X CONGRESO LATINOAMERICANO DE SOCIEDADES DE ESTADÍSTICA
CÓRDOBA, ARGENTINA. 16 A 19 DE OCTUBRE 2012

TAMAÑO DE MUESTRA PARA CORRELACIÓN

Carlos Henríquez-Roldán[1], Claudia Navarro[2], Alejandra Otárola[3], Bruno Barra[4]

[1] *Universidad de Valparaíso, Facultad de Ciencias, Departamento de Estadística, profesor, Centro de Estudios Estadísticos de la Universidad de Valparaíso, director* – carlos.henriquez@uvach.cl

[2] *Universidad de Valparaíso, Facultad de Ciencias, Departamento de Estadística, profesora* – claudia.navarro@uv.cl

[3] *Universidad de Valparaíso, Facultad de Ciencias, Departamento de Estadística, estudiante Ingeniería en Estadística* – alejandra.o.veas@gmail.com

[4] *Universidad de Valparaíso, Centro de Estudios Estadísticos de la Universidad de Valparaíso, sociólogo* – bruno.barra@uvach.cl

RESUMEN

Una de las preguntas más recurrente para los estadísticos es ¿cuál es el tamaño muestral mínimo requerido para estimar –en términos estadísticos– un parámetro? Los parámetros típicos son medias, totales y proporciones. ¿Qué ocurre con otros parámetros?, ¿el coeficiente β_1 en regresión (simple, logística u otra regresión), por ejemplo?, ¿o el coeficiente de correlación de Pearson, ρ, de una distribución normal bivariada?

Se revisó una cantidad de trabajos: artículos, tesis y trabajos de titulación donde se presentaba al menos una correlación (generalmente de Pearson). En prácticamente, ninguno de ellos se justificó si el tamaño muestral era el adecuado para lo que se estaba estudiando. Por ende, no se mencionaba ni el nivel de confianza ni el error de estimación tan requeridos en las fichas técnicas de los estudios o en materiales y métodos de las investigaciones.

Para determinar un tamaño muestral con el objetivo de estimar una correlación de Pearson, no se requiere una estimación de una o de las dos varianzas de las variables bajo estudio. A través de simulaciones de Montecarlo, se obtiene un tamaño de muestra para estimar una correlación de Pearson cuando se proporciona el nivel de confianza, el error de estimación y una idea de la correlación. La situación más conservadora se produce cuando se supone que la correlación poblacional es cero. En este trabajo se muestra, a través de simulaciones, cómo enseñar la determinación de tamaños muestrales para correlación de Pearson.

***PALABRAS CLAVE*:** tamaño muestral, correlación, educación.

INTRODUCCIÓN

¿De qué manera justificar un tamaño de muestra para un estudio de correlación? ¿Cómo lograr que se comprenda la necesidad de trabajar con tamaños de muestra apropiados? No es una pregunta frecuente para quienes hacen uso de la estadística. En general, desde el punto de vista de la correlación de Pearson, en muchos trabajos no se menciona el nivel de confianza ni el error de estimación (ee) para determinar si el tamaño muestral trabajado fue el apropiado. ¿Qué ocurre con el tamaño muestral cuando el parámetro es la correlación (de Pearson), ρ, entre dos variables (aleatorias)? La correlación muestral está acotada entre –1 y 1, por ende la distribución muestral presentará una asimetría positiva cuando la correlación poblacional, ρ, esté próxima a –1 y una asimetría negativa cuando ρ esté próxima a 1. Kareev (1995) mostró empírcamente que la determinación de un tamaño muestral para correlación está relacionado con la distribución muestral de las correlaciones. Esto produce como ventaja que, a través de tamaños muestrales pequeños, se puede detectar tempranamente la correlación poblacional. Luego, Anderson *et al.*(2005) mostraron, a través de simulaciones, que tamaños muestrales pequeños para determinación de correlación son ventajosos pero solo cuando existe correlación alta. Este trabajo propone, a través de simulaciones de Montecarlo, determinar tamaños muestrales aproximados para estimar una correlación de Pearson.

MARCO TEÓRICO

Godino y Batanero (1994), afirman que los objetos matemáticos clasificados como conceptos, procedimientos, teorías, demostraciones, entre otros, surgen de la necesidad de dar respuestas a situaciones internas o externas de matemática. Cuando una clase de situaciones-problemas comparten soluciones, se consideran que están agrupadas en un campo de problemas. A partir de las nociones de situaciones-problemas, campo de problemas y práctica, con el fin de estudiar procesos cognitivos y didácticos, se desarrollan las nociones derivadas de práctica significativas y el significado de un objeto, para las cuales se postulan dos dimensiones interdependientes, una personal y otra institucional. Una práctica es significativa para una persona o para una institución, si cumple con la función de resolver el problema, comunicar, validar o entender su solución.

Es en esta práctica significativa que este trabajo pretende resolver la situación-problema de enfrentarse a la determinación de tamaños muestrales para cuando se trabaja con correlación de Pearson. Con esto se pretende también obtener la comprensión intuitiva de la necesidad de trabajar con tamaños de muestra apropiados para cálculo de correlaciones.

METODOLOGÍA

Se trabaja bajo el supuesto de que se tiene una población finita, generada a partir de una distribución normal bivariada con cinco parámetros: las dos esperanzas (μ_1 y μ_2), las dos varianzas (σ_1^2 y σ_2^2) y la covarianza entre las dos variables de interés (σ_{12}). Indistintamente, se puede utilizar la covarianza o la correlación como quinto parámetro (ya que, ρ_{12} o $\rho = \sigma_{12}/\sigma_1\sigma_2$). Generalmente se enseña a través de la siguiente fórmula, cómo se puede determinar el tamaño muestral *n*

$$P\{|\rho - \widehat{\rho_n}| \leq ee\} \geq 1 - \alpha$$

"solo" se debe despejar *n*. Es así como, motivados en la presentación dinámica para estimaciones de π con el software dedicado a educación estadística Fathom (2005), se logró realizar simulaciones con el software Stata (StataCorp, 2011)

Si ρ está en torno a cero se pueden utilizar resultados asintóticos (Lehmann, 1999) para obtener una expresión explícita para el cálculo de *n*. Paradójicamente, al ser un resultado asintótico, el resultado sería válido solo cuando las muestras fuesen grandes.

Si ρ estuviese próximo a –1 o a +1 la distribución muestral de $\widehat{\rho_n}$ sería asimétrica.

A diferencia del caso univariado, para estimar la media o una proporción (que se puede ver como una media de una variable aleatoria Bernoulli), se requiere hacer supuestos sobre una medida de dispersión de la variable de interés. En el caso bivariado se requiere tener una idea de la correlación entre las variables. El tamaño muestral más conservador será cuando ρ esté en torno de cero. Mientras que los tamaños de muestra más pequeños se obtendrán cuando la correlación poblacional esté próxima a los valores extremos (–1 o +1).

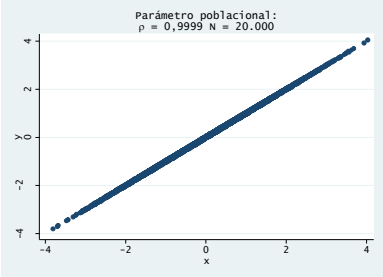
Escenarios de simulación: se generan tres poblaciones finitas lo suficientemente grandes de 20.000 datos con tres coeficientes de correlación: 0,9999; 0,80 y 0. Se obtienen mil muestras aleatorias de tamaño *n* de cada una de estas poblaciones. No siempre se utilizan los mismos valores de *n*; pues cuando ρ es 0,9999 el tamaño muestral debiera ser pequeño (no más grande que 10). Mientras que cuando ρ es cercano a cero el tamaño muestral debiera ser muchísimo más grande (por sobre 1.500 si el error de estimación es 0,05 con un nivel de confianza de 0,95). Se propone la siguiente estrategia: simular para *n* = 100(100)1000. Dependiendo de los resultados que se obtengan de acuerdo al nivel de confianza acordado, acotar el espectro de valores de *n.* Por ejemplo, si se satisfacen el error de estimación y el nivel de confianza establecidos, con *n* entre 300 y 400; realizar nuevas simulaciones para *n* = 300(10)400. Si funciona con *n* entre 360 y 370; simular luego para *n* = 360(2)370. Así por ensayo y error se puede tener un tamaño muestral aproximado a través de las simulaciones. Se trabaja con los *ee* de 0,01(0,01)0,06. Los niveles de significación empíricos que se buscan en los datos

simulados son los típicos: 0,99, 0,95 y 0,90. Esto llevará a tener un tamaño muestral aproximado por medio de simulaciones de Montecarlo.

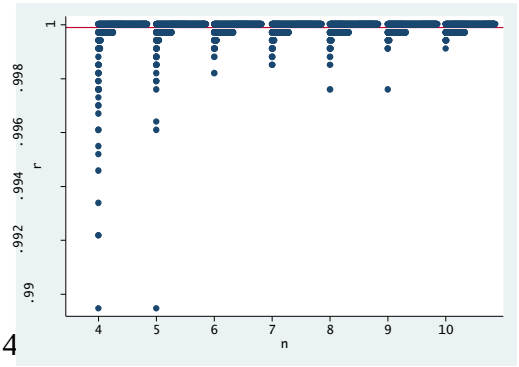
RESULTADOS

Se presenta los resultados de la simulación de las 1.000 m.a. de los tamaños indicados. Específicamente se presenta el promedio, el valor mínimo y el valor máximo.

ρ = 0,9999



n	media	mínimo	máximo
4	.9997754	.9894554	1
5	.9998448	.9895632	.9999997
6	.9998735	.9981711	.9999978
7	.9998767	.9984607	.9999981
8	.9998763	.9975671	.9999952
9	.9998801	.9974917	.9999931
10	.9998870	.9990374	.9999951



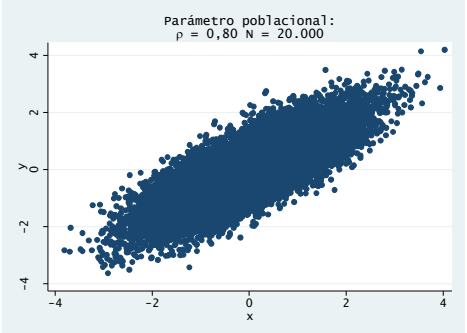
Nótese que prácticamente desde muestras de tamaño 4 de Pearson es insesgado.

Nivel de Confianza empírico con diferentes errores de estimación: ee = $|\rho - \hat{\rho}|$ y ρ = 0,9999.

n	ee 0,01	0,02	0,03	0,04	0,05
4	.999	1	1	1	1
5	.999	1	1	1	1
6	1	1	1	1	1
7	1	1	1	1	1
8	1	1	1	1	1
9	1	1	1	1	1
10	1	1	1	1	1

Esto indica que con un error de estimación de 0,01, un tamaño muestral de 4 está por sobre el nivel de confianza típicamente requerido de 0,95.

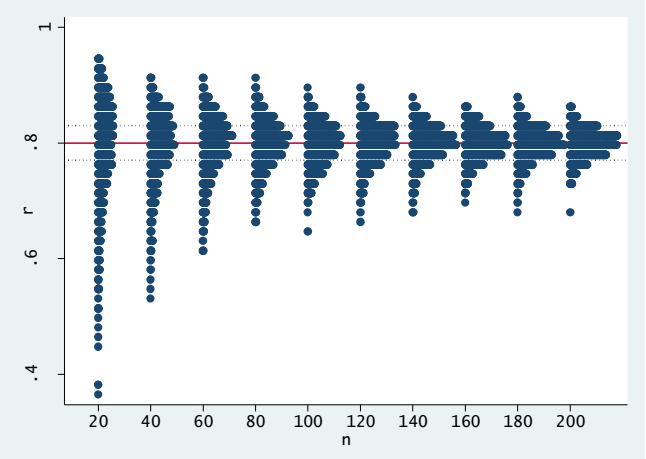
ρ = 0,80



Cuando ρ = 0,80 para los tamaños muestrales el coeficiente de correlación es levemente sesgado.

n	media	mínimo	máximo
20	.7900354	.3728603	.9535018
40	.7944921	.5305964	.9203593
60	.7973925	.6112576	.9071255
80	.7985517	.6680735	.9051461
100	.7968661	.6498818	.8888336
120	.7993504	.6571376	.8931972
140	.7995955	.6778954	.8837649
160	.7987943	.7012782	.8697453
180	.7991040	.6854054	.8816796
200	.7988093	.6846005	.8697265
220	.8002356	.7117779	.8722258
240	.7992160	.7225423	.8616627
260	.7981580	.7162411	.8678038
280	.7996707	.7149944	.8626329
300	.7997760	.7322009	.8587447
320	.7984860	.7367775	.8499449
340	.7991266	.7284509	.8527838
360	.7991005	.7287892	.8514519
380	.7998530	.7229125	.8517954
400	.7995737	.7411988	.8476196
420	.7993455	.7397757	.8511006
440	.7993878	.7442213	.8508561
460	.7989365	.7457233	.8521703
480	.7990113	.7410896	.8488842
500	.7996864	.7382683	.8468214
520	.7989039	.7421258	.8434726
540	.7995090	.7433167	.8425120
560	.7992653	.7328041	.8471696
580	.7988776	.7441533	.8388008
600	.7989543	.7461928	.8410357
620	.7993814	.7373608	.8338401
640	.7996098	.7377715	.8403177
660	.7992149	.7488344	.8374450
680	.7991313	.7293555	.8380374
700	.8002318	.7568400	.8348502
720	.7990090	.7580538	.8410199
740	.7997884	.7496807	.8336023
760	.7993602	.7580241	.8364706
780	.7992650	.7485521	.8390257
800	.7992327	.7461261	.8386211
820	.7990713	.7570381	.8357986
840	.7992887	.7624724	.8354997
860	.7994653	.7548546	.8296733
880	.7992178	.7580241	.8322426
900	.7996954	.7604823	.8321649
920	.8000314	.7517299	.8373899
940	.7981776	.7530555	.8311018
960	.7999011	.7621303	.8280833
980	.7991599	.7629917	.8323323
1000	.7996436	.7616993	.8384855

Se presenta en la gráfica las 1.000 correlaciones generadas en muestras aleatorias de 20(20) 200. La línea roja refiere a la correlación poblacional. Las líneas punteadas marcan los errores de estimación de 0,03.



Nivel de Confianza empírico con diferentes errores de estimación: ee = $|\rho - \hat{\rho}|$ y ρ = 0,80.

	ee				
n	0,01	0,02	0,03	0,04	0,05
20	.087	.177	.269	.369	.454
40	.14	.267	.402	.524	.624
60	.158	.331	.489	.618	.725
80	.189	.386	.55	.678	.795
100	.202	.391	.572	.714	.823
120	.236	.435	.617	.775	.871
140	.271	.513	.69	.83	.909
160	.279	.533	.708	.836	.927
180	.289	.533	.741	.855	.933
200	.307	.577	.786	.904	.954
220	.346	.603	.791	.909	.963
240	.363	.641	.826	.922	.97
260	.369	.627	.817	.916	.972
280	.364	.661	.833	.935	.978
300	.38	.673	.851	.956	.992
320	.386	.684	.874	.952	.989
340	.375	.707	.884	.963	.989
360	.425	.726	.896	.964	.99
380	.438	.739	.899	.968	.993
400	.436	.77	.917	.972	.995
420	.442	.744	.908	.977	.995
440	.429	.764	.914	.974	.995
460	.441	.76	.929	.975	.995
480	.475	.79	.941	.988	.998
500	.483	.793	.946	.989	.999
520	.484	.792	.935	.984	.994
540	.511	.842	.955	.992	.999
560	.48	.809	.951	.991	.998
580	.529	.841	.954	.994	.996
600	.542	.852	.974	.993	.997
620	.532	.871	.969	.994	.997
640	.506	.846	.968	.993	.998
660	.534	.861	.975	.994	.999
680	.538	.878	.976	.998	.999
700	.571	.865	.981	.998	1
720	.536	.852	.962	.996	1
740	.603	.886	.984	.997	.999
760	.578	.889	.977	.998	1
780	.547	.885	.981	.997	.999
800	.586	.906	.986	.997	.999
820	.573	.883	.975	.999	1
840	.594	.898	.99	1	1
860	.598	.908	.987	.998	1
880	.593	.898	.989	.999	1
900	.606	.911	.988	1	1
920	.607	.935	.992	.997	1
940	.616	.917	.984	.996	1
960	.639	.932	.996	1	1
980	.643	.939	.996	1	1
1000	.62	.906	.991	1	1

Si el nivel de confianza fuese 0,95 para un error de estimación de 0,05; 0,04 y 0,03, los tamaños muestrales estarían en torno de 200, 300 y 540, respectivamente. Para errores de estimación de 0,02 y 0,01 el tamaño muestral para un nivel de confianza de 0,95 supera las 1.000 unidades. En la tabla previa se destaca en amarillo cuando empíricamente a través de simulaciones se logra el nivel de confianza empírico de 0,95.

CONCLUSIONES

Se concluye que, a través de las simulaciones de Montecarlo, se puede determinar el tamaño de muestra adecuado para cuando se requiere estimar el coeficiente de correlación de

Pearson y esto será muy ventajoso para el entendimiento de la necesidad de trabajar con el tamaño de muestra apropiado para cuando se trabaja con correlaciones.

Además, se demuestra empíricamente que el tamaño de la muestra no depende de la variabilidad de las dos variables de interés. El tamaño de la muestra depende solo de la correlación poblacional. En otras palabras, el tamaño de muestra será mayor cuando ρ esté próximo a cero y disminuirán los tamaños muestrales a medida que $|\rho|$ tienda a 1.

Al igual que lo expuesto por Saldanha (2004), se pretende que los alumnos de manera espontánea sean capaces de concebir el significado tanto de correlación como del manejo apropiado de tamaños muestrales para cuando se trabaja con el coeficiente.

REFERENCIAS

Anderson, R., Doherty, M., Berg, N., and Friedrich, J. (2005). Sample size and the detection of correlation—A signal detection account: Comment on Kareev (2000) and Juslin and Olson (2005). *Psychological Review, 12*(1), 268-279.

Bhattracharyya, G. and Johnson, R. (1977). *Statistical concepts and methods*. NY, New York: John Wiley and Sons.

Godino J. y Batanero, C. (1994). Significado institucional y personal de los objetos matemáticos. *Recherches en Didactique de Mathematiques, 14*(3), 325-355.

Kareev, Y. (1995). Through a narrow window: Working memory capacity and the detection of covariation. *Cognition, 56,* 263–269.

Kish, L. (1965). *Survey sampling.* NY, New York: John Wiley and Sons.

Fathom (2005). *Dynamic data software.* Key Curriculum Press.

Lehmann, E. (1999). *Elements of large-sample theory.* NY, New York: Springer-Verlag.

Saldanha, L. (2004). *"Is this sample unusual?": An investigation on students exploring connections between sampling distributions and statistical inference*. Tesis Doctoral. Vanderbilt University.

StataCorp (2011). *Stata: Release 12. Statistical software.* TX, College Station: StataCorp LP.