InnOvaciOnes de NegOciOs 11(22): 235 – 268
© 2014 UANL, Impreso en México (ISSN 2007-1191)
Fecha de recepción: 26 de septiembre de 2014. Fecha de aceptación: 17 de noviembre de 2014.

Los tamaños de las muestras en encuestas de las ciencias sociales y su repercusión en la generación del conocimiento (Sample sizes for social science surveys and impact on knowledge generation)

Juan Rositas Martínez
Universidad Autónoma de Nuevo León, Facultad de Contaduría Pública y Administración, San Nicolás de los Garza, N.L., México.
Email: jrositasm@yahoo.com

Keywords: confidence intervals, Cronbach's alpha, effect size, factor analysis, hypothesis testing, sample size, structural equation modeling

Abstract. The purpose of this paper is to contribute to fulfilling the objectives of social sciences research such as proper estimation, explanation, prediction and control of levels of social reality variables and their interrelationships, especially when dealing with quantitative variables. It was shown that the sample size or the number of observations to be collected and analyzed is transcendental for the adequacy of the method of statistical inference selected and for the impact degree achieved in its results, especially for complying with reports guidelines issued by the American Psychological Association. Methods and formulations were investigated to determine the sample sizes that contribute to have good levels of estimation when establishing confidence intervals, with reasonable wide and relevant and significative magnitudes of the effects. Practical rules suggested by several researchers when determining samples sizes were tested and as a result it was integrated a guide for determining sample sizes for dichotomous, continuous, discrete and Likert variables, correlation and regression methods, factor analysis, Cronbach's alpha, and structural equation models. It is recommended that the reader builds scenarios with this guide and be aware of the implications and relevance in scientific research and decision making of the sample sizes in trying to meet the aforementioned objectives.

Palabras clave: análisis factorial, intervalo de confianza, alpha de Cronbach, modelación mediante ecuaciones estructurales, pruebas de hipótesis, tamaño de muestra, tamaño del efecto

Resumen. El propósito del presente documento es contribuir al cumplimiento de los objetivos de la investigación en las ciencias sociales de estimar, explicar, predecir y controlar niveles

de variables de la realidad social y sus interrelaciones, en investigaciones de tipo cuantitativo. Se demostró que el tamaño de la muestra o la cantidad de observaciones que hay que recolectar y analizar es trascendente tanto en la pertinencia del método de inferencia estadístico que se utilice como en el grado de impacto que se logre en sus resultados, sobre todo de cara a cumplir con lineamientos emitidos por la Asociación Americana de Psicología que es la que da la pauta en la mayoría de las publicaciones del área social. Se investigaron métodos y formulaciones para determinar los tamaños de muestra que contribuyan a tener buenos niveles de estimación al momento de establecer los intervalos de confianza, con aperturas razonables y con magnitudes de los efectos que sean de impacto y se pusieron a prueba reglas prácticas sugeridas por varios autores lográndose integrar una guía tanto para variables dicotómicas, continuas, discretas, tipo Likert y para interrelaciones en ellas, ya se trate de análisis factorial, alpha de Cronbach, regresiones o ecuaciones estructurales. Se recomienda que el lector crear escenarios con esta guía y se sensibilice y se convenza de las implicaciones y de trascendencia tanto en la investigación científica como en la toma de decisiones de los tamaños de muestra al tratar de cumplir con los objetivos de la que hemos mencionado.

Introducción

Un problema que se le presenta a todo investigador de las ciencias sociales es que sus estudios además de ser documentales (revisión de literatura), por ser también observacionales, incluyen un estudio de campo. Un problema que es crucial en estos estudios de campo, es que una vez que ha desarrollado el instrumento (cuestionario) debe determinar la muestra apropiado a la que lo aplicará. Su instrumento de investigación, generalmente contará con apartados de diversos tipos, en los que se presentan reactivos que equivalen a una diversidad de variables que serán sujetas posteriormente a varios procedimientos estadísticos de los que se espera arrojen resultados que sean considerados una contribución al conocimiento.

En los diversos apartados del cuestionario hay generalmente variables continuas, dicotómicas, tipo Likert, y generalmente en las etapas de análisis de datos, estas variables serán sujetas a análisis factoriales, de confiabilidad mediante el alfa de Cronbach, a análisis de regresión y correlación múltiple, e incluso a modelación mediante ecuaciones estructurales.

La cuestión crítica es que tradicionalmente en varias áreas de investigación, en el apartado de Resultados, se han estado presentando solamente análisis descriptivos y pruebas de hipótesis, sin dársele la

importancia debida al tamaño de muestra, ni presentando análisis estadísticos en los que el tamaño de muestra si llega a ser concluyente. En la actualidad por estarse exigiendo la estimación de parámetros poblacionales mediante intervalos en la publicación de resultados, que además de un buen nivel de confianza, presenten una apertura no muy amplia para ser útiles, la determinación del tamaño de la muestra llega a ser crucial. Esto es, la muestra no debe ser ni demasiado escaza de tal forma que reste trascendencia a los resultados de los distintos apartados del cuestionario, ni demasiado abundante que llegue a poner en peligro la viabilidad del proyecto.

Para darnos cuenta de la magnitud del problema, en el último par de décadas antes de la llegada del milenio 2000, en una muestra de 594 artículos del American Journal of Public Health, y 110 artículos de la revista Epidemiology, el 63% hacían referencia solo a pruebas de hipótesis y al valor-p, mientras que solamente el 10% presentaban intervalos de confianza. Ante la presión de los editores, la referencia del 63% bajó al 5% y el reporte de intervalos subió del 10% al 54%. Como veremos más adelante, mediante simulaciones o generación de escenarios a partir de datos reales, además de la variabilidad y el nivel de confianza, un tamaño de muestra adecuado hace la diferencia para que el intervalo realmente resulte útil o lo suficientemente preciso para que pueda ser considerado una contribución al conocimiento Schrout (1997).

El estado de la literatura, en cuanto a metodologías para estimar tamaños de muestras para las diversas variables y métodos de análisis estadísticos a los que se sujetaran los datos, además de no ser muy abundantes sobre todo las fácilmente manejables por el investigador, están algo dispersas, y no hemos encontrado un documento que sintetice en forma clara y guíe al investigador en una manera sencilla a determinar los distintos tamaños de muestra para cada apartado, y apegarse al tamaño de muestra que cubra el mínimo de confianza deseada para todo el instrumento.

El presente documento es un esfuerzo de integrar esta guía que apoye a su vez a los investigadores para que sus resultados lleguen a ser considerados metodológicamente y estadísticamente contribuciones al conocimiento.

En el siguiente apartado se hace una revisión de la literatura de metodologías para determinar la trascendencia de la determinación de

tamaños de muestras para los apartados típicos de los instrumentos de investigación social, y para los análisis de datos más frecuentes como son el análisis factorial, el alfa de Cronbach, la regresión y correlación y los sistemas de ecuaciones estructurales. Más adelante se detalla en el apartado de resultados la aplicación de estas metodologías enfocada a casos concretos, para finalmente terminar con una síntesis y conclusiones y recomendaciones de las metodologías aquí analizadas e interpretadas.

Revisión de literatura y método

En el presente apartado hacemos una revisión de la literatura en la que se realza la trascendencia de la determinación de tamaños de muestra. El método que hemos usado es hacer una investigación de literatura reciente en la que se aborde en una forma clara metodologías para determinar tamaños de muestras para diversos tipos de variables (dicotómicas, continuas, Likert) y los varios métodos de análisis estadísticos que hemos mencionado como más utilizados (Análisis Factorial, alfa de Cronbach, correlación y regresión, y sistemas de ecuaciones estructurales).

Trascendencia de la determinación de un buen tamaño de muestra

El objetivo de toda investigación social, observacional y de tipo cuantitativo es estimar (a partir de muestras inferir valores y relaciones en la población), explicar (cómo es la realidad y cómo se relacionan sus elementos), predecir (prever el funcionamiento futuro) e incluso controlar (tomar decisiones y actuar) una variable o una interrelación de variables Sierra-Bravo, R. (1983, pág. 41). Para ello podemos recurrir a los dos métodos básicos de la inferencia: la prueba de hipótesis y la estimación de parámetros.

La Asociación Psicológica Americana (APA, por sus siglas en inglés), desde hace tiempo ha indicado a través de un grupo de trabajo (taskforce), que en el apartado de Resultados, de los papers o artículos del campo de la psicología que se envíen a journals para su publicación, no deben incluir solamente la prueba de hipótesis, sino también el tamaño del efecto y la estimación de parámetros mediante intervalos.

En el reporte inicial de este grupo de trabajo de APA (1996) formado por una docena de especialistas en Inferencia Estadística, hay una declaración sobre la no prohibición de la utilización de las pruebas de hipótesis ni de los valores-p en la investigación y publicación del área de la psicología; prohibición que efectivamente se observó en otras disciplinas. El objetivo de este grupo de trabajo fue específicamente evaluar el papel de las pruebas de hipótesis nulas en la investigación psicológica, y la popularización de los procedimientos cuantitativos cada vez más al alcance con el uso de la computadora (pág. 2)

Respecto a recomendaciones relativas al lugar que ocupan las pruebas de hipótesis y otras formas de análisis de resultados, el grupo de trabajo recomienda un mejoramiento en la caracterización de los resultados, que vaya más allá del simple cálculo del valor-p para contrastar la hipótesis nula. Estas otras formas de análisis incluyen tanto la magnitud como la dirección del tamaño del efecto, por ejemplo en la diferencia de medias, o en los coeficientes de regresión y correlación, las razones de momio, e indicadores de mayor complejidad para el tamaño del efecto. El grupo de trabajo indica también que debe presentarse rutinariamente los intervalos de confianza de estos tamaños del efecto.

Posteriormente Wilkinson, en 1999, a nombre de ese grupo de trabajo, publica un reporte detallado relativo a guías y explicaciones sobre los métodos estadísticos que se sugieren para las publicaciones de investigaciones psicológicas. Indica que no solamente se haga la prueba de hipótesis y se reporte el valor-p, sino que se presente el tamaño del efecto en términos tanto prácticos como teóricos, por lo que los coeficientes de regresión y las diferencias de medias, se reporten no solamente en términos estandarizados, sino que también en términos no-estandarizados para apreciar el grado de impacto de cada variable independiente sobre la dependiente. Adicionalmente, en la medida de lo posible, deben de reportarse intervalos de confianza para las correlaciones y para otros coeficientes de regresión (pág. 6).

Fiedler et al. (2004), reportan que en las publicaciones del área médica, esta reforma de la manera de analizar los resultados fue más drástica. Shrout (1997), citado en Fiedler et al. (2004) indica que algunos editores del área médica exigían a los autores que enviaban artículos, para su posible publicación, que quitaran cualquier referencia a pruebas de

hipótesis y valores-p, o de lo contrario que buscaran otras revistas en donde publicar sus artículos (pág. 3). En algunas revistas como en American Journal of Public Health y en otra revista de epidemiologia, en el lapso de alrededor de una década, los artículos que solo hacían referencia a pruebas de hipótesis y valores-p, bajó del 63% al 5% y los artículos que presentaban intervalos de confianza subió del 10% al 54%, aunque gran cantidad de autores no hacían referencia a estos intervalos en el apartado de análisis de resultados. Una conclusión de Shrout (1997) fue que habría que estudiar si los investigadores comprendían realmente la interpretación correcta de los intervalos de confianza o que quizás habría necesidad de mejor la educación estadística y las interpretaciones correctas y sus implicaciones prácticas.

Esto tiene mucho sentido, ya que al hacer pruebas de hipótesis, generalmente la hipótesis nula se declara en relación a un valor cero, o a una diferencia cero entre parámetros de dos variables o más. En la mayoría de los casos, la expectativa del investigador en cuanto al resultado de la prueba de hipótesis es en el sentido de que "se rechace la hipótesis nula" y que por lo tanto se concluya que, con la evidencia obtenida, el valor del parámetro no sea cero. Si hasta aquí llega el análisis de resultados, ninguno de los objetivos de la ciencia basada en evidencias estadísticas, mencionados arriba, de estimar, explicar, predecir o controlar se estaría cumpliendo. Incluso si a partir del estadístico de la muestra hacemos solamente una estimación puntual, ésta no conlleva metodológicamente ninguna indicación en relación a algún nivel de confiabilidad o porcentaje de confianza. Por lo anterior, es justificable la sugerencia de que las publicaciones incluyan no solo pruebas de hipótesis, sino estimaciones mediante el establecimiento de un intervalo que conlleve un cierto nivel de confianza.

La cuestión es que si no hacemos, en el diseño de la investigación, un cálculo previo del tamaño de muestra para que luego en la fase de resultados podamos hacer estimaciones adecuadas de intervalos de confianza, nos podemos llevar sorpresas desagradables. Un de estas sorpresas es que aunque la estimación de un coeficiente de correlación, resulte considerablemente alto, como mostramos en un ejemplo real más adelante, con un valor de 0.76, el intervalo de estimación, con un nivel de confianza del 95%, tenga como límite inferior 0.412 y como límite superior 0.917. Un intervalo así de amplio, parecería más bien un intervalo de alta incertidumbre más que un intervalo de confianza.

Necesidad de aplicar varios métodos para el tamaño de la muestra

De lo considerado en párrafos del apartado anterior, es con lo que se justifica, que dentro de la fase de planeación o diseño de una investigación, se determine de antemano el tamaño de muestra apropiado, además del tipo de muestreo que se realizará. Aquí hay que recalcar también que el muestreo debe ser aleatorio, para que tenga fundamento o validez, tanto cualquier prueba de hipótesis, como cualquier estimación mediante inferencia estadística. Si la investigación social, por la problemática que lleva implícita en sus estudios de campo, no permite que la muestra sea estrictamente aleatoria, debe apegarse lo más posible a este requisito para que los resultados no se tomen con reservas o como un mero ejercicio del "mundo académico", sin aplicación confiable al "mundo real". Por lo tanto, si en una investigación social se menciona que el muestreo es no-probabilístico y de conveniencia, - que si fuera así, lo correcto es mencionarlo - todos los resultados que se presenten derivados de inferencias estadística con los datos recolectados, van a estarse tomando con las debidas reservas.

En gran cantidad de investigaciones, se cumple con la determinación de un tamaño de la muestra, pero solamente para una cierta pregunta, o ítem, digamos, por ejemplo, para el género del ocupante del puesto o del respondiente. En este caso, por tratarse de una variable dicotómica, se aplica únicamente la determinación de la muestra para una variable binomial, que en algunos softwares se le denomina variable atributo. Pero si mediante otros ítems del cuestionario se pretende medir variables cuantitativas, variables tipo Likert, u obtener alfas de Cronbach, determinación de constructos mediante análisis factorial, correlaciones y regresiones entre variables o incluso integrar un modelo de ecuaciones estructurales, y para estos ítems o métodos no se determinaron previamente los tamaños de muestra adecuados y no se seleccionó el valor máximo de ellos, para que en todos estos ítems cumplan con el nivel mínimo de confianza previsto, luego al hacer estimaciones mediante intervalos, estos intervalos serán demasiado abiertos para que puedan considerarse una contribución al conocimiento o para que con base en ellos pueda tomarse una decisión fundamentada racionalmente. Por otra parte, si el tamaño de muestra no se considera el adecuado, se cuestionará la validez o pertinencia de uso de los métodos de análisis mencionados.

Resultados sobre determinación de tamaños de muestra

En el presente apartado se presentan y se ilustran con aplicaciones concretas las metodologías encontradas para determinar los tamaños de muestras para variables y procedimientos de análisis diversos.

Para determinar el tamaño adecuado de la muestra que se recolectará en la investigación de campo, esto es, el número de encuestas, que permita que para todos los ítems relevantes del cuestionario, correspondientes a diferentes tipos de variables, o de interrelaciones entre ellas, lleguen a obtenerse intervalos de estimación no muy abiertos y con buenos niveles de confianza, es necesario determinar el tamaño de muestra para cada uno de estos tipos de variables o interrelaciones, y seleccionar el tamaño de muestra mayor. Por ello, en este apartado vamos a presentar la fórmula y el procedimiento para el tamaño de muestra de preguntas relacionadas con las diversas variables que hemos mencionado: dicotómicas, variables continuas, y de escalas Likert; y en cuanto a grupos de ítems interrelacionados: tamaños de muestras para alfas de Cronbach, análisis factorial, regresión múltiple y ecuaciones estructurales.

Una buena estimación de un parámetro mediante un intervalo de confianza, es aquel que sin ser muy abierto, tiene un buen nivel de confianza, de al menos 95%. El tamaño de la muestra estará en función además de la apertura tolerada del intervalo y del nivel de confianza, de la varianza del parámetro para el que se quiera hacer tal tipo de estimación. Intuitivamente podemos predecir que al querer estimar un parámetro para dos poblaciones, dado un mismo nivel de confianza y un mismo tamaño de población, si la primera población tiene una varianza menor en comparación con la segunda, la primera requerirá de tamaño de una muestra menor.

Disculpando el lenguaje coloquial de este párrafo, al referirnos al dicho del folklore mexicano que afirma que “para muestras, con solo un botón basta”, esto es obviamente cierto para un rosal o una prenda de vestir en que todos los botones son prácticamente iguales, y que por lo tanto la varianza es prácticamente cero. De igual forma, intuitivamente podemos anticipar que si en una población dos variables tienen una alta correlación, digamos alrededor de 0.95, la muestra será menor para detectar tal correlación en comparación con la muestra necesaria para detectar una baja correlación, digamos menor al 0.50 en otro par de variables en otra población.

Preguntas relacionadas con atributos dicotómicos

La fórmula más simple para calcular el tamaño de muestra, de una variable dicotómica, se presenta cuando la población es prácticamente infinita y es la siguiente.

$$n = \frac{Z_{\frac{\alpha}{2}}^{2} P(1-P)}{e^2} \quad (1)$$

En la ecuación (1),

P es el porcentaje, estimado tentativamente antes del muestreo, de elementos de la población en los que esperamos que se presente un cierto atributo; por ejemplo, que el 90% de los gerentes sean de género masculino. Automáticamente *Q,* por ser el valor "contrario", está determinado como complemento a 100%; esto es, Q = 1 – P.

e es el error tolerado, o los puntos porcentuales con que se construirán los límites del intervalo de estimación. Continuando con el ejemplo, del caso de los gerentes, si esperamos que el porcentaje poblacional p o *P* corresponda a un valor igual al 90% y la estimación la queremos hacer con *p* ± 5% de tolerancia, por ejemplo, el valor *e* vale, precisamente, esos 5 puntos porcentuales que estaremos sumando y restando a la proporción muestral (p) que nos resulte para obtener el límite superior e inferior del intervalo de estimación. Podríamos llamarle, a ese valor 5, precisión deseada o distancia en puntos porcentuales entre P y los límites del intervalo, y asignarle el símbolo *d* y no confundir el concepto con el error propio del muestreo que usa este mismo símbolo *e* pero que es una variable aleatoria.

$Z_{a/2}$ es un valor de la distribución normal estandarizada, que está en función de la confianza que queramos tener en que el intervalo de estimación que vayamos a formar prediga o estime correctamente el parámetro poblacional. Generalmente, la confianza se fija en 95%, aunque en algunos casos se intenta con el 99%. Como el tamaño de muestra generalmente sube con costos prohibitivos al acercarnos al 99% de confianza, se opta por el 95%. Un 95% significa que se esperaría, o se confía con bases estadísticas, que de cada 100 intervalos que se formaran a partir de 100 muestras de tamaño *n*, 95 de

> estos intervalos acertarían en contener dentro de sus límites al porcentaje poblacional. Por lo anterior, el investigador tiene la confianza que el intervalo que forme a partir de su muestra sea uno de los 95 de cada 100 intervalos que contiene, o predice correctamente el parámetro poblacional y que desafortunadamente puede ser uno de los 5 de cada 100 intervalos que no contienen el parámetro poblacional. Para un nivel de confianza del 95%, $Z_{a/2}$ = 1.96, y para 99% $Z_{a/2}$ = 2.54 Lind et al. (2012, p. 782).

La fórmula de la ecuación (1) se deriva del conocimiento previo del Teorema del Límite Central que nos dice que los errores propios del muestreo se comportan como una distribución normal con media cero y desviación estándar igual a $\sigma/\sqrt[2]{n}$. Partiendo de las siguientes tres elementos de juicio: a) Para una variable dicotómica al no conocer s utilizamos, utilizamos *S*, b) la desviación estándar muestral, *S*, es igual a p x q, que son los porcentajes muestrales, y c) que para un 95% de confianza el $e = 1.96\ s\ /\sqrt[2]{n}$, es sencillo el ejercicio algebraico de obtener *n,* a partir de esta expresión del error, y que se muestra en ecuación (1). La tarea se torna laboriosa, para poblaciones finitas, y para las que queremos elaborar varios escenarios de tamaños de muestra.

Usamos el término confianza, porque por basarnos en un valor estadístico de una muestra "p", no sabemos a ciencia cierta, si el intervalo que formemos, mediante la ecuación (2), contendrá, estimará o predecirá correctamente el porcentaje poblacional. Aquí ya no usamos el término probabilidad, porque esta ecuación (2), no contiene ninguna variable aleatoria si no que hace referencia a dos valores concretos; el primero (p) el porcentaje de la muestra que presenta el atributo de interés, y el segundo (d) el valor en puntos porcentuales con los que se formarán tanto el límite superior como el inferior del intervalo de estimación, valor que fue fijado por el investigador.

$$P = \ p \pm d \qquad (2)$$

Para una población finita de tamaño N, el cálculo de tamaño de muestra se deriva de la siguiente ecuación (3),

$$e = \frac{Z\frac{\alpha}{2}P(1-P)}{\mathrm{n}}\sqrt{\frac{(N-1)}{(N-n)}} \quad (3)$$

donde:

e es el valor del error tolerado, o la apertura del intervalo, en puntos porcentuales.

$Z_{a/2}$ es un valor de la distribución normal estandarizada correspondiente a un cierto nivel de confianza, que se explicó en ecuación 1,

P es la proporción preliminar de la proporción preliminar, fijada por el investigado, y explicada en ecuación 1, y en la que espera que se presente un cierto atributo de interés

(1- P) también llamada Q, es la proporción que espera el investigador en la que no se presente el atributo de interés.

N es el tamaño de la población

n es el tamaño de la muestra

Al despejar algebraicamente el tamaño de la muestra "n" es

$$n = \frac{NPQ}{(N-1)\left(\frac{e}{z}\right)^2 + PQ} \quad (4)$$

Si quisiéramos tener una confianza del 100%, en la ecuación 2, z tendería a infinito, el valor $\left(\frac{e}{z}\right)^2$ se convertiría prácticamente en cero y por consecuencia el primer sumando del denominador se hace cero, y por consecuencia:

$$n = \frac{NPQ}{PQ} \quad (5)$$

Por lo que, al eliminar PQ del numerador y denominador, quedaría que para una confianza del 100% *n = N,* por lo que tendríamos que aplicar el cuestionario a la población total.

Se ha determinado que para poblaciones mayores a 160,000, las dos fórmulas de determinación de "n" tanto para poblaciones finitas como infinitas, arrojan el mismo tamaño de muestra.

Por reflexión en todo lo argumentado en este apartado, el tamaño de la muestra estimado para una variable dicotómica depende o está en función de cinco valores:

(1) Tamaño de la población igual a N,
(2) Valor estimado de la proporción poblacional igual a P, derivado de una muestra piloto o de un valor encontrado en un estudio similar previo
(3) Varianza estimada de la población igual a P x Q, tomando los valores del punto 2,
(4) Precisión deseada *e*, también referida como *d*, que es fijada por el investigador y
(5) Confianza deseada, reflejada en valor Z, que también es fijada por el investigador.

Cuando, en el cálculo del tamaño de muestra, tenemos la mayor incertidumbre en cuanto los valores de P y Q, porque no hemos hecho una muestra piloto, o no tenemos resultados de un estudio previo, ni la menor sospecha de su valor, P= 050, y además nos ponemos muy ambiciosos y nos fijamos una precisión (e o d) muy alta, digamos del 1% y una confianza también muy alta, digamos 99%, como consecuencia el tamaño de muestra resulta considerablemente grande, por lo que generalmente el investigador termina bajándole tanto a la precisión como a la confianza, y en la medida que se vaya teniendo más información de la varianza (PQ), se va alimentando ésta para recalcular el tamaño de la muestra. Por todo lo anterior, se sugiere hacer simulaciones o crear escenarios sobre el tamaño n, con aplicaciones en Excel que existen para tal efecto.

El tamaño de muestra calculado, a partir de una primera estimación de la desviación estándar poblacional, no es el definitivo ya que en la medida en que en el desarrollo de la encuesta de campo, la desviación estándar vaya difiriendo de la desviación estándar prevista, consecuentemente el tamaño muestral que iremos recalculando también irá cambiando aunque convergiendo a un valor. Una última puntualización en este apartado es que si se tienen varias preguntas de tipo dicotómico, el cálculo de tamaño de muestra, para este tipo de preguntas debe hacerse para la pregunta que presente o de la que se estime una mayor varianza, reflejada como dijimos en P x Q.

Preguntas relacionadas con variables cuantitativas: continuas o discretas

La fórmula para el cálculo del tamaño de muestra, para este tipo de variables, al tratarse de una población prácticamente infinita es:

$$n = \frac{Z\ \alpha/2\ s^2}{\mathrm{d}^2} \tag{6}$$

Mientras que para una población finita es:

$$n = \frac{Ns^2}{(N-1)\left(\frac{d}{z}\right)^2 + s^2} \tag{7}$$

Similarmente con las ecuaciones (6) y (7) del apartado anterior,

Z es un valor de la distribución normal estandarizada, que está en función de la confianza que queramos tener en que el intervalo de estimación que vayamos a formar al predecir o estimar el parámetro poblacional. Generalmente, la confianza se fija en 95%, aunque en algunos casos se intenta con el 99%.

s es la desviación estándar muestral estimada. Puede estimarse ya sea: a) con base en un estudio previo similar, b) Mediante un muestreo piloto, digamos de 30 elementos, inicialmente, c) A partir de una idea que se tenga del RANGO TOTAL (Valor máximo menos mínimo de los valores poblaciones), y dividiendo este rango entre 6. El valor seis se deriva del supuesto de que la población sigue una distribución normal y que dentro de la media ± 3 σ, está prácticamente el 100% de todos los valores de la población, o sea de que el límite máximo es igual al valor mínimo más 6 σ.

d es el error tolerado, o la distancia en que se situará cada uno de los límites del intervalo de estimación en relación a la media muestral y expresada en la mismas unidades que tengan los valores observados Continuando con el ejemplo de una investigación sobre gerentes, si esperamos que el sueldo promedio poblacional se encuentre o se sitúe a partir de la media muestral, y dentro de los límites de un intervalo formado por $\bar{x}$ ± d, por ejemplo, el valor d, pudiera ser $ 2,500, y es fijado por el investigador, aunque implícitamente d inicialmente se

exprese como porcentaje de la media; esto es, si el sueldo promedio se estima en $50,000, y queremos un ±5% de precisión , d resulta ser $ 2,500.

Z, como hemos dicho es fijado por los investigadores generalmente con el valor 1.96, que se deriva de la distribución normal para una confiabilidad del 95%.

N, corresponde al tamaño de la población.

El tamaño de muestra calculado, a partir de la estimación de la desviación estándar poblacional, no es el definitivo en la medida en que en el desarrollo de la encuesta de campo, la desviación al irla recalculando vaya difiriendo de la desviación estimada inicialmente y consecuentemente el tamaño muestral también vaya cambiando aunque convergiendo a un valor. Al igual que en el apartado anterior, una última puntualización es que si se tienen varias preguntas de tipo cuantitativo, el cálculo de tamaño de muestra, para este tipo debe hacerse para la pregunta que presente o se estime tenga una mayor varianza, dentro de las que se consideren más relevantes.

Escalas de Likert o de intensidad

En años recientes se ha acentuado la inquietud respecto a que si es correcto o no es correcto el procesamiento estadístico que se ha estado dando en las investigaciones a ítems de instrumentos de investigación formulados como escalas Likert (Norman, 2010, Moral de la Rubia, 2006). Esto debido al uso cada vez más frecuente de las escalas Likert y su procesamiento con softwares cada vez más accesibles para los investigadores, que incluyen análisis factorial, alfa de Cronbach y ecuaciones estructurales que tienen su base en constructos formados precisamente a partir de escalas tipo Likert o de intensidad, que al final de cuentas se siguen considerando como ordinales.

Por el lado de los que desalientan su procesamiento con los softwares mencionados tenemos a revisores de *journals o* publicaciones científicas, o investigadores de metodologías estadísticas que opinan que las escalas Likert por ser escalas de nivel ordinal, no deben ser sujetas a procedimientos de análisis estadísticos a los que sí pueden someterse las variables de

intervalo o de razón; procedimientos que van desde el simple cálculo de promedios, hasta análisis factoriales o de regresión múltiple (Moral de la Rubia, 2006; Jamieson, 2004). Estas críticas son salvables cuando la Escala Likert cumple con ciertos requisitos.

De acuerdo a Moral de la Rubia (2006) "Los reactivos de una escala Likert, que suelen tener menos de nueve puntos de recorrido, aun asumiendo un supuesto de continuidad, no se ajustan a una curva normal por su escasa amplitud, por lo que son variables ordinales y no deberían considerarse de intervalo" , puntualizado que, si las escalas "tienen una amplitud mayor a nueve unidades y si se ajustan a una curva normal, asumiendo un supuesto de continuidad, se les trata, a nivel estadístico, como escalas de intervalo", añadiendo que al no cumplir con este requisito no son aptas ni para el análisis factorial ni para el análisis de correlación y regresión. En opinión del autor citado, la consecuencia de no respetar este supuesto, es que se resta potencia a la prueba y los resultados deben tomarse como aproximados, aunque menciona que las escalas Likert de cinco opciones ha sido parte de la rutina y de la tradición psicométricas (pág. 395).

Otro crítico es Jamieson (2004,) citado por Norman (2010) quien declara que los intervalos entre los valores de las respuestas ordinales no podemos tomarlos como iguales, y por consecuencia si el investigador utiliza algún método paramétrico de la estadística descriptiva o inferencial inadvertidamente en datos ordinales, pero que haya sido diseñado solo para variables de intervalo (o de razón), corre un alto riesgo de llegar a una conclusión equivocada. Lo cuestionable, como afirma Norman (2010) es que aunque estrictamente la crítica de Jamieson es cierta, la duda es si realmente habrá un alto riesgo de llegar a conclusiones estadísticas erróneas dado que estos métodos son métodos robustos, en el sentido que dan respuestas correctas incluso cuando los supuestos sean violados (pág. 3). Respecto a la robustez de algunos métodos estadísticos multivariados "La r [coeficiente de correlación] es efectivamente insensible a violaciones extremas de los supuestos de normalidad y de tipos de escalas" Havlicek y Peterson (1976), citado en Norman (2010).

Norman (2010) menciona además que la atención a estas críticas es importante porque gran cantidad de estudios utilizan escalas tipo Likert; y de cinco opciones, como afirma Moral de la Rubia (2006) en un párrafo anterior. Esto no es un asunto trivial, porque si las consecuencias adversas del uso de

métodos estadísticos en investigaciones con variables ordinales o de tipo Likert para las que no se haya verificado el supuesto de normalidad, son ciertas, entonces el 75% de la investigación educacional y médica, y quizás de otras áreas, no tendrían validez científica.

Algunos autores opinan que además de tratarse de variables ordinales de origen, no importa de dónde provengan los números, en tanto que los números se distribuyan razonablemente bien. Incluso si utilizamos la escala Likert: 1= Definitivamente en desacuerdo, 2=En desacuerdo, 3=Sin opinión, 4=De acuerdo, 5=Moderadamente de acuerdo, y teóricamente no haya garantía de que las distancias entre estos valores sean iguales, y por lo tanto implícitamente no sea una variable de intervalo, a la computadora le da lo mismo en referencia a las descripciones que se le hayan puesto a las opciones después del símbolo de igual Gaito (1980), y en tanto que los valores se distribuyan razonablemente bien.

La cuestión es que de hacer correlaciones con dos variables de este tipo, y tratamos de ver el impacto de una variable sobre otra, en el apartado previo del reporte de investigación referente a la caracterización de los participantes, no tendría mucho significado el valor del promedio que resultara por ejemplo 3.0, o incluso 3.2. En mi opinión las descripciones verbales, después del signo de igual deben corresponder a lo que pudiéramos llamar escala semántica de intensidad, en la que los valores no solo hagan referencias a categorías, si no a un aumento de tono en la fuerza con que se presenta la variable, y haya una mayor consistencia en las respuestas por contarse con esa guía semántica. Incluso algo que pudiera considerase subjetivo como *0= sin problema, 2= problema leve, 4= problema moderado, 6= problema severo, 8= problema muy serio, 10= problema lo más serio posible*, pudiendo responderse con valores intermedios, es una mejor escala que la tradicional escala de Likert, cuando no se trata de meras opiniones, sino de una intensidad que puede ser ordenadamente creciente. Otra sugerencia importante es que el número de opciones, sean números pares como 6, 8 y 10, para evitar que por cuestiones de comodidad del entrevistado y no esforzarse en pensar al responder, haya la tendencia subjetiva al centralismo, esto es, a contestar la mayor de las veces el valor 4, por tratarse de una escala impar, por ejemplo que vaya del 1 al 7.

Para apoyar la robustez de los métodos paramétricos, Norman (2010) ha hecho una demostración, de que da lo mismo aplicar una técnica paramétrica como el coeficiente de Pearson a un grupo de datos obtenidos de escalas del 0 al 10 (considerada como continua) y luego aplicar una técnica no-paramétrica como lo es el coeficiente de correlación de Spearman a los mismos datos pero reagrupados en una escala de 4 opciones, en las que 1={1}, 2={2}, 3={3,4 5}, 4= {6,7,8,9, 10}. Con las anteriores indicaciones a los datos de una muestra de 93 pacientes, les calculo 64 correlaciones tanto de Pearson como de Spearman. La nueva correlación de Pearson resultante de este grupo de 64 correlaciones le dio un coeficiente de correlación de 0.99 con una pendiente de 1.001 y un intercepto de -0.007.

Otra manera de darle salida a esta crítica de que no deben usarse escalas tipo Likert por ser ordinales, es promediar grupos de ítems que aludan a una misma variable, lo cual convierte a estas escalas Likert en una escala de intervalo. Esto es similar a la obtención de calificaciones de estudiantes a los que se les aplican exámenes de elección múltiple y en los que el resultado de cada ítem es variable de tipo binario, ya que es considerado correcto o incorrecto (Norman, 2010, pág. 5).

Royer y Zarlowski, aportan en Thiéthart (1999, p. 159) una tabla (ver Tabla 1) tomada de Churchill (1991, p. 159) que presenta las varianzas típicas según el número de puntos de una escala Likert.

Tabla 1. *Varianzas de los datos dependiendo el número de puntos de la escala Likert*

Puntos en la escala Likert	Media	Varianza en distribución normal	Varianza en distribución uniforme
4	2.5	0.7	1.3
5	3.0	1.2	2.0
6	3.5	2.0	3.0
7	4.0	2.5	4.0
10	5.5	3.0	7.0

Fuente: Thiétart (1999, p. 159)

Partiendo del supuesto (que puede verificarse) que por cuestiones de tendencia central la media se sitúa en el centro de cada escala, podemos fijar la precisión deseada como un cierto porcentaje de este promedio (digamos el 5%) y tomando la varianza adecuada de esta tabla, podemos alimentar estos dos datos en ecuación (7) y obtener el tamaño de muestra adecuado.

Una recomendación de Churchill que llama la atención de una nota de pie de su tabla, es que usemos las varianzas para datos uniformemente distribuidos ya que en su opinión son los que predominan sobre la distribución normal en los datos de campo

Una recomendación de Churchill que llama la atención de una nota de pie de su tabla, es que usemos las varianzas para datos uniformemente distribuidos ya que en su opinión son los que predominan sobre la distribución normal en los datos de campo.

Un ejemplo de aplicación de esta información para determinar el tamaño de muestra es el siguiente en una escala de Likert de 10 puntos. La media tiende a 5.5 entonces, la varianza tiende a 7. Tomando esto como referencia, y con un 95% de confianza y un error tolerable de $d \pm 0.275$ (el 5% de 5.5.) que utilizaríamos como valor de la precisión "d" cuando se haga la estimación por intervalo del parámetro poblacional, entonces para tamaños de población grandes, digamos varios miles de personas o elementos el tamaño de muestra se sitúa cuando mucho alrededor de 368 encuestas.

Tamaño de muestra en Análisis Factorial

De acuerdo Moral de la Rubia (2006, p. 448), idealmente, el análisis factorial confirmatorio (AFC) debería aplicarse sobres muestras probabilísticas y de tamaño mayor a 200 elementos muestrales. Las variables deben ser de escala numérica (intervalo o razón) con instrumentos fiables y válidos.

Cuando en un instrumento de investigación o cuestionario de una encuesta hablamos de variable, cada variable equivale a un ítem, pregunta o reactivo. Ahora, en un AFC, un concepto o constructo investigado, a-priori, está formado por un conjunto de ítems. Lo que logramos con el AFC, es verificar que cada conjunto de ítems que presentamos en el cuestionario para cada constructo, coincide con cada uno de los factores que nos arroja el AFC mediante algún software, como SPSS o Minitab, por mencionar algunos.

De acuerdo a Hair et al. (1999) los investigadores no deben usar el análisis factorial para muestras de tamaño inferior a 50 observaciones. Preferiblemente el tamaño muestral debería ser de un mínimo de 100 observaciones. La regla propuesta es que un tamaño aceptable depende del número de variables o ítems. El tamaño debe ser un múltiplo de 10 observaciones por variable, e incluso 20. Si la proporción o múltiplo de observaciones en relación a variables es bajo, según este autor "los resultados deben interpretarse con cautela" (pág. 88).

Según De la Garza-García et al. (2013) la técnica de factores no debe usarse para los casos en los que el número de entrevistas sea menor a 50, preferentemente debe aplicarse en investigaciones que comprendan 100 o más entrevistas o encuestas. Como regla se considera que el tamaño de la muestra debe ser 4 o 5 veces el número de variables que se pretenda agrupar con la técnica (pág. 340)

Pero de acuerdo a Hair et al. (1999, p. 99) el tamaño muestral apropiado, no depende solo del número de variables o factores, sino también de la carga factorial entre el ítem y el factor o constructo. "La carga es la correlación entre la variable y el factor, el cuadrado de la carga es la cuantía de la varianza total de la variable de la que da cuenta el factor". Conviene recalcar aquí, que la variable, ítem o pregunta equivale a la variable dependiente y el factor equivale a una variable independiente.

Se muestra enseguida, Tabla 2 el tamaño de muestra necesario para lograr una significancia del 5% o intervalos de confianza del 95% y una nivel de potencia del 80%, en relación al valor de la carga, según la compañía BMPD Statistical Software, Hair et al. (1999, p. 100). Por nivel de potencia se entiende "la probabilidad de rechazar correctamente la hipótesis nula" (Sánchez, et al., 1992, p. 19).

En la Figura 1, hemos elaborado la relación funcional del tamaño de muestra respeco a la carga factorial que hemos derivado de los datos de la Tabla 2. La ecuación que resume esta relación en números aproximados y fáciles de recordar, a manera de regla para determinar el tamaño de la muestra, para un análisis factorial, es

$$Tamaño = \frac{30}{Carga^2} \tag{8}$$

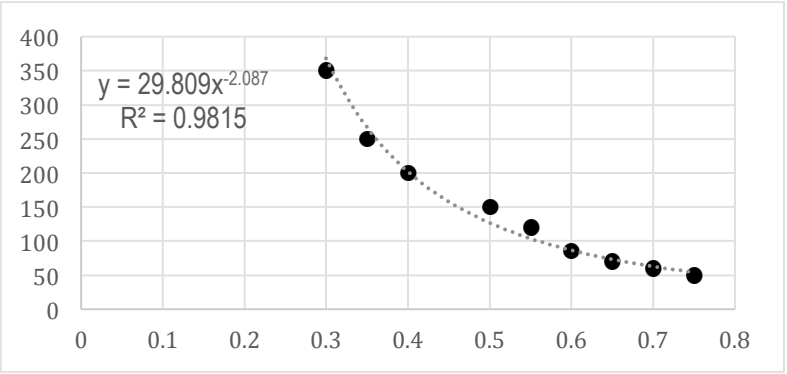
Tabla 2. *Tamaños muestrales necesarios asociados a la significancia de las Cargas Factoriales*

Carga Factorial	Tamaño muestral necesario para la significación [a]
0.3	350
0.35	250
0.4	200
0.5	150
0.55	120
0.6	85
0.65	70
0.7	60
0.75	50

[a] La significancia se basa en un nivel alfa de 0.05, un nivel de potencia de 80 por ciento y los errores estándar supuestamente dos veces mayores que los coeficientes convencionales de correlación

Fuente: Hair, et al. (1999, p. 100) con base en BMPD Statistical Software, Inc.

Figura 1: *Tamaño muestral en función de cargas factoriales significativas*



Fuente: Tabla 2

La correlación de Pearson entre los valores de esta regla y los valores de la tabla es del 0.994 con un coeficiente de regresión de 1.0.

Inicialmente, presentamos opiniones de expertos en el sentido de que el tamaño de muestra conveniente es de más de 200 encuestas, o un múltiplo de se sugiere desde 5, hasta 20 encuestas por variable para que los resultados de un análisis factorial no tengan “que interpretarse con cautela”. Una pregunta muy válida, dados los costos y tiempo que en la investigación social se lleva el trabajo de campo ¿Son necesarias muestras tan grandes?

Aunque Garza-García (2003) afirma que los análisis factorial no deben de hacerse con menos de 50 encuestas, también afirma que la muestra debe ser de 4 o 5 veces el número de variables que se pretende agrupar con la técnica. Es típico que en cursos de Estadística Multivariable, se recurra al archivo wage.sav de SPSS, disponible en internet, para encontrar mediante análisis factorial, qué factores pudieran integrarse a partir de 7 variables con 474 observaciones.

Aplicando la regla de un múltiplo de 5, el tamaño de muestra es de 35. Con la expectativa de que los valores de las cargas anden en 0.85 y aplicando la regla de que el tamaño de muestra es de 30 sobre el cuadrado de la carga factorial, el tamaño sugerido es 42.

En seguida presentamos, en la Figura 2, los resultados de análisis factoriales con el total de 474 datos, así como con 42 en la Figura 3, generados por el procedimiento de ANALISIS FACTORAL de SPSS aplicados a la base de datos wage.sav.

Figura 2. *Matriz de correlaciones y prueba KMO con 473 datos*

ExpPre08	95.95	104.880	473

Matriz de correlaciones

		CatSal04	NivEdu03	Salario05	SalarioInicial	Meses Contratado	Edad	ExpPre08
Correlación	CatSal04	1.000	.515	.780	.755	.004	.009	.062
	NivEdu03	.515	1.000	.661	.633	050	-.281	-.252
	Salario05	.780	.661	1.000	.880	084	-.144	-.097
	SalarioInicial	.755	.633	.880	1.000	-.018	-.009	.045
	MesesContratado	.004	.050	.084	-.018	1.000	.054	.002
	Edad	.009	-.281	-.144	-.009	054	1.000	.800
	ExpPre08	.062	-.252	-.097	.045	.002	.800	1.000
Sig. (Unilateral)	CatSal04		.000	.000	.000	.468	.420	.088
	NivEdu03	.000		.000	.000	.137	.000	.000
	Salario05	.000	.000		.000	.033	.001	.017
	SalarioInicial	.000	.000	.000		.344	.423	.162
	MesesContratado	.468	.137	.033	.344		.121	.485
	Edad	.420	.000	.001	.423	.121		.000
	ExpPre08	.088	.000	.017	.162	.485	.000	

KMO y prueba de Bartlett

Medida de adecuación muestral de Kaiser-Meyer-Olkin.		.724
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	2071.703
	gl	21
	Sig.	.000

Fuente: Procedimiento ANÁLSIS FACTORIAL de SPSS aplicado a base de datos wage.sav

Figura 3. *Matriz de correlaciones y prueba KMO con 42 datos*

ExpPre08	113.36	118.208	42

a. Sólo aquellos casos para los que Selección = 1, serán utilizados en la fase de análisis.

Matriz de correlaciones[a,b]

		NivEdu03	Salario05	SalarioInicial	Meses Contratado	Edad	CatSal04	ExpPre08
Correlación	NivEdu03	1.000	.797	.749	.171	-.433	.631	-.328
	Salario05	.797	1.000	.937	.245	-.329	.773	-.288
	SalarioInicial	.749	.937	1.000	.148	-.274	.736	-.178
	MesesContratado	.171	.245	.148	1.000	.030	.209	-.069
	Edad	-.433	-.329	-.274	.030	1.000	-.150	.743
	CatSal04	.631	.773	.736	.209	-.150	1.000	-.194
	ExpPre08	-.328	-.288	-.178	-.069	.743	-.194	1.000
Sig. (Unilateral)	NivEdu03		.000	.000	.139	.002	.000	.017
	Salario05	.000		.000	.059	.017	.000	.032
	SalarioInicial	.000	.000		.175	.040	.000	.130
	MesesContratado	.139	.059	.175		.425	.092	.332
	Edad	.002	.017	.040	.425		.172	.000
	CatSal04	.000	.000	.000	.092	.172		.109
	ExpPre08	.017	.032	.130	.332	.000	.109	

a. Sólo aquellos casos para los que Selección = 1, serán utilizados en la fase de análisis.

b. Determinante = .005

KMO y prueba de Bartlett[a]

Medida de adecuación muestral de Kaiser-Meyer-Olkin.		.736
Prueba de esfericidad de Bartlett	Chi-cuadrado aproximado	202.389

Fuente: Procedimiento ANÁLSIS FACTORIAL de SPSS aplicado a base de datos wage.sav

Como podemos observar el coeficiente KMO, es aceptable para 42 datos de la Figura 3 al igual que lo fue para 473 de la Figura 2, de acuerdo a Tabla 3. El número de correlaciones altas y significativas son similares.

Tabla 3: *Valor del KMO*

Valor del coeficiente KMO	Adecuación de los datos
0.91 en adelante	Excelente
0.81 a 0.90	Bueno
0.71 a 0.80	Aceptable
0.61 a 0.70	Regular
0.51 - 60.0	Bajo
Menor a 0.50	Inaceptable

Fuente: Adaptación de De-la-Garza-Morales (2013, p. 344)

En las siguientes Figuras (4 y 5) presentamos los factores generados y las cargas factoriales tanto para el total de 473 datos, como para 42 datos. Observamos también que los resultados son similares.

Figura 4. *Cargas factoriales con 473 datos*

Varianza total explicada

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción			Suma de las saturaciones al cuadrado de la rotación		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	3.167	45.242	45.242	3.167	45.242	45.242	3.122	44.598	44.598
2	1.855	26.502	71.744	1.855	26.502	71.744	1.898	27.112	71.710
3	1.008	14.406	86.150	1.008	14.406	86.150	1.011	14.440	86.150
4	.429	6.125	92.275						
5	.247	3.522	95.797						
6	.196	2.801	98.598						
7	.098	1.402	100.000						

Método de extracción: Análisis de Componentes principales.

Matriz de componentes[a]

	Componente		
	1	2	3
Salario05	.944		
SalarioInicial	.910	.232	
CatSal04	.843	.260	
NivEdu03	.806	-.173	
ExpPre08	-.179	.927	
Edad	-.232	.913	
MesesContratado			.996

Método de extracción: Análisis de componentes principales.

a. 3 componentes extraídos

Fuente: Procedimiento ANÁLSIS FACTORIAL de SPSS aplicado a base de datos wage.sav

Figura 5. *Cargas factoriales con 42 datos*

Varianza total explicada[a]

Componente	Autovalores iniciales			Sumas de las saturaciones al cuadrado de la extracción			Suma de las saturaciones al cuadrado de la rotación		
	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado	Total	% de la varianza	% acumulado
1	3.686	52.655	52.655	3.686	52.655	52.655	3.248	46.395	46.395
2	1.495	21.363	74.018	1.495	21.363	74.018	1.934	27.623	74.018
3	.939	13.420	87.438						
4	.382	5.457	92.895						
5	.246	3.518	96.413						
6	.203	2.900	99.314						
7	.048	.686	100.000						

Método de extracción: Análisis de Componentes principales.

a. Sólo aquellos casos para los que Selección = 1, serán utilizados en la fase de análisis.

Matriz de componentes[a,b]

	Componente	
	1	2
Salario05	.944	.203
SalarioInicial	.895	.266
NivEdu03	.879	
CatSal04	.807	.318
Edad	-.523	.780
ExpPre08	-.478	.777
MesesContratado	.260	.263

Método de extracción: Análisis de componentes principales.

a. 2 componentes extraídos

b. Sólo aquellos casos para los que Selección = 1, serán utilizados en la

Fuente: Procedimiento ANÁLSIS FACTORIAL de SPSS aplicado a base de datos wage.sav

Alfa de Cronbach

En un artículo ya clásico, Cortina (1993) reportó que para darnos cuenta de la gran aceptación del alfa de Cronbach en diferentes áreas del conocimiento, en el lapso de 25 años, de 1966 a 1990, se mencionó en 278 *journals* y se citó aproximadamente 60 veces por año. No obstante su uso tan difundido, continuaba afirmando Cortina, que seguía existiendo confusión sobre el verdadero significado y la adecuada interpretación de este coeficiente

Para comprender su significado y hacer una correcta interpretación de este coeficiente, conviene contextualizar su aplicación en la construcción de escalas en referencia a constructos, especialmente en cuanto a confiabilidad o fiabilidad.

Este coeficiente nos permite tener una estimación de la fiabilidad de una escala aditiva formada por varios indicadores cuyo promedio o combinación lineal representará los niveles de un constructo o concepto. En la construcción de una escala aditiva intervienen cuatro aspectos básicos: la definición conceptual, la unidimensionalidad, la fiabilidad y la validación. En seguida hacemos una síntesis y puntualizaciones de estos aspectos con base en Hair et al. (1999, pp. 104-106).

La definición conceptual, que es el punto de partida, en la investigación académica se basa en investigaciones previas (marco teórico) que definen lo característico y la naturaleza de un concepto; en la toma de decisiones (gestión) se refiere a objetivos como satisfacción del consumidor, imagen, etc. Lo que podríamos llamar la definición operacional, es la validación de contenido o validación aparente, en la que nos aseguramos de la correspondencia de los ítems de cada apartado del cuestionario (concepto o constructo) precisamente con el constructo al que pertenecen. Una forma de hacer esta validación es a través de evaluaciones de expertos.

La *unidimensionalidad* se refiere justamente a que los ítems de un apartado del cuestionario estén asociados fuertemente unos con otros y representen al constructo o concepto al que pertenecen. Un caso en el que no habría unidimensionalidad, sería un apartado relativo al desempeño organizacional en un cuestionario y dentro de este apartado hubiera un grupo de preguntas que hicieran referencia al desempeño financiero, otro grupo al desempeño operacional, etc. Para asegurar la unidimensionalidad nos

apoyamos previamente en el análisis factorial, ya sea exploratorio o confirmatorio, que hemos que hemos analizado en el apartado previo.

Fiabilidad. Aunque hay medidas de fiabilidad que se refieren a cada ítem aislado, lo más común es que se evalúe la escala entera (todos los indicadores al mismo tiempo). Aquí es donde se aplica el alfa de Cronbach, que es la más utilizada. El acuerdo general, para considerar que hay consistencia interna, es que el límite inferior se situé en 0.70, aunque para investigaciones exploratorias el resultado sea de 0.60 hacia arriba. Más adelante, al presentar las fórmulas alternativas para su cálculo, veremos que no aparece el tamaño de muestra, aunque nos daremos cuenta que a mayor número de ítems el alpha de Cronbach se incrementa, por lo que para un alpha de Cronbach alto, digamos mayor de 0.90 debemos asegurarnos que algunos de los ítems, no sean sino mero parafraseos de otros ítems.

La *validación,* es para asegurarnos que el conjunto de medidas representa el concepto de interés. Es adicional a la validación de contenido. La validación de este apartado puede hacerse en tres formas: convergente, discriminante y nomológica. Aquí nos referiremos solo a la validación convergente; considerándose que existe validez convergente en dos mediciones, del mismo concepto, cuando encontramos correlacionadas estas dos mediciones (Hair et al. 1999, pp. 104-106).

En la mayoría de las investigaciones cuantitativas, es muy importante el tamaño de muestra, ya que reducen el error en la estimación del parámetro; esto no es del todo cierto en el alpha de Cronbach, ya que en la fórmula que presentamos, no aparece el número de observaciones, sino más bien el número de ítems que forman la escala. Cervantes (2005) nos presenta la siguiente fórmula en ecuación (9).

$$\alpha = \frac{n\overline{\rho}_{kh}}{1+\overline{\rho}_{kh}(n-1)} \tag{9}$$

Es posible aumentar el alpha de Cronbach reduciendo el número de ítems, cuando se eliminan ítems con correlaciones bajas respecto de la puntuación total del constructo, SPSS, y Hair et al (1999, p. 835)

Cervantes (2005) presenta la siguiente Tabla, en la que sugiere tamaños de muestra en relación a número de ítems.

Tabla 4. *Relación entre número de ítems por constructo y tamaño de muestra*

Ítems en el test, o constructo en un cuestionario. ("n")	Regla en cuanto a cuestionarios por ítem	Tamaño de la muestra. Usaremos símbolo "m", ya que la "n" ya se usó en ecuación previa
20 ítems	Entre 5 y 20 sujetos, observaciones o encuestas por ítem	Entre 100 y 400 sujetos o encuestas.
10 ítems o menos	10 sujetos por ítem (Tamaño similar a un análisis factorial exploratorio	Máximo 100 encuestas sería el tamaño ideal

Fuente: Cervantes (2005)

Una conclusión práctica es que para cada uno de los constructos del cuestionario o test, se desarrollen 10 ítems, y se aplique análisis factorial para que se dejen los 6 o 7 que más carguen; esto es, que más correlacionen con el constructo que se esté formando. El SPSS nos puede ayudar en esto, además de la guía que nos de la teoría que se esté aplicando.

Tamaños de muestras para regresiones y correlaciones

Hanke y Wichern (2010) afirman que "Algunos especialistas sugieren que deben existir por lo menos diez observaciones por cada variable independiente" (página 310). Aplicando esta regla, cuando vamos a hacer una regresión simple, con solo diez observaciones sería suficiente. En la Figura 6, de datos reales relativos a un auto, la correlación entre Precio y Km, con 15 observaciones, es considerable -0.76, y aparentemente el haber sobrepasado el mínimo sugerido por Hanke y Wichern nos da buenos resultados; el valor t es -4.3 por lo que se rechaza la hipótesis nula de que el coeficiente de correlación poblacional **r** =0.0.

La cuestión es que observando el intervalo de confianza, éste va de Ah -0.917 a -0.412.con una apertura o rango de 50 puntos porcentuales. En la correlación Precio vs Uso, con una correlación de -0.96, además de rechazarse la hipótesis nula **r** = 0.0, el intervalo va de 0.89 a 0.99, con un rango de solo 10 puntos porcentuales.

Figura 6. *Correlaciones y sus intervalos de confianza*

Núm. línea	DATOS USO (años)	KM	PRECIO ($)
1	12	85,000	97,000
2	9	98,000	115,000
3	9	129,000	120,000
4	9	100,000	125,000
5	7	113,000	160,000
6	7	88,000	164,900
7	7	91,000	165,000
8	6	80,000	162,000
9	6	88,000	168,000
10	3	26,000	247,900
11	11	121,000	105,000
12	11	130,000	95,500
13	5	78,000	215,000
14	5	80,000	216,900
15	10	10,000	95,000

		Precio vs Uso B	Precio vs Km C	Núm. línea	
	R2	0.925	0.582		
ZONA DE DATOS	R MUESTRAL =	-0.962	-0.763	1	
	n=	15	15	2	
	Nivel de Conf.	95	95	3	**FÓRMULAS UTILIZADAS para columna B**
CÁLCULOS RESULTANTES	rtransf	-1.972	-1.004	4	'=0.5*(LN((1+B1)/(1-B1)))
	EE(rtransf)	0.289	0.289	5	'=1/RAIZ(B2-3)
	alfa medio (%)	2.5	2.5	6	'=(100-B3)/2
	Z	1.960	1.960	7	'=DISTR.NORM.ESTAND.INV(B6/100)
	Límite Inferior (r transf)	-2.538	-1.570	8	'=B4-(B7*B5)
	Límite Superior (r transf)	-1.406	-0.438	9	=B4+(B7*B5)
	Límite Inferior de Confianza (95%)=	-0.988	-0.917	10	'=(EXP(2*B8)-1)/(EXP(2*B8)+1)
	Límite Superior de Confianza (95%)	-0.887	-0.412	11	'=(EXP(2*B9)-1)/(EXP(2*B9)+1)
	t	-12.703	-4.258	12	'=B1*((B2-2)/(1-(B1^2)))^0.5
	p (colas)=	0.000	0.001	13	'=DISTR.T(ABS(B12),B2-2,2)

Fuente: Generado por el autor en EXCEL y replicable por el lector al usar las fórmulas de la columna extrema derecha.

Este ejercicio nos muestra, que para tener intervalos de confianza que sean útiles debemos de tener tamaños adecuados de muestra, principalmente en función del número de parámetros, y del nivel del coeficiente de correlación.

Para una correlación simple, podemos tener una mejor estimación del tamaño de muestra con la fórmula de ecuación 10, en la que a es el riesgo de cometer un error tipo I, y lo ubicamos en 5%, equivalente a un nivel de confianza del 95%, luego entonces $Z_{1-a\ /2}$ = 1.96; (1- b) se iguala a un poder estadístico del 80%, por lo que Z_{1-b} = 0.84, mientras que r es el coeficiente de correlación esperado. En nuestro caso, cuando el valor r lo situamos en 90%, el tamaño de muestra es 7 (Pértegas-Díaz y Pita-Fernández, S.):

$$n = \left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)} \right)^2 + 3 \qquad (10)$$

Cuando el coeficiente de correlación *r* anda en niveles de alrededor de 0.6 el tamaño de la muestra arrojado por la fórmula es de 19.

Para la correlación múltiple esta fórmula no se aplica, ya que no toma en cuenta el número de parámetros.

En cuanto a correlación múltiple, continuando con los datos presentados, si de lo que se tratara es de hacer una correlación múltiple de precio en función de Uso y Kms, una muestra de 15 observaciones, según Hanke y Witchern sería insuficiente, ya que al tratarse de dos predictores, necesitaríamos 20 observaciones.

De las reflexiones derivadas de párrafos previos tomamos conciencia de que los tamaños de muestra dependen también del nivel de r o r2, por lo que de la investigación bibliográfica, hemos buscado y encontrado la Tabla 5 y 6 que nos muestran el tamaño de muestra recomendado tanto en función del número de predictores, como del r2 estimado en forma tentativa, antes de la investigación.

Este tamaño de muestra de 20 observaciones recomendado por Hanke y Witchern como podemos verificar en la Tabla 5, estaría sobrado en caso que el coeficiente de determinación poblacional r^2 fuera 0.70 o más, pero sería insuficiente en investigaciones en las que r^2 fuera menor a 0.50, para llegar a tener buenos niveles de predicción.

Tabla 5. *Tamaño de muestra en función de r^2 esperado y de número de predictores para un BUEN nivel de PREDICCIÓN*

r^2	Número de variables predictoras					
	2	**3**	**4**	**5**	**7**	**9**
0.10	240	380	440	550	700	900
0.15	160	220	280	340	440	550
0.2	110	170	200	260	320	400
0.25	85	120	150	180	240	300
0.30	65	95	130	150	190	240
0.40	45	65	80	95	120	150
0.50	35	45	55	65	85	100
0.70	15	21	25	35	40	50
0.90	7	9	10	11	14	16

Fuente: Adaptado de Knofczynski y Mundfrom (2008, p. 440)

Si el r^2 que esperáramos fuera alrededor del 0.90, el tamaño de muestra de 20 observaciones, recomendado por Hanke y Witchern, en este caso de dos variables, estaría muy por encima del 7 para tener *un buen nivel*

de predicción, según Tabla 5 y todavía por encima del 15, según Tabla 6 para tener un *excelente nivel de predicción.*

Tabla 6. Tamaño de muestra en función de r^2 esperado y de la cantidad de predictores para lograr un EXCELENTE nivel de PREDICCIÓN

r^2	Número de variables predictoras					
	2	**3**	**4**	**5**	**7**	**9**
0.10	950	1,500	1,800	2,200	2,800	-
0.15	600	850	1,200	1,400	1,800	2,200
0.2	420	650	800	950	1,300	1,500
0.25	320	460	600	750	950	1,200
0.30	260	360	480	600	800	1,000
0.40	160	260	300	380	480	600
0.50	110	130	220	230	320	400
0.70	50	70	95	110	140	170
0.90	15	21	29	35	40	50

Fuente: Adaptado de Knofczynski y Mundfrom (2008, p. 440)

Por reflexión e intuición, podemos darnos cuenta que entre más alto sea el coeficiente de correlación población (mayores a 0.90), menos observaciones se requieren para captar mediante una muestra tal correlación y que la regla de 10 observaciones por cada predictor queda sobreestimada.

Ahora, aunque según la tabla de Knofzynski y Mundfrom (2008) el nivel de predicción sería excelente con un muestra superior a 15 para un coeficiente de determinación de dos variables predictoras, en este caso kilometraje y años de uso como predictoras del precio del auto. De nuevo, la cuestión es que esto no es garantía de que el modelo de regresión sea el adecuado, ya que como podemos observar en Tabla 7, el impacto del kilometraje no pasa la prueba de hipótesis de que el coeficiente de regresión poblacional no sea cero.

Para el coeficiente de regresión del predictor USO, en la regresión múltiple de Precio en función de Km y Uso, la muestra no apoya la hipótesis nula de que el coeficiente de regresión poblacional sea cero, y el tamaño del efecto se sitúa con un 95% entre 12 mil y 20 mil pesos, en números redondos, como lo podemos observar en los resultados presentados en Tabla 7. La conclusión es que más vale retomar, para este caso, el modelo de regresión simple, en el que el coeficiente de determinación sigue siendo

alto (0.93), y el tamaño del efecto se sitúa, en forma un poco más precisa, entre 15 mil y 21 mil pesos por año, con un nivel de confianza del 95%, como observamos en Tabla 8.

Tabla 7a. *Resumen de la regresión de precio en función de Uso y Kms*

Estadísticas de la regresión	Valores
Coeficiente de correlación múltiple	0.96740385
Coeficiente de determinación	0.93587022
R^2 ajustado	0.92518192
Error típico	13255.5662
Observaciones	15

Fuente: Elaboración propia con base en datos de la Figura 6.

Tabla 7b. *Resumen de la regresión de precio en función de Uso y Kms*

	Coeficientes	Error típico	Estadístico t	Probabilidad	Inferior 95%	Superior 95%
Intercepción	300772.546	13576.1337	22.1545068	4.2112E-11	271192.692	330352.4
Uso	-16011.7367	1968.74479	-8.13296719	3.1747E-06	-20301.2631	-11722.2104
Km	-0.27435323	0.20016261	-1.37065173	0.19557682	-0.71047009	0.16176363

Fuente: Elaboración propia con base en datos de la Figura 6.

Tabla 8. Resultados de la regresión de Precio en función de uso

	Coeficientes	*Error típico*	*Estadístico t*	*Probabilidad*	*Inferior 95%*	*Superior 95%*
Intercepción	290212.853	11550.0508	25.1265435	2.0981E-12	265260.485	315165.221
Uso	-17957.2034	1409.66133	-12.7386649	1.0175E-08	-21002.5916	-14911.8152

Fuente: Elaboración propia con base en datos de la Figura 6.

En síntesis para este apartado de tamaño de muestra en regresiones y correlaciones, podemos decir que es importante, en la fase de diseño tomar en cuenta tanto el número de predictores como el coeficiente de regresión esperado y prepararnos con un tamaño de muestra que aspire a un nivel excelente de predicción y ajustar tal tamaño a como vayan saliendo los resultados pero sin perder aleatoriedad, a un tamaño de muestra que nos dé un buen nivel de predicción, con coeficientes de regresión correctos.

Aquí la lección es que si en la población el coeficiente de regresión es diferente de cero, siempre habrá un tamaño de muestra lo suficientemente grande que capte tal valor, y que "rechace" la hipótesis nula de que es diferente de cero, aunque para propósitos de toma de decisiones un valor de R^2 muy por debajo de 0.50 no sería útil.

Si R^2 anda alrededor del 0.50, tratándose de investigaciones pioneras todavía puede ser de interés el modelo cuando se tiene la esperanza que en futuras investigaciones resulte con un valor algo considerablemente por encima del 0.50, y supere perceptiblemente a los puntos porcentuales no explicados por el modelo.

En ecuaciones estructurales

El tamaño de muestra en sistemas de ecuaciones estructurales, obedece a ciertas reglas cuando el enfoque es de covarianza y a otras reglas cuando el enfoque es de mínimos cuadrados parciales.

En cuanto a consideraciones sobre el tamaño de muestra para modelos de ecuaciones estructurales con el enfoque de covarianza, software de modelación AMOS, por ejemplo, Hair et al. (1999) afirman que aunque el tamaño de la muestra impactará la significancia estadística y el índice de bondad del ajuste, no existe una regla única para determinar el tamaño de muestra para un modelo de ecuaciones estructurales. No obstante, Hair et al., recomiendan tamaños de muestra que se ubiquen entre 100 y 200 unidades muestrales. Este tamaño depende de:

Aseguramiento del error de especificación

a. El número de covarianzas y correlaciones de la matriz de entrada
b. El número de parámetros a estimar, recomendándose de 5 a 10 encuestas por parámetro.
c. La no-normalidad de los datos de entrada, en cuyo caso el múltiplo del inciso c se eleva a 15 encuestas por parámetro.

Cuando el enfoque del modelo de ecuaciones estructurales tiene el enfoque de mínimos cuadrados parciales, como los softwares PLSGraph o SmartPLS, el investigador debe observar su modelo gráfico (diagrama de

flechas) y comparar cuál de las siguientes dos mediciones resulta más grande:

a. El máximo número de indicadores formativos que impactan a una variable, ya que se realizan regresiones múltiples de cada variable y sus indicadores formativos. En el caso de indicadores reflexivos, por realizarse regresiones simples, no es necesario buscar un máximo ya que el valor en este caso es igual 1.
b. El mayor número de variables latentes independientes que impacten a otra variable latente dependiente; esto es, la ecuación estructural más grande en términos de variables independientes.

De lo anterior, una regla heurística para determinar el tamaño de muestra cuando se tienen solo indicadores reflexivos, según el profesor Chin, es que el número de encuestas recomendado sea 10 veces el número de variables independientes encontradas en la medición del inciso "b", mencionado anteriormente.

Conclusiones y recomendaciones

En la introducción del presente documento nos propusimos integrar una guía práctica con recomendaciones y ejemplos concretos para determinar tamaños de muestras para las diversas variables de un cuestionario y sus interrelaciones, en los estudios de campo en la investigación social, tanto para propósitos de investigación académica como para toma de decisiones.

En la determinación de estos tamaños de muestra se investigaron métodos y formulaciones que contribuyeran a tener buenos niveles de estimación al momento de establecer los intervalos de confianza, y con aperturas razonables.

Se logró el objetivo al ubicar y presentar detalladamente estos métodos y formulaciones tanto para variables dicotómicas, continuas y discretas y de tipo Likert y para interrelaciones entre ellas ya se trate de análisis factorial, alpha de Cronbach, regresiones o ecuaciones estructurales.

Se presentaron valores de observaciones de casos reales, y se crearon escenarios alternativos en cada uno de ellos. Con estos valores

reales se pusieron a prueba reglas prácticas sugeridas por varios autores para determinar tamaños de muestras e incluso se propone una nueva para el caso del análisis factorial. El aprendizaje aquí es que si hay más de una regla práctica para alguno de los casos mencionados en el párrafo anterior, se tomen en cuenta todas ellas y se elija la que nos dé mayor protección sin poner en riesgo la viabilidad del proyecto de investigación, por cuestiones de costos y tiempo.

La recomendación aquí es que el lector, siga creando estos escenarios y se sensibilice y se convenza de sus implicaciones y su trascendencia práctica al tratar de cumplir con los objetivos de la actividad científica como son la estimación, la explicación, la predicción y el control, y no solo cumplir con un mero requisito de "rechazar hipótesis nulas".

Referencias

APA, Board of Scientific Affairs. (1996). *Initial Report of The Task Force on Statistical Inference*. Recuperado de http://www.apa.org/science/leadership/bsa/statistical/tfsi-initial-report.pdf el 18 de Septiembre del 2014

Cervantes, V.H. (2005). *Interpretaciones del coeficiente alpha de Cronbach*. Avances en Medición, *3*, 9-28.

Chin, W.W. (1988). Issues and opinions on structural equation modeling. *Management Information Systems Quarterly*, *22*(1), 7-16.

Chin, W.W. (1998) Chapter 10: "The Partial Least Squares Approach for Structural Equation Modeling", *in* Maroulides, G. A. (ed.). *Modern Methods for Business Research,* Hillsdate,N.J. : Lawrence Erlbaum Associates.

Churchill, G.A. (1991), *Marketing research: Methodological foundations*, 5th ed. Hinsdale, Il: Dryden Press

Cortina, J.M. (1993). What is coefficient alpha? An Examination of Theory and Applications. *Journal of Applied Psychology*, *78(*1), 98-104

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297-334.

De-la-Garza-García, J., Morales-Serrano, B.N. y González-Cavazos, B.A.(2013). *Análisis Estadístico-Multivariado: Un enfoque teórico y práctico.* Monterrey, México: McGraw-Hill.

Fidler, F., Thomason, Neil, Cumming, G. Finch, S. & Leeman, J. (2004). Editors Can Lead Researchers to Confidence Intervals, but Can't Make Then Think: Statistical Lessons From Medicine. *Psychological Science* 2004 15:119 Recuperado de http://pss.sagepub.com/content/15/2/119

Hair, J.F, Anderson, R. E., Tatham, R. L. y Black, W. C. (1999). *Análisis Multivariante*, 5ª. Edición. Madrid: Prentice-Hall

Hanke, J. E. & Wichern, D. W. (2010). *Pronósticos en los Negocios*, 9ª. Edición. México: Prentce-Hall

Havlicek, L. & Peterson, N. (1976). Robustness of the Pearson correlation against violations of assumptions, *Perceptual and Motor Skills*, *43*(3f), 319-1334.

Jamieson, S. (2004). Likert scales: How to (ab)use them. *Medical Education*, *38*(12), 1217-1218.

Knofczyinski, G. T. & Mundford, D. (2008). Sample sizes when using multiple linear regression for prediction. *Educational and psychological measurement*, *68*(3), 431-442.

Lind, D. A., Marchal, W. C. & Wathen, S.A. (2012). *Estadística aplicada a los negocios y la economía 15ª. Edición*. México: McGraw-Hill.

Moral de la Rubia, José. (2009). Capítulo 13. Análisis factorial y su aplicación al desarrollo de escalas, en Landeros-Hernández, R. & González-Ramírez (eds.). *Estadística con SPSS y metodología de la investigación*, Monterrey: Trillas.

Pértegas-Díaz, S & Pita-Fernández, S. (2002). *Metodología de investigación*. Coruña: Fisterra.

Royer, I. & Zarlowski, P. (1999). Chap. 8. Sampling, en Thiétart, Raymond-Alain. *Doing Management Research: A Comprehensive Guide*. London: Sagen

Shrout, P. (1997). Should significance test be banned? Introduction to a special section exploring the pros and cons. *Psychological Science*, *8*(1), 1-2.

Sierra-Bravo, R. (1983). *Ciencias sociales. Espistemología, lógica y metodología. Teoría y ejercicios.* Madrid: Paraninfo

Thiétart, A. (1999). *Doing management research*. London: Sage.

Wilkinson, L. (1999). The task force on statistical inference. *American Psychologist, 54*(8), 594-604.