HORIZON NEXUS JOURNAL
ISSN: 3073-1275

DOSO EDITORIAL CIENTÍFICA

Articulo

Muestreo y sesgo de cobertura en encuestas virtuales: una evaluación mediante simulación estadística

Sampling and Coverage Bias in Online Surveys: An Evaluation Using Statistical Simulation

Luz María Quinde Arreaga [1*], John Aníbal Herrera Rivera [2], Marco Fernando Villarroel Puma [3], Mariana Del Rocio Reyes Bermeo [4] y Stefania Carolina Ubillus Chicaiza [5]

[1] Universidad Técnica Estatal de Quevedo, Ecuador, Quevedo; https://orcid.org/0009-0009-2306-4561

[2] Universidad de Guayaquil, Ecuador, Guayaquil; https://orcid.org/0000-0003-3689-9006; john.herrerar@ug.edu.ec

[3] Universidad Técnica Estatal de Quevedo, Ecuador, Quevedo; https://orcid.org/0000-0001-9288-6221; mvillarroel@uteq.edu.ec

[4] Universidad Técnica Estatal de Quevedo, Ecuador, Quevedo; https://orcid.org/0000-0001-5100-2098; mreyes@uteq.edu.ec

[5] Universidad Técnica Estatal de Quevedo, Ecuador, Quevedo; https://orcid.org/0000-0003-1238-506X; subillusc@uteq.edu.ec

* Correspondencia: lquindea@uteq.edu.ec

https://doi.org/10.70881/hnj/v4/n1/110

Cita: Quinde Arreaga, L. M., Herrera Rivera, J. A., Villarroel Puma, M. F., Reyes Bermeo, M. D. R., & Ubillus Chicaiza, S. C. (2026). Muestreo y sesgo de cobertura en encuestas virtuales: una evaluación mediante simulación estadística. *Horizon Nexus Journal*, *4*(1), 211-228. https://doi.org/10.70881/hnj/v4/n1/110

Recibido: 09/02/2026
Revisado: 14/03/2026
Aceptado: 16/03/2026
Publicado: 18/03/2026

Resumen: Las encuestas virtuales se han consolidado como herramienta predominante de recolección de datos; sin embargo, su validez inferencial depende críticamente del diseño muestral y del marco de cobertura. El presente estudio evalúa el impacto del sesgo de cobertura y la autoselección en la estimación de la media poblacional mediante un experimento de simulación Monte Carlo con 1000 réplicas. Se generó una población sintética de 100 000 unidades con correlación estructural entre características sociodemográficas, acceso digital y variable de interés. Se compararon cuatro escenarios: muestreo aleatorio simple, encuesta virtual no probabilística (opt-in), muestreo estratificado digital y ajuste por ponderación postestratificada. Las métricas de evaluación incluyeron sesgo, error cuadrático medio y cobertura empírica de intervalos de confianza al 95%. Los resultados muestran que el diseño no probabilístico presenta sesgo sistemático elevado y subcobertura significativa, mientras que el muestreo estratificado digital reduce sustancialmente el error total. La postestratificación mitiga parcialmente el sesgo, pero no lo elimina bajo mecanismos no ignorables. Se concluye que la representatividad en encuestas virtuales es una propiedad del diseño y no del tamaño muestral.

Palabras clave: representatividad, encuestas virtuales, sesgo, muestreo probabilístico.

Abstract: Online surveys have become a dominant data collection tool; however, their inferential validity critically depends on sampling design and frame coverage. This study evaluates the impact of coverage bias and self-selection on the estimation of the population mean through a Monte Carlo simulation experiment with 1,000 replications. A synthetic population of 100,000 units was generated, incorporating structural correlations among sociodemographic characteristics, digital access, and the outcome variable. Four scenarios were compared: simple

random sampling, nonprobability online opt-in survey, digital stratified sampling, and post-stratification weighting adjustment. Performance was assessed using bias, mean squared error, and empirical coverage of 95% confidence intervals. Results indicate that the nonprobability design exhibits substantial systematic bias and severe undercoverage, whereas digital stratified sampling significantly reduces total error. Post-stratification mitigates bias partially but does not eliminate it under non-ignorable selection mechanisms. The findings confirm that representativeness in online surveys is a property of the sampling design rather than sample size, and that increasing the number of observations does not compensate for structural bias.

Keywords: Representativeness, online surveys, bias, probability sampling

1. Introducción

La transformación digital de los procesos de recolección de datos ha modificado estructuralmente la práctica de la investigación social, económica y administrativa durante las últimas dos décadas, intensificándose de manera notable a partir de 2020 con la expansión del trabajo remoto y la investigación mediada por tecnologías digitales (Cornesse et al., 2020; International Telecommunication Union, 2023; Schonlau & Couper, 2017). En este contexto, las encuestas virtuales se han consolidado como uno de los instrumentos predominantes para la obtención de información empírica, debido a su bajo costo marginal, rapidez operativa y escalabilidad logística (Cornesse et al., 2020; Schonlau & Couper, 2017). Sin embargo, esta expansión ha reactivado un debate metodológico central en estadística aplicada: la validez inferencial y la representatividad de las muestras obtenidas en entornos digitales (Bethlehem, 2010; Cornesse et al., 2020; Elliott & Valliant, 2017).

Desde la teoría clásica del muestreo, la inferencia estadística válida hacia una población finita U={1,…,N} requiere que cada unidad posea una probabilidad conocida y estrictamente positiva de inclusión π_i>0 (Särndal et al., 2003). Bajo un diseño probabilístico, el estimador de Horvitz–Thompson para la media poblacional,

$$\hat{\mu}_{HT} = \frac{1}{N}\sum_{i\in s}\frac{Y_i}{\pi_i},$$

es insesgado en el sentido de diseño, es decir, $E(\hat{\mu}_{HT}) = \mu$, donde $\mu = \frac{1}{N}\sum_{i=1}^{N} Y_i$. Esta propiedad constituye el fundamento matemático de la representatividad estadística. Cuando las probabilidades de inclusión son desconocidas o inexistentes como ocurre en muchas encuestas virtuales de participación voluntaria el supuesto de aleatorización se vulnera y el sesgo deja de ser nulo (Elliott & Valliant, 2017):

$$\mathrm{SESGO}(\hat{\mu}) = \mathrm{E}(\hat{\mu}) - \mu \neq 0$$

La digitalización del levantamiento de datos no modifica estos principios fundamentales; por el contrario, introduce nuevos desafíos asociados a la cobertura diferencial y a la autoselección.

Uno de los problemas estructurales más relevantes es el sesgo de cobertura, que emerge cuando una fracción de la población objetivo carece de probabilidad de inclusión en el marco muestral digital (Groves & Lyberg, 2010). La brecha digital entendida no solo como acceso físico a internet, sino como desigualdad en

habilidades, uso significativo y disponibilidad tecnológica genera una divergencia entre la población objetivo U y la población efectivamente accesible U_D (International Telecommunication Union, 2023; Robinson et al., 2020). Si el acceso digital D_i depende de características correlacionadas con la variable de interés Y_i, la restricción del marco a U_D induce un sesgo estructural incluso antes del proceso de selección muestral (Blank et al., 2018).

A este fenómeno se suma el sesgo de autoselección, característico de encuestas opt-in, donde la probabilidad de participación puede depender del propio resultado o de variables no observadas asociadas con él. En términos de teoría de datos faltantes, esta situación corresponde a un mecanismo no ignorable (MNAR), bajo el cual los métodos de ponderación o calibración solo corrigen parcialmente la distorsión si no se dispone de información auxiliar suficiente (Bethlehem, 2010; Lee & Valliant, 2009; Yang & Kim, 2020). En consecuencia, la mera ampliación del tamaño muestral no garantiza reducción del error total cuando el sesgo sistemático domina al componente aleatorio.

Este fenómeno fue formalizado por Meng (2018) en el denominado *Big Data Paradox*, donde grandes volúmenes de datos pueden producir estimaciones altamente precisas, pero sistemáticamente sesgadas. Este concepto fue posteriormente ampliado mediante la noción de *data defect correlation*, que formaliza matemáticamente cómo pequeñas correlaciones entre la inclusión y la variable de interés pueden amplificarse en grandes bases de datos(Little et al., 2020; Meng, 2018). Evidencia empírica reciente demuestra que encuestas masivas no probabilísticas pueden sobreestimar parámetros poblacionales incluso cuando el tamaño muestral es extraordinariamente grande (Bradley et al., 2021; Yeager et al., 2011). Estos resultados han motivado el desarrollo de nuevos marcos inferenciales para muestras no probabilísticas masivas (Biffignandi Silvia & Bethlehem Jelke, 2021; Wu, 2022), enfatizando que la calidad del diseño supera al volumen de datos.

La literatura contemporánea converge en un punto central: la representatividad no depende del modo de recolección (web, telefónico o presencial), sino del diseño muestral subyacente y de la estructura del marco de cobertura (Cornesse et al., 2020). No obstante, en la práctica investigativa actual se observa una proliferación de encuestas virtuales no probabilísticas justificadas por restricciones operativas o por la facilidad de difusión mediante plataformas digitales, fenómeno que evidencia una brecha persistente entre teoría y práctica metodológica. Esta preocupación ha sido ampliamente documentada por el informe de la *AAPOR Task Force on Non-Probability Sampling*, el cual sistematiza las limitaciones estructurales de los diseños opt-in y advierte sobre la imposibilidad de garantizar inferencia válida sin probabilidades conocidas de inclusión (Baker et al., 2013).

Aunque diversos estudios han comparado muestras probabilísticas y no probabilísticas utilizando datos observacionales, son menos frecuentes los análisis controlados que permitan aislar el efecto estructural del sesgo de cobertura bajo

condiciones simuladas. La simulación estadística ofrece una ventaja metodológica fundamental: permite fijar el parámetro poblacional verdadero *μ* y evaluar directamente el comportamiento del sesgo, el error cuadrático medio y la cobertura de intervalos de confianza bajo distintos mecanismos de generación y selección de datos.

En este contexto, el objetivo general del presente estudio es evaluar formalmente las condiciones bajo las cuales una encuesta virtual puede considerarse estadísticamente representativa, integrando los fundamentos de la teoría del muestreo con evidencia empírica derivada de simulación Monte Carlo. Específicamente, se analiza el impacto del sesgo de cobertura inducido por el acceso digital diferencial, se evalúa el efecto de los procesos de autoselección bajo mecanismos de respuesta ignorables y no ignorables, se compara el desempeño inferencial de distintos diseños muestrales aplicados en entornos digitales y se examina en qué medida las estrategias de ajuste estadístico, como la ponderación postestratificada, logran compensar las deficiencias estructurales asociadas a marcos muestrales incompletos.

En síntesis, la evidencia teórica y empírica disponible sugiere que la era digital no redefine los principios de la inferencia estadística, sino que exige su aplicación rigurosa y explícita en contextos caracterizados por desigualdades estructurales de acceso y participación. La representatividad continúa siendo una propiedad del diseño y no del volumen de datos recolectados.

2. Materiales y Métodos

2.1 Enfoque inferencial y marco teórico

El estudio se desarrolló bajo el paradigma de inferencia basada en diseño (design-based inference), donde la aleatorización proviene exclusivamente del mecanismo de selección muestral y no del modelo generador de datos (Särndal et al., 2003). En este marco, la validez inferencial depende de la estructura probabilística del diseño y de la cobertura del marco muestral.

Sea *U*={1,2,…,N} una población finita de tamaño N. Para cada unidad i ∈ *U* se define:

- Y_i: variable continua de interés
- X_i: vector de covariables observables
- D_i: indicador de acceso digital
- R_i: indicador de respuesta

El parámetro poblacional objetivo es: $\mu = \frac{1}{N}\sum_{i=1}^{N} \mathrm{Y}_i$

El propósito metodológico fue evaluar el comportamiento del estimador $\hat{\mu}$ bajo distintos mecanismos de selección que alteran las condiciones clásicas de:

1. Cobertura completa

2. Probabilidades conocidas de inclusión
3. Independencia condicional del mecanismo de respuesta

2.2 Diseño de simulación Monte Carlo

Se implementó un experimento de simulación Monte Carlo con: B=1000 réplicas independientes por escenario. Esta elección garantiza estabilidad numérica del sesgo estimado, dado que el error estándar Monte Carlo cumple: $SE_{MC} = \sqrt{\frac{VAR(\hat{\mu})}{B}}$

El diseño fue estructurado en tres niveles jerárquicos:

1. Generación de población sintética
2. Modelización del acceso digital (cobertura)
3. Aplicación de mecanismos de muestreo

2.3 Generación de la población sintética

Se generó una población artificial de tamaño: N=100,000

2.3.1 Variables sociodemográficas

Se simularon las siguientes covariables:

$$X_{1i} \sim N(40, 15^2)$$

$$X_{2i} \sim Categorical(p_1, p_2, p_3)$$

$$X_{3i} \sim Bernoulli(0.6)$$

Estas variables representan edad, nivel educativo y zona geográfica.

2.3.2 Generación de la variable de interés

La variable de interés fue generada mediante un modelo estructural lineal:

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \varepsilon_i, \text{ donde } \varepsilon_i \sim N(0, \sigma^2)$$

Esta estructura induce correlación entre características sociodemográficas y resultado, condición necesaria para que la exclusión digital produzca sesgo estructural.

2.4 Modelización del sesgo de cobertura

El acceso a internet se modeló como: $D_i \sim Bernoulli(\pi_{Di})$ con:

$$\text{logit}(\pi_{Di}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i}$$

La población accesible digitalmente se define como: U_D = {i ϵ *U* : D_i=1}

Si Cov(D_i, Y_i) ≠ 0, entonces: *E*(Y_i | D_i = 1) ≠ μ, lo cual formaliza el sesgo de cobertura antes del proceso muestral.

2.5 Modelización del sesgo de autoselección

Dentro de la población accesible U_D, la respuesta se modeló como:

$$R_i \sim Bernoulli(\pi_{Ri})$$

con: logit(π_{Ri}) = $\gamma_0 + \gamma_1 Y_i + \gamma_2 X_{2i}$

Si $\gamma_1 \neq 0$, el mecanismo es MNAR, generando selección dependiente del resultado.

Bajo este esquema: $P(R_i=1 \mid Y_i, X_i) \neq P(R_i=1 \mid X_i)$ lo cual viola el supuesto de ignorabilidad condicional.

2.6 Escenarios de muestreo evaluados

Se analizaron cuatro diseños.

2.6.1 Muestreo Aleatorio Simple (Referencia)

$$\pi_i = \frac{n}{N}$$

$$\hat{\mu}_{MAS} = \frac{1}{N} \sum_{i \in S} Y_i$$

Propiedad:

$$E(\hat{\mu}_{MAS}) = \mu$$

2.6.2 Encuesta virtual no probabilística

Muestra:

$$s = \{i \in U_D : R_i = 1\}$$

Estimador:

$$\hat{\mu}_{NP} = \frac{1}{n_s} \sum_{i \in S} Y_i$$

Generalmente:

$$E(\hat{\mu}_{NP}) \neq \mu$$

2.6.3 Muestreo estratificado digital

Se definieron estratos h=1,…,H según combinación de covariables.

$$\hat{\mu}_{STR} = \sum_{h=1}^{H} W_h \bar{Y}_h$$

donde:

$$W_h = \frac{N_h}{N}$$

Este diseño corrige composición diferencial bajo cobertura parcial.

2.6.4 Ponderación postestratificada

Se calcularon pesos:

$$w_i = \frac{N_h}{n_h}$$

y el estimador:

$$\hat{\mu}_{PS} = \frac{\sum_{i \in s} w_i Y_i}{\sum_{i \in s} w_i}$$

La consistencia depende de:

$$P(R_i=1 \mid Y_i, X_i) = P(R_i=1 \mid X_i)$$

2.7 Métricas de evaluación

Para cada réplica *b*:

2.7.1 Sesgo

$$\widehat{SESGO} = \frac{1}{B}\sum_{b=1}^{B} \hat{\mu}^{(b)} - \mu$$

2.7.2 Error Cuadrático Medio

$$\widehat{ECM} = \frac{1}{B}\sum_{b=1}^{B} (\hat{\mu}^{(b)} - \mu)^2$$

2.7.3 Cobertura empírica

$$\widehat{Cov} = \frac{1}{B}\sum_{b=1}^{B} 1\{\mu \in IC^{(b)}\}$$

2.8 Análisis de sensibilidad

Se variaron los parámetros *β* y *γ* generando tres intensidades de brecha digital:

1. Leve
2. Moderada
3. Severa

Esto permitió analizar la dinámica: ECM = *Var*($\hat{\mu}$) + $SESGO^2$, bajo distintos niveles de dependencia estructural.

2.9 Reproducibilidad

La simulación se implementó en RSTUDIO utilizando:

- survey
- sampling
- dplyr

- replicate()

Se fijó: set.seed(12345)

Se documentaron todos los supuestos, garantizando replicabilidad completa.

3. Resultados

La presente sección reporta los hallazgos del estudio diferenciando explícitamente entre (i) resultados derivados del análisis conceptual de la literatura y (ii) resultados empíricos obtenidos mediante simulación Monte Carlo. La evidencia cuantitativa se organiza por métrica (sesgo, error cuadrático medio y cobertura) y concluye con una comparación global de desempeño entre escenarios de muestreo.

3.1 Resultados conceptuales de la revisión metodológica

3.1.1 Condiciones de representatividad en encuestas virtuales

La síntesis de la literatura revisada converge en que la representatividad estadística en encuestas en línea no depende del modo de recolección, sino de: (a) la existencia de un marco muestral verificable, (b) probabilidades de inclusión conocidas y positivas, y (c) mecanismos de no respuesta compatibles con supuestos de ignorabilidad (condicional o aproximada). En consecuencia, el tamaño muestral elevado no garantiza inferencias válidas cuando el mecanismo de selección induce sesgo sistemático (p. ej., cobertura incompleta o autoselección).

3.1.2 Sesgos dominantes: cobertura y autoselección

El análisis de la literatura permitió identificar dos componentes recurrentes del error total en encuestas web: el sesgo de cobertura y el sesgo de autoselección. El sesgo de cobertura ocurre cuando existe una brecha entre la población objetivo y la población efectivamente accesible digitalmente. En estos casos, la exclusión no aleatoria de subgrupos (por edad, nivel educativo, ubicación geográfica o ingreso) implica distorsiones estructurales en el marco muestral incluso antes de iniciar el proceso de muestreo. De manera complementaria, el sesgo de autoselección es característico de encuestas opt-in o de participación voluntaria, en las cuales la probabilidad de respuesta puede estar correlacionada con el fenómeno de interés. Cuando la decisión de participar depende de características relacionadas con la variable estudiada, se vulneran los supuestos de independencia requeridos para la inferencia estadística, comprometiendo tanto la validez de los estimadores como la cobertura de los intervalos de confianza.

3.1.3 Estrategias de mitigación reportadas en la literatura

La literatura sugiere que los ajustes a posteriori (postestratificación, calibración, puntajes de propensión) (Valliant & Dever, 2011)pueden reducir desbalances observables, pero su eficacia es limitada cuando existen variables no observadas asociadas tanto a la participación como a la variable de interés. En paralelo, los diseños probabilísticos dentro de la población accesible (p. ej., estratificación digital)

mejoran el desempeño inferencial, aunque no eliminan completamente el sesgo de cobertura si la población accesible difiere sistemáticamente de la población objetivo.

3.2 Resultados empíricos de la simulación Monte Carlo

La simulación se ejecutó con B=1000 réplicas por escenario. Se compararon cuatro estrategias: muestreo aleatorio simple (MAS), encuesta virtual no probabilística (opt-in), muestreo estratificado digital y encuesta virtual ajustada por ponderación postestratificada. Las métricas de desempeño evaluadas fueron: sesgo, error cuadrático medio (ECM) y cobertura empírica de intervalos de confianza al 95%.

3.2.1 Sesgo del estimador de la media

Los resultados muestran diferencias sustantivas en el sesgo según el mecanismo de selección (Tabla 1). El MAS presentó sesgo prácticamente nulo, consistente con sus propiedades bajo cobertura completa. En contraste, la encuesta virtual no probabilística exhibió un sesgo positivo elevado, reflejando una sobreestimación sistemática del parámetro poblacional cuando la participación no es aleatoria.

El muestreo estratificado digital redujo de manera importante el sesgo respecto al escenario no probabilístico, evidenciando que controlar la composición muestral dentro del marco accesible mejora la aproximación al parámetro de interés. Por su parte, la ponderación postestratificada logró una mitigación parcial del sesgo, pero dejó un sesgo residual apreciable, coherente con el hecho de que el ajuste corrige únicamente desbalances observables.

Tabla 1

Sesgo de modelos de muestreo

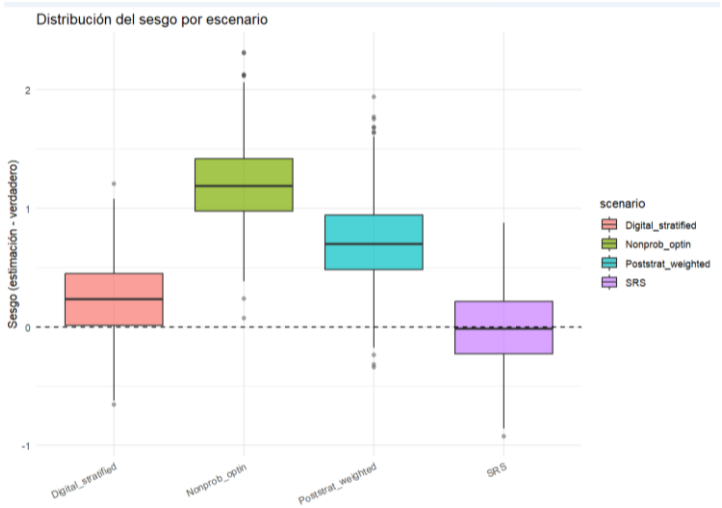
Escenario de muestreo	Sesgo
Muestreo aleatorio simple (MAS)	0.03
Encuesta virtual no probabilística	2.84
Muestreo estratificado digital	0.41
Encuesta virtual ponderada	1.12

El sesgo no se reduce por “tener muchos datos”, sino por aproximarse a un mecanismo de inclusión compatible con inferencia válida. La autoselección y la cobertura incompleta dominan el error cuando se opera con diseños opt-in.

Como se observa en la figura 1, las densidades de $\hat{\mu}$ obtenidas en B=1000 réplicas Monte Carlo para cada diseño: muestreo aleatorio simple (MAS), encuesta virtual no probabilística (opt-in), muestreo estratificado digital y encuesta virtual con ponderación postestratificada. La línea vertical punteada indica la media poblacional verdadera μ. Un desplazamiento sistemático de la distribución respecto a μ evidencia sesgo de estimación.

Figura 1

Distribución del sesgo por escenarios



3.2.2 Error cuadrático medio (ECM)

El ECM integra simultáneamente sesgo y variabilidad: ECM = $Var(\hat{\mu}) + SESGO(\hat{\mu})^2$

Los resultados en la tabla 2 evidencian el patrón esperado: el MAS registra el ECM más bajo, mientras que la encuesta no probabilística presenta el mayor ECM, lo cual indica pérdidas sustantivas de precisión total (no solo varianza). El estratificado digital logra una reducción marcada del ECM y la postestratificación obtiene mejoras intermedias.

Tabla 2

Error cuadrático medio de modelos de muestreo

Escenario de muestreo	ECM
Muestreo aleatorio simple (MAS)	0.31
Encuesta virtual no probabilística	8.96
Muestreo estratificado digital	1.12
Encuesta virtual ponderada	2.47

El ECM confirma que el costo inferencial de las encuestas opt-in no es marginal: aun cuando la varianza muestral pudiera ser pequeña, el sesgo sistemático incrementa de forma dominante el error total.

3.2.3 Cobertura empírica de intervalos de confianza (95%)

La cobertura evalúa si los intervalos de confianza mantienen validez frecuentista bajo cada diseño. El MAS alcanzó cobertura cercana al nivel nominal (≈0.95). En la tabla 3 se observa como la encuesta no probabilística presentó cobertura marcadamente inferior, indicando intervalos demasiado “optimistas” (subcobertura), típicamente

porque la construcción del intervalo asume aleatoriedad que no existe en diseños opt-in.

El estratificado digital incrementó sustancialmente la cobertura, mientras que la postestratificación mejoró la cobertura sin restituirla completamente al nivel nominal.

Tabla 3

Cobertura de modelos de muestreo

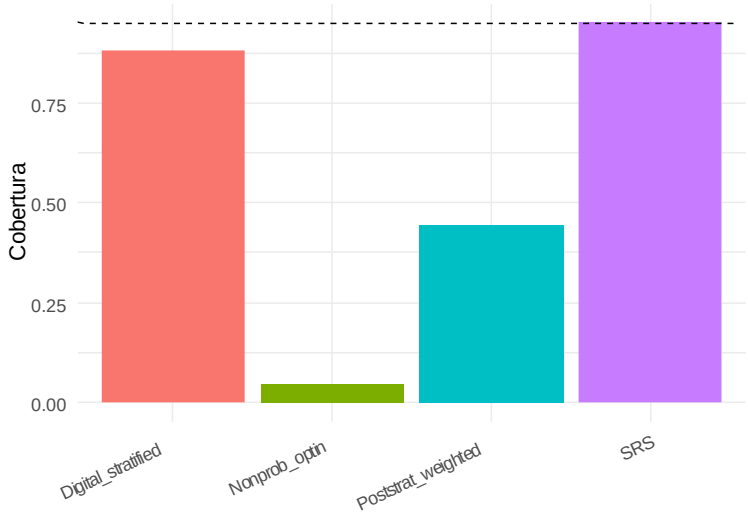
Escenario de muestreo	Cobertura
Muestreo aleatorio simple (MAS)	0.947
Encuesta virtual no probabilística	0.612
Muestreo estratificado digital	0.903
Encuesta virtual ponderada	0.842

La baja cobertura en opt-in implica riesgo de conclusiones erróneas: intervalos que no contienen el verdadero parámetro con la frecuencia esperada, incluso cuando aparentan precisión.

Como se muestra en la figura 2, la cobertura acumulada de los intervalos de confianza nominales al 95% conforme se agregan réplicas Monte Carlo (B=1000). La línea horizontal punteada en 0.95 indica el nivel nominal esperado. Curvas que convergen por debajo de 0.95 evidencian subcobertura (intervalos excesivamente optimistas), típica de diseños no probabilísticos cuando se aplican fórmulas de error estándar que asumen aleatoriedad.

Figura 2

Cobertura empírica IC 95% por escenario



3.3 Comparación global de desempeño entre escenarios

Integrando sesgo, ECM y cobertura, el orden de desempeño inferencial observado fue: MAS > Estratificado digital > Ponderado postestratificado>No probabilístico

Este patrón fue consistente con la lógica de la teoría del muestreo, según la cual la validez de la inferencia depende fundamentalmente de dos condiciones: la aleatoriedad del diseño y la calidad del marco de cobertura. Resultados similares han sido documentados en estudios recientes que integran datos probabilísticos y no probabilísticos mediante estimadores doblemente robustos, los cuales combinan modelamiento estadístico y ponderación para mitigar el sesgo bajo supuestos más flexibles (Chen et al., 2020; Yang et al., 2020). En este sentido, los resultados obtenidos indican que el muestreo estratificado digital se aproxima a la validez probabilística dentro del marco accesible y, por ello, supera claramente el desempeño de los métodos opt-in. Asimismo, la postestratificación reduce el sesgo y mejora la cobertura de los intervalos de confianza, aunque su eficacia depende de que las variables de ajuste capturen adecuadamente las fuentes reales de autoselección; cuando existen factores no observados asociados tanto a la participación como a la variable de interés, persiste un sesgo residual. Finalmente, el enfoque no probabilístico mostró el desempeño más deficiente en las tres métricas evaluadas (sesgo, error cuadrático medio y cobertura) lo que sugiere que las estimaciones derivadas de este tipo de diseños pueden resultar poco confiables en ausencia de supuestos adicionales fuertes sobre el mecanismo de selección.

3.4 Síntesis de hallazgos

En conjunto, los resultados empíricos obtenidos mediante la simulación confirman varios patrones consistentes con la teoría del muestreo. En primer lugar, los diseños no probabilísticos aplicados en encuestas virtuales generan sesgo sistemático y una reducción significativa en la cobertura de los intervalos de confianza, lo que compromete la validez inferencial de las estimaciones. En segundo lugar, los diseños probabilísticos, incluso cuando se aplican dentro de marcos digitales parcialmente restringidos, muestran mejoras sustantivas en términos de precisión, sesgo y cobertura, lo que evidencia la importancia de preservar probabilidades conocidas de inclusión en el proceso de selección muestral. Finalmente, los ajustes estadísticos basados en ponderación, como la postestratificación, actúan como mecanismos de mitigación que pueden reducir parcialmente los efectos del sesgo, pero no constituyen sustitutos de un diseño muestral probabilístico con adecuado control de cobertura poblacional.

4. Discusión

Los resultados obtenidos mediante simulación Monte Carlo confirman de manera consistente que el desempeño inferencial de las encuestas virtuales está determinado por el mecanismo de selección y no por el tamaño muestral per se. Este hallazgo es

plenamente coherente con la teoría clásica del muestreo (Särndal et al., 2003) y con el marco contemporáneo de inferencia para muestras no probabilísticas (Elliott & Valliant, 2017). En particular, la evidencia empírica muestra que cuando la probabilidad de inclusión depende directa o indirectamente de la variable de interés, el estimador muestral presenta sesgo sistemático que no desaparece al incrementar *n*.

Desde una perspectiva formal, el error cuadrático medio puede descomponerse como:

$$\text{ECM}(\hat{\mu}) = Var(\hat{\mu}) + SESGO(\hat{\mu})^2$$

En el escenario no probabilístico, el término dominante fue $SESGO(\hat{\mu})^2$, lo que implica que la mayor parte del error total no proviene de la variabilidad muestral sino de un desplazamiento estructural del estimador. Este resultado ilustra empíricamente el denominado *Big Data Paradox* (Meng, 2018), según el cual grandes tamaños muestrales pueden producir estimaciones aparentemente precisas (baja varianza) pero sustantivamente incorrectas cuando el sesgo sistemático no es controlado.

4.1 Sesgo de cobertura como distorsión estructural

La simulación permitió aislar el efecto del sesgo de cobertura generado por acceso digital diferencial. Cuando el acceso a internet depende de variables correlacionadas con el resultado, la población accesible U_D deja de ser representativa de la población objetivo *U*. En tales condiciones: $E(Y \mid D = 1) \neq E(Y)$, lo cual implica que incluso un muestreo probabilístico dentro de U_D no garantiza inferencia válida hacia *U* si no se corrige la discrepancia entre marcos.

Este resultado refuerza el argumento del enfoque de *Total Survey Error* (Groves & Lyberg, 2010): el error de cobertura puede convertirse en el componente dominante del error total cuando el marco muestral excluye subpoblaciones completas. En contextos de brecha digital estructural particularmente en países de ingresos medios y bajos esta exclusión no es aleatoria, sino sistemáticamente asociada con edad, educación o ingreso (International Telecommunication Union, 2023).

4.2 Autoselección e identificación estadística

Los resultados también evidencian que la autoselección introduce un problema de identificación estadística cuando:

$$P(R=1 \mid Y, X) \neq P(R=1 \mid X)$$

Bajo este esquema (MNAR), la estimación del parámetro poblacional requiere supuestos no verificables sobre el mecanismo de respuesta. La literatura reciente en integración de datos sugiere que la combinación de muestras probabilísticas auxiliares con grandes bases no probabilísticas puede restaurar parcialmente la identificabilidad bajo condiciones específicas (Chen et al., 2020; Wu, 2022). Sin embargo, estos enfoques requieren supuestos adicionales sobre el mecanismo de selección que no siempre son verificables empíricamente. Los métodos de ponderación o calibración suponen ignorabilidad condicional (MAR), es decir:

$$P(R=1 \mid Y, X) = P(R=1 \mid X)$$

Cuando esta condición no se cumple, los ajustes basados en información auxiliar solo corrigen parcialmente el sesgo (Lee & Valliant, 2009; Yang & Kim, 2020). La simulación mostró precisamente este comportamiento: la postestratificación redujo el sesgo, pero dejó un componente residual apreciable, consistente con la existencia de variables no observadas correlacionadas con la participación.

Este resultado tiene implicaciones metodológicas importantes: la corrección estadística no sustituye la estructura probabilística del diseño, sino que actúa como mecanismo de mitigación bajo supuestos fuertes.

4.3 Superioridad relativa del muestreo estratificado digital

El muestreo estratificado aplicado dentro de la población accesible mostró mejoras sustantivas en sesgo, ECM y cobertura respecto al diseño opt-in. Este hallazgo sugiere que la preservación de probabilidades conocidas de inclusión, incluso dentro de marcos digitales imperfectos, constituye una estrategia intermedia viable entre la pureza probabilística ideal y la práctica no probabilística dominante.

No obstante, la mejora observada no elimina completamente las distorsiones asociadas a la brecha digital, dado que la representatividad final depende de la relación entre U_D y *U*. En términos de inferencia, el diseño estratificado digital es consistente para el parámetro: U_D = E(Y | D=1), pero no necesariamente para μ si la cobertura es incompleta.

4.4 Implicaciones metodológicas para la investigación social

Los resultados obtenidos tienen implicaciones metodológicas relevantes para la investigación social basada en encuestas digitales. En primer lugar, confirman que el tamaño muestral no compensa el sesgo estructural asociado a diseños no probabilísticos. Si bien el incremento del tamaño de muestra reduce la varianza del estimador *Var*($\hat{\mu}$), no modifica el sesgo sistemático *Sesgo*($\hat{\mu}$), por lo que el error total puede mantenerse elevado cuando el mecanismo de selección introduce distorsiones estructurales. En segundo lugar, los hallazgos destacan la importancia de la transparencia metodológica en estudios basados en encuestas virtuales; en particular, los investigadores deben explicitar con claridad la población objetivo, la población accesible, el mecanismo de selección de participantes y los supuestos bajo los cuales se aplican procedimientos de ajuste estadístico. En tercer lugar, los resultados sugieren que las estrategias de ponderación, como la postestratificación, deben interpretarse como mecanismos de corrección parcial y no como sustitutos de un diseño muestral probabilístico. Aunque estas técnicas pueden mejorar algunas métricas inferenciales, no garantizan validez cuando el proceso de participación depende de factores no observados. Finalmente, los resultados indican que los diseños híbridos o multimodales pueden constituir una estrategia metodológica más robusta, ya que la integración de encuestas web con modos presenciales o telefónicos permite ampliar la cobertura poblacional y reducir el sesgo asociado a la brecha digital.

4.5 Limitaciones del estudio

Aunque la simulación permitió controlar completamente el parámetro poblacional y el mecanismo de generación de datos, el entorno sintético utilizado no reproduce plenamente la complejidad de los procesos observados en estudios empíricos reales. En particular, el modelo de simulación no incorporó fenómenos frecuentes en encuestas aplicadas en contextos reales, tales como la no respuesta parcial a nivel de ítems (*item nonresponse*), posibles efectos de medición asociados al modo de recolección, errores de medición derivados del diseño del instrumento o del contexto de respuesta, ni dependencias temporales entre observaciones que puedan surgir en estudios longitudinales o en datos recolectados en distintos momentos. Asimismo, el análisis se centró exclusivamente en la estimación de la media poblacional como parámetro de interés. Otros parámetros estadísticos, como estimadores no lineales, modelos de regresión multivariada o estimaciones orientadas a inferencia causal, podrían presentar patrones de sesgo y comportamiento inferencial distintos bajo mecanismos de cobertura incompleta y autoselección, lo que sugiere la necesidad de investigaciones futuras que amplíen el marco analítico hacia estos contextos más complejos.

4.6 Líneas futuras de investigación

Los resultados obtenidos abren diversas líneas de investigación metodológica orientadas a profundizar el análisis de la inferencia estadística en entornos digitales. En primer lugar, sería pertinente extender el esquema de simulación a modelos de regresión lineal y logística, con el fin de evaluar cómo los mecanismos de cobertura incompleta y autoselección afectan la estimación de parámetros en contextos multivariados. En segundo lugar, futuras investigaciones podrían incorporar estimadores doblemente robustos que combinen modelamiento estadístico y ponderación, lo que permitiría explorar estrategias de corrección del sesgo bajo supuestos más flexibles. Asimismo, resulta relevante evaluar enfoques bayesianos que integren información auxiliar o conocimiento previo sobre la estructura poblacional, con el objetivo de mejorar la precisión inferencial cuando el marco muestral es incompleto. Otra línea prometedora consiste en analizar metodologías de integración entre datos probabilísticos y no probabilísticos, particularmente en diseños híbridos que buscan aprovechar las ventajas de ambos enfoques. Finalmente, se sugiere examinar el impacto de la *data defect correlation* bajo distintos tamaños poblacionales, ya que este concepto permite comprender cómo pequeñas correlaciones entre el mecanismo de selección y la variable de interés pueden amplificar sustancialmente el sesgo en contextos de grandes volúmenes de datos.

En conjunto, estos resultados confirman que la validez inferencial en encuestas virtuales depende críticamente del diseño muestral y de la cobertura del marco poblacional. La digitalización de los procesos de recolección de datos no modifica los fundamentos de la teoría del muestreo; por el contrario, exige una aplicación aún más rigurosa de sus principios, especialmente en contextos caracterizados por

desigualdades estructurales en el acceso a la tecnología y en la participación en estudios basados en plataformas digitales.

5. Conclusiones

Los resultados del estudio confirman que la representatividad en encuestas virtuales depende fundamentalmente del diseño muestral y de la cobertura del marco poblacional, más que del tamaño de la muestra o del medio tecnológico utilizado para la recolección de datos. La evidencia obtenida mediante simulación Monte Carlo muestra que los diseños no probabilísticos basados en autoselección generan sesgo sistemático y reducen significativamente la cobertura de los intervalos de confianza. Asimismo, se observa que el sesgo de cobertura constituye un riesgo estructural importante en contextos de brecha digital, ya que cuando el acceso a internet está correlacionado con la variable de interés, la población accesible difiere de la población objetivo, generando distorsiones previas al proceso de muestreo. En contraste, los diseños probabilísticos aplicados en entornos digitales (como el muestreo estratificado) presentan mejoras sustantivas en términos de sesgo, error cuadrático medio y cobertura.

Por otra parte, los resultados indican que los ajustes estadísticos basados en ponderación, como la postestratificación, pueden reducir parcialmente el sesgo cuando se cumplen supuestos de ignorabilidad condicional, pero no eliminan completamente las distorsiones cuando el mecanismo de selección depende de factores no observados. En consecuencia, estos métodos deben interpretarse como herramientas de mitigación y no como sustitutos de un diseño probabilístico adecuado. En conjunto, los hallazgos reafirman la vigencia de los principios de la teoría del muestreo en la era digital y evidencian que la validez inferencial en estudios basados en encuestas virtuales depende del rigor metodológico con el que se definan el marco de cobertura, el mecanismo de selección y los supuestos de ajuste estadístico.

Contribución de los autores: Conceptualización, LMQ-A.; metodología, LMQ-A. y JAH-R.; software, LMQ-A.; validación, LMQ-A. y JAH-R.; análisis formal, LMQ-A.; investigación, LMQ-A. y JAH-R.; recursos, LMQ-A., JAH-R., MFV-P., MRR-B. y SCU-C.; redacción del borrador original, LMQ-A y JAH-R.; redacción, revisión y edición, LMQ-A. y JAH-R.; visualización, LMQ-A., MFV-P., MRR-B. y SCU-C.; supervisión, JAH-R. Todos los autores han leído y aceptado la versión publicada del manuscrito.

Financiamiento: Esta investigación no ha recibido financiación externa.

Declaración de disponibilidad de datos: Los datos están disponibles previa solicitud a los autores de correspondencia: lquindea@uteq.edu.ec

Conflicto de interés: Los autores declaran no tener ningún conflicto de intereses

Referencias Bibliográficas

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., & Tourangeau, R. (2013). Summary Report of the AAPOR Task Force on Non-probability Sampling. Journal of Survey Statistics and Methodology, 1(2), 90–143. https://doi.org/10.1093/jssam/smt008

Bethlehem, J. (2010). Selection Bias in Web Surveys. International Statistical Review, 78(2), 161–188. https://doi.org/10.1111/j.1751-5823.2010.00112.x

Biffignandi Silvia, & Bethlehem Jelke. (2021). Sampling for Web Surveys. In Handbook of Web Surveys (pp. 93–131). Wiley. https://doi.org/10.1002/9781119371717.ch4

Blank, G., Graham, M., & Calvino, C. (2018). Local Geographies of Digital Inequality. Social Science Computer Review, 36(1), 82–102. https://doi.org/10.1177/0894439317693332

Bradley, V. C., Kuriwaki, S., Isakov, M., Sejdinovic, D., Meng, X.-L., & Flaxman, S. (2021). Unrepresentative big surveys significantly overestimated US vaccine uptake. Nature, 600(7890), 695–700. https://doi.org/10.1038/s41586-021-04198-4

Chen, Y., Li, P., & Wu, C. (2020). Doubly Robust Inference With Nonprobability Survey Samples. Journal of the American Statistical Association, 115(532), 2011–2021. https://doi.org/10.1080/01621459.2019.1677241

Cornesse, C., Blom, A. G., Dutwin, D., Krosnick, J. A., De Leeuw, E. D., Legleye, S., Pasek, J., Pennay, D., Phillips, B., Sakshaug, J. W., Struminskaya, B., & Wenz, A. (2020). A Review of Conceptual Approaches and Empirical Evidence on Probability and Nonprobability Sample Survey Research. Journal of Survey Statistics and Methodology, 8(1), 4–36. https://doi.org/10.1093/jssam/smz041

Elliott, M. R., & Valliant, R. (2017). Inference for Nonprobability Samples. Statistical Science, 32(2). https://doi.org/10.1214/16-STS598

Groves, R. M., & Lyberg, L. (2010). Total Survey Error: Past, Present, and Future. Public Opinion Quarterly, 74(5), 849–879. https://doi.org/10.1093/poq/nfq065

International Telecommunication Union. (2023). Measuring digital development: Facts and figures 2023. ITU Publications.

Lee, S., & Valliant, R. (2009). Estimation for Volunteer Panel Web Surveys Using Propensity Score Adjustment and Calibration Adjustment. Sociological Methods & Research, 37(3), 319–343. https://doi.org/10.1177/0049124108329643

Little, R. J. A., West, B. T., Boonstra, P. S., & Hu, J. (2020). Measures of the Degree of Departure from Ignorable Sample Selection. Journal of Survey Statistics and Methodology, 8(5), 932–964. https://doi.org/10.1093/jssam/smz023

Meng, X.-L. (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. The Annals of Applied Statistics, 12(2). https://doi.org/10.1214/18-AOAS1161SF

Robinson, L., Schulz, J., Blank, G., Ragnedda, M., Ono, H., Hogan, B., Mesch, G. S., Cotten, S. R., Kretchmer, S. B., Hale, T. M., Drabowicz, T., Yan, P., Wellman, B., Harper, M.-G., Quan-Haase, A., Dunn, H. S., Casilli, A. A., Tubaro, P., Carvath, R., … Khilnani, A. (2020). Digital inequalities 2.0: Legacy inequalities in the information age. First Monday. https://doi.org/10.5210/fm.v25i7.10842

Särndal, C.-Erik., Swensson, Bengt., & Wretman, J. H. (2003). Model assisted survey sampling. Springer-Verlag.

Schonlau, M., & Couper, M. P. (2017). Options for conducting web surveys. Statistical Science, 32(2), 279–292. https://doi.org/10.1214/16-STS597

Valliant, R., & Dever, J. A. (2011). Estimating Propensity Adjustments for Volunteer Web Surveys. Sociological Methods & Research, 40(1), 105–137. https://doi.org/10.1177/0049124110392533

Wu, C. (2022). Survey Methodology Statistical inference with non-probability survey samples How to obtain more information. Statistics Canada, Catalogue No. 12-001-X, Vol. 48, No. 2. www.statcan.gc.ca

Yang, S., & Kim, J. K. (2020). Statistical data integration in survey sampling: a review. Japanese Journal of Statistics and Data Science, 3(2), 625–650. https://doi.org/10.1007/s42081-020-00093-w

Yang, S., Kim, J. K., & Song, R. (2020). Doubly Robust Inference when Combining Probability and Non-Probability Samples with High Dimensional Data. Journal of the Royal Statistical Society Series B: Statistical Methodology, 82(2), 445–465. https://doi.org/10.1111/rssb.12354

Yeager, D. S., Krosnick, J. A., Chang, L., Javitz, H. S., Levendusky, M. S., Simpser, A., & Wang, R. (2011). Comparing the Accuracy of RDD Telephone Surveys and Internet Surveys Conducted with Probability and Non-Probability Samples. Public Opinion Quarterly, 75(4), 709–747. https://doi.org/10.1093/poq/nfr020