

Técnica y Tamaño de la Muestra: Principal Fuente de Confiabilidad de los Datos

Technique and Sample Size: Main Source of Data Reliability

Iván Guillermo González Palomo^{1*} & Azucena Minerva García León²

^{1y 2} Universidad Autónoma de Nuevo León, Facultad de Ciencias Químicas, Nuevo León, México.

* ivan.gonzalezpl@uanl.edu.mx

Abstract

The results and conclusions of a research study constitute key aspects of the structure of a scientific work, as they demonstrate the degree to which the objectives established at the beginning of the project have been achieved. The theoretical framework provides the conceptual foundation of the study by integrating empirical evidence and contributions from the specialized scientific literature. However, the methodological section also requires a detailed description of the experimental design, the procedures followed, as well as the sampling technique and sample size. The sampling technique determines the method used to select the study participants or experimental units, while the sample size must be statistically significant in order to ensure valid and reliable estimates of the population. Both elements are fundamental in social science research, as they directly influence the validity, reliability, and generalizability of the results. The objective of this study is to present a review of the literature on the most commonly used sampling techniques and sample sizes, as well as the procedures for validation testing, reliability estimation, and the construction of confidence intervals, with the aim of addressing researchers' methodological challenges in providing statistical and mathematical support for the selected sampling technique and sample size.

Keyword

Sampling technique, Sample size, Validity, Reliability, and Confidence interval.

Resumen

Los resultados y conclusiones de una investigación constituyen aspectos clave en la estructura de un trabajo científico, dado que evidencian el grado de cumplimiento de los objetivos planteados al inicio del proyecto. El marco teórico proporciona el sustento conceptual del estudio al integrar evidencia empírica y aportes de la literatura científica especializada. Sin embargo, la sección metodológica también requiere una descripción detallada del diseño experimental, del procedimiento seguido, así como de la técnica de selección y el tamaño de la muestra. La técnica muestral determina el tipo de selección de los participantes de estudio o de las unidades experimentales, mientras que el tamaño de la muestra debe ser estadísticamente significativo para garantizar estimaciones válidas y confiables de la población. Ambos elementos son fundamentales en las investigaciones en ciencias sociales, dado que influyen directamente en la validez, confiabilidad y capacidad de generalización de los resultados. El objetivo de este trabajo es presentar una revisión de la literatura con respecto a las técnicas de muestreo y tamaños muestrales más utilizados, así como los procedimientos de las pruebas de validación, estimación de la confiabilidad y de intervalos de confianza; esto para resolver la problemática de los investigadores de darle soporte metodológico y matemático al tamaño y técnica de muestro aplicado.

Palabras clave

Técnica de muestreo, Tamaño muestral, Validez, Confiabilidad, Intervalo de confianza.

1. Introducción

Uno de los desafíos frecuentes en la investigación en ciencias sociales radica en la inclusión de componentes observacionales que requieren trabajo de campo. En esta etapa, un aspecto crítico en esta etapa es la definición de la muestra adecuada a la cual se aplicará el instrumento de recolección de datos diseñado, como un cuestionario. Dicho instrumento suele estructurarse en diferentes secciones que contienen reactivos que representan diversas variables, las cuales serán posteriormente sometidas a análisis estadísticos. Se espera que los resultados derivados de éstos análisis contribuyan de manera significativa al desarrollo del conocimiento científico. En ese sentido, la determinación

adecuada del tamaño de la muestra y la aplicación rigurosa de una técnica de muestreo apropiada son fundamentales para garantizar la validez de los estimadores poblacionales derivados de la muestra. Decisiones metodológicas incorrectas, ya sea en la estrategia de muestreo o en la selección de los sujetos de estudio, pueden introducir sesgos que comprometan la confiabilidad y la capacidad de generalización de resultados.

La técnica de selección define cómo se eligen los participantes o unidades experimentales, su relevancia radica en la representatividad, ya que garantiza que la muestra refleje de manera adecuada a la población de estudio; la reducción del sesgo, ya que un muestreo correcto (probabilístico) evita que las conclusiones estén influenciadas por decisiones subjetivas. También la técnica de selección aporta a la validez externa, debido a que entre más representativa sea la muestra, mayor será la capacidad de generalizar los hallazgos a toda la población.

El tamaño de la muestra se refiere al número de casos incluidos en el estudio. Su importancia se centra en la precisión estadística, porque un tamaño adecuado reduce el error muestral y mejora la confiabilidad de las estimaciones. Además, se aporta a la potencia del estudio, dado que una muestra demasiado pequeña puede no detectar efectos reales (error tipo II), mientras que una demasiado grande puede detectar diferencias irrelevantes. Elegir el tamaño óptimo evita desperdicio de tiempo, esfuerzo y presupuesto.

A continuación, se presenta una recopilación de literatura clásica, que profundizan en la relevancia del tema de la técnica y tamaño muestral:

- Cochran (1977), escribió una de las obras clásicas en estadística de muestreo; contiene capítulos dedicados tanto a los distintos tipos de muestreo (aleatorio simple, estratificado, por conglomerados, etc.) como a la estimación del tamaño de muestra.
- Desu & Raghavarao (1990), los autores se centran explícitamente en cómo determinar el tamaño adecuado de muestra para diferentes tipos de diseños y problemas (estimación, pruebas de hipótesis, entre otros).
- Henry (1990) aborda teoría y práctica del muestreo de encuestas, con casos prácticos, análisis complejo, variantes modernas de muestreo e incluye reflexiones sobre diseño y representatividad.
- Chaudhuri & Pal (2023) abordan tanto teoría como práctica del muestreo de encuestas, con casos prácticos, análisis complejo, variantes modernas de muestreo. Incluye reflexiones sobre diseño y representatividad.

Hernández Sampieri et al. (2014) mencionan que, en investigaciones cuantitativas, la encuesta es una de las herramientas principales para recolectar información, su elección depende de los objetivos, el diseño de investigación, el presupuesto y el tipo de población que se desea estudiar. La diferencia entre la encuesta cuantitativa con la encuesta cualitativa, es que ésta última utiliza análisis más descriptivos de manera textual o de entrevista; mientras que la cualitativa busca obtener un métrico que pueda ser utilizado para poder concluir alguna hipótesis o afirmación de la población, de acuerdo a Buckingham & Saunders (2004).

2. Metodología

Este estudio empleó una metodología de revisión narrativa de la literatura centrada en el análisis de referencias relacionadas con las técnicas de muestreo y la determinación del tamaño de muestra. La revisión narrativa se consideró adecuada dado que el objetivo fue describir, comparar y sintetizar conceptos, métodos y enfoques teóricos reportados en la literatura especializada. El proceso metodológico incluyó un análisis descriptivo de la literatura, la síntesis teórica y la comparación conceptual de enfoques metodológicos. Para el análisis de las fuentes seleccionadas se aplicaron técnicas cualitativas, entre las que se incluyeron el análisis de contenido, el análisis temático, el mapeo conceptual y el análisis comparativo, lo que permitió la identificación, clasificación e interpretación sistemática de patrones y tendencias metodológicas relevantes en los estudios revisados.

3. Variables

Kerlinger & Lee (2002) mencionan que una variable es un constructo o característica que puede tomar diferentes valores. Asimismo, Namakforoosh (2005) concluye que las variables son características o atributos que se miden en los sujetos de estudio y que pueden adoptar diferentes modalidades o valores. Además, una variable es una propiedad que puede fluctuar y cuya variación es susceptible de medirse u observarse, de acuerdo a los trabajos de Hernández Sampieri et al. (2014). De acuerdo a la pregunta o afirmación realizada, la variable a considerar puede tener principalmente las siguientes escalas en los trabajos de investigación cuantitativa, esto de acuerdo a la literatura revisada de autores como Hernández Sampieri et al. (2014), Gravetter & Wallnau (2017) y Likert (1932); la información se resume en la Tabla 1.

Tabla 1. Variables en Investigación Cuantitativa en Ciencias Sociales.

Tipo	Naturaleza	Ejemplos de pregunta	Tipos de análisis
Dicotómica	Cualitativa nominal: sólo dos posibles valores	¿Trabaja actualmente? (Sí / No)	Porcentajes, tablas de frecuencia, Chi-cuadrada

Numérica	Cuantitativa: número que puede ser discreto o con decimal.	¿Cuántos trabajos tiene actualmente?, ¿Cuál es su ingreso mensual?	Media, desviación estándar, regresión, ANOVA
Likert	Cualitativa ordinal: a veces tratada como intervalo.	Califique su satisfacción: 1 a 5, donde 1 es insuficiente y 5 es excelente.	Medias de escala, análisis de confiabilidad (Alfa de Cronbach), análisis factorial

Casos con Variable Dicotómica

De acuerdo a Rositas Martínez (2014), si se tiene una población prácticamente infinita y se desea calcular el tamaño de la muestra para concluir acerca de proporciones (porcentajes) se deberá utilizar la ecuación [1].

$n = \frac{(Z_{\alpha})^2(p)(1 - p)}{e^2}$	[1]
--	-----

La variable p será la proporción previa de un estudio; por ejemplo, si en la validación del instrumento se encontró que 80% de las mujeres sufren de *burnout*, éste será el valor correspondiente a p utilizado como un estimador puntual. Aunque, si no se tiene el dato o se desea desprestigiar el de la muestra por algún sesgo o diseño experimental, el valor p recomendado deberá ser 0.5, lo que lleva a la cantidad máxima de encuestas posibles en vista de que tanto $p=0.5$ como $1-p=0.5$, en donde se considera balanceado el éxito y fracaso de la variable a medir. Por otra parte, e es el error máximo tolerable en la proporción. Si se tiene una idea de comprobación, por ejemplo, el demostrar que el 80% de las mujeres sufren *burnout*, ¿cuál es el porcentaje de error máximo que el investigador considera aceptable para la conclusión? Por ejemplo, si es 3%, la $e=0.03$. El valor de Z_{α} se obtendrá con la tabla de la distribución normal, la cual corresponde al nivel de confianza que se quiere mostrar en el cálculo de la muestra, por lo general si se desea el 95% de confianza de la, el $\alpha=0.05$ y por lo tanto $z=1.96$; mientras que para 99% de confianza $z=2.54$.

De acuerdo a González Palomo et al. (2023), en la Tabla 2, se muestran más valores para esta constante proveniente de la distribución normal; e representa el error del muestreo aceptable, error que el investigador acepta y tolera para que la muestra no sea significativa, habitualmente $0.05 < e < 0.10$.

Tabla 2. Valores de la constante Z_{α} según el nivel de confianza.

Z_{α}	1.645	1.96	2.24	2.576
Nivel de confianza	0.90	0.95	0.975	0.99

A continuación, se presenta la Figura 1 con el comparativo para valores de p . En ésta se demuestra que cuando $p=0.5$, al mantener el resto de las variables constantes ésta generará el mayor número de muestras requeridas. Se consideró $z=1.96$ (confianza del 95%) y $e=0.03$.

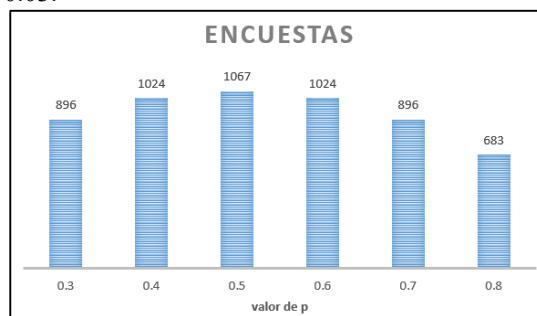


Figura 1. Número de encuestas modificando p , con N infinito.

Si se tiene el tamaño de la población (N), éste será tomado en cuenta, por lo que la expresión para el cálculo de la muestra deberá ser la ecuación [2]. Además, en la Figura 2, se presenta la expresión suponiendo una población $N=1,000$.

$n = \frac{Z_{\alpha}^2 N p (1 - p)}{e^2 (N) + Z_{\alpha}^2 p (1 - p)}$	[2]
---	-----

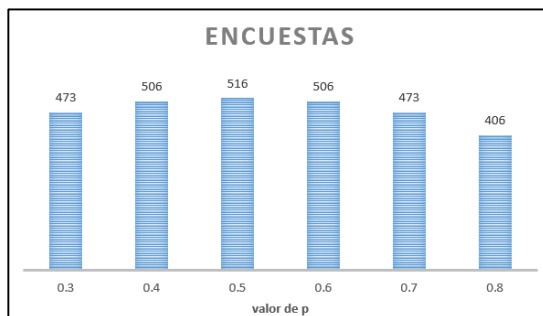


Figura 2. Número de encuestas modificando *p*, con valor de *N*=1000.

Casos con Variable Numérica

En el ámbito de la investigación, cuando se analiza una variable de tipo numérica, ésta puede clasificarse como discreta o continua, dependiendo de la naturaleza de los valores que puede asumir. Para obtener la mejor aproximación a la muestra representativa de la población se utilizan las siguientes expresiones, de igual manera separadas si se conoce o no la población. Aquí se incluye un nuevo término que habla de la desviación estándar de la muestral piloto o conocida. Esta dimensión representa la lejanía promedio que hay de un dato aleatorio con el promedio obtenido, es decir, la desviación estándar. En el caso anterior, de proporciones la $\sigma = pq = p(1-p)$. Enseguida se muestran las dos ecuaciones [3] y [4] apropiadas para el calcular el tamaño muestral dependiendo del conocimiento del tamaño de la población.

Se desconoce la población (población infinita)		Se conoce la población (población finita)	
$n = \frac{Z_{\alpha}^2 \sigma^2}{e^2}$	[3]	$n = \frac{Z_{\alpha}^2 N \sigma^2}{(N - 1)e^2 + Z_{\alpha}^2 \sigma^2}$	[4]

Casos con Variable Likert

La desviación estándar, como se comentó anteriormente, puede estimarse por algún estudio previo, mediante los resultados de la prueba piloto o a partir del rango total entre seis. Esto último aplica en caso de escala Likert, donde si se tienen valores de 1 a 5, donde el rango es 4 y la desviación $\sigma=4/6=0.66$. El seis proviene del supuesto de normalidad donde la regla empírica menciona que en 6 desviaciones estándar ($\bar{x} \pm 3s$) se encuentran todos los datos. Para escalas Likert largas (más de 10 reactivos), se recomienda tener 5 a 10 participantes por reactivo (regla de Nunnally & Bernstein, 1994) para tomarse de preferencia como muestra piloto. Por otro lado, Hernández Sampieri et al. (2014) sugieren utilizar fórmulas de muestreo y considerar el poder estadístico, por lo que recomienda muestras mayores a 30 sujetos por grupo para análisis comparativos.

4. Consideraciones Adicionales

Diversos autores concluyen sobre algunos de los elementos base que se deberán tomar en cuenta en todo momento al realizar su muestreo, dado que una muestra equivocada conlleva una mala conclusión acerca de lo que se piensa de la población.

- La población o universo se define según la unidad de análisis: personas, organizaciones, productos, etc.
- La muestra es un subconjunto de la población, que debe reflejar sus principales características para que los resultados puedan generalizarse.
- Para el nivel de confianza, suelen tomarse valores de 95% y 99%; cuanto más alto sea este nivel, mayor deberá ser la muestra.
- El error máximo tolerable o margen de error, será la desviación del verdadero valor poblacional que aceptará el investigador, considerando que un error menor requiere una muestra más grande.

De acuerdo al tipo de estudio a realizar, el tamaño muestral recomendado varía. Esto se basa en recomendaciones de varios autores reconocidos en metodología de investigación, psicometría y análisis estadístico. En la Tabla 3, se muestra qué autores respaldan qué consideración.

Tabla 3. Tipo de estudio y muestra recomendada.

Muestreo	Justificación
Uso de fórmula de muestreo para poblaciones grandes	Cochran (1977) explica y justifica el uso de la fórmula clásica de muestreo para poblaciones grandes y por qué aproximadamente 384 casos es suficiente con 95% de confianza y $\pm 5\%$ de error.

Al menos 30 sujetos por grupo para comparaciones	Hernández Sampieri et al. (2014) recomiendan al menos una muestra de 30 por grupo para análisis paramétricos comparativos (<i>t</i> -test, ANOVA). También apoyado por Roscoe (1975).
5–10 sujetos por reactivo de la escala (análisis factorial / Likert)	Nunnally & Bernstein (1994) proponen mínimo 5 participantes por reactivo, idealmente 10, para garantizar estabilidad de cargas factoriales.
Al menos 200 casos para modelos multivariados complejos	Hair et al. (2019) recomiendan entre 200 y 300 casos para análisis multivariados robustos.
Guía práctica de interpretación del tamaño de muestra	Comrey & Lee (1992) proponen la clasificación cualitativa del tamaño de muestra para análisis factorial, donde 50 es pobre y 300 es adecuada.
Tablas de Sudman para estudios regionales y nacionales	Sudman (1976) sugiere que muestras nacionales suelen rondar entre 1,500–2,500 casos, mientras que estudios regionales pueden trabajar con 400–600 casos, dependiendo del nivel de precisión deseado.

5. Técnicas de Muestreo

La técnica del muestro permite que los resultados sean representativos de la población de interés, lo que proporciona una mayor validez externa a los hallazgos. Además, esto ayuda a reducir sesgos de selección, ya que una muestra bien diseñada evita que ciertos grupos queden sub-representados, mejora la precisión estadística dado que una muestra adecuada reduce el error muestral, aumenta la confiabilidad de las estimaciones como medias, proporciones, correlaciones. Se debe tener en cuenta que seleccionar una muestra eficaz permite ahorrar tiempo y costos, evitando tener que estudiar toda la población y facilitando la generalización de resultados, lo que es esencial en estudios cuantitativos con objetivos de inferencia.

5.1 Técnicas Probabilísticas y No Probabilísticas

Una muestra probabilística significa que todos los posibles sujetos de la población tienen la misma probabilidad de ser elegidos, en contrario de la no probabilística, en la que por algún tema se preseleccionó, se otorgó prioridad o existe una característica a cumplir por el sujeto para que éste sea seleccionado. Cabe mencionar que el muestreo probabilístico algunos investigadores optan por no seleccionarlo, ya sea por su alto costo o porque el sujeto debe cumplir con ciertas características para ser considerado en la muestra. En la Tabla 4 se presenta un breve contraste de estas técnicas. Además, se presentan posteriormente algunas de las técnicas específicas, detallando cómo podrían aplicarse en una investigación orientada a evaluar la percepción de la calidad del servicio bancario en el Área Metropolitana de Monterrey (AMM).

Tabla 4. Características del Muestreo Probabilístico y No Probabilístico.

Característica	Muestreo Probabilístico	Muestreo No Probabilístico
Definición	Cada unidad de la población tiene una probabilidad conocida y no nula de ser seleccionada.	Las unidades se seleccionan sin usar un mecanismo de selección al azar, muchas veces basado en conveniencia, juicio, cuotas, etc.
Ventajas	Representatividad, permite estimar el error muestral, inferior riesgo de sesgo, generalización de los hallazgos.	Menor costo, rapidez, practicidad, más fácil de aplicar cuando el muestreo probabilístico no es viable.
Desventajas	Requiere marco muestral completo, puede ser costoso, más complejo de ejecutar y procesar los datos.	Falta de representatividad, mayor riesgo de sesgo, dificultad para generalizar, error muestral desconocido.

5.1.1 Probabilística: Aleatorio Simple (Simple Random Sampling)

Este tipo de muestro selecciona unidades al azar de la población, de modo que cada unidad tenga la misma probabilidad de ser elegida, para ello se puede utilizar un generador de números aleatorios o sorteos. Dentro de sus ventajas está el alto grado de representatividad cuando el muestreo es grande, es sencillo de entender y de explicar, además, permite estimaciones con confianza estadística y margen de error bien definidos (McCombes, 2023). Por otra parte, requiere un marco muestral completo, no siempre práctico si la población está muy dispersa geográficamente y puede ser costoso en términos de tiempo y recursos si la población es grande.

En el caso del estudio de calidad del servicio bancario, se tendría que obtener una base de datos de todos los clientes que tengan cuenta bancaria en el AMM. En dicho listado, los sujetos deben ser enumerados o codificados de alguna forma, posteriormente, seleccionados de manera aleatoria mediante un sorteo, con el fin de obtener la cantidad de muestra solicitada. Enseguida, se busca al cliente seleccionado para que complete la encuesta.

5.1.2 Probabilística: Sistemático (Systematic Sampling)

Se debe elegir una unidad inicial al azar i ($1 \leq i \leq N/n$) y enseguida seleccionar cada k -ésima unidad de la población ($k = N/n$), según un intervalo fijo; por ejemplo, cada décimo individuo de una lista. Una de las principales ventajas de este muestro es que es más simple que el aleatorio simple, especialmente si la población está listada, es menos costoso en términos de organizar la selección y puede ofrecer buen grado de representatividad si la lista no tiene sesgos. Por el contrario, si hay algún patrón en la lista (por ejemplo, orden por sección, por nivel socioeconómico etc.) puede introducir sesgos; asimismo, es menos flexible si el tamaño del marco o el orden de la lista no permiten un intervalo uniforme.

En el caso del estudio de calidad del servicio bancario, se inicia seleccionando aleatoriamente un individuo y , a partir de éste, se procede a identificar cada la k -ésimo participante conforme a un intervalo previamente definido. Por ejemplo, suponiendo que la población es de 10,000 clientes en el AMM y que el tamaño de muestra debe ser 250, entonces, se selecciona un cliente inicial al azar entre el número 1 y el 40 ($k=10,000/250=40$) y después se seleccionarán los clientes cada 40 personas en el listado.

5.1.3 Probabilística: Estratificado (Stratified Sampling)

Esta técnica divide la población en subgrupos homogéneos (estratos), tales como edad, género, nivel educativo, etc., y enseguida tomar muestras aleatorias (o sistemáticas) dentro de cada estrato. Esto permite aumentar la precisión de las estimaciones (menos error) si los estratos son internamente homogéneos y distintos entre ellos; asegura que cada subgrupo importante de la población esté representado y mejora la eficiencia si cierto estrato requiere más casos para análisis (Elfil & Negida, 2017). En cambio, existe mayor complejidad en el diseño, se requiere conocimiento previo de la población para estratificar adecuadamente, puede requerir tamaños de muestra mayores si hay muchos estratos pequeños y si los estratos están mal definidos o poco homogéneos, la ventaja puede perderse. Por otra parte, el tamaño muestral dentro de cada estrato, puede seleccionarse de manera proporcional o desproporcional en relación con el tamaño poblacional de dicho estrato, dependiendo de los objetivos del estudio y de la variabilidad esperada en cada subgrupo. La selección de los elementos de la muestra en cada estrato, se puede utilizar el muestreo aleatorio simple o sistemático.

En el caso supuesto de la calidad del servicio bancario, la población es de 10,000 clientes en el AMM, pero el investigador quiere mantener balanceado la distribución de clientes, debido a que se conoce que el 45% de los clientes son de tarjeta de crédito, 35% de débito y 20% de aplicación. Entonces, por estratificación deberá tener el tamaño y técnica de muestro que no dé prioridad a ninguno de este sector y que el tamaño se distribuya de manera igualitaria. Suponiendo que el tamaño muestral requerido total es de 250 clientes, entonces $250(0.45)=113$ clientes de tarjeta de crédito, $250(0.35)=88$ clientes de tarjeta de débito y $250(0.20)=50$ clientes de aplicación deberán ser encuestados. En este momento, se selecciona la técnica de aleatoria simple o sistemática para obtener los sujetos de la muestra.

5.1.4 Probabilística: Por Conglomerados (Cluster Sampling)

Esta técnica es a gran escala, dado que distribuye la población en grupos naturales (conglomerados), como escuelas, barrios, comunidades, etc., enseguida, selecciona aleatoriamente un cierto número de conglomerados, ya sea todos los elementos o una parte de ellos de forma aleatoria simple o sistemática. También, este tipo de muestreo puede realizarse siguiendo un procedimiento en múltiples etapas. Es económico y práctico cuando la población está muy dispersa geográficamente, menos costoso en transporte y logística si se visita conglomerados y útil en estudios grandes nacionales o regionales. Aunque la variabilidad entre conglomerados puede aumentar el error muestral (mayor error comparado con muestreo estratificado si los conglomerados no son homogéneos), requiere corrección en el cálculo del error estándar (efecto del diseño) y puede necesitar muestras mayores para compensar la homogeneidad dentro de conglomerados.

En el caso supuesto de calidad del servicio, se deberá dividir todos los clientes en sectores, tal vez por municipio o región, después de cada uno de ellos se deberá seguir un muestreo sistemático o aleatorio, aunque la muestra sigue siendo complicada de conseguir.

5.1.5 No Probabilística: Muestreo por Cuotas (Quota Sampling)

La población se divide en grupos o categorías (similares a estratos), pero la selección dentro de esos grupos se hace sin azar; se establecen cuotas a cumplir en cada categoría (*Publications*, ICI UMN). Es un muestreo no probabilístico, aunque permite controlar la representación de ciertos grupos importantes de la población, es más rápido y menos costoso que métodos probabilísticos y es útil cuando no se tiene un marco muestral completo. Deberá tenerse en cuenta que existe riesgo de sesgo, ya que quienes llenan la cuota pueden no ser representativos del grupo, es imposible calcular margen de error confiable asociado a la muestra y hay generalización limitada de los resultados.

En el caso supuesto de la calidad del servicio, el investigador tiene una "cuota" que es una característica que deberá conseguir. Por ejemplo, si se desea encuestar a clientes que hayan ido a la sucursal por alguna reclamación a la institución, la cuota a cumplir será que la muestra contenga solo clientes "reclamando". Entonces, el investigador seleccionará aquellos clientes que cumplan con esta característica únicamente.

5.1.6 No Probabilística: Muestreo Bola de Nieve (Snowball Sampling)

Se comienza con unos pocos sujetos, que cumplen los criterios y quienes recomiendan o refieren a otros participantes y así sucesivamente. Es muy utilizado cuando la población es difícil de localizar. Dentro de las **ventajas está que es útil** para poblaciones ocultas o difíciles de acceder (por ejemplo, grupos marginales, redes clandestinas), además, permite construir la muestra aun cuando no haya marco muestral. Aunque deberá tener en cuenta el alto riesgo de sesgo de red (los sujetos tienden a referir a otros como ellos), existe falta de representatividad, es difícil estimar el error y la confiabilidad estadística y presenta dependencia de los contactos iniciales, lo que puede limitar la diversidad de la muestra.

En el caso supuesto de la calidad del servicio bancario, el investigador encuesta a un cliente, pidiendo referenciarlo con otro para que éste sea el que continúe con otra encuesta y, así sucesivamente, hasta completar el tamaño muestral sugerido.

5.1.7 No Probabilística: Muestreo Intencional o de Conveniencia (Purposive / Convenience Sampling)

Se seleccionan los participantes según criterios de facilidad o pertinencia, porque están disponibles o cumplen algunas condiciones específicas de interés. Es la técnica de muestreo más común en las ciencias sociales. Se encuesta a quienes más cercanos estén, más fácil sea y que, sobre todo, cumplan con alguna característica que el investigador desee. Dentro de las ventajas está que es muy económico, rápido de implementar, útil en fases exploratorias de investigación o cuando los recursos son limitados y funciona para obtener datos preliminares o generar hipótesis. Debe tomarse en cuenta su alta probabilidad de sesgo, muy baja representatividad, que es imposible generalizar los resultados a toda la población con confianza estadística y puede omitir grupos importantes que no estén disponibles o visibles.

En el caso del servicio bancario supuesto, el investigador acude a diversas instituciones y encuesta a los clientes que permiten aplicarles la encuesta hasta conseguir el tamaño muestral sugerido.

5.1.8 No Probabilística: Muestreo por juicio o discrecional (Judgmental / Purposive Sampling)

El investigador selecciona deliberadamente casos que considera representativos o clave para el estudio, basándose en su conocimiento, experiencia o criterio. Esta técnica de muestreo es muy cercana al muestreo intencional, permite focalizarse en casos con información rica o relevante, es útil cuando se busca profundidad, no necesariamente generalización y puede ser más eficiente cuando se tienen expertos, casos especiales o muestras pequeñas. Dentro de las **desventajas está la subjetividad** del investigador, dado que puede introducir sesgo, la representatividad es dudosa y de generalización limitada, difícil de estimar confiabilidad inferencial. Este muestreo se utiliza típicamente cuando el investigador busca una respuesta en específico, lo cual puede sesgar o dirigir el resultado, por lo que está muy lejos de representar los principios de una encuesta.

En el caso del estudio bancario, el investigador desea comprobar en su investigación que el servicio bancario es deficiente, de ahí que éste solo busca encuestar a clientes que cumplan con un perfil de salir enojados de la sucursal o que tengan malas opiniones de ésta. No significa que esta técnica busque cosas negativas, pero sí trata de encontrar la manera de comprobar lo que el investigador desea que se concluya.

6. Validez

La validez de un instrumento de investigación, como la encuesta, constituye un aspecto clave en la metodología, ya que garantiza que las mediciones realizadas correspondan efectivamente a los constructos que se pretende evaluar. Si un cuestionario no es válido, los resultados pueden ser engañosos, conducir a conclusiones equivocadas y toma de decisiones erróneas. La validez de una encuesta se refiere al grado en que sus preguntas representan de manera precisa la variable o el concepto teórico que se pretende estudiar. Sin una adecuada validez, la confiabilidad (entendida como la consistencia de los resultados) pierde relevancia metodológica, de acuerdo a Anastasi & Urbina (1997). Según Carmines & Zeller (1979), la validación del instrumento debe abordarse cumpliendo tres dimensiones fundamentales:

1. Validez de contenido, implica la revisión de expertos para asegurar que los reactivos representen adecuadamente el constructo teórico.
2. Validez de criterio, se evalúa mediante la comparación del instrumento con una medida externa (criterio) reconocida, para observar la correlación entre ambas.
3. Validez de constructo, se determina a través de métodos estadísticos, como análisis factorial, para confirmar que los reactivos se alineen con la dimensión teórica que se pretende medir.

Pedrosa et al. (2013) hacen una revisión literaria con algunos métodos de validación del instrumento de investigación. Algunos de ellos validan precisamente la coherencia, relevancia y claridad de cada reactivo, cómo éste ayuda a formar el constructor y a medirlo. Aquí también se realizan adecuaciones de la redacción, ortografía y sintaxis de los preguntas o afirmaciones. Estas revisiones de validación y construcción de reactivos se recomiendan realizar por un grupo de expertos, investigadores y profesionistas de campo.

Por ejemplo, si se está investigando sobre la percepción de la calidad del servicio bancario, una buena práctica es reunir al menos 3 expertos en el campo de investigación y al menos 3 expertos en el sector bancario. A continuación, se presentan técnicas de validación de contenido mediante juicios de expertos. Cada uno tiene un enfoque ligeramente diferente, pero su propósito común es garantizar que los reactivos de un instrumento de investigación (tal como, una encuesta tipo Likert) representen adecuadamente el constructo que se quiere medir.

Análisis Factorial – Tucker (1961)

Se utiliza para identificar la estructura subyacente de las variables y verificar si los reactivos se agrupan de manera coherente en los factores o dimensiones teóricas que el investigador planteó. Se realiza un análisis factorial exploratorio con el propósito de examinar las cargas factoriales asociadas a cada reactivo. Un reactivo se considera válido cuando presenta una carga significativa en el factor correspondiente, lo que indica que contribuye adecuadamente a la dimensión teórica que se pretende medir. Esta técnica es ampliamente utilizada para validar constructos en cuestionarios aplicados en ciencias sociales y psicometría. Para facilitar la reducción de dimensiones y mejorar la interpretación de los factores, se recomienda emplear el método de componentes principales en conjunto con la rotación VARIMAX.

Índice de Validez de Contenido (CVI) – Lawshe (1975)

El procedimiento consiste en solicitar a un panel de expertos que clasifiquen cada reactivo como “esencial”, “útil pero no esencial” o “no esencial”. Posteriormente, se aplica el índice de razón de contenido (CVR), la cual considera la proporción de jueces que calificaron el reactivo como esencial en relación con el total de evaluadores. Este valor se compara contra los valores críticos establecidos por Lawshe; aquellos reactivos que no alcanzan el umbral mínimo son susceptibles de ser eliminados o modificados.

Índice de Congruencia Reactivo-Objetivo – Rovinelli y Hambleton (1977)

Este procedimiento se centra en evaluar la congruencia de cada reactivo con el objetivo de aprendizaje o el constructo que se desea medir. Para ello, los jueces asignan una puntuación a de “+1” si consideran que el reactivo es congruente, “0” si no están seguros o “-1” si lo consideran no congruente. A partir de estas valoraciones, se calcula un coeficiente que oscila entre -1 a +1; en donde, los valores cercanos a +1 indican un alto nivel de congruencia entre el reactivo y el constructo evaluado.

Índice de Congruencia – Hambleton (1980)

Este enfoque amplía trabajos previos, enfocándose específicamente en la correspondencia entre los reactivos y las dimensiones del constructo. Los jueces deben indicar a qué dimensión o categoría pertenece cada reactivo y, posteriormente, se calcula la proporción de coincidencia entre sus clasificaciones y la asignación teórica propuesta. Este procedimiento permite evaluar la coherencia entre la estructura conceptual del instrumento y la percepción experta sobre la ubicación de cada reactivo.

V de Aiken – Aiken (1980)

Se utiliza para evaluar la relevancia o pertinencia de cada reactivo en una escala ordinal, comúnmente de 1 a 5. El cálculo del coeficiente considera la diferencia entre la puntuación asignada y la mínima posible, el número de jueces y el número de categorías de respuesta disponibles. El valor resultante, que oscila entre 0 y 1, refleja el grado de validez de contenido, siendo los valores cercanos a 1 indicativos de una alta pertinencia del reactivo evaluado.

Escalamiento Multidimensional y Análisis de Clúster – Sireci & Geisenger (1992)

El enfoque propuesto prioriza la utilización de técnicas estadísticas avanzadas, como el escalamiento multidimensional y el análisis de clúster, para representar gráficamente las relaciones entre los reactivos y agruparlos en clúster o conjuntos según su similitud. El objetivo principal es verificar si los reactivos se organizan en agrupamientos coherentes con la estructura teórica: por ejemplo, si los reactivos de “motivación” se agrupan de forma distinta a los reactivos de “satisfacción”. Esta metodología resulta especialmente útil cuando se tiene un gran número de reactivos y múltiples dimensiones teóricas.

Coefficiente de Validez de Contenido – Hernández-Nieto (2002)

Este coeficiente representa una actualización del índice de Aiken, incorporando además la evaluación de aspectos como la claridad, coherencia y relevancia de cada reactivo. Este ampliamente considerado como una opción óptima si se está cuando se trabaja con escalas tipo Likert, debido a su aplicabilidad y robustez. El coeficiente obtenido oscila entre 0 y 1, siendo valores ≥ 0.80 indicativos de una validez aceptable. Su uso se ha extendido en investigaciones recientes en Latinoamérica, gracias a su sencillez metodológica y amplia documentación.

7. Confiabilidad

La confiabilidad se refiere a la estabilidad y **consistencia interna** de un instrumento de medición; es decir, su capacidad para generar resultados similares cuando se aplica en condiciones equivalentes. Medir la confiabilidad es fundamental, dado que sin ella no se puede garantizar que los datos obtenidos reflejen con precisión el constructo que se pretende estudiar; en cambio, los resultados podrían estar distorsionados por errores de medición. La confiabilidad permite distinguir entre la variación real del fenómeno observado y el “ruido” o error inherente al instrumento. Cuanto mayor sea la confiabilidad del instrumento, mayor será la precisión de las estimaciones estadísticas, como medias, correlaciones, entre otras. Además, un instrumento confiable fortalece la interpretación de los resultados, aporta credibilidad al estudio y facilita tanto su replicación como la comparación con estudios similares. Cabe destacar que, sin confiabilidad, la validez entendida como la capacidad del instrumento para medir lo que realmente se propone, tampoco puede ser evaluada apropiadamente.

El coeficiente Alfa de Cronbach es una medida estadística utilizada para evaluar la consistencia interna de un conjunto de reactivos que se supone están diseñados para medir un mismo constructo teórico (Nunnally & Bernstein, 1994). Su cálculo se basa en el número de reactivos, la varianza individual de cada reactivo y la varianza total del puntaje compuesto. En términos generales, cuanto mayor sea la correlación entre los reactivos y mayor sea el número de reactivos, más elevado será el valor de alfa. Esta métrica resulta especialmente útil en instrumentos que emplean escalas tipo Likert u otros formatos con reactivos relacionados, donde se espera que todos reflejen una misma dimensión latente. En la Tabla 5 se presentan algunos criterios ampliamente aceptados y consideraciones relevantes para la interpretación del alfa de Cronbach, respaldados por autores como González & Pazmiño (2015), Oviedo & Campo-Arias (2005) y Frías-Navarro (2005).

Tabla 5. Interpretación del Alfa de Cronbach.

Rango de Alfa de Cronbach	Interpretación usual	Consideraciones
≥ 0.90	Excelente	Un valor muy alto de alfa, podría sugerir que existe redundancia de reactivos en la escala. Sin embargo, esto no garantiza necesariamente que la escala sea unidimensional, ni que los reactivos estén evaluando con precisión el constructo deseado (validez). Por ello, es recomendable complementar el análisis con técnicas como el análisis factorial, que permite explorar la estructura subyacente de la escala y aportar evidencia adicional de validez.
0.80 – 0.89	Bueno	Este nivel se considera generalmente aceptable o adecuado para investigaciones o estudios concluyentes que persiguen resultados definitivos.
0.70 – 0.79	Aceptable	Este nivel puede considerarse suficiente, especialmente en el desarrollo de nuevas escalas o en investigaciones de carácter exploratorio.
0.60 – 0.69	Cuestionable / Mínimo aceptable	En investigaciones exploratorias, este nivel puede ser considerado aceptable; no obstante, es fundamental examinar si ciertos reactivos están afectando negativamente la consistencia interna del instrumento.
< 0.60	Bajo / Poco confiable	El instrumento requiere una revisión integral, que podría incluir la reformulación de reactivos, la eliminación de aquellos que presentan baja calidad psicométrica y, en caso contrario, la ampliación del número de reactivos para mejorar su consistencia y representatividad.

El Alfa de Cronbach es el indicador de consistencia interna más usado, pero no es el único. De hecho, hay críticas a su uso exclusivo porque puede sobreestimar o subestimar la confiabilidad dependiendo del número de reactivos y su dimensionalidad (véase Tabla 6). Si el cuestionario es unidimensional y con reactivos tipo Likert, el coeficiente Alfa de Cronbach suele ser suficiente. Si el instrumento es multidimensional o hay duda sobre la homogeneidad de los reactivos, se recomienda Omega de McDonald o consistencia compuesta. Para reactivos dicotómicos (sí/no, verdadero/falso, ...), usa KR-20 en lugar de alfa; esta información es sustentada por Campo-Arias & Oviedo (2008) y Dunn et al. (2014).

Tabla 6. Indicadores de consistencia alternos.

Indicador	Descripción	Ventajas	Limitaciones
Omega de McDonald (ω)	Basado en modelos de ecuaciones estructurales, estima la confiabilidad considerando cargas factoriales de cada reactivo del instrumento.	Más preciso que el alfa, no asume tau-equivalencia (igual carga factorial para todos los reactivos).	Requiere análisis factorial y <i>software</i> estadístico más avanzado (R, SPSS AMOS, Mplus).
Coficiente Guttman Lambda-2 (λ_2)	Alternativa clásica que puede dar una estimación más alta que alfa.	Más robusto que el alfa en ciertas condiciones.	Poco utilizado en investigación aplicada por complejidad de cálculo.
Coficiente de Spearman-Brown	Empleado para pruebas divididas en dos mitades (<i>split-half reliability</i>).	Útil cuando el cuestionario es largo y se requiere comprobar consistencia entre mitades.	No evalúa consistencia reactivo por reactivo.
Coficiente KR-20 y KR-21 (Kuder-Richardson)	Usado para reactivos dicotómicos (sí/no, correcto/incorrecto, ...). KR-20 es la versión más precisa.	Ideal para cuestionarios de conocimiento o reactivos dicotómicos.	Sólo válido para variables dicotómicas.
Índice de Consistencia Interna de Hoyt	Similar al análisis de varianza para medir consistencia entre reactivos.	Útil en contextos educativos y psicométricos.	Menos conocido y menos disponible en <i>software</i> comercial.
Consistencia Compuesta (<i>Composite Reliability</i> , CR)	Basada en análisis factorial confirmatorio, calcula la proporción de varianza verdadera frente a la varianza total.	Muy recomendada en modelos SEM (<i>Structural Equation Modeling</i>).	Requiere datos y <i>software</i> para análisis factorial confirmatorio.
Coficiente H de Hancock y Mueller	Estima la confiabilidad del constructo latente de forma más realista que alfa.	Recomendado en estudios de validación avanzada.	Poco implementado en investigación aplicada por desconocimiento.

Se dispone de métodos estadísticos que permiten calcular el intervalo de confianza del Alfa de Cronbach, lo cual resulta especialmente valioso, dado que proporciona no sólo una medida puntual de la confiabilidad, sino también un rango dentro del cual es probable que se ubique el valor verdadero de dicha confiabilidad en la población (Bonett, 2010; Charter, 1999). Una forma común de estimar los límites del intervalo de confianza para el coeficiente Alfa de Cronbrach son las expresiones [5] y [6]; donde α el nivel de significación, k el número de reactivos del instrumento y n el número de participantes; entonces, $v_1 = n-1$ y $v_2 = (n-1)(k-1)$ son los grados de libertad de la distribución F de Fisher en la cual se basa.

Límite inferior		Límite superior	
$IC_i = 1 - \frac{(1 - \alpha)F_{1-\frac{\alpha}{2},v_1,v_2}}{1}$	[5]	$IC_s = 1 - \frac{(1 - \alpha)}{F_{\frac{\alpha}{2},v_1,v_2}}$	[6]

Aunque la estimación exacta del intervalo de confianza del Alfa de Cronbach puede realizarse mediante la distribución F de Fisher, en la práctica, la mayoría de los paquetes estadísticos (como SPSS, R, Jamovi, Stata, entre otros) emplean métodos de aproximación o técnicas de re-muestreo como el *bootstrap*. Estos enfoques suelen ser más robusto, dado que no requieren asumir estrictamente la normalidad de los datos. El intervalo de confianza (IC) proporciona un rango dentro del cual es probable que se encuentre el verdadero valor de confiabilidad de la escala de la población. Por ejemplo, si el alfa estimada es 0.82 y el IC al 95% es [0.78, 0.86], se puede afirmar con un 95% de confianza que la consistencia interna de la escala en la población se encuentra dentro de ese rango. Un IC excesivamente amplio podría indicar que la muestra es insuficiente o una baja homogeneidad entre los reactivos en relación con el constructo que se pretende medir.

8. Conclusiones

En el desarrollo de investigaciones en ciencias sociales, la calidad metodológica depende en gran medida de la rigurosidad con la que se establecen y aplican los criterios de selección de la muestra, así como los mecanismos de validación de los instrumentos de medición. La elección de la técnica de muestreo es un componente crítico, dado que determina el grado de representatividad de los participantes o unidades de análisis y, en consecuencia, incide directamente en la posibilidad de generalizar los hallazgos a la población objetivo. Asimismo, el tamaño muestral debe ser suficiente para garantizar una adecuada potencia estadística y minimizar el error de estimación, permitiendo obtener resultados estadísticamente significativos y confiables.

Por otra parte, la validez del instrumento asegura que éste mida de manera efectiva el constructo teórico de interés propuesto, mientras que la confiabilidad garantiza la consistencia de las respuestas ante diferentes condiciones de aplicación. En este sentido, indicadores como el Alfa de Cronbach se consolidan como herramientas clave para evaluar la consistencia interna de escalas y cuestionarios. Su utilidad se potencia al complementarse con el intervalo de confianza, el cual proporciona un rango estimado en el que probablemente se encuentra el verdadero valor de la confiabilidad, incrementando la precisión y la transparencia en la interpretación de los resultados.

En síntesis, la articulación de una adecuada técnica de muestreo, el cálculo pertinente del tamaño de muestra y la comprobación sistemática de la validez y confiabilidad de los instrumentos, complementada con la interpretación de intervalos de confianza, constituye un eje fundamental para asegurar la solidez científica de una investigación. Estos elementos no sólo fortalecen la credibilidad del estudio, sino que también posibilitan que sus conclusiones aporten de manera significativa al avance del conocimiento en el ámbito de las ciencias sociales.

9. Referencias

- Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement*, 40(4), 955–959.
- Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (7th ed.). Prentice Hall.
- Bonett, D. G. (2010). *Varying coefficient alpha reliability: A simple and practical alternative to Cronbach's alpha*. *Organizational Research Methods*, 13(2), 186–202. <https://doi.org/10.1177/1094428109356478>
- Buckingham, A., & Saunders, P. (2004). *The Survey Methods Workbook: From Design to Analysis*. Wiley.
- Campo-Arias, A., & Oviedo, H. C. (2008). Propiedades psicométricas de una escala: la consistencia interna. *Revista de Salud Pública*, 10(5), 831-839. <https://doi.org/10.1590/S0124-00642008000500015>
- Carmine, E. G., & Zeller, R. A. (1979). *Reliability and validity assessment*. SAGE Publications.
- Charter, R. A. (1999). *Sample size requirements for precise estimates of reliability, generalizability, and validity coefficients*. *Journal of Clinical and Experimental Neuropsychology*, 21(4), 559–566. <https://doi.org/10.1076/jcen.21.4.559.889>
- Chaudhuri, A., & Pal, S. (2023). *A comprehensive textbook on sample surveys*. Springer. <https://doi.org/10.1007/978-981-19-1418-8>
- Cochran, W. G. (1977). *Sampling techniques* (3rd ed.). Wiley.
- Comrey, A. L., & Lee, H. B. (1992). *A first course in factor analysis* (2nd ed.). Lawrence Erlbaum Associates.
- Desu, M. M., & Raghavarao, D. (1990). *Sample size methodology*. Academic Press.
- Dunn, T. J., Baguley, T., & Brunnsden, V. (2014). From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation. *British Journal of Psychology*, 105(3), 399–412. <https://doi.org/10.1111/bjop.12046>
- Elfil, M., & Negida, A. (2017). *Sampling Methods in Clinical Research; an Educational Review*. *Journal of Clinical and Translational Research*, PMC.
- Frías-Navarro, D. (2021). *Apuntes de consistencia interna de las puntuaciones de un instrumento de medida*. Universidad de Valencia.
- González Alonso, J., & Pazmiño Santacruz, M. (2015). Cálculo e interpretación del alfa de Cronbach para el caso de validación de la consistencia interna de un cuestionario, con dos posibles escalas tipo Likert. *Revista Publicando*, 2(1), 62-77.
- González Palomo, I. G., Treviño Hernández, N. J., García León, A. M., & González Trejo, E. S. (2023). *Probabilidad y estadística* (1ª ed.). Universidad Autónoma de Nuevo León.
- Gravetter, F. J., & Wallnau, L. B. (2017). *Statistics for the Behavioral Sciences* (10th ed.). Cengage Learning.
- Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2019). *Multivariate data analysis* (8th ed.). Cengage.
- Hambleton, R. K. (1980). Test score validity and standard setting. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art* (pp. 80–123). Johns Hopkins University Press.
- Henry, G. T. (1990). *Practical sampling*. SAGE Publications.

- Hernández Sampieri, R., Fernández Collado, C., & Baptista Lucio, P. (2014). *Metodología de la investigación* (6ª ed.). McGraw-Hill Education.
- Hernández-Nieto, R. A. (2002). *Contribuciones al análisis estadístico de la validez de contenido*. Universidad de Los Andes.
- Kerlinger, F. N., & Lee, H. B. (2002). *Foundations of Behavioral Research* (4th ed.). Harcourt College Publishers.
- Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28(4), 563–575.
- Likert, R. (1932). A Technique for the Measurement of Attitudes. *Archives of Psychology*, 22(140), 1–55.
- McCombes, S. (2023, 22 de junio). *Sampling methods | Types, Techniques & Examples*. Scribbr. Retrieved from <https://www.scribbr.com/methodology/sampling-methods/> (Scribbr)
- Namakforoosh, M. N. (2005). *Metodología de la investigación* (3ª ed.). Limusa.
- Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.
- Oviedo, H. C., & Campo-Arias, A. (2005). Aproximación al uso del coeficiente alfa de Cronbach. *Acta Colombiana de Psicología*, 8(1), 103-118. SciELO Colombia.
- Pedrosa, I., Suárez-Álvarez y García-Cueto, E. (2013). Evidencias sobre la Validez de Contenido: Avances Teóricos y Métodos para su Estimación [Content Validity Evidences: Theoretical Advances and Estimation Methods]. *Acción Psicológica*, 10(2), x-xx. <http://dx.doi.org/10.5944/ap.10.2.11820>
- Publications, ICI UMN. (n.d.). Advantages and Disadvantages of Non-Probability Sampling. RTC/OM HCBS Measurement Education Modules. (publications.ici.umn.edu)
- Roscoe (1975): *Fundamental Research Statistics for the Behavioral Sciences* (2nd ed.).
- Rositas Martínez, J. (2014). Los tamaños de las muestras en encuestas de las ciencias sociales y su repercusión en la generación del conocimiento. *Innovaciones De Negocios*, 11(22), 235–268. <https://doi.org/10.29105/rinn11.22-4>
- Rovinelli, R. J., & Hambleton, R. K. (1977). On the use of content specialists in the assessment of criterion-referenced test item validity. *Dutch Journal of Educational Research*, 2, 49–60.
- Sireci, S. G., & Geisinger, K. F. (1992). Using subject-matter experts to assess content validity. *Educational Assessment*, 1(3), 277–290.
- Sudman, S. (1976). *Applied sampling*. Academic Press.
- Tucker, L. R. (1961). A suggested alternative formulation of the method of paired comparisons. *Psychometrika*, 26, 173–199.