transf. ArcTang		I.C L.R	
ρ	Long. 1	Nivel 1	Long. 2	Nivel 2	Long. 3	Nivel 3
			$n = 10$			
0	1.2074	0.937	1.2177	0.947	1.1416	0.941
0.1	1.1984	0.932	1.2126	0.947	1.1370	0.94
0.2	1.178	0.926	1.1925	0.943	1.1194	0.938
0.3	1.1402	0.928	1.1553	0.942	1.0868	0.933
0.4	1.0813	0.923	1.1025	0.938	1.0403	0.933
0.5	0.9858	0.933	1.0193	0.945	0.9668	0.942
0.6	0.8603	0.936	0.9014	0.952	0.8620	0.942
0.7	0.71087	0.938	0.7619	0.963	0.7365	0.956
0.8	0.5191	0.947	0.57417	0.961	0.5643	0.95
0.9	0.27952	0.935	0.32570	0.953	0.32834	0.944
			$n = 20$			
0	0.8584	0.945	0.8699	0.954	0.8411	0.953
0.1	0.8502	0.925	0.8620	0.936	0.8338	0.933
0.2	0.8340	0.953	0.8430	0.962	0.8163	0.957
0.3	0.8050	0.942	0.8112	0.952	0.7870	0.949
0.4	0.7399	0.939	0.7510	0.949	0.7313	0.946
0.5	0.6729	0.943	0.6840	0.949	0.6687	0.943
0.6	0.5887	0.932	0.6017	0.945	0.5914	0.944
0.7	0.4711	0.929	0.4832	0.948	0.4787	0.949
0.8	0.33573	0.94	0.34900	0.95	0.34905	0.948
0.9	0.18365	0.947	0.19441	0.951	0.19669	0.948

	I.C Bootstrap		I.C Transf. ArcTang		I.C L.R	
ρ	Long. 1	Nivel 1	Long. 2	Nivel 2	Long. 3	Nivel 3
			$n = 50$			
0	0.5467	0.942	0.5514	0.949	0.5436	0.949
0.1	0.5435	0.953	0.5490	0.953	0.5413	0.952
0.2	0.5312	0.958	0.5361	0.963	0.5289	0.963
0.3	0.5056	0.948	0.5079	0.949	0.5017	0.949
0.4	0.4681	0.951	0.4726	0.949	0.4675	0.947
0.5	0.4170	0.935	0.4211	0.939	0.4175	0.94
0.6	0.3592	0.945	0.3641	0.951	0.3619	0.948
0.7	0.2863	0.958	0.2911	0.963	0.2903	0.961
0.8	0.20072	0.942	0.20514	0.948	0.20540	0.955
0.9	0.10630	0.951	0.10980	0.957	0.11040	0.954
			$n = 100$			
0	0.3870	0.948	0.3912	0.951	0.3883	0.95
0.1	0.3856	0.957	0.3890	0.962	0.3860	0.96
0.2	0.3751	0.943	0.3778	0.945	0.3751	0.944
0.3	0.3557	0.932	0.3589	0.937	0.3565	0.935
0.4	0.3300	0.925	0.3327	0.928	0.3308	0.929
0.5	0.2953	0.955	0.2972	0.958	0.2958	0.956
0.6	0.2508	0.944	0.2535	0.949	0.2526	0.948
0.7	0.2013	0.949	0.2035	0.952	0.2032	0.952
0.8	0.14311	0.944	0.14448	0.942	0.14454	0.943
0.9	0.07486	0.948	0.07615	0.952	0.07637	0.948

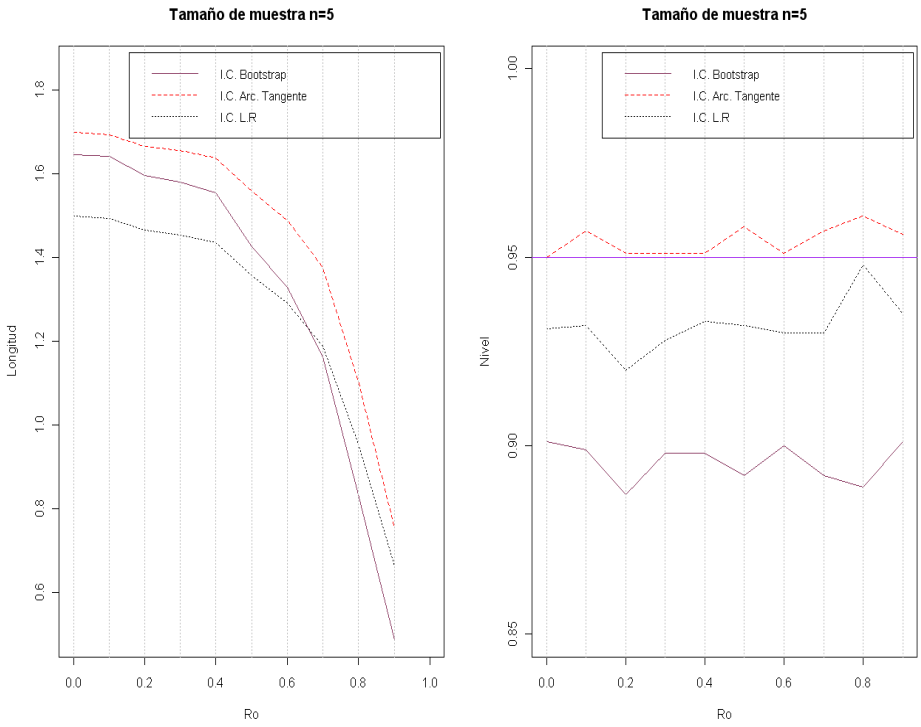

Figura 1: Longitud y Nivel real para I.C a un tamaño de muestra n=5

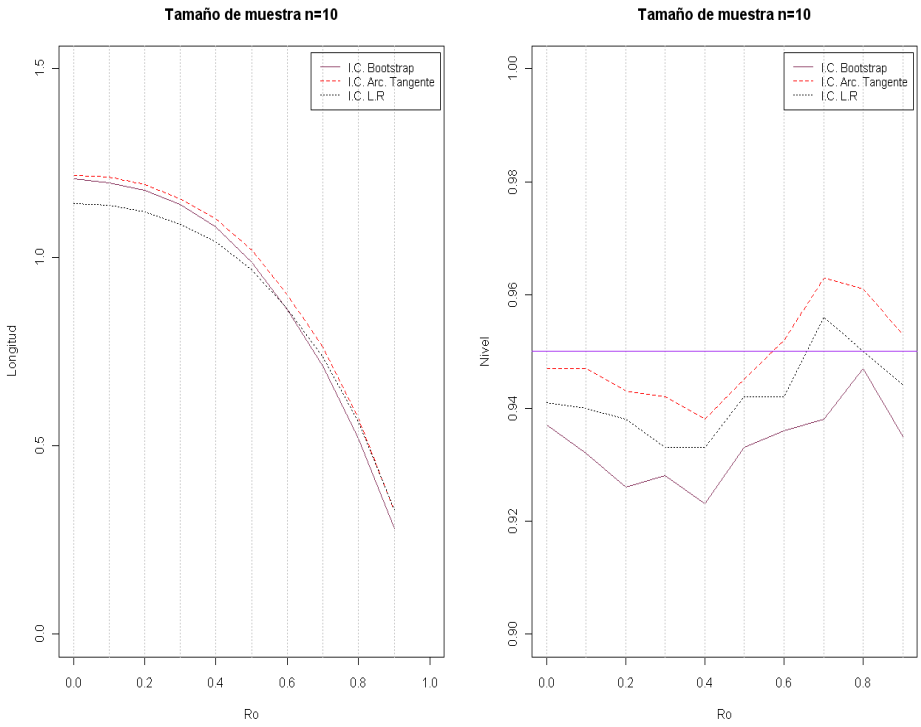

Figura 2: Longitud y Nivel real para I.C a un tamaño de muestra n=10

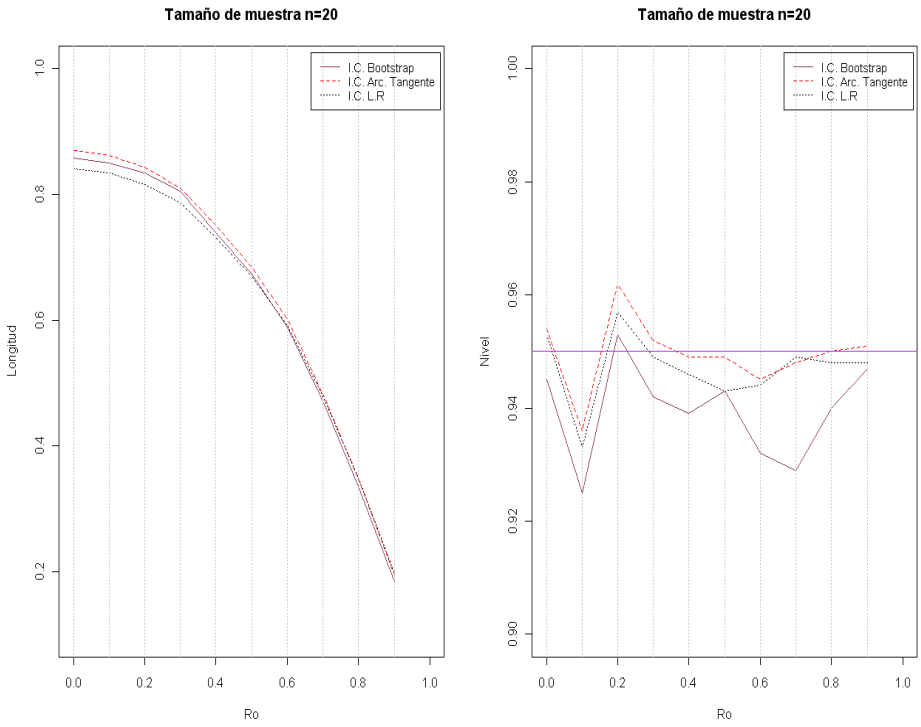

Figura 3: Longitud y Nivel real para I.C a un tamaño de muestra n=20

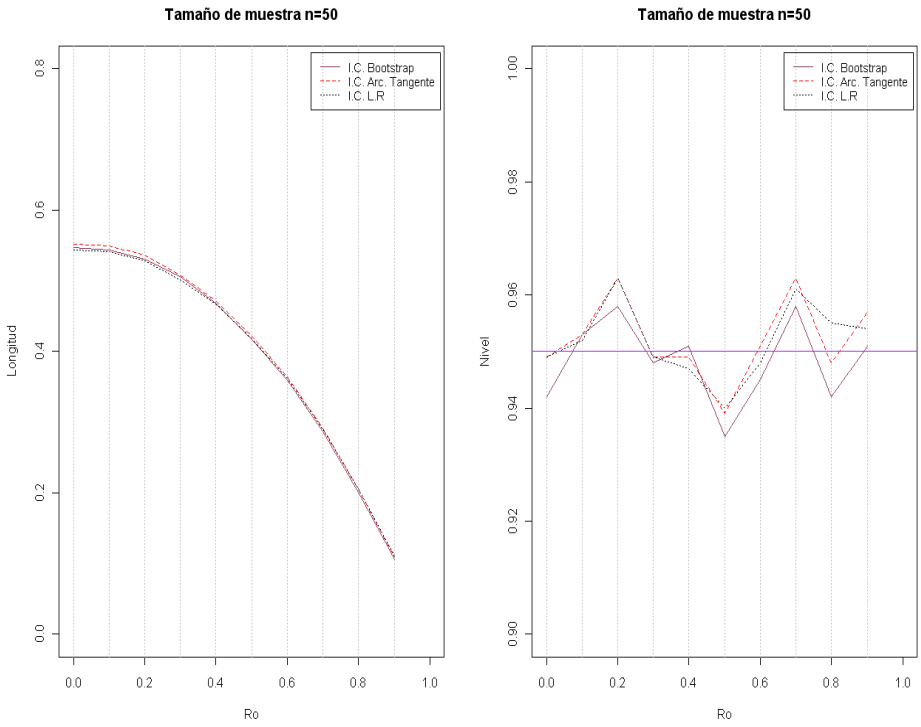

Figura 4: Longitud y Nivel real para I.C a un tamaño de muestra n=50

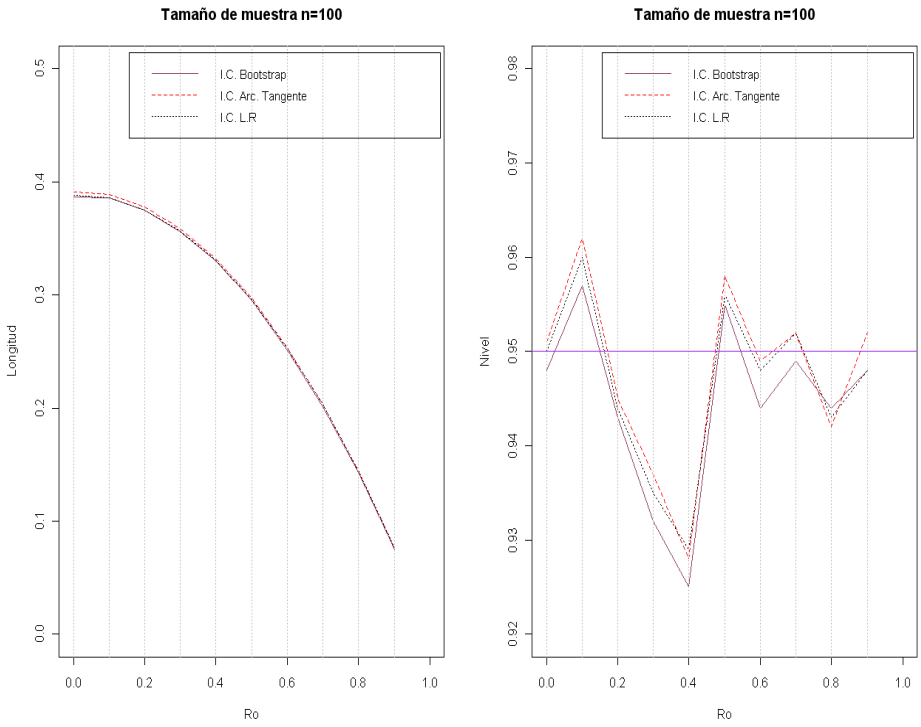

Figura 5: Longitud y Nivel real para I.C a un tamaño de muestra n=100

0.4. Discusión

De las anteriores tablas y gráficas se observa como el método III, es decir, el Bootstrap, tiene los niveles de confianza reales más bajos en casi todos los tamaños de muestra considerados. Sin embargo hay una relación inversa entre el nivel real y la longitud del intervalo, por lo que es preferible considerarlos conjuntamente. Un método será preferible si su nivel de confianza es al menos igual al nivel nominal deseado, y en caso de que ningún método lo cumpla, será aquel que esté lo más cerca.

El comportamiento del nivel real del método I, Transformación Arco Tangente, en muestras pequeñas lo presentan como el método preferible. Como comple-

mento a esto, el método I de Fisher, considerado el tradicional y que aparece en la mayoría de textos básicos de estadística, tiene niveles reales que están muy cerca al nominal, en la mayoría de los casos considerados. Luego, en términos de nivel de confianza, este método es quizá el mejor. Pero si el criterio de decisión se basara solo en la longitud del intervalo calculado, se podría concluir que en casi todos los casos de muestras de tamaño pequeño, el mejor método es el método II, el de Razón de verosimilitud. No se observan grandes diferencias en la longitud de los intervalos calculados mediante los tres métodos cuando el tamaño de muestra es grande.

De lo anterior se puede concluir que no se puede considerar longitud o nivel real aisladamente para seleccionar el mejor método, ya que no necesariamente los intervalos de confianza que posean menor longitud son los que tienen niveles de confianza reales más cercanos al nominal. Es importante anotar que el concepto de nivel de confianza real es poco manejado en la práctica, desconociendo su importancia cuando se trabaja con procedimientos asintóticos. Hemos encontrado que el método I es un método que resulta confiable cuando se considera el nivel de confianza real.

Bibliografía

[1] B.Efron. (1979). Computers and Theory of Statistics: Thinking the unthinkable. *SIAM Review*. 21 :460-480.

[2] R. Falk, A.D Well. (1997). Many Faces of the Correlation Coefficient. *Journal of Statistics Education*. 5. No. 3.

[3] F.A Graybill. (1976). *Theory and Application of the Linear Model*. Duxbury Press: Boston.

[4] J. G. Kalbfleish. (1985). *Probability and Statistical Inference*. 2. Segunda Edición. Springer-Verlag: New York.

[5] Y. Pawitan. (2001). *In All Likelihood*. Clarendon Press: Oxford.

[6] A.M. Polansky. (1999). Upper Bounds on the True Coverage of Bootstrap Percentile Type Confidence Intervals. *The American Statistician*. 53. Nro. 4 : 362-369.

[7] A. Stuart & J.K. Ord. (1987). *Kendall's Advanced Theory of Statistics*. Quinta Edición. Vol 1. Oxford University Press: New York.

[8] Q. Zheng & J. H. Matis. (1994). Correlation Coefficient Revisited. *The American Statistician*. 48. Nro. 3 :240-241.